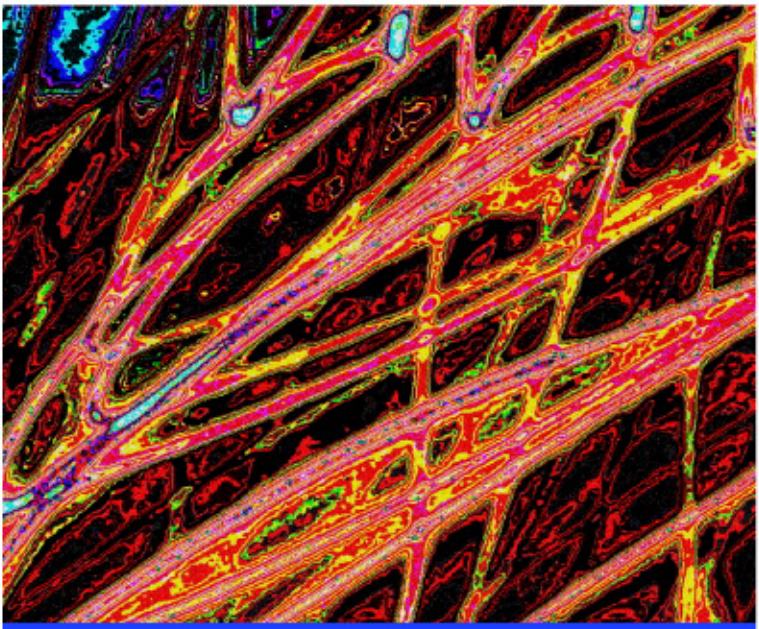


Photonic **Devices**

Jia-ming Liu

CAMBRIDGE

This page intentionally left blank

Photonic Devices

Photonic devices lie at the heart of the communications revolution, and have become a large and important part of the electronic engineering field, so much so that many colleges now treat this as a subject in its own right. With this in mind, the author has put together a unique textbook covering every major photonic device, and striking a careful balance between theoretical and practical concepts. The book assumes a basic knowledge of optics, semiconductors, and electromagnetic waves; many of the key background concepts are reviewed in the first chapter. Devices covered include optical fibers, couplers, electro-optic devices, magneto-optic devices, acousto-optic devices, nonlinear optical devices, optical amplifiers, lasers, light-emitting diodes, and photodetectors. Problems are included at the end of each chapter and a solutions set is available. The book is ideal for senior undergraduate and graduate courses, but being device-driven it is also an excellent reference for engineers.

Jia-Ming Liu is Professor of Electrical Engineering at the University of California, Los Angeles. He received his Ph.D. degree in applied physics from Harvard University in 1982. His research interests are in the areas of nonlinear optics, ultrafast optics, photonic devices, optical wave propagation, nonlinear laser dynamics, and chaotic communications. Dr. Liu has written about 200 scientific publications and holds eight US patents. He is a fellow of the Optical Society of America and the American Physical Society.

Photonic Devices

Jia-Ming Liu
Professor of Electrical Engineering
University of California, Los Angeles

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521551953

First published in print format 2005

ISBN-13 978-0-511-08152-1 eBook (EBL)
ISBN-10 0-511-08152-9 eBook (EBL)

ISBN-13 978-0-521-55195-3 hardback
ISBN-10 0-521-55195-1 hardback

To my family

Contents

Figures

Tables

Preface

Over the past two decades, photonics, the use of photons for engineering applications, has gradually become established as a well-defined engineering discipline. Photonics has developed from studies in crystal optics, guided-wave optics, nonlinear optics, lasers, and semiconductor optoelectronics. Though many excellent books exist on each of these subjects, and several have been written specifically to address photonics, it is still difficult to find one book where the diverse core subjects that are central to the study of photonic devices are presented with a good balance of breadth and depth of coverage. Through my teaching of undergraduate courses, I have found it very effective to introduce the field of photonics to undergraduate students using the rigorous, systematic approach of this book. Through my experience of working with graduate students in research, I have found that such a book is very much needed to prepare a solid foundation for graduate students who intend to major, or minor, in photonics. Through my teaching experience, I have found it highly desirable and beneficial for both instructors and students to have ample examples and problems that are well thought out and fully integrated with the subjects covered in the text. This book is written to address these needs.

I began this project in early 1994 after many years of teaching undergraduate and graduate courses in lasers, nonlinear optics, quantum electronics, and quantum mechanics. Though I had already accumulated a large collection of classnotes and problem sets when I started this project, it still took me exactly nine years to finish writing this book, with fully one-third of that time devoted to the work on examples and problems. Then, it took another year to prepare the figures. My students, both those in my classes and those in my research group, have been highly collaborative with the writing of this book. Throughout this process, I have taught various parts in different undergraduate and graduate courses to several hundred students. These students range from junior undergraduates to second-year graduates majoring in the diverse fields of photonics, solid-state electronics, electromagnetics, materials engineering, mechanical engineering, bioengineering, physics, chemistry, and many other disciplines. Many of their suggestions and feedback have been incorporated. All of the equations, examples, and problem solutions have been checked by several highly capable students. All of the

figures were produced, originally, by my graduate students. The manuscript underwent three major and numerous minor revisions before the book was finalized.

Objectives

This book is written for advanced undergraduate students and new graduate students who are interested in studying photonics as an engineering subject. A novice graduate student who plans to major in photonics can study this book thoroughly over a one-year period to lay a very solid foundation. It is also intended for practicing engineers and scientists who wish to broaden or deepen their knowledge in the principles of photonic devices. The objectives of this book are for a student (1) to obtain a good understanding of the core theory of photonic devices through coherent coverage of the subject, (2) to develop a deep physical insight into the principles of photonic devices through descriptive and illustrative approaches, (3) to gain realistic concepts of the functions of practical devices through numerical examples and discussions, and (4) to lay a solid foundation for further study and research in the photonics field through rigorous analytical treatment of the subject.

Guiding principles

To fulfill the objectives through a consistent approach, I followed several guidelines that I laid down for myself at the beginning of this project:

1. To address the subject at the device level, as the book title suggests. The physics and principles of devices are treated in depth, but the fabrication and processing of devices are not touched. The functions and characteristics of devices are also emphasized, but specific applications in subsystems and systems are not discussed for the reason that they are too diverse and vary quickly as time goes on.
2. To cover both bulk and guided-wave devices, with sufficient emphasis on guided-wave devices to reflect the development of photonics into integrated photonics.
3. To use a macroscopic treatment with two central approaches: (a) to treat the optical properties of materials through reference to the susceptibility tensor, χ, and permittivity tensor, ϵ; and (b) to treat the interaction of optical waves using coupled-wave theory for bulk devices and coupled-mode theory for guided-wave devices. With these approaches, it is possible to treat the majority of devices in great depth without ever touching quantum mechanics. For topics that necessitate an understanding of quantum concepts, I have adopted an approach that requires as little quantum mechanics background from the students as possible.

4. To balance both physics and engineering aspects with descriptive and analytical approaches to a significant, and consistent, depth throughout the entire book.
5. To concentrate on selected key topics and address them with sufficient rigor and thoroughness. On the one hand, analytical formulations and results that can be used at the level of practical applications and research are obtained. On the other, detailed and tedious mathematical derivations are avoided in favor of developing physical insight through an emphasis on the physical meanings of the analytical results.
6. To make the tables and figures useful and informative by using real data if possible while avoiding tedious details. Thus, the majority of the figures depicted in the book can be generated by the reader with realistic data using the analytical formulations obtained in the text.
7. To develop the concepts and data of working devices into realistic examples and problems.

Scope and structure

Photonics is a diverse field that can be addressed at various levels from many different perspectives. The scope and structure of this book are basically set by the guiding principles delineated above. This book focuses on the core topics of photonics at the device level covering both bulk and guided-wave devices. The entire book, as well as each chapter, is highly structured. Except for the general prerequisites described below, this book is written to be self-contained. General background and formulations that are needed for more than one chapter are provided in a few properly located individual chapters. Specific background needed only for the topics addressed within a particular chapter is provided at the start of each chapter. This arrangement allows the chapters and sections covering advanced topics to be treated as modules that can be added or dropped independently in a course or a study plan. Thus a minimum number of prerequisites are needed of the reader to begin studying any part of this book.

This book is divided into five parts. The first part consists of only one chapter that provides the relevant background in electromagnetics and optics for the entire book. This part also introduces χ and ϵ as the central concept for describing optical properties of materials. Part II covers four chapters on waveguides and couplers and lays the foundation for guided-wave devices. This part also develops coupled-wave and coupled-mode theories, which are used to formulate optical interactions throughout the entire book. Part III consists of four chapters covering devices based on electro-optics, magneto-optics, acousto-optics, and nonlinear optics. The fourth part contains two chapters on general discussions of laser amplifiers and laser oscillators. Fiber amplifiers and fiber lasers are specifically discussed in depth. Part V covers optoelectronic devices in three chapters. One chapter, i.e., Chapter 12, provides the background on semiconductors

relevant to optoelectronics. The other two chapters in Part V cover semiconductor lasers, LEDs, and photodetectors.

All chapters are organized in a consistent manner that mirrors the structure of the book. Basically, each begins with a general introduction of the underlying fundamental physics of the topics covered in the chapter, followed by general formulations of the physical effects. The principles and functions of bulk devices are then discussed. In the final section, or sections, of a chapter, guided-wave devices are addressed.

Symbols and units

Consistent symbols and notations are used throughout the entire book. The symbols and notations are chosen based on two criteria: (1) they are the same as those commonly used in the literature, whenever possible; and (2) they are intuitive to recognize and easy to distinguish. I also choose not to use many special fonts; thus, *script* is the only special font used. However, in a book like this that covers a diverse range of topics, it is inevitable that one quickly runs into a situation that a particular symbol is commonly used in the literature to represent two or more different meanings on different occasions. Whenever there is no confusion, I still choose to use the common symbol for different meanings. Otherwise, I choose to use subscripts and superscripts to clarify the meaning of the symbols. The system of symbols and notations followed throughout this book is described in Appendix A, and a partial list of symbols is presented later among these preliminary pages.

The SI metric system, which is summarized in Appendix C, is used. The values of some important fundamental physical constants in SI units are listed in Appendix D. Values of all the parameters listed in the tables throughout the chapters in this book are commonly given in SI units. On some rare occasions when the value of a parameter is not quoted in an SI unit, a conversion to the SI unit is given in the text.

Examples and problems

There are a total of 164 examples and more than 600 problems in the book. The examples and problems justly take up about one-third of the volume of this book as they took me about one-third of the time spent on this entire project. All examples and problems are originally generated and they are evenly distributed across the entire book. To illustrate the concepts developed in the text, most examples are realistic numerical problems based on working devices. Problems are tied closely to the text and examples. There are four types of problems: (1) qualitative questions on general concepts, (2) analytical steps leading to important results presented in the text because filling such steps by the reader enhances understanding, (3) further development of certain concepts covered in the text

into an advanced level beyond the general depth of the text, and (4) practical numerical problems reduced from realistic working devices. The problems are collected at the end of each chapter and are identified with the relevant section. They are not grouped by type, but are arranged in an order that parallels the presentation of the text. This arrangement, though not what I prefer, facilitates adding or dropping a particular topic module in a course syllabus or study plan.

Bibliography and reading lists

Though this book is intended to be self-contained, a reader always gains a deeper understanding and a different perspective of a topic by reading other books and journal articles. To maintain the coherence of the presentation in the text and to avoid unnecessarily distracting a reader, references and footnotes are rarely used. Instead, a bibliography containing reference books and a list of useful journal articles for advanced reading are placed at the end of each chapter. The reference books in a bibliography are meant to help a reader obtain a different perspective or further information on a particular topic. The journal articles listed in a reading list are meant for a reader to go beyond the level of the presentation in this book. The bibliographies and reading lists are rather extensive, but are carefully selected to limit their sizes to a manageable level.

Prerequisites and use of the book

The prerequisites of this book include background knowledge in optics covered in a college-level general physics course, a foundation in electromagnetic waves preferably in an electromagnetics course, and some background in semiconductors and quantum physics obtained in an introductory solid-state electronics course. In my experience, it is possible for a student who has only minimal background in these areas to succeed in an undergraduate course using this book if the background chapters of this book are studied thoroughly. Within the book, the prerequisites of each section are listed in a table in Appendix B.

This book can be used in a one-year undergraduate course by dropping advanced sections, and thus cutting about one-third of the material in the book, while covering every chapter. It can also be used in a one-year intensive graduate course covering all sections. I also envision this book as being used at different levels in different courses, including one-quarter or one-semester courses, depending on the interest and emphasis of a particular curriculum. The modular structure of this book and the table of prerequisites given in Appendix B make it very easy for an instructor to put together a specific course syllabus and for an independent reader to make up a study plan.

Acknowledgments

Before acknowledging the many people who have made direct contributions to this project, I would like to pay tribute first to Professor Nicolaas Bloembergen, who brought me into the fields of nonlinear optics and lasers and guided me through my graduate studies, which began 26 years ago. I would like to express my gratitude to Erich P. Ippen, Chi Hsiang Lee, Thomas B. Simpson, and Jeffery Y. Tsao for their friendship, support, and intellectual illumination over more than 20 years. I also thank my colleagues Tatsuo Itoh and Kung Yao for their encouragement during the course of writing this book.

At the end of this long project, I want to give my deepest appreciation to my wife, Vida, and my daughter, Janelle, for their support, patience, and sacrifices throughout the past ten years while I indulged myself in this most expensive and time-consuming hobby. My editors, Philip Meyler, who convinced me to start this project, and Eric Willner, who helped me to finish it, at Cambridge University Press deserve my special thanks not only for their input and assistance but particularly for their patience and understanding when this project dragged on longer than originally expected.

Many students have contributed to this project. Numerous students in my classes have given me valuable input and feedback. Two overlapping groups of graduate students have made direct contributions to this project. Andrew K. Newman, Juan C. Garcia, Kevin Geary, and Sze-Chun Chan formed a study group to read the text, check the equations, and check my solutions of the examples and problems. How-Foo Chen, Sheng-Kwang Hwang, Shuo Tang, Fan-Yi Lin, Sze-Chun Chan, Margaret C. Chiang, and Tyan-Lin Wang shared the efforts to produce all of the electronic files for the figures from my drafts. They applied their research abilities and skills to generate many original plots based on real data of materials and devices. Two of them, Sze-Chun Chan and Margaret C. Chiang, made the extraordinary efforts of finalizing all figures uniformly. I am truly blessed with these highly capable and supporting students. Their crucial contributions to the completion of this project are most appreciated. Thanks are also due to my copy editor, Lesley Thomas, and production editor, Joseph Bottrill, for their numerous valuable suggestions and professional efforts at the final stage of this project.

Partial list of symbols

Symbol	Unit	Meaning; derivatives	References[1]
a	m	fiber core radius	(3.1)
a	none	beam divergence ratio	(8.108)
a	none	round-trip intracavity field amplification factor	(11.5)
a_{E}, a_{M}	none	asymmetry factors for TE and TM modes	(2.48), (2.49)
A, $\tilde{A}$	$\mathrm{W}^{1/2}$	mode amplitude	(4.45), (4.48)
A_ν	$\mathrm{W}^{1/2}$	amplitude of mode ν; $A_{q,\nu}$	(4.23)
A_{21}	s^{-1}	Einstein A coefficient	(10.19)
A, A_{e}, A_{h}	s^{-1}	Shockley–Read recombination coefficients	(12.48)
$\mathcal{A}$	m^2	area; $\mathcal{A}_{\mathrm{eff}}$, $\mathcal{A}_q$, $\mathcal{A}_\nu^{\mathrm{eff}}$	(9.103)f, (12.118)
ABD	bit m^{-2}	areal bit density	(7.55)
b	m	confocal parameter of Gaussian beam, $b = 2z_{\mathrm{R}}$	(1.134)f
b	none	normalized guide index	(2.47)
b	none	linewidth enhancement factor	(13.61)
B, $\tilde{B}$	$\mathrm{W}^{1/2}$	mode amplitude	(4.45), (4.49)
B	Hz	bandwidth; B_{o}	(7.57), (14.11)
B	$\mathrm{m}^3\ \mathrm{s}^{-1}$	bimolecular carrier recombination coefficient	(12.49), (13.5)
B_{12}, B_{21}	$\mathrm{m}^3\ \mathrm{J}^{-1}\ \mathrm{s}^{-1}$	Einstein B coefficients	(10.17), (10.18)
$\boldsymbol{B}$	T	real magnetic induction in the time domain	(1.2)
$\mathbf{B}$, B	T	complex magnetic induction	(1.40)
c	$\mathrm{m\ s}^{-1}$	speed of light in free space	(1.38)
$c_{\nu\mu}$	none	overlap coefficient between modes ν and μ	(4.41)
c_{ijkl}, $c_\perp$, $c_\parallel$	$\mathrm{m}^2\ \mathrm{A}^{-2}$	quadratic magneto-optic coefficient	(7.13), (7.15)
$\mathrm{cn}(z, m)$	none	Jacobi elliptic function	(9.271)

[1] Suffixes, f "forward" and b "backward," on equation indicate symbols explained for the first time in the text immediately after or before the equation cited.

Symbol	Unit	Meaning; derivatives	References
C	F	capacitance; C_d, C_i, C_j, C_p, C_s	(12.119)
C, C_e, C_h	$\mathrm{m^6\ s^{-1}}$	Auger recombination coefficients	(12.49)
C_0	none	characteristic parameter of absorptive bistable device	(9.178)
$C_{\nu\mu\xi}$, C_Q	$\mathrm{s\ m^{-1}\ W^{-1/2}}$	effective second-order nonlinear coefficient for guided modes	(9.223), (9.239)
$C_{\nu\mu\xi\zeta}$	$\mathrm{s\ m^{-1}\ W^{-1}}$	effective third-order nonlinear coefficient for guided modes	(9.243)
d	m	thickness or distance; d_eff, d_g, d_QW, d_skin	(2.7)
d, d_0	m	beam spot size diameter, $d = 2w$, $d_0 = 2w_0$	(1.134)b
d_E, d_M	m	effective waveguide thickness for TE and TM modes	(2.59), (2.64)
$\mathbf{d}$, d_{ijk}, $d_{i\alpha}$	$\mathrm{m\ V^{-1}}$	nonlinear d coefficient tensor and elements, $\mathbf{d} = [d_{ijk}]$	(9.34)
d_eff, d_Q	$\mathrm{m\ V^{-1}}$	effective nonlinear d coefficients; $d_\mathrm{eff}^\mathrm{I}$, $d_\mathrm{eff}^\mathrm{II}$	(9.59)f
D	none	group-velocity dispersion; D_1, D_2, D_β	(1.166)
D	$\mathrm{W^{-1}}$	detectivity	(14.45)
D_e, D_h	$\mathrm{m^2\ s^{-1}}$	electron and hole diffusion coefficients	(12.58), (12.59)
D_λ	$\mathrm{s\ m^{-2}}$	group-velocity dispersion, $D_\lambda = -D/c\lambda$	(1.167)
D^*	$\mathrm{m\ Hz^{1/2}\ W^{-1}}$	specific detectivity	(14.46)
$\boldsymbol{D}$	$\mathrm{C\ m^{-2}}$	real electric displacement in the time domain	(1.1)
$\mathbf{D}$, D	$\mathrm{C\ m^{-2}}$	complex electric displacement; $\mathbf{D}_\mathrm{e}$, $\mathbf{D}_\mathrm{o}$, D_+, D_-	(1.41)
$\boldsymbol{\mathcal{D}}$, $\mathcal{D}$	$\mathrm{C\ m^{-2}}$	slowly varying amplitude of $\mathbf{D}(\mathbf{r}, t)$ and D; $\mathcal{D}_\mathrm{e}$, $\mathcal{D}_\mathrm{o}$	(1.126)
DR	dB	dynamic range	(14.49)
DS	none	normalized difference signal	(7.47)
e	C	electronic charge	(7.57)
$\hat{e}$	none	unit vector of electric field polarization; $\hat{e}_\mathrm{e}$, $\hat{e}_\mathrm{o}$, $\hat{e}_+$, $\hat{e}_-$	(1.58)
E_1, E_2	eV	energies of levels $\|1\rangle$ and $\|2\rangle$	(10.1)
E_c, E_v	eV	conduction- and valence-band edges; E_cn, E_cp, $E_{\mathrm{c},q}^\mathrm{QW}$, $E_{\mathrm{v},q}^\mathrm{QW}$	(12.2)
E_eff	$\mathrm{V\ m^{-1}}$	effective modulation electric field	(6.71)
E_F	eV	Fermi energy; E_Fc, E_Fi, E_Fv	(12.1)
E_g	eV	bandgap; E_gn, E_gp, E_g^QW	(12.2)
E_th	eV	threshold photon energy	(14.53)
$\boldsymbol{E}$	$\mathrm{V\ m^{-1}}$	real electric field in the time domain	(1.1)
$\boldsymbol{E}_0$, E_0	$\mathrm{V\ m^{-1}}$	static or low-frequency electric field	(6.1)

Symbol	Unit	Meaning; derivatives	References
$\boldsymbol{E}_\mathrm{e}$, $\boldsymbol{E}_\mathrm{h}$	V m^{-1}	electric field seen by electrons and holes	(12.58), (12.59)
$\mathbf{E}$, E	V m^{-1}	complex electric field; $\mathbf{E}_\mathrm{e}$, $\mathbf{E}_\mathrm{o}$, $\mathbf{E}_\mathrm{L}$, $\mathbf{E}_\mathrm{T}$	(1.39)
$\mathbf{E}_\nu$	V m^{-1}	complex electric field of mode ν	(2.1)
$\boldsymbol{\mathcal{E}}$, $\mathcal{E}$	V m^{-1}	slowly varying amplitude of $\mathbf{E}$ and E; $\mathcal{E}_\mathrm{i}$, $\mathcal{E}_\mathrm{r}$, $\mathcal{E}_\mathrm{t}$, $\mathcal{E}_+$, $\mathcal{E}_-$	(1.47)
$\boldsymbol{\mathcal{E}}_\nu$, $\mathcal{E}_\nu$	V m^{-1}	complex electric field profile of mode ν	(2.1)
$\hat{\boldsymbol{\mathcal{E}}}_\nu$	V m^{-1} W$^{-1/2}$	normalized electric mode field distribution, $\boldsymbol{\mathcal{E}}_\nu = A_\nu \hat{\boldsymbol{\mathcal{E}}}_\nu$	(2.41)
ER	dB	extinction ratio	(6.78)
f	m	focal length; f_K	(1.187), (9.144)
f	Hz	microwave or acoustic frequency, $f = \Omega/2\pi$; f_0, f_m, f_pk	(6.95), (8.3)
f_3dB	Hz	3-dB cutoff frequency; $f_\mathrm{m}^\mathrm{3dB}$	(6.92)
f_B, f_R	Hz	Brillouin and Raman frequencies, $f_\mathrm{B} = \Omega_\mathrm{B}/2\pi$, $f_\mathrm{R} = \Omega_\mathrm{R}/2\pi$	(9.185), (9.182)
f_{ijk}, f	m A^{-1}	linear magneto-optic coefficient, Faraday coefficient	(7.13)
f_r	Hz	relaxation resonance frequency	(13.134)
$f(E)$	none	Fermi–Dirac distribution function; $f_\mathrm{c}(E)$, $f_\mathrm{v}(E)$	(12.1)
F	none	excess noise factor	(14.25)
F, F_0	none	finesse of optical cavity, F_0 for lossless cavity	(9.168), (11.13)
F_o	none	optical noise figure	(10.116)
$F_{1/2}(\cdot)$	none	Fermi integral of order 1/2	(12.25)
$\mathbf{F}(z; z_0)$	none	forward-coupling matrix for codirectional coupling	(4.59)
g, $g(\nu)$, g_0	m^{-1}	gain coefficient, g_0 for unsaturated gain coefficient; g_max	(1.103)f, (10.44)
g_B, g_R	m^{-1}	Brillouin and Raman gain coefficients, $g_\mathrm{B} = \tilde{g}_\mathrm{B} I_\mathrm{p}$, $g_\mathrm{R} = \tilde{g}_\mathrm{R} I_\mathrm{p}$	(9.204), (9.191)
g_th	m^{-1}	threshold gain coefficient; g_{mn}^th	(11.69)
$\tilde{g}_\mathrm{B}$, $\tilde{g}_\mathrm{R}$	m W^{-1}	Brillouin and Raman gain factors; $\tilde{g}_\mathrm{B0}$, $\tilde{g}_\mathrm{R0}$	(9.186), (9.181)
$\hat{g}(\nu)$	s	lineshape function	(10.2)
g	none	degeneracy factor	(10.26)
g, g_0	s^{-1}	gain parameter, g_0 for unsaturated gain parameter; g_i	(11.73), (13.111)
g_n	m^3 s^{-1}	differential gain parameter	(13.129)
g_p	m^3 s^{-1}	nonlinear gain parameter	(13.129)
g_th	s^{-1}	threshold gain parameter	(11.70), (13.113)

Symbol	Unit	Meaning; derivatives	References
G	none	laser round-trip field gain; G_c, G_{mn}, G_{mn}^c	(11.5)
G	none	photodetector current gain	(14.23)
G, G_0	none	optical amplifier power gain, G_0 for unsaturated gain	(9.137), (10.97)
G_0	$\mathrm{m^{-3}\ s^{-1}}$	thermal generation rate; G_e^0, G_h^0	(13.5)
G_B, G_R	none	Brillouin and Raman amplifier gains; G_B^th, G_R^f, G_R^b, G_R^th	(9.205), (9.193)
h, $\hbar$	J s	Planck's constant, $\hbar = h/2\pi$	(1.178)
h, h_1	$\mathrm{m^{-1}}$	transverse spatial parameter of guided mode field	(3.11), (2.50)
H	m	height of acousto-optic transducer	(8.105)
H_c	$\mathrm{A\ m^{-1}}$	coercive magnetic field	Fig. 7.18
$H(\cdot)$	none	Heaviside function	(13.54)
$H_m(\cdot)$	none	Hermite function	(1.138)
$\boldsymbol{H}$	$\mathrm{A\ m^{-1}}$	real magnetic field in the time domain	(1.2)
$\boldsymbol{H}_0$, H_0	$\mathrm{A\ m^{-1}}$	static or low-frequency magnetic field	(7.4)
$\mathbf{H}$, H	$\mathrm{A\ m^{-1}}$	complex magnetic field	(1.41)
$\mathbf{H}_\nu$	$\mathrm{A\ m^{-1}}$	complex magnetic field of mode ν	(2.2)
$\boldsymbol{\mathcal{H}}$, $\mathcal{H}$	$\mathrm{A\ m^{-1}}$	slowly varying amplitude of $\mathbf{H}$ and H	
$\boldsymbol{\mathcal{H}}_\nu$	$\mathrm{A\ m^{-1}}$	complex magnetic field profile of mode ν	(2.2)
$\hat{\boldsymbol{\mathcal{H}}}_\nu$	$\mathrm{A\ m^{-1}\ W^{-1/2}}$	normalized magnetic mode field distribution, $\boldsymbol{\mathcal{H}}_\nu = A_\nu \hat{\boldsymbol{\mathcal{H}}}_\nu$	(2.41)
i	none	$\sqrt{-1}$	
i	A	current; i_b, i_d, i_da, i_dd, i_dk, i_n, i_ph, i_s	(7.94), (14.14)f
I	A	injection current; I_0, I_sat	(12.116)
I	$\mathrm{W\ m^{-2}}$	optical intensity; I_0, I_i, I_in, I_out, I_r, I_t, I_th, I_tr	(1.51)
$I(\nu)$	$\mathrm{W\ m^{-2}\ Hz^{-1}}$	optical spectral intensity distribution	(10.15)
I_a	$\mathrm{W\ m^{-2}}$	intensity of acoustic wave	(8.19)
I_p, I_s	$\mathrm{W\ m^{-2}}$	pump and signal intensities; $I_\mathrm{p}^\mathrm{sat}$, I_p^tr	(10.84), (10.94)
I_sat	$\mathrm{W\ m^{-2}}$	saturation intensity	(9.153), (10.73)
$\boldsymbol{J}$, J	$\mathrm{A\ m^{-2}}$	current density; $\boldsymbol{J}_\mathrm{diffusion}$, $\boldsymbol{J}_\mathrm{drift}$, J_th	(1.6), (12.109)
$\boldsymbol{J}_\mathrm{e}$, J_e	$\mathrm{A\ m^{-2}}$	electron current densities	(12.58), (12.109)
$\boldsymbol{J}_\mathrm{h}$, J_h	$\mathrm{A\ m^{-2}}$	hole current densities	(12.59), (12.109)
J_sat	$\mathrm{A\ m^{-2}}$	saturation current density	(12.113)
$J_m(\cdot)$	none	Bessel function of the first kind	(3.16)
k	$\mathrm{m^{-1}}$	propagation constant, wavenumber; k_0, k_i, k_p, k_r, k_t	(1.84)
k	none	impact ionization ratio	(14.104)
k_B	$\mathrm{J\ K^{-1}}$	Boltzmann constant	(3.95), (10.20)
k_e, k_o	$\mathrm{m^{-1}}$	propagation constant of extraordinary and ordinary waves	(1.126)

Symbol	Unit	Meaning; derivatives	References
k', k''	m^{-1}	real and imaginary parts of k, $k = k' + \mathrm{i}k''$	(1.100)
k^x, k^y, k^z	m^{-1}	propagation constants of x, y, and z polarized fields	(1.118)
k^X, k^Y, k^Z	m^{-1}	propagation constants of X, Y, and Z polarized fields	(6.37), (6.60)
k^+, k^-	m^{-1}	propagation constants of circularly polarized fields	(7.21)
$\hat{k}$	none	unit vector in the $\mathbf{k}$ direction	(1.92)
$\mathbf{k}$	m^{-1}	wavevector; $\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, $\mathbf{k}_\mathrm{s}$, $\mathbf{k}_\mathrm{t}$, $\mathbf{k}_q$	(1.47)
$\mathbf{k}_\mathrm{e}$, $\mathbf{k}_\mathrm{o}$	m^{-1}	wavevectors of extraordinary and ordinary waves	(1.127)b
$\mathbf{k}^x$, $\mathbf{k}^y$, $\mathbf{k}^z$	m^{-1}	wavevectors of x, y, and z polarized fields	(1.118)f
$\mathbf{k}^+$, $\mathbf{k}^-$	m^{-1}	wavevectors of left and right circularly polarized fields	(7.21)f
K	m^{-1}	wavenumber of grating or acoustic wave, $K = 2\pi/\Lambda$	(4.54), (8.3)
K	$\mathrm{lm\ W}^{-1}$	peak efficacy, photometric radiation equivalent	(13.65)
K	s	K factor of semiconductor laser	(13.136)
K_L, K_T	m^{-1}	wavenumbers of longitudinal and transverse acoustic waves	(8.23), (8.24)
$K_m(\cdot)$	none	modified Bessel function of the second kind	(3.17)
$\mathbf{K}$	m^{-1}	wavevector of grating, acoustic wave, or material wave	(8.1), (9.180)
l	m	length or distance; l_eff, l_F, l_g, l_p, l_opt, l_RT	(3.89), (11.33)
l_a	m	aperture distance	(9.90)
l_c	m	coupling length; l_c^PM	(4.67)
l_coh	m	coherence length	(9.70)
$l_{\lambda/2}$, $l_{\lambda/4}$	m	half-wave and quarter-wave lengths	(1.119), (1.120)
$\mathbf{l}$	m	length vector along a path	(7.49)
L	m	length of acousto-optic transducer	(8.105)
L_0	dB	background optical loss	(7.43)
L_e, L_h	m	electron and hole diffusion lengths	(12.105), (12.106)
m	none	transverse mode index associated with x or ϕ	(1.140)
m	none	modulation index	(13.73)
m	none	electron multiplication factor	(14.58)
m_0	kg	free electron rest mass	(12.15)f

Symbol	Unit	Meaning; derivatives	References
m_c, m_v	kg	conduction- and valence-band density of states effective masses	(12.15), (12.16)
m_{hh}, m_{lh}	kg	heavy-hole and light-hole density of states effective masses	(12.17)
m_e^*, m_h^*	kg	density of states effective masses of electrons and holes	(12.15), (12.16)
m_r^*	kg	reduced effective mass	(13.15)
M	none	number of guided modes; M_β	(2.75), (3.61)
M	kg	atomic or molecular mass	(10.13)
M_2	$m^2\ W^{-1}$	acousto-optic figure of merit	(8.22)
M_s	$A\ m^{-1}$	saturation magnetization	(7.27)
M_{QW}	none	number of quantum wells	(13.81)
$\boldsymbol{M}$	$A\ m^{-1}$	real magnetic polarization in the time domain	(1.2)
$\boldsymbol{M}_0$, M_0	$A\ m^{-1}$	static or low-frequency magnetization	(7.6)
n	none	transverse mode index associated with y or r	(1.140), (3.3)
n	none	index of refraction; n_d, n_i, n_m, n_p, n_β, $\bar{n}$	(1.95)
n	m^{-3}	electron concentration; n_n, n_p	(12.41)
n_0	m^{-3}	equilibrium concentration of electrons; n_{n0}, n_{p0}	(12.18)
n_1, n_2, n_3	none	refractive indices of waveguide layers, $n_1 > n_2 > n_3$	(2.5)
n_2	$m^2\ W^{-1}$	coefficient of intensity-dependent index change	(9.49)
n_i	m^{-3}	intrinsic carrier concentration; n_{in}, n_{ip}	(12.29)
n_e, n_o	none	extraordinary and ordinary indices of refraction	(1.125)
n_x, n_y, n_z	none	principal indices of refraction	(1.110)
n_X, n_Y, n_Z	none	new principal indices of refraction	(6.13)
n_+, n_-	none	principal indices of refraction for circularly polarized modes	(7.20)
$n_\perp$, $n_\parallel$	none	indices of second-order magneto-optic effect	(7.16)
n', n''	none	real and imaginary parts of refractive index, $n = n' + in''$	(1.101)
$\hat{n}$	none	unit normal vector	(1.17)
N	none	some number; N_{DBR}, N_e, N_g	(8.131), (11.105)
N	none	group index; N_1, N_2, N_{high}, N_{low}, N_β	(1.170)
N	m^{-3}	excess carrier density; N_{th}, N_{tr}	(12.55)
N_1, N_2, N_t	m^{-3}	population densities in energy levels $\|1\rangle$, $\|2\rangle$, and all levels	(10.24), (10.70)

Symbol	Unit	Meaning; derivatives	References
N_{2D}	m^{-2}	two-dimensional excess carrier density; N_{tr}^{2D}	(13.57)
N_a, N_d	m^{-3}	concentrations of acceptors and donors; N_a^-, N_d^+	(12.30)
N_c, N_v	m^{-3}	conduction- and valence-band effective density of states	(12.22), (12.23)
N_{sp}	none	spontaneous emission factor	(10.114)
$\mathcal{N}$	none	number of charge carriers	(14.13)
NA	none	numerical aperture	(3.2)
NEP	W	noise equivalent power	(14.42)
p	none	probability	(14.1)
p	none	cross-section ratio for pumping, $p = \sigma_e^p/\sigma_a^p$	(10.67)
p	m^{-3}	hole concentration; p_n, p_p	(12.42)
p_0	m^{-3}	equilibrium concentration of holes; p_{n0}, p_{p0}	(12.19)
p_{ijkl}, $p_{\alpha\beta}$, p	none	elasto-optic coefficient	(8.7), (8.8)
$p(\nu_k)$	Hz^{-1}	probability density function	(10.9)
P	W	power; P_a, P_{av}, P_c, P_e, P_{in}, P_{out}, P_{pk}, P_{th}, $P_{n,th}$	(2.38)
P_p, P_s	W	pump and signal powers; P_p^{in}, P_p^{out}, P_p^{sat}, P_p^{th}, P_p^{tr}, P_s^{in}, P_s^{out}	(10.100), (10.95)
P_{sat}	W	saturation power	(10.95)
P_{sp}	W	spontaneous emission power	(10.90)f
P_{sp}^{tr}	W	critical fluorescence power	(10.93)f
$\hat{P}_{sp}$	$W\,m^{-3}$	spontaneous emission power density, $P_{sp} = \hat{P}_{sp}\mathcal{V}$	(10.90)
$\hat{P}_{sp}^{tr}$	$W\,m^{-3}$	critical fluorescence power density, $P_{sp}^{tr} = \hat{P}_{sp}^{tr}\mathcal{V}$	(10.93)
$\boldsymbol{P}$	$C\,m^{-2}$	real electric polarization in the time domain	(1.1)
$\mathbf{P}$, P	$C\,m^{-2}$	complex electric polarization	(1.45)
$\boldsymbol{P}^{(n)}$	$C\,m^{-2}$	nth-order nonlinear real electric polarization	(9.1)
$\mathbf{P}^{(n)}$, $P^{(n)}$	$C\,m^{-2}$	nth-order nonlinear complex electric polarization	(9.13)
$\mathbf{P}_{res}$	$C\,m^{-2}$	complex electric polarization from resonant transition	(10.56)
q	none	frequency index or longitudinal mode index	(4.5), (11.5)
q	none	grating order	(5.1)
q	none	quantum number	(13.49)

Symbol	Unit	Meaning; derivatives	References
$q(z)$	m	complex radius of curvature of Gaussian beam	(1.139)
Q	C	charge	(12.118)
Q	none	acousto-optic diffraction parameter	(8.49)
Q, Q_{mnq}	none	quality factor of resonator	(11.27), (11.31)
r	m	radial distance, radial coordinate	
r	none	reflection coefficient; r_1, r_2, r_p, r_s, r_pp, r_ps, r_sp, r_ss, r_+, r_-	(1.147)f
r	none	pumping ratio of a laser	(11.76)
r_{ijk}, $r_{\alpha k}$	m V^{-1}	linear electro-optic coefficients, Pockels coefficients	(6.14), (6.15)
$r(f)$, $r(\Omega)$	none	complex modulation response function	(13.75)
$\hat{r}$	none	unit vector of radial coordinate r	(3.70)
$\mathbf{r}$	m	spatial vector	(1.1)
R	none	reflectance, reflectivity; R_1, R_2, R_DBR, R_p, R_s, R_+, R_-	(1.149)f
R	Ω	resistance; R_eq, R_i, R_L, R_s	(14.29)
R	none	chromatic resolving power	(8.152)
R	m^{-3} s^{-1}	recombination rates; R_e, $\mathrm{R_h}$, R_nonrad, R_rad, R_SR	(12.49)f
$R_\mathrm{a}(\nu)$	m^{-3}	absorption spectrum	(13.18)
$R_\mathrm{e}(\nu)$	m^{-3}	stimulated emission spectrum	(13.19)
$R_\mathrm{sp}(\nu)$	m^{-3}	spontaneous emission spectrum; $R^0_\mathrm{sp}(\nu)$	(13.20)
R^0_sp	m^{-3} s^{-1}	spontaneous emission rate in thermal equilibrium	(13.45)
R_B	none	conversion efficiency of a Brillouin generator	(9.210)
R_1, R_2	m^{-3} s^{-1}	pumping rates for laser levels $\|1\rangle$ and $\|2\rangle$	(10.61), (10.62)
$R(f)$	none	electrical power spectrum of modulation response	(13.76)
$\mathbf{R}(z;0,l)$	none	reverse-coupling matrix for contradirectional coupling	(4.70)
$\mathbf{R}$, R_{ij}	none	rotation tensor and elements, $\mathbf{R} = [R_{ij}]$	(8.6)
$\mathcal{R}$	m	radius of curvature; $\mathcal{R}_1$, $\mathcal{R}_2$	(1.136)
$\mathcal{R}$	A W^{-1}	responsivity of photodetector with current output; $\mathcal{R}_0$	(14.37)
$\mathcal{R}$	V W^{-1}	responsivity of photodetector with voltage output	(14.38)
s	m	separation; s_e	(5.68), (6.69)
s	none	pumping ratio of an amplifier; s_th	(10.104)
s	none	signal; s_n	(14.1)

Symbol	Unit	Meaning; derivatives	References
s_{ijkl}, $s_{\alpha kl}$	$\mathrm{m^2\ V^{-2}}$	quadratic electro-optic coefficients, Kerr coefficients	(6.14), (6.15)
S	$\mathrm{m^{-3}}$	photon density	(11.71)
S_{sat}	$\mathrm{m^{-3}}$	saturation photon density	(11.74)
$\boldsymbol{S}$	$\mathrm{W\ m^{-2}}$	real Poynting vector	(1.28)
$\mathbf{S}$	$\mathrm{W\ m^{-2}}$	complex Poynting vector; $\mathbf{S}_{\mathrm{e}}$, $\mathbf{S}_{\mathrm{o}}$	(1.49)
$\mathbf{S}$, S_{ij}	none	strain tensor, $\mathbf{S} = [S_{ij}]$	(8.5)
$\mathbf{S}(z; z_0)$	none	forward-coupling matrix for contradirectional coupling	(4.95)
$\mathcal{S}$	none	amplitude of strain; $\mathcal{S}_{kl}$	(8.16)
$\mathcal{S}$	none	number of photons	(14.12)
SNR	none, dB	signal-to-noise ratio	(14.9)
t	s	time	
t	none	transmission coefficient; t_{p}, t_{s}	(1.148)
t_{r}, t_{f}	s	risetime and falltime	(8.114), (14.50)
T	K	temperature; T_{c}, T_{comp}, T_{PM}	(3.95)
T	s	time interval	(1.48), (14.11)
T	s	round-trip time of laser cavity	(11.1)
T	none	transmittance, transmissivity; T_{DBR}, T_{h}, T_{p}, T_{s}, $T_{\perp}$, $T_{\parallel}$	(1.150)f
$\mathbf{T}$	none	transformation matrix	(6.7)
$\tilde{\mathbf{T}}$	none	transpose of transformation matrix $\mathbf{T}$	(6.8)
u, u_0	$\mathrm{J\ m^{-3}}$	electromagnetic energy density	(1.33), (1.29)
$u(\nu)$	$\mathrm{J\ m^{-3}\ Hz^{-1}}$	spectral energy density	(10.14)
$\mathbf{u}$, u_i	m	elastic deformation wave and its components	(8.1)f
U	J	optical energy; U_{mode}	(11.77)
$\boldsymbol{\mathcal{U}}$	m	amplitude of elastic wave	(8.1)
v	V	voltage; v_{n}, v_{out}, v_{s}	(14.87)
v	$\mathrm{m\ s^{-1}}$	velocity	
v_{a}	$\mathrm{m\ s^{-1}}$	acoustic wave velocity; $v_{\mathrm{a,L}}$, $v_{\mathrm{a,T}}$	(8.3)
v_{g}	$\mathrm{m\ s^{-1}}$	group velocity	(1.164)
v_{p}	$\mathrm{m\ s^{-1}}$	phase velocity; $v_{\mathrm{p}}^{\mathrm{m}}$, $v_{\mathrm{p}}^{\mathrm{o}}$	(1.161)
v_{e}, v_{h}, v_{sat}	$\mathrm{m\ s^{-1}}$	electron and hole drift velocities and saturation velocity	(14.103)
V	none	normalized frequency and waveguide thickness, V number	(2.46)
V_{c}	none	cutoff V number; V_m^{c}	(2.70)
V	$\mathrm{rad\ A^{-1}}$	Verdet constant	(7.26)
V	V	voltage; V_{b}, V_{j}, V_{MC}, V_{pk}, V_{π}, $V_{\pi/2}$, $V_{\lambda/2}$, $V_{\lambda/4}$	(6.39)

Symbol	Unit	Meaning; derivatives	References
V_0	V	contact potential	(12.72)
$V(\lambda)$	none	normalized photopic luminous efficiency	Fig. 13.21
$\mathcal{V}$	m^3	volume; $\mathcal{V}_{active}$, $\mathcal{V}_{mode}$	(1.26), (11.2)
w	m	width	Fig. 2.15
w, w_0	m	Gaussian beam radius, spot size; w_{0K}; w_p; w_s; $w_\perp$; $w_\parallel$	(1.134)f
W	m	width of acousto-optic cell	(8.125)
W	m	depletion layer width; W_0	(12.94)
W	s^{-1}	transition probability rate; W_{12}, W_{21}, W_p, W_p^{tr}, W_{sp}	(10.21)–(10.23)
W_p, W_m	W m^{-3}	power densities expended by EM field on $\boldsymbol{P}$ and $\boldsymbol{M}$	(1.30), (1.31)
$W(\nu)$	none	transition rate per unit frequency; $W_{12}(\nu)$, $W_{21}(\nu)$, $W_{sp}(\nu)$	(10.17)–(10.19)
x	m	spatial coordinate	
$\hat{x}$	none	unit coordinate vector or principal dielectric axis	(1.59), (1.109)b
x_p, x_n	m	depletion layer penetration depths into p and n regions	(12.94)
X	m	spatial coordinate along $\hat{X}$	(6.9)
$\hat{X}$	none	new principal dielectric axis	(6.6)
y	m	spatial coordinate	
$\hat{y}$	none	unit coordinate vector or principal dielectric axis	(1.59), (1.109)b
Y	m	spatial coordinate along $\hat{Y}$	(6.9)
$\hat{Y}$	none	new principal dielectric axis	(6.6)
z	m	spatial coordinate	
$\hat{z}$	none	unit coordinate vector or principal dielectric axis	(2.37), (1.109)b
z_R	m	Rayleigh range of Gaussian beam	(1.134)
Z	m	spatial coordinate along $\hat{Z}$	(6.9)
$\hat{Z}$	none	new principal dielectric axis	(6.6)
Z	Ω	impedance; Z_s	(1.97), (6.107)
Z_0	Ω	characteristic impedance of free space	(1.93)
α	rad	field polarization angle; α_K	(1.61)
α	rad	walk-off angle of extraordinary wave	(1.131)
α	none	power-law fiber index profile parameter	(3.83)
α, $\alpha(\nu)$	m^{-1}	attenuation coefficient; α_{iq}, α_m, α_p, α_s, α_r, α_R	(1.100), (10.43)
α_0, $\alpha_0(\nu)$	m^{-1}	intrinsic or unsaturated absorption coefficient	(9.153), (13.32)

Symbol	Unit	Meaning; derivatives	References
α_c	m^{-1}	propagation parameter for contradirectional coupling	(4.72)
α_e, α_h	m^{-1}	electron and hole ionization coefficients	(14.104)
β	m^{-1}	propagation constant; β_B, β_{mn}, β_{TE}, β_{TM}, β_ν, β_ν^{eff}, β_ν^{NL}	(1.100), (2.1)
β	$m^3\ J^{-1}$	isothermal compressibility	(3.95)
β_c	m^{-1}	propagation parameter for codirectional coupling	(4.61)
γ	s^{-1}	relaxation rate, decay rate; γ_{21}, γ_B, γ_i, γ_{out}, γ_R	(1.174)
γ, γ_2, γ_3	m^{-1}	transverse spatial decay parameter of mode field	(3.12), (2.51)
γ_a	s^{-1}	acoustic decay rate	(8.127)
γ_c	s^{-1}	cavity decay rate, photon decay rate; γ_{cl}, γ_{mnq}^c	(11.26)
γ_n	s^{-1}	differential carrier relaxation rate	(13.130)
γ_p	s^{-1}	nonlinear carrier relaxation rate	(13.130)
γ_r	s^{-1}	total carrier relaxation rate	(13.133)
γ_s	s^{-1}	spontaneous carrier recombination rate	(13.3)
Γ	none	overlap or confinement factor; Γ_{EM}, Γ_{id}, Γ_p, Γ_s, Γ_ν, $\Gamma_{\nu\mu\xi}$	(2.77), (11.2)
δ	none	index dispersion parameter for optical fiber	(3.108)
δ	m^{-1}	phase mismatch parameter for phase mismatch of 2δ	(4.50)
$\delta\varphi$	rad	small variation of phase retardation $\Delta\varphi$	(6.66)
Δ	none	normalized index step of optical fiber	(3.48)
$\Delta_{ijk}^{(2)}$	$m\ V^{-1}$	Miller's constant	(9.276)
ΔE_F	eV	separation between quasi-Fermi levels	(12.47)
Δf_B, Δf_R	Hz	Brillouin and Raman spectral linewidths, $\Delta f = \Delta\Omega/2\pi$	(9.187), (9.182)f
$\Delta\mathbf{k}$, Δk	m^{-1}	phase mismatch; Δk_Q	(9.60)f
Δn	none	index step of waveguide structure or optical fiber	(3.48)
Δn, Δp	m^{-3}	excess electron and hole concentrations	(12.54)
$\Delta\mathbf{P}$	$C\ m^{-2}$	change in electric polarization	(4.3)
Δt	s	temporal broadening or pulsewidth; Δt_g, Δt_m, Δt_{ps}	(11.95)
$\Delta\beta$	m^{-1}	guided mode phase mismatch; $\Delta\beta_{\nu\mu\xi}$	(4.56)
$\Delta\beta$	m^{-1}	change in propagation constant; $\Delta\beta_{pk}$	(6.97)
$\Delta\epsilon$, $\Delta\epsilon$	$F\ m^{-1}$	variation or modulation of electric permittivity	(4.35)

Symbol	Unit	Meaning; derivatives	References
$\Delta\tilde{\boldsymbol{\epsilon}}$, $\Delta\tilde{\epsilon}$	F m^{-1}	amplitude of $\Delta\boldsymbol{\epsilon}$ and $\Delta\epsilon$	(8.17)
$\Delta\boldsymbol{\eta}$, $\Delta\eta$	none	variation or modulation of relative impermeability	(6.17)
$\Delta\theta$	rad	divergence angle of Gaussian beam; $\Delta\theta_{\mathrm{o}}$, $\Delta\theta_{\perp}$, $\Delta\theta_{\parallel}$	(1.137)
$\Delta\theta_{\mathrm{a}}$	rad	acoustic beam divergence	(8.108)
$\Delta\lambda$	m	spectral width	(8.151)
$\Delta\nu$	Hz	optical linewidth, bandwidth; $\Delta\nu_{\mathrm{D}}$, $\Delta\nu_{\mathrm{inh}}$, $\Delta\nu_{\mathrm{h}}$, $\Delta\nu_{\mathrm{p}}$, $\Delta\nu_{\mathrm{ps}}$	(10.4)
$\Delta\nu_{\mathrm{c}}$	Hz	longitudinal mode linewidth	(11.19)
$\Delta\nu_{\mathrm{L}}$	Hz	longitudinal mode frequency separation	(11.17)
$\Delta\nu_{\mathrm{SB}}$	Hz	stop band of DBR	(13.103)
$\Delta\nu_{\mathrm{ST}}$	Hz	Shawlow–Townes linewidth of laser mode	(11.65)
$\Delta\varphi$	rad	phase shift or retardation; $\Delta\varphi_0$, $\Delta\varphi_{\mathrm{b}}$, $\Delta\varphi_{\mathrm{NL}}$, $\Delta\varphi_{\mathrm{rec}}$	(6.49), (9.259)
$\Delta\varphi_{\mathrm{c}}$	rad	width of a resonance peak of passive cavity	(11.12)
$\Delta\varphi_{\mathrm{L}}$	rad	phase separation between neighboring longitudinal modes	(11.11)
$\Delta\boldsymbol{\chi}$, $\Delta\chi$	none	variation or modulation of electric susceptibility	(6.1)
$\Delta\omega$	rad s^{-1}	optical bandwidth, linewidth, $\Delta\omega = 2\pi\,\Delta\nu$; $\Delta\omega_{\mathrm{inh}}$, $\Delta\omega_{\mathrm{h}}$	(5.28), (10.3)f
$\Delta\Omega_{\mathrm{B}}$, $\Delta\Omega_{\mathrm{R}}$	rad s^{-1}	Brillouin and Raman spectral linewidths, $\Delta\Omega = 2\pi\,\Delta f$	(9.186)f, (9.182)f
ϵ	F m^{-1}	complex electric permittivity; ϵ_{n}, ϵ_{p}	(1.95)
ϵ_0	F m^{-1}	electric permittivity of free space	(1.1)
ϵ', ϵ''	F m^{-1}	real and imaginary parts of ϵ, $\epsilon = \epsilon' + \mathrm{i}\epsilon''$	(1.99)
ϵ_x, ϵ_y, ϵ_z	F m^{-1}	principal dielectric permittivities	(1.109)
ϵ_X, ϵ_Y, ϵ_Z	F m^{-1}	new principal dielectric permittivities	(6.12)
ϵ_+, ϵ_-	F m^{-1}	principal dielectric permittivities of circular polarizations	(7.17)
$\epsilon_{\mathrm{res}}(\omega)$	F m^{-1}	permittivity of resonant transition	(11.37)
$\boldsymbol{\epsilon}(\mathbf{r}, t)$	F m^{-4} s^{-1}	real electric permittivity tensor in the time domain	(1.16)
$\boldsymbol{\epsilon}(\omega)$, $\epsilon_{ij}(\omega)$	F m^{-1}	complex electric permittivity tensor in the frequency domain	(1.55)
ε	rad	ellipticity of polarization ellipse; ε_{F}, ε_{K}	(1.65), (7.30)
ζ	none	a mode parameter for multimode fiber	(3.107)
ζ	none	linear birefringence in magneto-optics	(7.74)

Symbol	Unit	Meaning; derivatives	References
$\zeta_{mn}(z)$	rad	phase variation of Gaussian mode field; ζ_{mn}^{RT}	(1.140)
ζ_{p}	none	fraction of pump power absorbed by gain medium; $\zeta_{\mathrm{p}}^{\mathrm{th}}$	(10.107)
η	none	characteristic constant for HE or EH fiber mode	(3.39)
η	none	coupling efficiency; η_{in}, η_{max}, η_{out}, η_{PM}	(4.66)
η_{c}	none	power conversion efficiency	(10.109)
η_{coll}	none	collection efficiency	(14.35)
η_{e}	none	external quantum efficiency	(11.91), (13.63)
η_{esc}	none	escape efficiency	(13.68)
η_{i}	none	internal quantum efficiency	(11.91)
η_{inj}	none	injection efficiency	(13.67)
η_{l}	lm W^{-1}	photometric efficiency, luminous efficiency	(13.65)
η_{p}	none	pump quantum efficiency	(10.84)
η_{q}	none	quantum efficiency of laser amplifier	(10.111)
η_{R}	none	Raman conversion efficiency	(9.199)
η_{s}	none	slope efficiency, differential power conversion efficiency	(10.110), (11.90)
η_{t}	none	conversion efficiency of transducer	(8.105)
η_{t}	none	extraction efficiency or transmission efficiency	(13.67), (14.35)
η_{SH}	none	second-harmonic conversion efficiency	(9.115)
$\hat{\eta}_{\mathrm{SH}}$	W^{-1}	normalized second-harmonic conversion efficiency	(9.117)
$\boldsymbol{\eta}$, η_{ij}, η_{α}	none	relative impermeability tensor and its elements, $\boldsymbol{\eta} = [\eta_{ij}]$	(1.111)
θ	rad	coordinate angle	(1.121)
θ	rad	orientation of the polarization ellipse	(1.66)
θ_{a}	rad	acceptance angle	(3.2)
θ_{B}	rad	Brewster angle or Bragg angle	(1.156), (8.65)
θ_{c}	rad	critical angle	(1.158)
θ_{d}	rad	angle of diffraction; $\theta_{\mathrm{d}}^{\mathrm{PM}}$	(8.60)
θ_{def}	rad	deflection angle	Fig. 8.5
θ_{F}, θ_{K}	rad	Faraday and Kerr rotation angles	(7.24), (7.36)
θ_{i}, θ_{r}, θ_{t}	rad	angles of incidence, reflection, and refraction (transmitted)	(1.144)
θ_{PM}	rad	phase-matching angle; $\theta_{\mathrm{PM}}^{\mathrm{I}}$, $\theta_{\mathrm{PM}}^{\mathrm{II}}$	Table 9.6
κ	m^{-1}	coupling coefficient; κ_{eff}, κ_{EE}, κ_{EM}, κ_{ME}, κ_{MM}, $\kappa_{\nu\mu}$	(4.33)

Symbol	Unit	Meaning; derivatives	References
$\tilde{\kappa}$	m^{-1}	coupling coefficient defined in (4.42); $\tilde{\kappa}_{\nu\mu}$	(4.42)
λ	m	optical wavelength in free space; λ_d, λ_p, λ_s	(1.85)
λ_B	m	free-space Bragg wavelength	(5.23)
λ_c	m	cutoff wavelength; λ_m^c	(2.74)
λ_g	m	wavelength for photon with bandgap energy E_g, $\lambda_\mathrm{g} = hc/E_\mathrm{g}$	(12.2)f
λ_th	m	threshold wavelength	(14.53)
Λ	m	grating period or acoustic wavelength, $\Lambda = 2\pi/K$; Λ_0	(4.54), (8.3)
$\boldsymbol{\mu}$	$\mathrm{H\ m}^{-1}$	magnetic permeability tensor	(7.3)
μ_0	$\mathrm{H\ m}^{-1}$	magnetic permeability of free space	(1.4)
μ_e, μ_h	$\mathrm{m}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1}$	electron and hole mobilities	(12.58), (12.59)
ν	Hz	optical frequency; ν_p, ν_{mnq}, ν_q, ν_s	(1.85)
ν_0	Hz	central or carrier optical frequency, $\nu_0 = \omega_0/2\pi$	(10.11)
ν_{21}	Hz	resonance frequency between energy levels $\|1\rangle$ and $\|2\rangle$	(10.1)
ν_B	Hz	Bragg frequency	(5.24)b
ν_c	Hz	characteristic frequency in McCumber relation	(10.47)
ξ	none	duty factor or splitting factor	(5.16), (9.261)
ξ, $\xi(M_{0z})$	none	electric permittivity tensor elements for magneto-optic effect	(7.16)
ρ	$\mathrm{C\ m}^{-3}$	charge density	(1.7)
ρ	$\mathrm{kg\ m}^{-3}$	density of mass	(8.19)
ρ	rad	walk-off angle between two beams	(9.90)
ρ	Ω m	resistivity, $\rho = 1/\sigma$	(12.70)f
ρ_F	$\mathrm{rad\ m}^{-1}$	specific Faraday rotation, rotatory power	(7.27)
$\rho_\mathrm{c}(E)$, $\rho_\mathrm{v}(E)$	$\mathrm{m}^{-3}\ \mathrm{J}^{-1}$	densities of states for conduction and valence bands	(12.15), (12.16)
$\rho(\nu)$	$\mathrm{m}^{-3}\ \mathrm{Hz}^{-1}$	density of states for band-to-band optical transitions	(13.16)
σ	none	extinction ratio of polarizer, $\sigma = T_\perp/T_\parallel$; σ_in, σ_out	(7.42)
σ	$\mathrm{m}^{-1}\ \mathrm{W}^{-1}$	nonlinear coefficient in self-phase modulation; $\sigma_{\nu\nu}$	(9.245), (9.266)
σ	$\Omega^{-1}\ \mathrm{m}^{-1}$	conductivity; σ_0	(12.70)
σ	m^2	gain cross section	(13.40), (13.59)

Symbol	Unit	Meaning; derivatives	References
σ_{12}, σ_{21}	m^2	transition cross sections	(10.34), (10.35)
σ_a, σ_e	m^2	absorption and emission cross sections; σ_{ap}, σ_e^h, σ_e^{inh}, σ_{ep}	(10.36), (10.37)
σ_s^2	none	variance of s	(14.2)
τ	s	lifetime, decay time, or time constant; τ_R, τ_{rad}, τ_{RC}, τ_{VM}	(6.92), (10.5)
τ_1, τ_2	s	fluorescence lifetime of laser levels $\|1\rangle$ and $\|2\rangle$	(10.5), (10.6)
τ_a	s	acoustic transit time	(8.109)
τ_c	s	photon lifetime, $\tau_c = 1/\gamma_c$; τ_{cl}, τ_{mnq}^c	(11.24)
τ_d	s	dielectric relaxation time	(14.75)
τ_e, τ_h	s	electron and hole lifetimes; τ_{e0}, τ_{h0}	(12.52), (12.53)
τ_r	s	relaxation time constant	(14.73)
τ_s	s	saturation lifetime or spontaneous carrier lifetime	(10.74), (12.56)
τ_{sp}	s	spontaneous radiative lifetime	(10.30)
τ_{tr}	s	transit time; τ_{tr}^e, τ_{tr}^h	(6.92)b, (14.102)
ϕ	rad	azimuthal angle, azimuthal angular coordinate	(1.122)
ϕ	V	work function potential, $e\phi$ = work function	(14.54)
Φ	s^{-1}	photon or electron flux; Φ_p, Φ_{out}, Φ_s^{in}, Φ_s^{out}	(10.111), (13.63)
Φ_l	lm	luminous flux	(13.66)
φ	rad	phase or phase shift; φ_0, φ_1, φ_2, φ_K, φ_{pk}, φ_{RT}, φ_χ	(1.60)
χ	none	complex electric susceptibility in the frequency domain	(1.54)
χ	V	electron affinity potential, $e\chi$ = electron affinity	(14.55)
χ_{eff}	$m\ V^{-1}$	effective second-order nonlinear susceptibility; χ_{eff}^I, χ_{eff}^{II}	(9.59)
χ_Q	$m\ V^{-1}$	χ_{eff} for quasi-phase matching	(9.97)
χ_{res}	none	resonant electric susceptibility, $\chi_{res} = \chi'_{res} + i\chi''_{res}$	(10.52), (11.50)f
χ_R	$m^2\ V^{-2}$	effective Raman susceptibility	(9.75)
χ_x, χ_y, χ_z	none	principal dielectric susceptibilities	(1.110)f
χ', χ''	none	real and imaginary parts of χ, $\chi = \chi' + i\chi''$	(1.176)
$\boldsymbol{\chi}(\mathbf{r}, t)$	$m^{-3}\ s^{-1}$	real electric susceptibility tensor in the time domain	(1.15)

Symbol	Unit	Meaning; derivatives	References
$\boldsymbol{\chi}(\omega)$, χ_{ij}	none	complex electric susceptibility tensor in the frequency domain	(1.54)
$\boldsymbol{\chi}^{(2)}$, $\chi^{(2)}_{ijk}$	m V^{-1}	second-order nonlinear susceptibility in the frequency domain	(9.19), (9.21)
$\boldsymbol{\chi}^{(3)}$, $\chi^{(3)}_{ijkl}$	m^2 V^{-2}	third-order nonlinear susceptibility in the frequency domain	(9.20), (9.22)
$\boldsymbol{\chi}_{\mathrm{m}}$	none	magnetic susceptibility tensor	(7.1)
ψ	rad	spatial phase of mode field distribution	(2.55)
ψ_{e}	rad	angle between $\mathbf{S}_{\mathrm{e}}$ and optical axis of crystal	(1.131)
ω	rad s^{-1}	optical angular frequency, $\omega = 2\pi\nu$; ω_{p}, ω_q, ω_{mnq}	(1.47)
ω_0	rad s^{-1}	central or carrier optical frequency, $\omega_0 = 2\pi\nu_0$	(1.162), (10.12)
ω_{B}	rad s^{-1}	Bragg frequency, $\omega_{\mathrm{B}} = 2\pi\nu_{\mathrm{B}}$	(5.24)b
ω_{c}	rad s^{-1}	cutoff frequency; ω^{c}_m	(2.74)
Ω	rad s^{-1}	microwave or acoustic angular frequency, $\Omega = 2\pi f$	(6.39), (8.1)
Ω_{B}, Ω_{R}	rad s^{-1}	Brillouin and Raman frequencies, $\Omega_{\mathrm{B}} = 2\pi f_{\mathrm{B}}$, $\Omega_{\mathrm{R}} = 2\pi f_{\mathrm{R}}$	(9.184), (9.182)
Ω_{esc}	rad	escape angle	(13.68)
Ω_{r}	rad s^{-1}	relaxation resonance frequency, $\Omega_{\mathrm{r}} = 2\pi f_{\mathrm{r}}$	(13.133)

Abbreviations

ABD	areal bit density
ADP	ammonium dihydrogen phosphate, $NH_4H_2PO_4$
APD	avalanche photodiode
AS	absorbing substrate
ASE	amplified spontaneous emission
BBO	beta barium borate, β-BaB_2O_4
BGO	bismuth germanate, $Bi_{12}GeO_{20}$
BSO	bismuth silicate, $Bi_{12}SiO_{20}$
CSP	channeled-substrate planar
CW	continuous wave
dB	decibel
dBm	decibel for power measured in milliwatts
dBμ	decibel for power measured in microwatts
dBn	decibel for power measured in nanowatts
DBR	distributed Bragg reflector
DC	direct current
DC-PBH	double-channel planar buried heterostructure
DFB	distributed feedback
DFG	difference-frequency generation
DH	double heterostructure
EDFA	erbium-doped fiber amplifier
EH	electric and magnetic, hybrid true fiber modes
ESA	excited-state absorption
FCSEL	folded-cavity surface-emitting laser
FP	Fabry–Perot
FWHM	full width at half maximum
GCSEL	grating-coupled surface-emitting laser
GGG	gadolinium gallium garnet, $Gd_3Ga_5O_{12}$
GRIN-SCH	graded-index separate confinement heterostructure
HE	magnetic and electric, hybrid true fiber modes

IDT	interdigital transducer
IR	infrared
KDP	potassium dihydrogen phosphate, KH_2PO_4
KTA	potassium titanyl arsenate, $KTiOAsO_4$
KTP	potassium titanyl phosphate, $KTiOPO_4$
LBO	lithium triborate, LiB_3O_5
LD	laser diode
LED	light-emitting diode
LiSAF	lithium strontium aluminium fluoride, $LiSrAlF_6$
LN	lithium niobate, $LiNbO_3$
LP	linearly polarized, approximate fiber modes
MQW	multiple quantum wells
MSM	metal–semiconductor–metal
NA	numerical aperture
NDFA	neodymium-doped fiber amplifier
NEA	negative electron affinity
NEP	noise equivalent power
OPA	optical parametric amplifier
OPG	optical parametric generator
OPO	optical parametric oscillator
PBH	planar buried heterostructure
PDFA	praseodymium-doped fiber amplifier
PMT	photomultiplier tube
PPKTP	periodically poled KTP
PPLN	periodically poled $LiNbO_3$
QW	quantum well
RC	resistance–capacitance
RE–TM	rare-earth transition-metal
RF	radio frequency
SAM	separate absorption and multiplication
SAW	surface acoustic wave
SBS	stimulated Brillouin scattering
SFG	sum-frequency generation
SI	international system of units
SH	single heterostructure
SHG	second-harmonic generation
SNR	signal-to-noise ratio
SPM	self-phase modulation
SQW	single quantum well
SRS	stimulated Raman scattering
TE	transverse electric

TEM	transverse electric and magnetic
TGG	terbium gallium garnet, $Tb_3Ga_5O_{12}$
THG	third-harmonic generation
TM	transverse magnetic
TPA	two-photon absorption
TS	transparent substrate
TWPD	traveling-wave photodiode
UV	ultraviolet
VCSEL	vertical-cavity surface-emitting laser
VIPD	vertically illuminated photodiode
VMDP	velocity-matched distributed photodetector
WGPD	waveguide photodetector
WKB	Wentzel–Kramers–Brillouin
XPM	cross-phase modulation
YAG	yttrium aluminum garnet, $Y_3Al_5O_{12}$
YIG	yttrium iron garnet, $Y_3Fe_5O_{12}$
YLF	yttrium lithium fluoride, $YLiF_4$

Part I

Background

1 General background

Photonics is an engineering discipline concerning the control of light, or photons, for useful applications, much as electronics has to do with electrons. Light is electromagnetic radiation of frequencies in the range from 1 THz to 10 PHz, corresponding to wavelengths between ~300 μm and ~30 nm in free space. This optical spectral range is generally divided into infrared, visible, and ultraviolet regions, as indicated in Table 1.1. The spectral range of concern in photonics is usually in a wavelength range between 10 μm and 100 nm. The primary interest in the applications of photonic devices is in an even narrower range of visible and near infrared wavelengths. As we shall see later, this spectral range of application is largely determined by the properties of materials used for photonic devices.

The wave nature of light is very important in the function of photonic devices. In particular, the propagation of light in a photonic device is completely characterized by its wave nature. However, in the spectral range of interest for practical photonic devices, the quantum energies of photons are in a range where the quantum nature of light is also important. For example, photons of visible light have energies between 1.7 and 3.1 eV, which are in the range of the bandgaps of most semiconductors. Photon energy is an important factor that determines the behavior of an optical wave in a semiconductor photonic device. The uniqueness of photonic devices is that both wave and quantum characteristics of light have to be considered for the function and applications of these devices. Generally speaking, the photon nature of light is important in the operation of photonic devices for generation, amplification, frequency conversion, or detection of light, while the wave nature is important in the operation of all photonic devices but is particularly so for devices used in transmission, modulation, or switching of light. In this chapter, we review some relevant wave and quantum properties of light as a general background for later chapters.

1.1 Optical fields and Maxwell's equations

When dealing with photonic devices, we consider in most situations optical fields in media of various electromagnetic properties. The electromagnetic field in a medium is

Table 1.1 *Electromagnetic spectrum*

Wave region	Frequency	Wavelength	Devices
Radio	kHz–MHz–GHz	km–m–cm	Electronic devices
Microwave	1 GHz–1 THz	300 mm–300 μm	Microwave devices
Optical			
Infrared	1 THz–430 THz	300 μm–700 nm	Photonic devices
Visible	430 THz–750 THz	700 nm–400 nm	
Ultraviolet	750 THz–10 PHz	400 nm–30 nm	
X-ray	10 PHz–10 EHz	30 nm–300 pm	
Gamma ray	10 EHz and above	300 pm and shorter	

generally characterized by the following four field quantities:

electric field $\boldsymbol{E}(\mathbf{r}, t)$ V m^{-1},
electric displacement $\boldsymbol{D}(\mathbf{r}, t)$ C m^{-2},
magnetic field $\boldsymbol{H}(\mathbf{r}, t)$ A m^{-1},
magnetic induction $\boldsymbol{B}(\mathbf{r}, t)$ T or Wb m^{-2}.

Note that $\boldsymbol{E}$ and $\boldsymbol{B}$ are fundamental microscopic fields, while $\boldsymbol{D}$ and $\boldsymbol{H}$ are macroscopic fields that include the response of the medium. The units given above and below for the field quantities are SI units consistent with the SI system used in this book for Maxwell's equations. Experimentally measured magnetic field quantities are sometimes given in Gaussian units, which are gauss for the $\boldsymbol{B}$ field and oersted (Oe) for the $\boldsymbol{H}$ field. The conversion relations between SI and Gaussian units are 1 T = 1 Wb m^{-2} = 10^4 gauss for $\boldsymbol{B}$ and 1 A m^{-1} = $4\pi \times 10^{-3}$ Oe for $\boldsymbol{H}$.

The response of a medium to an electromagnetic field generates the *polarization* and the *magnetization*:

polarization (electric polarization) $\boldsymbol{P}(\mathbf{r}, t)$ C m^{-2},
magnetization (magnetic polarization) $\boldsymbol{M}(\mathbf{r}, t)$ A m^{-1}.

They are connected to the field quantities through the following relations:

$$\boldsymbol{D}(\mathbf{r}, t) = \epsilon_0 \boldsymbol{E}(\mathbf{r}, t) + \boldsymbol{P}(\mathbf{r}, t) \tag{1.1}$$

and

$$\boldsymbol{B}(\mathbf{r}, t) = \mu_0 \boldsymbol{H}(\mathbf{r}, t) + \mu_0 \boldsymbol{M}(\mathbf{r}, t), \tag{1.2}$$

where

$$\epsilon_0 \approx \frac{1}{36\pi} \times 10^{-9} \text{ F m}^{-1} \text{ or A s V}^{-1} \text{ m}^{-1} \tag{1.3}$$

is the *electric permittivity* of free space and

$$\mu_0 = 4\pi \times 10^{-7} \text{ H m}^{-1} \text{ or V s A}^{-1} \text{ m}^{-1} \tag{1.4}$$

is the *magnetic permeability* of free space. In addition, independent charge or current sources may exist:

charge density $\rho(\mathbf{r}, t)$ C m^{-3},
current density $\boldsymbol{J}(\mathbf{r}, t)$ A m^{-2}.

In a medium, the behavior of a time-varying electromagnetic field is governed by the following space- and time-dependent macroscopic *Maxwell's equations*:

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \qquad \text{Faraday's law,} \tag{1.5}$$

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \qquad \text{Ampere's law,} \tag{1.6}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = \rho \qquad \text{Coulomb's law,} \tag{1.7}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \qquad \text{absence of magnetic monopoles.} \tag{1.8}$$

The current and charge densities are constrained by the following *continuity equation*:

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} + \frac{\partial \rho}{\partial t} = 0 \qquad \text{conservation of charge.} \tag{1.9}$$

In a medium free of sources, $\boldsymbol{J} = 0$ and $\rho = 0$. Then, Maxwell's equations are simply

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{1.10}$$

$$\boldsymbol{\nabla} \times \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t}, \tag{1.11}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{D} = 0, \tag{1.12}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0. \tag{1.13}$$

These are the equations normally used for optical fields because optical fields are usually not generated directly by free currents or free charges.

Transformation properties

Maxwell's equations and the continuity equation are the basic physical laws that govern the behavior of electromagnetic fields. They are invariant under the transformation of *space inversion*, in which the spatial vector $\mathbf{r}$ is changed to $\mathbf{r}' = -\mathbf{r}$, or $(x, y, z) \rightarrow (-x, -y, -z)$, and the transformation of *time reversal*, in which the time variable t is changed to $t' = -t$, or $t \rightarrow -t$. This means that the form of these equations is not changed when we perform the space-inversion transformation or the time-reversal transformation, or both together.

Different quantities in Maxwell's equations have different transformation properties. An understanding of these properties is important and leads to a fundamental appreciation of the difference between the characteristics of the electric and magnetic fields, which is the origin of the difference between the electric and magnetic symmetry properties of materials. It also helps in understanding many basic characteristics of the electro-optic, magneto-optic, and nonlinear optical properties of materials to be addressed in later chapters.

The electric field vectors, $\boldsymbol{E}$ and $\boldsymbol{D}$, have the same transformation properties as those of $\boldsymbol{P}$, while the transformation properties of the magnetic field vectors, $\boldsymbol{H}$ and $\boldsymbol{B}$, are the same as those of $\boldsymbol{M}$. The origin of the electric properties of a material is the charge-density distribution, $\rho(\mathbf{r}, t)$, at the atomic level in the material, whereas that of the magnetic properties stems from the current-density distribution, $\boldsymbol{J}(\mathbf{r}, t)$. The transformation properties of the scalar quantity ρ are such that *the sign of ρ remains unchanged under the transformation of either space inversion or time reversal*. In contrast, $\boldsymbol{J}$ is a *polar vector* because it is charge density times velocity, $\rho\mathbf{v}$, where velocity, $\mathbf{v}$, is a polar vector. Thus, *the vector $\boldsymbol{J}$ changes sign under the transformation of either space inversion or time reversal*. It changes sign under space inversion because a polar vector changes sign under space inversion, and it changes sign under time reversal because $\mathbf{v}$ is the first time derivative of $\mathbf{r}$. The electric polarization $\boldsymbol{P}$ is a polar vector because it is the volume average of the electric dipole moment density defined by $\rho(\mathbf{r}, t)\mathbf{r}$, and the product of a scalar quantity ρ and a polar vector $\mathbf{r}$ is a polar vector. In contrast, magnetization $\boldsymbol{M}$ is an *axial vector* because it is the volume average of the magnetic dipole moment density defined by $\mathbf{r} \times \boldsymbol{J}(\mathbf{r}, t)$, and the cross product of two polar vectors, $\mathbf{r}$ and $\boldsymbol{J}$, is an axial vector. Therefore, we find the following transformation properties.

1. **Electric fields.** The electric field vectors, $\boldsymbol{P}$, $\boldsymbol{E}$, and $\boldsymbol{D}$, change sign under space inversion but not under time reversal.
2. **Magnetic fields.** The magnetic field vectors, $\boldsymbol{M}$, $\boldsymbol{H}$, and $\boldsymbol{B}$ change sign under time reversal but not under space inversion.

With these transformation properties understood, the invariance of Maxwell's equations and the continuity equation under the transformation of space inversion or time reversal or both can be easily verified.

Response of medium

Polarization and magnetization in a medium are generated, respectively, by the response of the medium to the electric and magnetic fields. Therefore, $\boldsymbol{P}(\mathbf{r}, t)$ depends on $\boldsymbol{E}(\mathbf{r}, t)$, while $\boldsymbol{M}(\mathbf{r}, t)$ depends on $\boldsymbol{B}(\mathbf{r}, t)$. *At optical frequencies, the magnetization vanishes*, $\boldsymbol{M} = 0$. Consequently, for optical fields, the following relation is always true:

$$\boldsymbol{B}(\mathbf{r}, t) = \mu_0 \boldsymbol{H}(\mathbf{r}, t). \tag{1.14}$$

This is not true at low frequencies, however. It is possible to change the properties of a medium through a magnetization induced by a DC or low-frequency magnetic field, leading to the functioning of magneto-optic devices. It should be noted that even for magneto-optic devices, magnetization is induced by a DC or low-frequency magnetic field that is separate from the optical fields. No magnetization is induced by the magnetic components of the optical fields.

Except for magneto-optic devices, most photonic devices are made of dielectric materials that have zero magnetization at all frequencies. The optical properties of such materials are completely determined by the relation between $\boldsymbol{P}(\mathbf{r}, t)$ and $\boldsymbol{E}(\mathbf{r}, t)$. This relation is generally characterized by an *electric susceptibility tensor*, $\boldsymbol{\chi}$, through the following definition for electric polarization:

$$\boldsymbol{P}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}(\mathbf{r} - \mathbf{r}', t - t') \cdot \boldsymbol{E}(\mathbf{r}', t'). \tag{1.15}$$

From (1.1), then

$$\begin{aligned} \boldsymbol{D}(\mathbf{r}, t) &= \epsilon_0 \boldsymbol{E}(\mathbf{r}, t) + \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}(\mathbf{r} - \mathbf{r}', t - t') \cdot \boldsymbol{E}(\mathbf{r}', t') \\ &= \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\epsilon}(\mathbf{r} - \mathbf{r}', t - t') \cdot \boldsymbol{E}(\mathbf{r}', t'), \end{aligned} \tag{1.16}$$

where $\boldsymbol{\epsilon}$ is the *electric permittivity tensor* of the medium.

Because $\boldsymbol{\chi}$ and, equivalently, $\boldsymbol{\epsilon}$ represent the response of a medium to the optical field and thus completely characterize the macroscopic electromagnetic properties of the medium, (1.15) and (1.16) can be regarded as the definitions of $\boldsymbol{P}(\mathbf{r}, t)$ and $\boldsymbol{D}(\mathbf{r}, t)$, respectively. A few remarks can be made:

1. Both $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ are generally tensors because the vectors $\boldsymbol{P}$ and $\boldsymbol{D}$ are, in general, not parallel to vector $\boldsymbol{E}$ due to material *anisotropy*. In the case of an *isotropic* medium, both $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ can be reduced to scalars χ and ϵ, respectively.
2. The relations in (1.15) and (1.16) are in the form of convolution integrals. The convolution in time accounts for the fact that the response of a medium to excitation of an electric field is generally not *instantaneous* or *local* in time and will not vanish for some time after the excitation is over. Because time is unidirectional, *causality* exists in physical processes. An earlier excitation can have an effect on the property of a medium at a later time, but not a later excitation on the property of the medium at an earlier time. Therefore, the upper limit in the time integral is t, not infinity. In contrast, the convolution in space accounts for the *spatial nonlocality* of the material response. Excitation of a medium at a location $\mathbf{r}'$ can result in a change

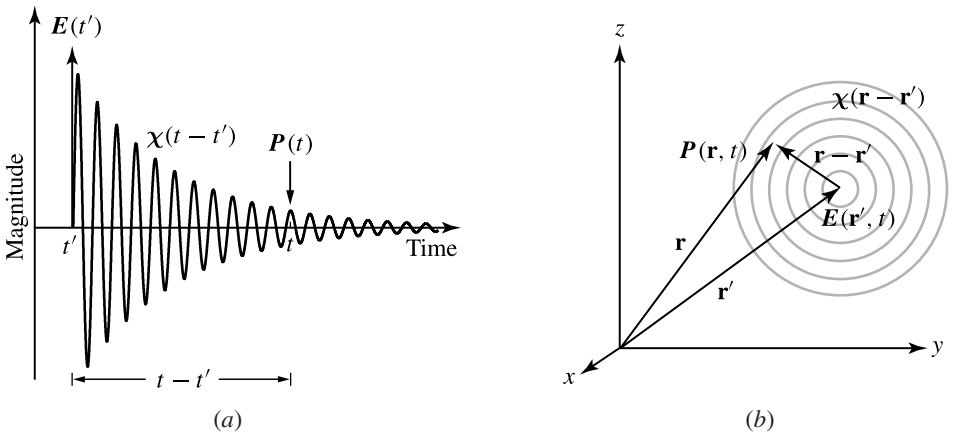


Figure 1.1 Nonlocal responses in (*a*) time and (*b*) space.

in the property of the medium at another location $\mathbf{r}$. For example, the property of a semiconductor at one location can be changed by electric or optical excitation at another location through carrier diffusion. Because space is not unidirectional, there is no spatial causality, in general, and spatial convolution is integrated over the entire space. Figure 1.1 shows the temporal and spatial nonlocality of responses to electromagnetic excitations. The temporal nonlocality of the optical response of a medium results in *frequency dispersion* of its optical property, while the spatial nonlocality results in *momentum dispersion*.

3. In addition to the dependence on space and time through the convolution relation with the optical field, χ and ϵ can also be functions of space or time independent of the optical field because of spatial or temporal *inhomogeneities* in the medium. Spatial inhomogeneity exists in all optical structures, such as optical waveguides, where the index of refraction is a function of space. Temporal inhomogeneity exists when the optical property of a medium varies with time, for example, because of modulation by a low-frequency electric field or by an acoustic wave.
4. In a linear medium, χ and ϵ do not depend on the optical field $\boldsymbol{E}$. In a nonlinear optical material, χ and ϵ are themselves also functions of $\boldsymbol{E}$.

Boundary conditions

At the interface of two media of different optical properties as shown in Fig. 1.2, the optical field components must satisfy certain boundary conditions. These boundary conditions can be derived from Maxwell's equations given in (1.10)–(1.13). From (1.10) and (1.11), the tangential components of the fields at the boundary satisfy

$$\hat{n} \times \boldsymbol{E}_1 = \hat{n} \times \boldsymbol{E}_2 \tag{1.17}$$

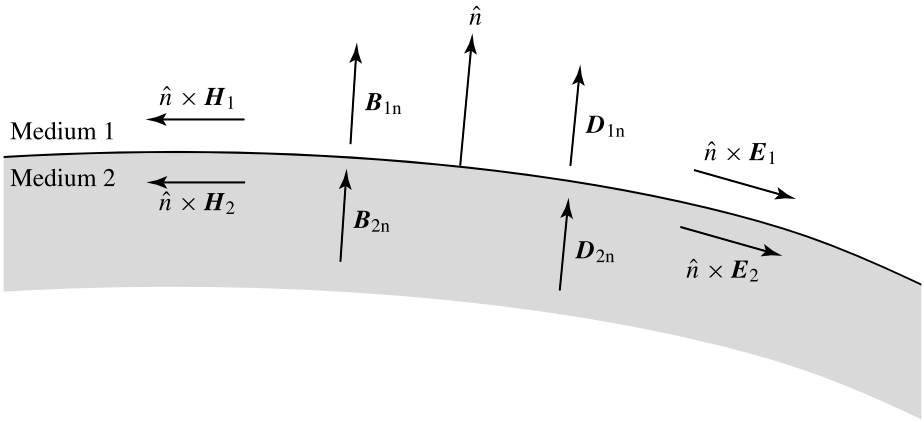


Figure 1.2 Boundary between two media of different optical properties.

and

$$\hat{n} \times \boldsymbol{H}_1 = \hat{n} \times \boldsymbol{H}_2, \tag{1.18}$$

where $\hat{n}$ is the unit vector normal to the interface as shown in Fig. 1.2. From (1.12) and (1.13), we have

$$\hat{n} \cdot \boldsymbol{D}_1 = \hat{n} \cdot \boldsymbol{D}_2 \tag{1.19}$$

and

$$\hat{n} \cdot \boldsymbol{B}_1 = \hat{n} \cdot \boldsymbol{B}_2 \tag{1.20}$$

for the normal components of the fields.

The tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$ must be continuous across an interface, while the normal components of $\boldsymbol{D}$ and $\boldsymbol{B}$ are continuous. Because $\boldsymbol{B} = \mu_0 \boldsymbol{H}$ for optical fields, as discussed above, (1.18) and (1.20) also imply that the tangential component of $\boldsymbol{B}$ and the normal component of $\boldsymbol{H}$ are also continuous. Consequently, *all of the magnetic field components in an optical field are continuous across a boundary. Possible discontinuities in an optical field exist only in the normal component of* $\boldsymbol{E}$ *or the tangential component of* $\boldsymbol{D}$.

Optical power and energy

By multiplying $\boldsymbol{E}$ by (1.6) and multiplying $\boldsymbol{H}$ by (1.5), we obtain

$$\boldsymbol{E} \cdot (\nabla \times \boldsymbol{H}) = \boldsymbol{E} \cdot \boldsymbol{J} + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t}, \tag{1.21}$$

$$\boldsymbol{H} \cdot (\nabla \times \boldsymbol{E}) = -\boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t}. \tag{1.22}$$

Using the vector identity

$$\mathbf{B} \cdot (\boldsymbol{\nabla} \times \mathbf{A}) - \mathbf{A} \cdot (\boldsymbol{\nabla} \times \mathbf{B}) = \boldsymbol{\nabla} \cdot (\mathbf{A} \times \mathbf{B}), \tag{1.23}$$

we can combine (1.21) and (1.22) to have

$$-\boldsymbol{\nabla} \cdot (\boldsymbol{E} \times \boldsymbol{H}) = \boldsymbol{E} \cdot \boldsymbol{J} + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t}. \tag{1.24}$$

Using (1.1) and (1.2) and rearranging (1.24), we obtain

$$\boldsymbol{E} \cdot \boldsymbol{J} = -\boldsymbol{\nabla} \cdot (\boldsymbol{E} \times \boldsymbol{H}) - \frac{\partial}{\partial t}\left(\frac{\epsilon_0}{2}|\boldsymbol{E}|^2 + \frac{\mu_0}{2}|\boldsymbol{H}|^2\right) - \left(\boldsymbol{E} \cdot \frac{\partial \boldsymbol{P}}{\partial t} + \mu_0 \boldsymbol{H} \cdot \frac{\partial \boldsymbol{M}}{\partial t}\right). \tag{1.25}$$

Recall that power in an electric circuit is given by voltage times current and has the unit of W = V A (watts = volts × amperes). In an electromagnetic field, we find similarly that $\boldsymbol{E} \cdot \boldsymbol{J}$ is the power density that has the unit of V A m^{-3} or W m^{-3}. Therefore, the total power dissipated by an electromagnetic field in a volume $\mathcal{V}$ is just

$$\int_{\mathcal{V}} \boldsymbol{E} \cdot \boldsymbol{J}\, \mathrm{d}\mathcal{V}. \tag{1.26}$$

Expressing (1.25) in an integral form, we have

$$\begin{aligned}\int_{\mathcal{V}} \boldsymbol{E} \cdot \boldsymbol{J}\, \mathrm{d}\mathcal{V} = &- \oint_{\mathcal{A}} \boldsymbol{E} \times \boldsymbol{H} \cdot \hat{n}\, \mathrm{d}\mathcal{A} - \frac{\partial}{\partial t}\int_{\mathcal{V}} \left(\frac{\epsilon_0}{2}|\boldsymbol{E}|^2 + \frac{\mu_0}{2}|\boldsymbol{H}|^2\right) \mathrm{d}\mathcal{V} \\ &- \int_{\mathcal{V}} \left(\boldsymbol{E} \cdot \frac{\partial \boldsymbol{P}}{\partial t} + \mu_0 \boldsymbol{H} \cdot \frac{\partial \boldsymbol{M}}{\partial t}\right) \mathrm{d}\mathcal{V},\end{aligned} \tag{1.27}$$

where the first term on the right-hand side is a surface integral over the closed surface $\mathcal{A}$ of volume $\mathcal{V}$ and $\hat{n}$ is the outward-pointing unit normal vector of the surface, as shown in Fig. 1.3.

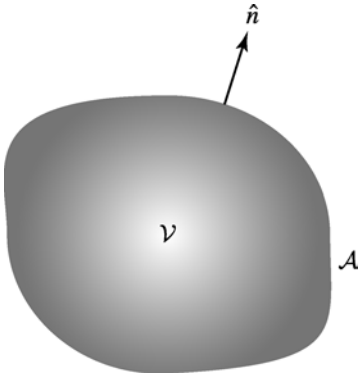


Figure 1.3 Boundary surface enclosing a volume element and the unit surface normal vector.

Clearly, each term in (1.27) has the unit of power. Each has an important physical meaning. The vector quantity

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{1.28}$$

is called the *Poynting vector* of the electromagnetic field. It represents the *instantaneous magnitude and direction of the power flow* of the field. The scalar quantity

$$u_0 = \frac{\epsilon_0}{2}|\boldsymbol{E}|^2 + \frac{\mu_0}{2}|\boldsymbol{H}|^2 \tag{1.29}$$

has the unit of energy per unit volume and is the *energy density stored in the propagating field*. It consists of two components, thus accounting for energies stored in both electric and magnetic fields at any instant of time. The last term in (1.27) also has two components associated with electric and magnetic fields, respectively. The quantity

$$W_{\mathrm{p}} = \boldsymbol{E} \cdot \frac{\partial \boldsymbol{P}}{\partial t} \tag{1.30}$$

is the *power density expended by the electromagnetic field on the polarization*. It is the rate of energy transfer from the electromagnetic field to the medium by inducing electric polarization in the medium. Similarly, the quantity

$$W_{\mathrm{m}} = \mu_0 \boldsymbol{H} \cdot \frac{\partial \boldsymbol{M}}{\partial t} \tag{1.31}$$

is the *power density expended by the electromagnetic field on the magnetization.* With these physical meanings attached to these terms, it can be seen that (1.27) simply states the law of conservation of energy in any arbitrary volume element $\mathcal{V}$ in the medium. The total energy in the medium equals that in the propagating field plus that in the electric and magnetic polarizations.

In the special case of a linear, nondispersive medium where $\boldsymbol{\epsilon}(\mathbf{r} - \mathbf{r}', t - t') = \boldsymbol{\epsilon}\delta(\mathbf{r} - \mathbf{r}')\delta(t - t')$, (1.16) simply reduces to $\boldsymbol{D}(\mathbf{r}, t) = \boldsymbol{\epsilon} \cdot \boldsymbol{E}(\mathbf{r}, t)$. Then, instead of (1.25), we have

$$\boldsymbol{E} \cdot \boldsymbol{J} = -\boldsymbol{\nabla} \cdot \boldsymbol{S} - \frac{\partial}{\partial t}\left(\frac{1}{2}\boldsymbol{E} \cdot \boldsymbol{D} + \frac{1}{2}\boldsymbol{H} \cdot \boldsymbol{B}\right) \tag{1.32}$$

from (1.24). In this situation, the total energy density stored in the medium, including that in the propagating field and that in the polarizations, is simply

$$u = \frac{1}{2}\boldsymbol{E} \cdot \boldsymbol{D} + \frac{1}{2}\boldsymbol{H} \cdot \boldsymbol{B}. \tag{1.33}$$

For an optical field, $\boldsymbol{J} = 0$ and $\boldsymbol{M} = 0$, as is discussed above. Then, (1.27) becomes

$$-\oint_{\mathcal{A}} \boldsymbol{S} \cdot \hat{n}\mathrm{d}\mathcal{A} = \frac{\partial}{\partial t}\int_{\mathcal{V}} u_0 \mathrm{d}\mathcal{V} + \int_{\mathcal{V}} W_{\mathrm{p}} \mathrm{d}\mathcal{V}, \tag{1.34}$$

which states that the total optical power flowing into volume $\mathcal{V}$ through its boundary surface $\mathcal{A}$ is equal to the rate of increase with time of the energy stored in the propagating fields in $\mathcal{V}$ plus the power transferred to the polarization of the medium in this volume. In a linear, nondispersive medium, we have

$$-\oint_{\mathcal{A}} \boldsymbol{S} \cdot \hat{n} \mathrm{d}\mathcal{A} = \frac{\partial}{\partial t} \int_{\mathcal{V}} u \mathrm{d}\mathcal{V}. \tag{1.35}$$

Wave equation

By applying $\boldsymbol{\nabla}\times$ to (1.10) and using (1.14) and (1.11), we have

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{E} + \mu_0 \frac{\partial^2 \boldsymbol{D}}{\partial t^2} = 0. \tag{1.36}$$

Using (1.1), (1.36) can be expressed as

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \boldsymbol{E} + \frac{1}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = -\mu_0 \frac{\partial^2 \boldsymbol{P}}{\partial t^2}, \tag{1.37}$$

where

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \approx 3 \times 10^8 \text{ m s}^{-1} \tag{1.38}$$

is the *speed of light* in free space. The *wave equation* in (1.37) describes the space-and-time evolution of the electric field of the optical wave. Its right-hand side can be regarded as the driving source for the optical wave. The polarization in a medium drives the evolution of an optical field. This wave equation can take on various forms depending on the characteristics of the medium, as will be seen on various occasions later. For now, we leave it in this general form.

1.2 Harmonic fields

Optical fields are harmonic fields that vary sinusoidally with time. The field vectors defined in the preceding section are all real quantities. For harmonic fields, it is always convenient to use *complex fields*. We define the space- and time-dependent complex electric field, $\mathbf{E}(\mathbf{r}, t)$, through its relation to the real electric field, $\boldsymbol{E}(\mathbf{r}, t)$:[1]

$$\boldsymbol{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \text{c.c.}, \tag{1.39}$$

[1] In some literature, the complex field is defined through a relation with the real field as $\boldsymbol{E}(\mathbf{r}, t) = 1/2(\mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t))$, which differs from the relation in (1.39) by the factor $1/2$. The magnitude of the complex field defined through this alternative relation is twice that of the complex field defined through (1.39). As a result, expressions for many quantities may be different under the two different definitions. An example is that of the optical intensity given in (1.98). We have chosen to define the complex field through the relation in (1.39) without the factor $1/2$ primarily because this definition is more convenient and less confusing in expressing the nonlinear polarizations discussed in Chapter 9.

where c.c. means the complex conjugate. In our convention, $\mathbf{E}(\mathbf{r}, t)$ contains the complex field components that vary with time as $\exp(-\mathrm{i}\omega t)$ with positive values of ω, while $\mathbf{E}^*(\mathbf{r}, t)$ contains those varying with time as $\exp(\mathrm{i}\omega t)$ with positive ω, or $\exp(-\mathrm{i}\omega t)$ with negative ω. The complex fields of other field quantities are similarly defined (see Appendix A).

With this definition for the complex fields, all of the linear field equations retain their forms. In particular, Maxwell's equations for the complex optical fields are

$$\boldsymbol{\nabla} \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{1.40}$$

$$\boldsymbol{\nabla} \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}, \tag{1.41}$$

$$\boldsymbol{\nabla} \cdot \mathbf{D} = 0, \tag{1.42}$$

$$\boldsymbol{\nabla} \cdot \mathbf{B} = 0. \tag{1.43}$$

The wave equation in terms of the complex electric field is

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \mathbf{E} + \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_0 \frac{\partial^2 \mathbf{P}}{\partial t^2}, \tag{1.44}$$

while

$$\mathbf{P}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}(\mathbf{r} - \mathbf{r}', t - t') \cdot \mathbf{E}(\mathbf{r}', t') \tag{1.45}$$

and

$$\begin{aligned} \mathbf{D}(\mathbf{r}, t) &= \epsilon_0 \mathbf{E}(\mathbf{r}, t) + \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}(\mathbf{r} - \mathbf{r}', t - t') \cdot \mathbf{E}(\mathbf{r}', t') \\ &= \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\epsilon}(\mathbf{r} - \mathbf{r}', t - t') \cdot \mathbf{E}(\mathbf{r}', t'). \end{aligned} \tag{1.46}$$

It is important to note that while $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{E}$ are complex, $\boldsymbol{\chi}(\mathbf{r} - \mathbf{r}', t - t')$ and $\boldsymbol{\epsilon}(\mathbf{r} - \mathbf{r}', t - t')$ in (1.45) and (1.46) are always real and are the same as those in (1.15) and (1.16).

For a harmonic optical field of wavevector $\mathbf{k}$ and angular frequency ω, its complex electric field can be further written as

$$\mathbf{E}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}(\mathbf{r}, t) \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t), \tag{1.47}$$

where $\boldsymbol{\mathcal{E}}(\mathbf{r}, t)$ is the space- and time-varying field envelope, such as that for a modulated field, a guided field, or an optical pulse. Other complex field quantities, such as $\mathbf{H}(\mathbf{r}, t)$, can be similarly expressed. The phase factor in (1.47) indicates the direction

of wave propagation:

$$\begin{aligned} &\mathrm{i}\mathbf{k}\cdot\mathbf{r}-\mathrm{i}\omega t, \quad \text{forward propagating in } \mathbf{k} \text{ direction;}\\ &-\mathrm{i}\mathbf{k}\cdot\mathbf{r}-\mathrm{i}\omega t, \quad \text{backward propagating in } -\mathbf{k} \text{ direction.}\end{aligned}$$

The light *intensity*, or *irradiance*, is the *power density* of the harmonic optical field. It can be calculated by time averaging of the Poynting vector over one wave cycle:

$$\overline{\boldsymbol{S}} = \frac{1}{T}\int_0^T \boldsymbol{E}\times\boldsymbol{H}\mathrm{d}t = 2\,\mathrm{Re}(\mathbf{E}\times\mathbf{H}^*), \tag{1.48}$$

where $\mathrm{Re}(\cdot)$ means taking the real part. We can define a *complex Poynting vector*:

$$\mathbf{S} = \mathbf{E}\times\mathbf{H}^* \tag{1.49}$$

so that

$$\overline{\boldsymbol{S}} = \mathbf{S}+\mathbf{S}^*, \tag{1.50}$$

which has the same form as the relation between the real and complex fields defined in (1.39) except that the real Poynting vector in this relation is time averaged. The light intensity, I, is simply the magnitude of the *real* time-averaged Poynting vector:

$$I = |\overline{\boldsymbol{S}}| = |\mathbf{S}+\mathbf{S}^*|, \tag{1.51}$$

where I is in watts per square meter.

For harmonic optical fields, it is often useful to consider the complex fields in the momentum space and frequency domain defined by the following Fourier-transform relations:

$$\mathbf{E}(\mathbf{k},\omega) = \int_{-\infty}^{\infty}\mathrm{d}\mathbf{r}\int_{-\infty}^{\infty}\mathrm{d}t\mathbf{E}(\mathbf{r},t)\exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{r}+\mathrm{i}\omega t), \quad \text{for } \omega>0, \tag{1.52}$$

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{(2\pi)^4}\int_{-\infty}^{\infty}\mathrm{d}\mathbf{k}\int_0^{\infty}\mathrm{d}\omega\mathbf{E}(\mathbf{k},\omega)\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}-\mathrm{i}\omega t). \tag{1.53}$$

Note that $\mathbf{E}(\mathbf{k},\omega)$ in (1.52) is defined for $\omega>0$ only, and the integral for the time dependence of $\mathbf{E}(\mathbf{r},t)$ in (1.53) extends only over positive values of ω. This is in accordance with the convention we used to define the complex field $\mathbf{E}(\mathbf{r},t)$ in (1.39). All other space- and time-dependent quantities, including other field vectors and the permittivity and susceptibility tensors, are transformed in a similar manner. Through the Fourier transform, the convolution integrals in real space and time become simple products in the momentum space and frequency domain. Consequently,

we have

$$\mathbf{P}(\mathbf{k}, \omega) = \epsilon_0 \boldsymbol{\chi}(\mathbf{k}, \omega) \cdot \mathbf{E}(\mathbf{k}, \omega) \tag{1.54}$$

and

$$\mathbf{D}(\mathbf{k}, \omega) = \boldsymbol{\epsilon}(\mathbf{k}, \omega) \cdot \mathbf{E}(\mathbf{k}, \omega). \tag{1.55}$$

1.3 Linear optical susceptibility

As mentioned above, the susceptibility tensor $\boldsymbol{\chi}(\mathbf{r}, t)$ and the permittivity tensor $\boldsymbol{\epsilon}(\mathbf{r}, t)$ of space and time are always real quantities although all field quantities, including both $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{k}, \omega)$, can be defined in a complex form. This is true even in the presence of an optical loss or gain in the medium. However, the susceptibility and permittivity tensors in the momentum space and frequency domain, $\boldsymbol{\chi}(\mathbf{k}, \omega)$ and $\boldsymbol{\epsilon}(\mathbf{k}, \omega)$, can be complex. If an eigenvalue, χ_i, of $\boldsymbol{\chi}$ is complex, the corresponding eigenvalue, ϵ_i, of $\boldsymbol{\epsilon}$ is also complex, and their imaginary parts have the same sign because $\boldsymbol{\epsilon} = \epsilon_0(1 + \boldsymbol{\chi})$. The signs of such imaginary parts of eigenvalues tell whether the medium has an optical gain or loss. In our convention, we write, for example, $\chi_i = \chi_i' + \mathrm{i}\chi_i''$ in the frequency domain. Then, $\chi_i''(\omega) > 0$ *corresponds to an optical loss or absorption*, while $\chi_i''(\omega) < 0$ *represents an optical gain or amplification*.

The fact that $\boldsymbol{\chi}(\mathbf{r}, t)$ and $\boldsymbol{\epsilon}(\mathbf{r}, t)$ are real quantities leads to the following symmetry relations for the tensor elements of $\boldsymbol{\chi}(\mathbf{k}, \omega)$ and $\boldsymbol{\epsilon}(\mathbf{k}, \omega)$:

$$\chi_{ij}^*(\mathbf{k}, \omega) = \chi_{ij}(-\mathbf{k}, -\omega) \tag{1.56}$$

and

$$\epsilon_{ij}^*(\mathbf{k}, \omega) = \epsilon_{ij}(-\mathbf{k}, -\omega), \tag{1.57}$$

which are called the *reality condition*. The reality condition implies that $\chi_{ij}'(\mathbf{k}, \omega) = \chi_{ij}'(-\mathbf{k}, -\omega)$ and $\chi_{ij}''(\mathbf{k}, \omega) = -\chi_{ij}''(-\mathbf{k}, -\omega)$. Similar relations also apply for the real and imaginary parts of ϵ_{ij}. Therefore, the real parts of χ_{ij} and ϵ_{ij} are even functions of $\mathbf{k}$ and ω, whereas the imaginary parts are odd functions of $\mathbf{k}$ and ω. Any constant contribution, independent of $\mathbf{k}$ and ω, in χ_{ij} and ϵ_{ij} is an even function of $\mathbf{k}$ and ω; hence it can appear only in the real parts. As a result, the imaginary parts, if they exist, are always functions of either $\mathbf{k}$ or ω, or both. The loss, or gain, in a medium is associated with the imaginary parts of the eigenvalues of $\boldsymbol{\chi}(\omega)$; consequently, it is inherently dispersive. Any other effects that can be described by the imaginary parts of the eigenvalues of $\boldsymbol{\chi}(\mathbf{k}, \omega)$ are also dispersive in either momentum or frequency, or both.

The momentum and frequency dependencies of an electric susceptibility, $\boldsymbol{\chi}(\mathbf{k}, \omega)$, are due to the spatial and temporal nonlocality properties of the underlying physical

mechanisms that contribute to χ. As discussed in the preceding section, spatial nonlocality causes spatially convoluted effects and results in momentum dependence of the susceptibility, and temporal nonlocality causes temporal convolution and results in frequency dispersion of the medium.

In addition to nonlocality, it is also important to consider inhomogeneity, in both space and time. *In a linear medium, changes in the wavevector of an optical wave, or coupling between waves of different wavevectors, can occur only if the optical property of the medium in which the wave propagates is spatially inhomogeneous such that* $\chi(\mathbf{k}, \omega)$ *is spatially dependent*. Likewise, *changes in the frequency of an optical wave, or coupling between waves of different frequencies, are possible in a linear medium only if the optical property of the medium is time varying such that* $\chi(\mathbf{k}, \omega)$ *varies with time.* Changes in the wavevector of an optical wave can take the form of changes in the wave propagation direction, as in reflection and diffraction, or in the optical wavelength, as in the case when a wave propagates from one part of the medium to another of different refractive index. Changes in the frequency of an optical wave result in the generation of other frequencies or the conversion of the optical wave to a completely different frequency. Consequently, for practical photonic devices, it is often necessary to consider both nonlocality and inhomogeneity in both space and time, thus writing $\chi(\mathbf{r}, t; \mathbf{k}, \omega)$ and, correspondingly, $\epsilon(\mathbf{r}, t; \mathbf{k}, \omega)$.

1.4 Polarization of light

Consider a monochromatic plane optical wave that has a complex field

$$\mathbf{E}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}} \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t) = \hat{e}\mathcal{E} \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t), \tag{1.58}$$

where $\boldsymbol{\mathcal{E}}$ is a constant independent of $\mathbf{r}$ and t, and $\hat{e}$ is its unit vector. The *polarization* of the optical field is characterized by the unit vector $\hat{e}$. The wave is *linearly polarized*, also called *plane polarized*, if $\hat{e}$ can be expressed as a constant, real vector. Otherwise, the wave is *elliptically polarized* in general, and is *circularly polarized* in some special cases. For the convenience of discussion, we take the direction of wave propagation to be the z direction so that $\mathbf{k} = k\hat{z}$ and assume that both $\mathbf{E}$ and $\mathbf{H}$ lie in the xy plane.[2] Then, we have

$$\boldsymbol{\mathcal{E}} = \hat{x}\mathcal{E}_x + \hat{y}\mathcal{E}_y = \hat{x}|\mathcal{E}_x|\mathrm{e}^{\mathrm{i}\varphi_x} + \hat{y}|\mathcal{E}_y|\mathrm{e}^{\mathrm{i}\varphi_y}, \tag{1.59}$$

where $\mathcal{E}_x$ and $\mathcal{E}_y$ are space- and time-independent complex amplitudes, with phases φ_x and φ_y, respectively.

[2] This assumption is generally true if the medium is isotropic. It is not necessarily true if the medium is anisotropic. Propagation and polarization in isotropic and anisotropic media are discussed in the following two sections. However, the general concept discussed here does not depend on the validity of this assumption.

The polarization of the wave depends only on the phase difference and the magnitude ratio between the two field components $\mathcal{E}_x$ and $\mathcal{E}_y$. It can be completely characterized by the following two parameters:

$$\varphi = \varphi_y - \varphi_x, \qquad -\pi < \varphi \leq \pi, \tag{1.60}$$

and

$$\alpha = \tan^{-1} \frac{|\mathcal{E}_y|}{|\mathcal{E}_x|}, \qquad 0 \leq \alpha \leq \frac{\pi}{2}. \tag{1.61}$$

Because only the relative phase φ matters, we can set $\varphi_x = 0$ and take $\mathcal{E}$ to be real in the following discussions. Then $\boldsymbol{\mathcal{E}}$ from (1.59) can be written as

$$\boldsymbol{\mathcal{E}} = \mathcal{E}\hat{e}, \quad \text{with} \quad \hat{e} = \hat{x}\cos\alpha + \hat{y}e^{i\varphi}\sin\alpha. \tag{1.62}$$

Using (1.39), the space- and time-dependent real field is

$$\boldsymbol{E}(z,t) = 2\mathcal{E}\left[\hat{x}\cos\alpha\cos(kz-\omega t) + \hat{y}\sin\alpha\cos(kz-\omega t+\varphi)\right]. \tag{1.63}$$

At a fixed z location, say $z = 0$, we see that the electric field varies with time as

$$\boldsymbol{E}(t) = 2\mathcal{E}\left[\hat{x}\cos\alpha\cos\omega t + \hat{y}\sin\alpha\cos(\omega t-\varphi)\right]. \tag{1.64}$$

In general, $\mathcal{E}_x$ and $\mathcal{E}_y$ have different phases and different magnitudes. Therefore, the values of φ and α can be any combination. At a fixed point in space, both the direction and the magnitude of the field vector $\boldsymbol{E}$ in (1.64) can vary with time. Except when the values of φ and α fall into one of the special cases discussed below, the tip of this vector generally describes an ellipse, and the wave is said to be elliptically polarized. Note that we have assumed that the wave propagates in the positive z direction. When we view the ellipse by facing *against* this direction of wave propagation, we see that the tip of the field vector rotates *counterclockwise*, or *left handedly*, if $\varphi > 0$, and *clockwise*, or *right handedly*, if $\varphi < 0$. Figure 1.4 shows the ellipse traced by the tip of the rotating field vector at a fixed point in space. Also shown in the figure are the relevant parameters that characterize elliptic polarization.

In the description of the polarization characteristics of an optical wave, it is sometimes convenient to use, in place of α and φ, a set of two other parameters, θ and ε, which specify the *orientation* and *ellipticity* of the ellipse, respectively. The orientational parameter θ is the directional angle measured from the x axis to the major axis of the ellipse. Its range is taken to be $0 \leq \theta < \pi$ for convenience. Ellipticity ε is defined as

$$\varepsilon = \pm\tan^{-1}\frac{b}{a}, \qquad -\frac{\pi}{4} \leq \varepsilon \leq \frac{\pi}{4}, \tag{1.65}$$

where a and b are the major and minor semiaxes, respectively, of the ellipse. The plus

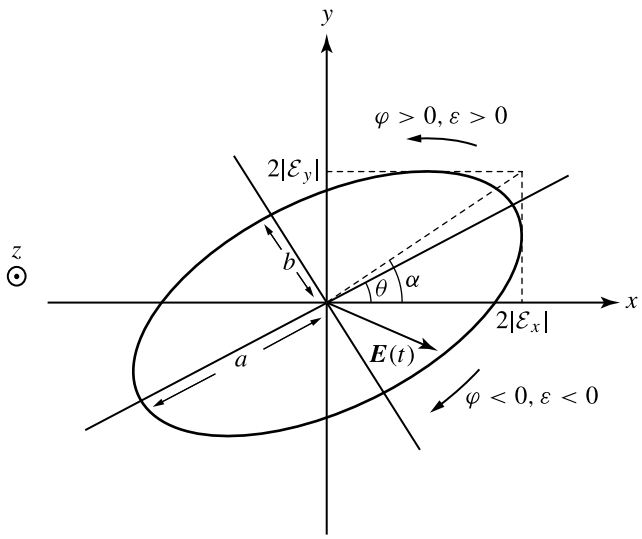


Figure 1.4 Ellipse described by the tip of the field of an elliptically polarized optical wave at a fixed point in space. Also shown are relevant parameters characterizing the state of polarization. The propagation direction is assumed to be the positive z direction, and the ellipse is viewed by facing against this direction.

sign for $\varepsilon > 0$ is taken to correspond to $\varphi > 0$ for left-handed polarization, whereas the minus sign for $\varepsilon < 0$ is taken to correspond to $\varphi < 0$ for right-handed polarization. The two sets of parameters (α, φ) and (θ, ε) have the following relations:

$$\tan 2\theta = \tan 2\alpha \cos\varphi, \tag{1.66}$$

$$\sin 2\varepsilon = \sin 2\alpha \sin\varphi. \tag{1.67}$$

Either set is sufficient to characterize the polarization state of an optical wave completely.

The following special cases are of particular interest.

1. **Linear polarization.** This happens when $\varphi = 0$ or π for any value of α. It is also characterized by $\varepsilon = 0$, and $\theta = \alpha$, if $\varphi = 0$, or $\theta = \pi - \alpha$, if $\varphi = \pi$. Clearly, the ratio $\mathcal{E}_x/\mathcal{E}_y$ is real in this case; therefore, linear polarization is described by a constant, real unit vector as

   $$\hat{e} = \hat{x}\cos\theta + \hat{y}\sin\theta. \tag{1.68}$$

   It follows that $\boldsymbol{E}(t)$ described by (1.64) reduces to

   $$\boldsymbol{E}(t) = 2\mathcal{E}\hat{e}\cos\omega t, \tag{1.69}$$

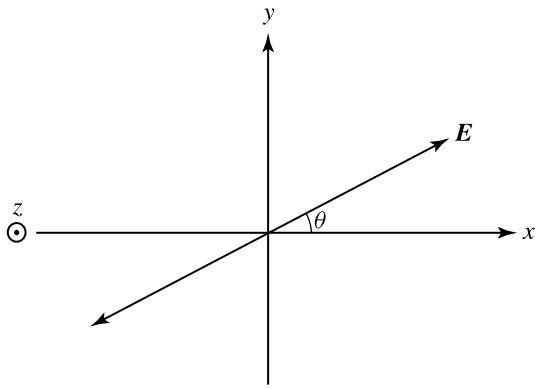


Figure 1.5 Field of a linearly polarized optical wave.

The tip of this vector traces a line in space at an angle θ with respect to the x axis, as shown in Fig. 1.5.

2. **Circular polarization.** This happens when $\varphi = \pi/2$ or $-\pi/2$, and $\alpha = \pi/4$. It is also characterized by $\varepsilon = \pi/4$ or $-\pi/4$, and $\theta = 0$. Because $\alpha = \pi/4$, we have $|\mathcal{E}_x| = |\mathcal{E}_y| = \mathcal{E}/\sqrt{2}$. There are two different circular polarization states:

   a. **Left-circular polarization.** For $\varphi = \pi/2$, also $\varepsilon = \pi/4$, the wave is *left-circularly polarized* if it propagates in the positive z direction. The complex field amplitude in (1.62) becomes

   $$\boldsymbol{\mathcal{E}} = \mathcal{E}\frac{\hat{x} + \mathrm{i}\hat{y}}{\sqrt{2}} = \mathcal{E}\hat{e}_+, \tag{1.70}$$

   and $\boldsymbol{E}(t)$ described by (1.64) reduces to

   $$\boldsymbol{E}(t) = \sqrt{2}\mathcal{E}(\hat{x}\cos\omega t + \hat{y}\sin\omega t). \tag{1.71}$$

   As we view against the direction of propagation $\hat{z}$, we see that the field vector $\boldsymbol{E}(t)$ rotates *counterclockwise* with an angular frequency ω. The tip of this vector describes a circle. This is shown in Fig. 1.6(a). This left-circular polarization is also called *positive helicity*. Its eigenvector is

   $$\hat{e}_+ \equiv \frac{\hat{x} + \mathrm{i}\hat{y}}{\sqrt{2}}. \tag{1.72}$$

   b. **Right-circular polarization**. For $\varphi = -\pi/2$, also $\varepsilon = -\pi/4$, the wave is *right-circularly polarized* if it propagates in the positive z direction. We then have

   $$\boldsymbol{\mathcal{E}} = \mathcal{E}\frac{\hat{x} - \mathrm{i}\hat{y}}{\sqrt{2}} = \mathcal{E}\hat{e}_-, \tag{1.73}$$

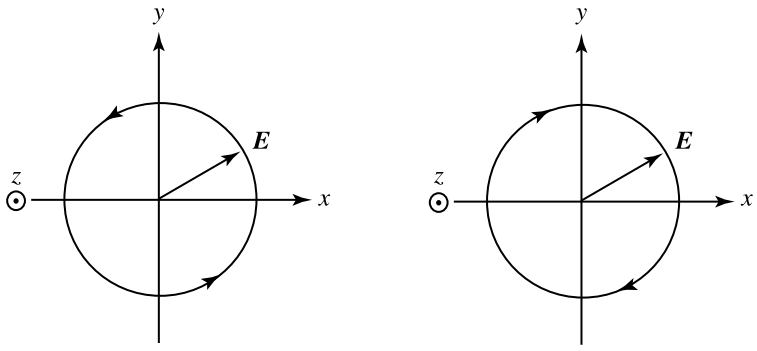


Figure 1.6 (*a*) Field of a left-circularly polarized wave. (*b*) Field of a right-circularly polarized wave.

and

$$\boldsymbol{E}(t) = \sqrt{2}\mathcal{E}(\hat{x}\cos\omega t - \hat{y}\sin\omega t). \tag{1.74}$$

The tip of this field vector rotates *clockwise* in a circle, as shown in Fig. 1.6(*b*). This right-circular polarization is also called *negative helicity*. Its eigenvector is

$$\hat{e}_- \equiv \frac{\hat{x} - \mathrm{i}\hat{y}}{\sqrt{2}}. \tag{1.75}$$

As can be seen, neither $\hat{e}_+$ nor $\hat{e}_-$ is a real vector. Note that the identification of $\hat{e}_+$, defined in (1.72), with left-circular polarization and that of $\hat{e}_-$, defined in (1.75), with right-circular polarization are based on the assumption that the wave propagates in the positive z direction. For a wave that propagates in the negative z direction, the handedness of these unit vectors changes: $\hat{e}_+$ becomes right-circular polarization, while $\hat{e}_-$ becomes left-circular polarization.

Linearly polarized light can be produced from unpolarized light using a *polarizer*. A polarizer can be of *transmission type*, which often utilizes the phenomenon of double refraction in an anisotropic crystal, discussed in Section 1.6, or of *reflection type*, which takes advantage of the polarization-sensitive reflectivity of a surface, discussed in Section 1.7. A very convenient transmission-type polarizer is the *Polaroid* film, which utilizes a material with *linear dichroism*, having low absorption for light linearly polarized in a particular direction and high absorption for light polarized orthogonally to this direction. The output is linearly polarized in the direction defined by the polarizer irrespective of the polarization state of the input optical wave. A polarizer can also be used to analyze the polarization of a particular optical wave. When so used, a polarizer is also called an *analyzer*.

1.5 Propagation in an isotropic medium

The propagation of an optical wave is governed by the wave equation. It depends on the optical property and physical structure of the medium. It also depends on the makeup of the optical wave, such as its frequency contents and its temporal characteristics. In this section, we consider the basic characteristics of the propagation of a monochromatic plane optical wave in an infinite homogeneous medium. For such a monochromatic wave, there is only one value of $\mathbf{k}$ and one value of ω. Its complex electric field is that given by (1.58), in which the field amplitude $\mathcal{E}$ is independent of $\mathbf{r}$ and t. Thus,

$$\mathbf{P}(\mathbf{r}, t) = \epsilon_0 \boldsymbol{\chi}(\mathbf{k}, \omega) \cdot \mathbf{E}(\mathbf{r}, t) \tag{1.76}$$

and

$$\mathbf{D}(\mathbf{r}, t) = \boldsymbol{\epsilon}(\mathbf{k}, \omega) \cdot \mathbf{E}(\mathbf{r}, t). \tag{1.77}$$

Also, in this section, we shall assume no spatial nonlocality in the media thus neglecting the $\mathbf{k}$ dependence of $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$. Then,

$$\mathbf{P}(\mathbf{r}, t) = \epsilon_0 \boldsymbol{\chi}(\omega) \cdot \mathbf{E}(\mathbf{r}, t) \tag{1.78}$$

and

$$\mathbf{D}(\mathbf{r}, t) = \boldsymbol{\epsilon}(\omega) \cdot \mathbf{E}(\mathbf{r}, t). \tag{1.79}$$

For a monochromatic wave of a frequency ω, the wave equation is simply

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \mathbf{E} + \mu_0 \boldsymbol{\epsilon}(\omega) \cdot \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \tag{1.80}$$

For an isotropic medium, $\boldsymbol{\epsilon}(\omega)$ is reduced to a scalar $\epsilon(\omega)$ and

$$\boldsymbol{\nabla} \cdot \mathbf{E} = \frac{1}{\epsilon(\omega)} \boldsymbol{\nabla} \cdot \mathbf{D} = 0. \tag{1.81}$$

Then, by using the vector identity $\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times = \boldsymbol{\nabla}\boldsymbol{\nabla} \cdot - \nabla^2$, the wave equation in (1.80) is reduced to the following simple form:

$$\nabla^2 \mathbf{E} - \mu_0 \epsilon(\omega) \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \tag{1.82}$$

For an anisotropic medium, (1.82) is generally not valid because (1.81) does not hold.

Note that with $\mathcal{E}$ in (1.58) being independent of $\mathbf{r}$ and t, we can make the following replacement for the operators when operating on $\mathbf{E}$ of the form in (1.58) or $\mathbf{H}$ of the same form:

$$\boldsymbol{\nabla} \longrightarrow \mathrm{i}\mathbf{k}, \qquad \frac{\partial}{\partial t} \longrightarrow -\mathrm{i}\omega. \tag{1.83}$$

Free space

In free space, $\mathbf{P} = 0$ and $\boldsymbol{\epsilon}$ is reduced to the scalar ϵ_0. Substitution of (1.58) in (1.82) then yields

$$k^2 = \omega^2 \mu_0 \epsilon_0. \tag{1.84}$$

The *propagation constant* in free space is

$$k = \frac{\omega}{c} = \frac{2\pi\nu}{c} = \frac{2\pi}{\lambda}, \tag{1.85}$$

where ν is the frequency of the optical wave and λ is its wavelength. Because k is proportional to $1/\lambda$, it is also called the *wavenumber*.

Using (1.83) and noting that $\mathbf{B} = \mu_0 \mathbf{H}$ and $\mathbf{D} = \epsilon_0 \mathbf{E}$, Maxwell's equations in (1.40)–(1.43) become

$$\mathbf{k} \times \mathbf{E} = \omega \mu_0 \mathbf{H}, \tag{1.86}$$

$$\mathbf{k} \times \mathbf{H} = -\omega \epsilon_0 \mathbf{E}, \tag{1.87}$$

$$\mathbf{k} \cdot \mathbf{E} = 0, \tag{1.88}$$

$$\mathbf{k} \cdot \mathbf{H} = 0. \tag{1.89}$$

From (1.86) and (1.87), we also have

$$\mathbf{E} \cdot \mathbf{H} = 0. \tag{1.90}$$

Therefore, the three vectors **E**, **H**, and **k** are orthogonal. These relationships also imply that

$$\mathbf{S} \parallel \mathbf{k}. \tag{1.91}$$

The relationships among the directions of these vectors are shown in Fig. 1.7.

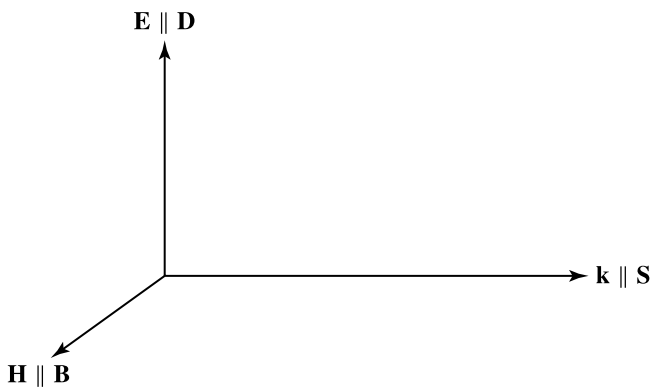


Figure 1.7 Relationships among the directions of **E**, **D**, **H**, **B**, **k**, and **S** in free space or in an isotropic medium.

Using (1.85), we can also write (1.86) and (1.87) in the following form:

$$\mathbf{H} = \frac{1}{Z_0}\hat{k} \times \mathbf{E}, \qquad \mathbf{E} = Z_0\mathbf{H} \times \hat{k}, \tag{1.92}$$

where $\hat{k} = \mathbf{k}/k$ is the unit vector in the $\mathbf{k}$ direction and

$$Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \approx 120\pi\ \Omega \approx 377\ \Omega \tag{1.93}$$

is the free-space *impedance*. The concept of this impedance is not that of the impedance of a resistor but is analogous to the concept of the impedance of a tranmission line.

Because $\mathbf{S} \parallel \mathbf{k}$, the light intensity in free space can be expressed as

$$I = \hat{k} \cdot \overline{\boldsymbol{S}} = 2\frac{|\mathbf{E}|^2}{Z_0} = 2Z_0|\mathbf{H}|^2. \tag{1.94}$$

Lossless medium

In this case, $\boldsymbol{\epsilon}(\omega)$ is reduced to a positive real scalar $\epsilon(\omega)$, which is different from ϵ_0. All of the results obtained for free space remain valid, except that ϵ_0 is replaced by $\epsilon(\omega)$. This change of the electric permittivity from a vacuum to a material is measured by the *relative electric permittivity*, ϵ/ϵ_0, which is a dimensionless quantity also known as the *dielectric constant* of the material. Therefore, the propagation constant in the medium is

$$k = \omega\sqrt{\mu_0\epsilon} = \frac{n\omega}{c} = \frac{2\pi n\nu}{c} = \frac{2\pi n}{\lambda}, \tag{1.95}$$

where

$$n = \sqrt{\frac{\epsilon}{\epsilon_0}} = (\text{dielectric constant})^{1/2} \tag{1.96}$$

is the *index of refraction*, or *refractive index*, of the medium.

In a medium that has an index of refraction n, the optical frequency is still ν, but the optical wavelength is λ/n, and the speed of light is $v = c/n$. Because $n(\omega)$ in a medium is generally frequency dependent, the speed of light in a medium is also frequency dependent. This results in various dispersive phenomena such as the separation of different colors by a prism and the broadening or shortening of an optical pulse traveling through a medium. We also find that

$$Z = \frac{Z_0}{n} \tag{1.97}$$

in a medium. The light intensity is then

$$I = 2\frac{|\mathbf{E}|^2}{Z} = 2Z|\mathbf{H}|^2 = \frac{2k}{\omega\mu_0}|\mathbf{E}|^2 = \frac{2k}{\omega\epsilon}|\mathbf{H}|^2. \tag{1.98}$$

Medium with a loss or gain

As discussed in the preceding section, χ and ϵ become complex when a medium has an optical loss or gain. Therefore,

$$k^2 = \omega^2 \mu_0 \epsilon = \omega^2 \mu_0 (\epsilon' + \mathrm{i}\epsilon''), \tag{1.99}$$

and the propagation constant k becomes complex:

$$k = k' + \mathrm{i}k'' = \beta + \mathrm{i}\frac{\alpha}{2}. \tag{1.100}$$

The index of refraction also becomes complex:

$$n = \sqrt{\frac{\epsilon' + \mathrm{i}\epsilon''}{\epsilon_0}} = n' + \mathrm{i}n''. \tag{1.101}$$

The relation between k and n in (1.95) is still valid. Meanwhile, the impedance Z of the medium also becomes complex. Therefore, $\mathbf{E}$ and $\mathbf{H}$ are no longer in phase, as can be seen from (1.92) by replacing Z_0 with a complex Z, and I is not simply given by (1.98) but is given by the real part of it.

It can be shown that if we choose β to be positive, the sign of α is the same as that of ϵ''. In this case, k' and n' are also positive and k'' and n'' also have the same sign as ϵ''. If we consider as an example an optical wave propagating in the z direction, then $\hat{k} = \hat{z}$ and, from (1.58) and (1.100), the complex electric field is

$$\mathbf{E}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}} \mathrm{e}^{-\alpha z/2} \exp(\mathrm{i}\beta z - \mathrm{i}\omega t). \tag{1.102}$$

It can be seen that the wave has a phase that varies sinusoidally with a period of $1/\beta$ along z. In addition, its amplitude is not constant but varies exponentially with z. Thus, light intensity is also a function of z:

$$I \propto \mathrm{e}^{-\alpha z}. \tag{1.103}$$

Clearly, β is the *wavenumber* in this case, and the sign of α determines the attenuation or amplification of the optical wave:

1. If $\chi'' > 0$, then $\epsilon'' > 0$ and $\alpha > 0$. As the optical wave propagates, its field amplitude and intensity decay exponentially along the direction of propagation. Therefore, α is called the *absorption coefficient* or *attenuation coefficient*.
2. If $\chi'' < 0$, then $\epsilon'' < 0$ and $\alpha < 0$. The field amplitude and intensity of the optical wave grow exponentially. Then, we define $g = -\alpha$ as the *gain coefficient* or *amplification coefficient*.

The unit of both α and g is per meter, often also quoted per centimeter.

EXAMPLE 1.1 The complex susceptibility of GaAs at an optical wavelength of $\lambda = 850\,\text{nm}$ is $\chi = 12.17 + \text{i}0.49$. Therefore, at this wavelength, GaAs has a complex refractive index of

$$n = (\epsilon/\epsilon_0)^{1/2} = (1+\chi)^{1/2} = (13.17 + \text{i}0.49)^{1/2} = 3.63 + \text{i}0.0676$$

and an absorption coefficient of

$$\alpha = 2k'' = \frac{4\pi n''}{\lambda} = \frac{4\pi \times 0.0676}{850 \times 10^{-9}}\,\text{m}^{-1} = 10^6\,\text{m}^{-1}.$$

An optical beam at 850 nm wavelength can travel in GaAs only for a distance of $l = -\ln(1-0.99)/\alpha = 4.6\,\mu\text{m}$ before losing 99% of its energy to absorption, which is obtained by solving $1 - \text{e}^{-\alpha l} = 0.99$ with $\alpha = 10^6\,\text{m}^{-1}$.

1.6 Propagation in an anisotropic medium

In an anisotropic medium, the tensors $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ do not reduce to scalars. Therefore, $\mathbf{P} \nparallel \mathbf{E}$ and $\mathbf{D} \nparallel \mathbf{E}$. As a result, (1.81) is not true any more, and, in general,

$$\boldsymbol{\nabla} \cdot \mathbf{E} \neq 0. \tag{1.104}$$

Consequently, (1.82) cannot be used for propagation of a monochromatic wave in an anisotropic medium. Instead, (1.80) has to be used.

Anisotropic $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$

In a linear anisotropic medium, both $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ are second-rank tensors. They can be expressed in the following matrix forms:

$$\boldsymbol{\chi} = \begin{bmatrix} \chi_{11} & \chi_{12} & \chi_{13} \\ \chi_{21} & \chi_{22} & \chi_{23} \\ \chi_{31} & \chi_{32} & \chi_{33} \end{bmatrix} \tag{1.105}$$

and

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix}. \tag{1.106}$$

The relationships $\mathbf{P} = \epsilon_0 \boldsymbol{\chi} \cdot \mathbf{E}$ and $\mathbf{D} = \boldsymbol{\epsilon} \cdot \mathbf{E}$ are carried out as products between a tensor and a column vector. For example,

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}. \tag{1.107}$$

In general, the matrix in (1.106) representing the tensor $\boldsymbol{\epsilon}$ is not diagonal. It can be diagonalized by a proper choice of the coordinate system, yielding

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{bmatrix}, \tag{1.108}$$

where ϵ_i, for $i = 1, 2, 3$, are the *eigenvalues* of $\boldsymbol{\epsilon}$ with their corresponding *eigenvectors*, $\hat{u}_i$, being the axes of the coordinate system chosen to diagonalize $\boldsymbol{\epsilon}$. The characteristics of ϵ_i and $\hat{u}_i$ depend on the symmetry properties of $\boldsymbol{\epsilon}$. The two matrices representing $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ have the same symmetry properties because $\boldsymbol{\epsilon} = \epsilon_0(1 + \boldsymbol{\chi})$, where 1 has the form of a 3×3 identity matrix in its addition to the tensor $\boldsymbol{\chi}$. Therefore, $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ are diagonalized by the same set of eigenvectors that represent the axes of the chosen coordinate system.

The symmetry properties of $\boldsymbol{\epsilon}$, as well as those of $\boldsymbol{\chi}$, are determined by the properties of the medium.

1. **Reciprocal media.** Nonmagnetic materials in the absence of an external magnetic field are *reciprocal media*. In a reciprocal medium, the *Lorentz reciprocity theorem* of electromagnetics holds; consequently, the source and the detector of an optical signal can be interchanged. If such a material is not *optically active*, its optical properties are described by a symmetric $\boldsymbol{\epsilon}$ tensor: $\epsilon_{ij} = \epsilon_{ji}$. *For a symmetric tensor, the eigenvectors $\hat{u}_i$ are always real vectors.* They can be chosen to be $\hat{x}$, $\hat{y}$, and $\hat{z}$ of a rectangular coordinate system in real space. This is true even when $\boldsymbol{\epsilon}$ is complex. (a) If a nonmagnetic medium does not have an optical loss or gain, its $\boldsymbol{\epsilon}$ tensor is Hermitian. A symmetric Hermitian tensor is real and symmetric: $\epsilon_{ij}^* = \epsilon_{ij} = \epsilon_{ji} = \epsilon_{ji}^*$. In this case, the eigenvalues ϵ_i have real values. (b) If a nonmagnetic medium has an optical loss or gain, its $\boldsymbol{\epsilon}$ tensor is not Hermitian but is complex and symmetric: $\epsilon_{ij} = \epsilon_{ji}$ but $\epsilon_{ij} \neq \epsilon_{ji}^*$. Then, the eigenvalues ϵ_i are complex. (c) If a nonmagnetic medium is optically active, it is still reciprocal although its $\boldsymbol{\epsilon}$ tensor is not symmetric. In this case, the eigenvectors are complex but the eigenvalues can be real if the medium is lossless.
2. **Nonreciprocal media.** Magnetic materials, and nonmagnetic materials subject to an external magnetic field, are *nonreciprocal media*. In such a medium, no symmetry exists when the source and the detector of an optical signal are interchanged. The $\boldsymbol{\epsilon}$ tensor describing the optical properties of such a material is not symmetric: $\epsilon_{ij} \neq \epsilon_{ji}$. *The eigenvectors $\hat{u}_i$ are complex vectors.* Therefore, they are not ordinary coordinate axes in real space, as seen later in the discussion on magneto-optic devices. (a) For a lossless magnetic medium, $\boldsymbol{\epsilon}$ is Hermitian: $\epsilon_{ij} = \epsilon_{ji}^*$. In this case, the eigenvalues ϵ_i are real even though the eigenvectors are complex. (b) For a magnetic medium

that has an optical loss or gain, $\boldsymbol{\epsilon}$ is neither symmetric nor Hermitian. Both the eigenvectors and the eigenvalues are complex.

Most materials used for photonic devices are nonmagnetic dielectric materials that are not optically active. The properties of magnetic materials are of interest to us only in consideration of magneto-optic devices, discussed in Chapter 7. Similarities and differences between magnetic and optically active materials are also briefly mentioned in Section 7.2. The discussion in the rest of this section is specific to nonmagnetic dielectric materials that are not optically active.

According to the above, in a dielectric material the axes of the coordinate system in which $\boldsymbol{\epsilon}$ is diagonal are real in space and can be labeled $\hat{x}$, $\hat{y}$, and $\hat{z}$. Noncrystalline materials are generally isotropic, for which the choice of the orthogonal coordinate axes $\hat{x}$, $\hat{y}$, and $\hat{z}$ is arbitrary. In contrast, many crystalline materials that are useful for photonic device applications are anisotropic. For any given anisotropic crystal, there is a unique set of coordinate axes for $\boldsymbol{\epsilon}$ to be diagonal. These unique $\hat{x}$, $\hat{y}$, and $\hat{z}$ coordinate axes are called the *principal dielectric axes*, or simply the *principal axes*, of the crystal. In the coordinate system defined by these principal axes, $\boldsymbol{\epsilon}$ is diagonalized with eigenvalues ϵ_x, ϵ_y, and ϵ_z. The components of $\mathbf{D}$ and $\mathbf{E}$ along these axes have the following simple relations:

$$D_x = \epsilon_x E_x, \qquad D_y = \epsilon_y E_y, \qquad D_z = \epsilon_z E_z. \tag{1.109}$$

The values ϵ_x/ϵ_0, ϵ_y/ϵ_0, and ϵ_z/ϵ_0 are the eigenvalues of the *dielectric constant tensor*, $\boldsymbol{\epsilon}/\epsilon_0$, and are called the *principal dielectric constants*. They define three *principal indices of refraction*:

$$n_x = \sqrt{\frac{\epsilon_x}{\epsilon_0}}, \qquad n_y = \sqrt{\frac{\epsilon_y}{\epsilon_0}}, \qquad n_z = \sqrt{\frac{\epsilon_z}{\epsilon_0}}. \tag{1.110}$$

Note that when $\boldsymbol{\epsilon}$ is diagonalized, $\boldsymbol{\chi}$ is also diagonalized along the same principal axes with corresponding *principal dielectric susceptibilities*, χ_x, χ_y, and χ_z. The principal dielectric susceptibilities of any lossless dielectric material always have positive values; therefore, the principal dielectric constants of such a material are always larger than unity.

Because $\mathbf{D} \perp \mathbf{k}$ due to the fact that $\nabla \cdot \mathbf{D} = 0$, there is no $\mathbf{D}$ component along the direction of wave propagation. In general, $\mathbf{D}$ can be decomposed into two mutually orthogonal components, each of which is also orthogonal to $\mathbf{k}$. In an anisotropic crystal, these two components generally have different indices of refraction, and thus different propagation constants. This phenomenon is called *birefringence*. Such a crystal is a *birefringent crystal*.

EXAMPLE 1.2 At an optical wavelength of 1 μm, the permittivity tensor of the KDP crystal represented in a rectangular coordinate system defined by $\hat{x}_1$, $\hat{x}_2$, and $\hat{x}_3$ is found to be

$$\boldsymbol{\epsilon} = \epsilon_0 \begin{bmatrix} 2.28 & 0 & 0 \\ 0 & 2.25 & -0.051\,96 \\ 0 & -0.051\,96 & 2.19 \end{bmatrix}.$$

Find the principal axes and the corresponding principal indices for this crystal.

Solution Note that $\boldsymbol{\epsilon}$ is represented by a symmetric matrix because KDP is a nonmagnetic dielectric crystal. Diagonalization of this matrix yields the following eigenvalues and corresponding eigenvectors:

$$\epsilon_x = 2.28\epsilon_0,\ \hat{x} = \hat{x}_1,$$
$$\epsilon_y = 2.28\epsilon_0,\ \hat{y} = 0.866\hat{x}_2 - 0.500\hat{x}_3,$$
$$\epsilon_z = 2.16\epsilon_0,\ \hat{z} = 0.500\hat{x}_2 + 0.866\hat{x}_3.$$

Therefore, the principal axes of the crystal are $\hat{x}$, $\hat{y}$, and $\hat{z}$, given above, and the principal indices of refraction are $n_x = \sqrt{2.28} = 1.51$, $n_y = \sqrt{2.28} = 1.51$, and $n_z = \sqrt{2.16} = 1.47$.

Index ellipsoid

The inverse of the dielectric constant tensor mentioned above is the *relative impermeability tensor*:

$$\boldsymbol{\eta} = [\eta_{ij}] = \left(\frac{\boldsymbol{\epsilon}}{\epsilon_0}\right)^{-1}, \tag{1.111}$$

where i and j are spatial coordinate indices. In a general rectangular coordinate system (x_1, x_2, x_3), the ellipsoid defined by

$$\sum_{i,j} x_i \eta_{ij} x_j = 1 \tag{1.112}$$

is called the *index ellipsoid* or the *optical indicatrix*. In a nonmagnetic dielectric medium, $\boldsymbol{\eta}$ is a symmetric tensor, i.e., $\eta_{ij} = \eta_{ji}$, because $\boldsymbol{\epsilon}$ is symmetric. Therefore, (1.112) can be written as

$$\eta_{11}x_1^2 + \eta_{22}x_2^2 + \eta_{33}x_3^2 + 2\eta_{23}x_2x_3 + 2\eta_{31}x_3x_1 + 2\eta_{12}x_1x_2 = 1. \tag{1.113}$$

This equation is usually written as

$$\eta_1 x_1^2 + \eta_2 x_2^2 + \eta_3 x_3^2 + 2\eta_4 x_2 x_3 + 2\eta_5 x_3 x_1 + 2\eta_6 x_1 x_2 = 1 \tag{1.114}$$

using the following *index contraction* rule to reduce the double index ij of η_{ij} to the

single index α of η_α:

$$\begin{array}{lcccccc} ij: & 11 & 22 & 33 & 23, 32 & 31, 13 & 12, 21 \\ \text{or } ij: & xx & yy & zz & yz, zy & zx, xz & xy, yx \\ \alpha: & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \tag{1.115}$$

The index ellipsoid equation is invariant with respect to coordinate rotation. When a coordinate system with axes aligned with the principal dielectric axes of the crystal is chosen, $\boldsymbol{\epsilon}$ is diagonalized. Thus the tensor $\boldsymbol{\eta}$ is also diagonalized with the following eigenvalues:

$$\eta_x = \frac{\epsilon_0}{\epsilon_x} = \frac{1}{n_x^2}, \qquad \eta_y = \frac{\epsilon_0}{\epsilon_y} = \frac{1}{n_y^2}, \qquad \eta_z = \frac{\epsilon_0}{\epsilon_z} = \frac{1}{n_z^2}. \tag{1.116}$$

In this coordinate system, the index ellipsoid takes the following simple form:

$$\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1. \tag{1.117}$$

Comparing (1.117) with (1.114), we find that the terms containing cross products of different coordinates are eliminated when the coordinate system of the principal dielectric axes is used. The principal axes of the index ellipsoid now coincide with the principal dielectric axes of the crystal, and the principal indices of refraction of the crystal are given by the semiaxes of the index ellipsoid. This is illustrated in Fig. 1.8. Therefore, a coordinate transformation by rotation to eliminate cross-product terms in the index ellipsoid equation is equivalent to diagonalization of the $\boldsymbol{\epsilon}$ tensor. The

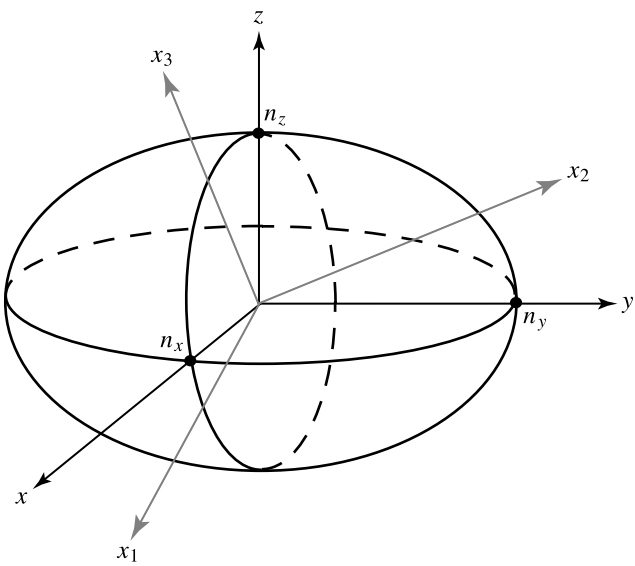


Figure 1.8 Index ellipsoid and its relationship with the coordinate system. Here (x, y, z) is the coordinate system aligned with the principal axes of the crystal, while (x_1, x_2, x_3) is an arbitrary coordinate system.

principal dielectric axes and their corresponding principal indices of refraction can be found through either approach. Between the two approaches, however, diagonalization of the ϵ tensor is better because it is more systematic and is easier to carry out.

EXAMPLE 1.3 Find the principal axes and their corresponding principal indices for the KDP crystal given in Example 1.2 by using the index ellipsoid instead of diagonalizing the ϵ tensor as done in Example 1.2. Compare the two approaches.

Solution The relative impermeability tensor in the (x_1, x_2, x_3) coordinate system can be found by inverting the ϵ tensor:

$$\eta = \left(\frac{\epsilon}{\epsilon_0}\right)^{-1} = \begin{bmatrix} 2.28 & 0 & 0 \\ 0 & 2.25 & -0.051\,96 \\ 0 & -0.051\,96 & 2.19 \end{bmatrix}^{-1} \approx \begin{bmatrix} \dfrac{1}{2.28} & 0 & 0 \\ 0 & \dfrac{1}{2.25} & 0.010\,55 \\ 0 & 0.010\,55 & \dfrac{1}{2.19} \end{bmatrix}.$$

In the (x_1, x_2, x_3) coordinate system, the index ellipsoid is thus described by the following equation:

$$\frac{x_1^2}{2.28} + \frac{x_2^2}{2.25} + \frac{x_3^2}{2.19} + 0.0211 x_2 x_3 = 1.$$

To find the principal axes and their principal indices of refraction, the cross-product term has to be eliminated by rotating the coordinates. From Example 1.2, we know that this can be done by taking

$$x_1 = x, \quad x_2 = 0.866y + 0.500z, \quad x_3 = -0.500y + 0.866z.$$

Substitution of these relations into the above index ellipsoid equation transforms it into the following equation for the index ellipsoid in the (x, y, z) coordinate system:

$$\frac{x^2}{2.28} + \frac{y^2}{2.28} + \frac{z^2}{2.16} = 1.$$

Thus the principal indices are $n_x = \sqrt{2.28} = 1.51$, $n_y = \sqrt{2.28} = 1.51$, and $n_z = \sqrt{2.16} = 1.47$.

Comparing the two approaches illustrated in this example and in Example 1.2, it is clear that they are equivalent to one another. It is also clear that the method of diagonalizing ϵ described in Example 1.2 is more systematic and straightforward than that of eliminating the cross-product terms in the equation for the index ellipsoid, particularly when there is more than one cross-product term.

Propagation along a principal axis

We first consider the simple case when an optical wave propagates along one of the principal axes, say $\hat{z}$. Then the field can be decomposed into two *normal modes*, each of which is polarized along one of the other two principal axes, $\hat{x}$ or $\hat{y}$. We see from (1.109) and (1.110) that each field component along a principal axis has a characteristic index of refraction n_i, meaning that it has a characteristic propagation constant of $k^i = n_i\omega/c$, which is determined by the polarization of the field but not by the direction of wave propagation. For a wave propagating along $\hat{z}$, the electric field can be expressed as

$$\begin{aligned}\mathbf{E} &= \hat{x}\mathcal{E}_x e^{ik^x z - i\omega t} + \hat{y}\mathcal{E}_y e^{ik^y z - i\omega t} \\ &= \left[\hat{x}\mathcal{E}_x + \hat{y}\mathcal{E}_y e^{i(k^y - k^x)z}\right] e^{ik^x z - i\omega t}. \end{aligned} \tag{1.118}$$

Because the wave propagates in the z direction, the wavevectors are $\mathbf{k}^x = k^x\hat{z}$ for the x-polarized field and $\mathbf{k}^y = k^y\hat{z}$ for the y-polarized field. Note that $k^x = n_x\omega/c$ and $k^y = n_y\omega/c$ are the propagation constants of the x- and y-polarized fields, respectively, not to be confused with the x and y components of a wavevector $\mathbf{k}$, which are normally expressed as k_x and k_y. The field expressed in (1.118) has the following propagation characteristics.

1. If it is originally linearly polarized along one of the principal axes, it remains linearly polarized in the same direction.
2. If it is originally linearly polarized at an angle $\theta = \tan^{-1}(\mathcal{E}_y/\mathcal{E}_x)$ with respect to the x axis, its polarization state varies periodically along z with a period of $2\pi/|k^y - k^x|$. In general, its polarization follows a sequence of variations from linear to elliptical to linear in the first half-period and then reverses the sequence back to linear in the second half-period. At the half-period position, it is linearly polarized at an angle θ on the other side of the x axis. Thus the polarization is rotated by 2θ from the original direction. This is shown in Fig. 1.9(a). In the special case when $\theta = 45^\circ$, the wave is circularly polarized at the quarter-period point and is linearly polarized with its polarization rotated by 90° from the original direction at the half-period point. This is shown in Fig. 1.9(b).

These characteristics have very useful applications. A plate of an anisotropic material that has a quarter-period thickness of

$$l_{\lambda/4} = \frac{1}{4}\cdot\frac{2\pi}{|k^y - k^x|} = \frac{\lambda}{4|n_y - n_x|} \tag{1.119}$$

is called a *quarter-wave plate*. It can be used to convert a linearly polarized wave to circular or elliptic polarization, and vice versa. A plate of thickness $3l_{\lambda/4}$ or $5l_{\lambda/4}$ or any odd integral multiple of $l_{\lambda/4}$ also has the same function. In contrast, a plate of a

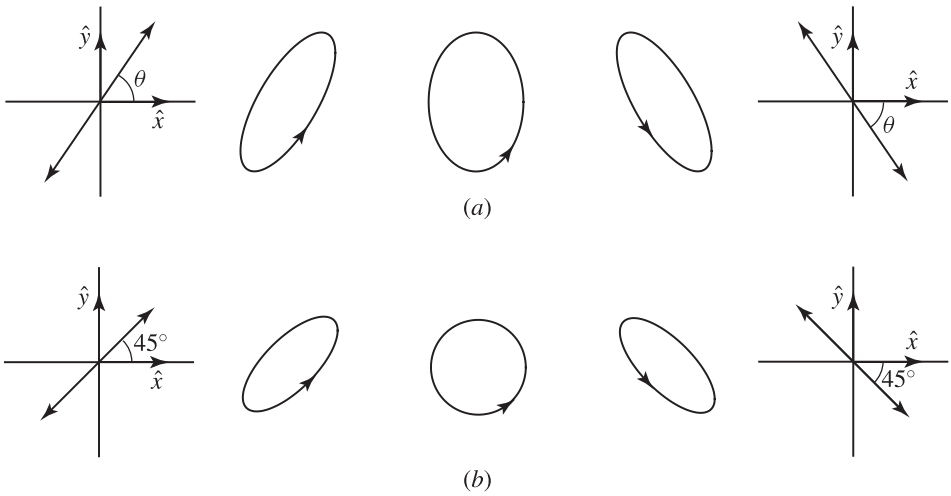


Figure 1.9 Evolution of the polarization state of an optical wave propagating along the principal axis $\hat{z}$ of an anisotropic crystal that has $n_x \neq n_y$. Only the evolution over one half-period is shown here. (*a*) The optical wave is initially linearly polarized at an arbitrary angle θ with respect to the principal axis $\hat{x}$. (*b*) The optical wave is initially polarized at 45° with respect to $\hat{x}$.

half-period thickness of

$$l_{\lambda/2} = \frac{\lambda}{2|n_y - n_x|} \tag{1.120}$$

is called a *half-wave plate*. It can be used to rotate the polarization direction of a linearly polarized wave by any angular amount by properly choosing the angle θ between the incident polarization with respect to the principal axis $\hat{x}$, or $\hat{y}$, of the crystal. A plate of a thickness that is any odd integral multiple of $l_{\lambda/2}$ has the same function. Note that though the output from a quarter-wave or half-wave plate can be linearly polarized, the wave plates are not polarizers. They are based on different principles and have completely different functions.

For the quarter-wave and half-wave plates discussed here, $n_x \neq n_y$. Between the two crystal axes $\hat{x}$ and $\hat{y}$, the one with the smaller index is called the *fast axis* while the other, with the larger index, is the *slow axis*.

EXAMPLE 1.4 KDP can be used to make quarter-wave and half-wave plates. Find the thicknesses of the quarter-wave and half-wave plates made of KDP for 1 μm wavelength.

Solution From Example 1.3, we know that $n_x = n_y = 1.51$ and $n_z = 1.47$ for KDP at 1 μm wavelength. Because $n_x = n_y$, we cannot use n_x and n_y to make a wave plate that allows the beam to propagate in the z direction. Instead, the beam can propagate in any direction on the xy plane so that the difference between n_z and $n_x = n_y$ can

be used for the function of a wave plate. Assuming that the wave propagates in the x direction, then the thickness of a quarter-wave plate for $\lambda = 1$ μm is

$$l_{\lambda/4} = \frac{\lambda}{4|n_y - n_z|} = \frac{1\ \mu\text{m}}{4 \times |1.51 - 1.47|} = 6.25\ \mu\text{m}.$$

A quarter-wave plate at 1 μm wavelength can have a thickness of any odd integral multiple, such as 18.75 μm, 31.25 μm, . . . , of 6.25 μm. A half-wave plate for the 1 μm wavelength has a thickness of

$$l_{\lambda/2} = \frac{\lambda}{2|n_y - n_z|} = \frac{1\ \mu\text{m}}{2 \times |1.51 - 1.47|} = 12.5\ \mu\text{m}.$$

A plate of a thickness that is an odd multiple, such as 37.5 μm, 62.5 μm, . . . , of 12.5 μm also functions as a half-wave plate at 1 μm wavelength. For these wave plates, $\hat{z}$ is the fast axis and $\hat{y}$ is the slow axis because $n_z < n_y$.

Optical axes

The state of polarization of an optical wave generally varies along its path of propagation through an anisotropic crystal unless it is linearly polarized in the direction of a principal axis. However, in an anisotropic crystal with $n_x = n_y \neq n_z$, a wave propagating in the z direction does not see the anisotropy of the crystal because in this situation the x and y components of the field have the same propagation constant. This wave will maintain its original polarization as it propagates through the crystal. Evidently, this is true only for propagation along the z axis in such a crystal. Such a unique axis in a crystal along which an optical wave can propagate with an index of refraction that is independent of its polarization direction is called the *optical axis* of the crystal.

For an anisotropic crystal that has only one distinctive principal index among its three principal indices, there is only one optical axis, which coincides with the axis of the distinctive principal index of refraction. Such a crystal is called a *uniaxial crystal*. It is customary to assign $\hat{z}$ to this unique principal axis. The identical principal indices of refraction are called the *ordinary index*, n_o, and the distinctive index of refraction is called the *extraordinary index*, n_e. Thus, $n_x = n_y = n_\text{o}$ and $n_z = n_\text{e}$. The crystal is called *positive uniaxial* if $n_\text{e} > n_\text{o}$ and is *negative uniaxial* if $n_\text{e} < n_\text{o}$.

For a crystal that has three distinct principal indices of refraction, there are two optical axes, neither of which coincides with any one of the principal axes. Such a crystal is called a *biaxial crystal* because of the existence of two optical axes.

Ordinary and extraordinary waves

When an optical wave propagates in a direction other than that along an optical axis, the index of refraction depends on the direction of its polarization. In this situation, there exist two normal modes of linearly polarized waves, each of which sees a unique

index of refraction. One of them is the polarization perpendicular to the optical axis. This normal mode is called the *ordinary wave*. We use $\hat{e}_o$ to indicate its direction of polarization. The other normal mode is clearly one that is perpendicular to $\hat{e}_o$ because the two normal-mode polarizations are orthogonal to each other. This normal mode is called the *extraordinary wave*, and its direction of polarization is indicated by $\hat{e}_e$. *Note that these are the directions of* **D** *rather than those of* **E**. For the ordinary wave, $\hat{e}_o \parallel \mathbf{D}_o \parallel \mathbf{E}_o$. For the extraordinary wave, $\hat{e}_e \parallel \mathbf{D}_e \not\parallel \mathbf{E}_e$ except when $\hat{e}_e$ is parallel to a principal axis. Both $\hat{e}_o$ and $\hat{e}_e$, being the unit vectors of $\mathbf{D}_o$ and $\mathbf{D}_e$, are perpendicular to the direction of wave propagation, $\hat{k}$. From this understanding, both $\hat{e}_o$ and $\hat{e}_e$ can be found if both $\hat{k}$ and the optical axis are known. For a uniaxial crystal with optical axis $\hat{z}$, this means that

$$\hat{e}_o = \frac{1}{\sin\theta}\hat{k} \times \hat{z}, \qquad \hat{e}_e = \hat{e}_o \times \hat{k} \tag{1.121}$$

if the vector $\hat{k}$ is in a direction that is at an angle θ with respect to $\hat{z}$ and an angle ϕ with respect to the axis $\hat{x}$. Therefore, we have (see Problem 1.6.12)

$$\hat{k} = \hat{x}\sin\theta\cos\phi + \hat{y}\sin\theta\sin\phi + \hat{z}\cos\theta, \tag{1.122}$$

$$\hat{e}_o = \hat{x}\sin\phi - \hat{y}\cos\phi, \tag{1.123}$$

$$\hat{e}_e = -\hat{x}\cos\theta\cos\phi - \hat{y}\cos\theta\sin\phi + \hat{z}\sin\theta. \tag{1.124}$$

The relationships among these vectors are illustrated in Fig. 1.10.

The indices of refraction associated with the ordinary and extraordinary waves can be found by using the index ellipsoid given in (1.117), as is shown in Fig. 1.11. The

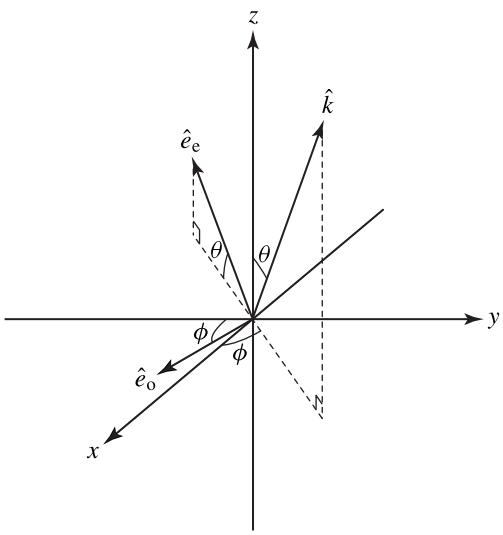


Figure 1.10 Relationships among the direction of wave propagation and the polarization directions of the ordinary and extraordinary waves.

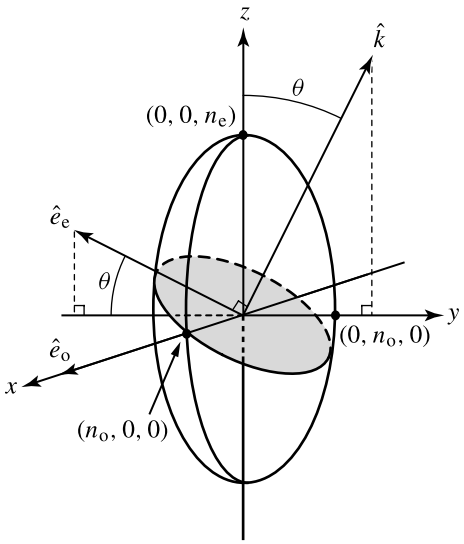


Figure 1.11 Determination of the indices of refraction for the ordinary and extraordinary waves in a uniaxial crystal using index ellipsoid.

intersection of the index ellipsoid and the plane normal to $\hat{k}$ at the origin of the ellipsoid defines an index ellipse. The principal axes of this index ellipse are in the directions of $\hat{e}_o$ and $\hat{e}_e$, and their half-lengths are the corresponding indices of refraction. For a uniaxial crystal, the index of refraction for the ordinary wave is simply n_o. The index of refraction for the extraordinary wave depends on the angle θ and is given by (see Problem 1.6.12)

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2}, \tag{1.125}$$

which can be seen from Fig. 1.11. Because $\mathbf{D}$ is orthogonal to $\mathbf{k}$ and can be decomposed into $\mathbf{D}_o$ and $\mathbf{D}_e$ components, we have

$$\mathbf{D} = \hat{e}_o \mathcal{D}_o e^{ik_o \hat{k}\cdot\mathbf{r} - i\omega t} + \hat{e}_e \mathcal{D}_e e^{ik_e \hat{k}\cdot\mathbf{r} - i\omega t}, \tag{1.126}$$

where $k_o = n_o\omega/c$ and $k_e = n_e(\theta)\omega/c$. In general, $\mathbf{E}$ cannot be written in the form of (1.126) because its longitudinal component along the wave propagation direction $\mathbf{k}$ does not vanish except when $\theta = 0^\circ$ or 90°. We see that $n_e(0^\circ) = n_o$ and $n_e(90^\circ) = n_e$. The special case when the wave propagates along one of the principal axes discussed earlier belongs to one of these situations.

The normal-mode polarizations for an optical wave propagating in a biaxial crystal can be found following a similar, albeit more complicated, procedure.

EXAMPLE 1.5 From the preceding three examples, we find that KDP is a uniaxial crystal with $\hat{z}$ being its optical axis because $n_x = n_y \neq n_z$. At 1 μm wavelength, we have

$n_o = 1.51$ and $n_e = 1.47$. KDP is negative uniaxial because $n_o > n_e$. For an optical wave propagating in KDP along a direction $\hat{k}$ that makes an angle θ with respect to the optical axis $\hat{z}$, the refractive index for the extraordinary wave is a function of θ. For $\theta = 0^\circ, n_e(0^\circ) = n_o = 1.51$. For $\theta = 90^\circ, n_e(90^\circ) = n_e = 1.47$. For $0^\circ < \theta < 90^\circ$, $1.47 < n_e(\theta) < 1.51$. For example,

$$n_e(30^\circ) = \left(\frac{\cos^2 30^\circ}{n_o^2} + \frac{\sin^2 30^\circ}{n_e^2}\right)^{-1/2} = 1.50,$$

$$n_e(60^\circ) = \left(\frac{\cos^2 60^\circ}{n_o^2} + \frac{\sin^2 60^\circ}{n_e^2}\right)^{-1/2} = 1.48.$$

Spatial beam walk-off

Each of the normal modes has a well-defined propagation constant. Therefore, the fields of monochromatic ordinary and extraordinary waves in an anisotropic medium can be separately written in the form of (1.47), with $\mathbf{k} = \mathbf{k}_o$ for the ordinary way and $\mathbf{k} = \mathbf{k}_e$ for the extraordinary way. By using (1.83), Maxwell's equations for a normal mode, either ordinary or extraordinary, reduce to the following:

$$\mathbf{k} \times \mathbf{E} = \omega\mu_0 \mathbf{H}, \tag{1.127}$$

$$\mathbf{k} \times \mathbf{H} = -\omega \mathbf{D}, \tag{1.128}$$

$$\mathbf{k} \cdot \mathbf{D} = 0, \tag{1.129}$$

$$\mathbf{k} \cdot \mathbf{H} = 0. \tag{1.130}$$

Note that because $n_o \neq n_e$, these relations apply to the ordinary and the extraordinary normal mode *separately* with different values for $\mathbf{k}$ but not to a wave mixing the two modes. At optical frequencies, $\mathbf{B} = \mu_0 \mathbf{H}$ is also true in an anisotropic medium. Therefore, (1.127) and (1.130) have the same forms as (1.86) and (1.89), respectively. Because (1.88) for a wave in an isotropic medium is now replaced by (1.129), we have $\mathbf{D} \perp \mathbf{k}$ for both ordinary and extraordinary waves. For an ordinary wave, $\mathbf{E}_o \perp \mathbf{k}_o$ because $\mathbf{D}_o \parallel \mathbf{E}_o$. Therefore, the relationships shown in Fig. 1.12(a) among the field vectors for an ordinary wave in an anisotropic medium are the same as those shown in Fig. 1.7 for a wave in an isotropic medium. However, $\mathbf{E}_e \not\perp \mathbf{k}_e$ for an extraordinary wave in general, and $\mathbf{S}_e$ is not necessarily parallel to $\mathbf{k}_e$ because $\mathbf{D}_e \nparallel \mathbf{E}_e$. The only exception is when $\hat{e}_e$ is parallel to a principal axis. As a result, the direction of power flow, which is that of $\mathbf{S}_e$, is not the same as the direction of wavefront propagation, which is normal to the planes of constant phase and is that of $\mathbf{k}_e$. This is shown in Fig. 1.12(b) together with the relationships among the directions of the field vectors. Note that $\mathbf{E}_e$, $\mathbf{D}_e$, $\mathbf{k}_e$, and $\mathbf{S}_e$ lie in a plane normal to $\mathbf{H}_e$ because $\mathbf{B}_e \parallel \mathbf{H}_e$. Though (1.90) is still true according to (1.127), the relations between $\mathbf{E}$ and $\mathbf{H}$ in (1.92) are no longer valid for an extraordinary wave.

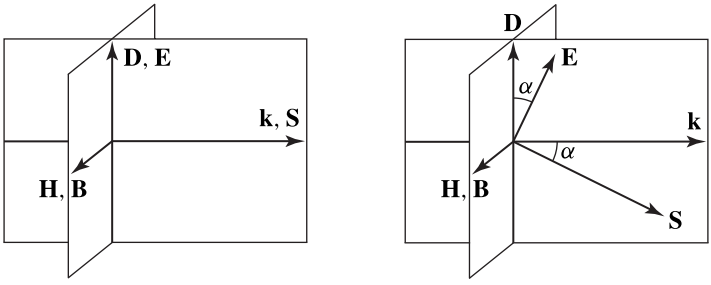


Figure 1.12 Relationships among the directions of **E**, **D**, **H**, **B**, **k**, and **S** in an anisotropic medium for (*a*) an ordinary wave and (*b*) an extraordinary wave. In both cases, the vectors **E**, **D**, **k**, and **S** lie in a plane normal to **H**.

If the electric field of an extraordinary wave is not parallel to a principal axis, its Poynting vector is not parallel to its propagation direction because $\mathbf{E}_e$ is not parallel to $\mathbf{D}_e$. As a result, its energy flows away from the direction of its wavefront propagation. This phenomenon is known as *spatial beam walk-off*. If this characteristic appears in one of the two normal modes of an optical wave propagating in an anisotropic crystal, the optical wave will split into two beams of parallel wavevectors but separate, nonparallel traces of energy flow.

For simplicity, let us consider the propagation of an optical wave in a uniaxial crystal with $\hat{k}$, for both ordinary and extraordinary waves, at an angle θ with respect to the optical axis $\hat{z}$. Clearly, there is no walk-off for the ordinary wave because $\mathbf{E}_o \parallel \mathbf{D}_o$ and $\mathbf{S}_o \parallel \hat{k}$. For the extraordinary wave, $\mathbf{S}_e$ is not parallel to $\hat{k}$ but points in a direction at an angle ψ_e with respect to the optical axis. Figure 1.13(*a*) shows the relationships among these vectors. The angle α between $\mathbf{S}_e$ and $\hat{k}$, which is defined as $\alpha = \psi_e - \theta$, is called the *walk-off angle* of the extraordinary wave. Note that α is also the angle between $\mathbf{E}_e$ and $\mathbf{D}_e$, as can be seen from Fig. 1.13(*a*). Because neither $\mathbf{E}_e$ nor $\mathbf{D}_e$ is parallel to any principal axis, their relationship is found through their projections on the principal axes: $D_z^e = n_e^2 \epsilon_0 E_z^e$ and $D_{xy}^e = n_o^2 \epsilon_0 E_{xy}^e$. Using these two relations and the definition of α in Figs. 1.12(*b*) and 1.13(*a*), it can be shown that the walk-off angle is given by (see Problems 1.6.14 and 1.6.15)

$$\alpha = \psi_e - \theta = \tan^{-1}\left(\frac{n_o^2}{n_e^2}\tan\theta\right) - \theta. \tag{1.131}$$

If the crystal is positive uniaxial, α as defined in Fig. 1.13(*a*) is negative. This means that $\mathbf{S}_e$ is between $\hat{k}$ and $\hat{z}$ for a positive uniaxial crystal. If the crystal is negative uniaxial, α is positive and $\hat{k}$ is between $\mathbf{S}_e$ and $\hat{z}$. No walk-off appears if an optical wave propagates along any of the principal axes of a crystal.

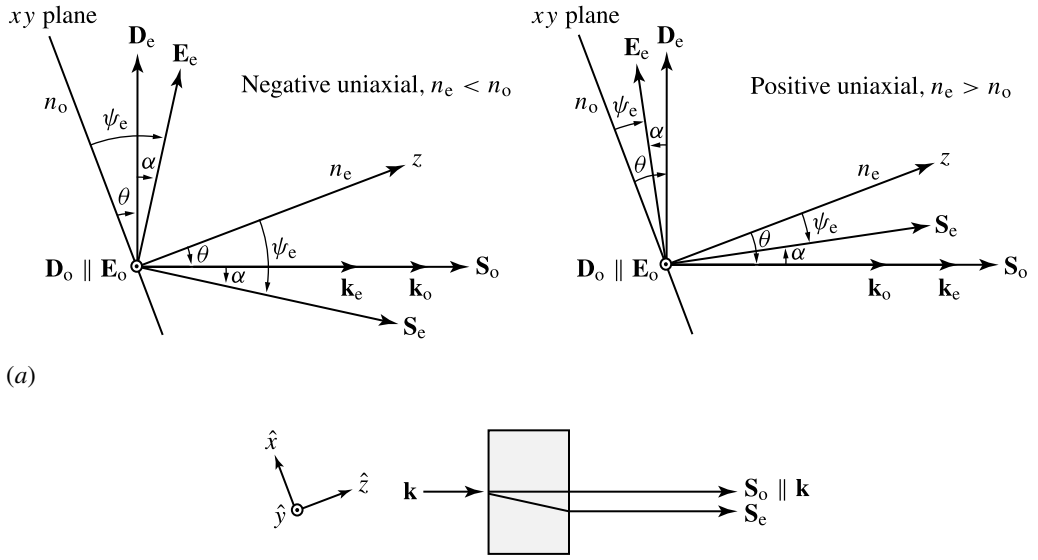


Figure 1.13 (*a*) Wave propagation and walk-off in a uniaxial crystal. (*b*) Birefringent plate acting as a polarizing beam splitter for a normally incident wave. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors indicate the principal axes of the birefringent plate.

A birefringent crystal can be used to construct a very simple *polarizing beam splitter* by taking advantage of the walk-off phenomenon. For such a purpose, a uniaxial crystal can be cut into a plate whose surfaces are at an oblique angle with respect to the optical axis, as is shown in Fig. 1.13(*b*). When an optical wave is normally incident upon the plate, it splits into ordinary and extraordinary waves in the crystal if its original polarization contains components of both polarizations. The extraordinary wave is then separated from the ordinary wave because of spatial walk-off, creating two orthogonally polarized beams. However, because of normal incidence, both $\mathbf{k}_e$ and $\mathbf{k}_o$ are parallel to the direction of $\hat{k}$ although they have different magnitudes. When both beams reach the other side of the plate, they are separated by a distance of $d = l \tan\alpha$, where l is the thickness of the plate. After leaving the plate, the two spatially separated beams propagate parallel to each other along the same direction $\hat{k}$ because the directions of their wavevectors have not changed, as is also shown in Fig. 1.13(*b*).

EXAMPLE 1.6 Find the spatial walk-off angle at 1 μm wavelength at a few representative propagation directions in KDP. Design a polarizing beam splitter at this wavelength using a KDP crystal.

Solution For a KDP crystal, $n_o = 1.51$ and $n_e = 1.47$ at 1 μm wavelength. The spatial walk-off angle α of an extraordinary wave is a function of the angle θ between the wave

propagation direction $\hat{k}$ and the optical axis $\hat{z}$ of the crystal. For example,

$$\alpha = \tan^{-1}\left(\frac{1.51^2}{1.47^2}\tan 30^\circ\right) - 30^\circ = 1.35^\circ, \qquad \text{for } \theta = 30^\circ,$$

$$\alpha = \tan^{-1}\left(\frac{1.51^2}{1.47^2}\tan 45^\circ\right) - 45^\circ = 1.54^\circ, \qquad \text{for } \theta = 45^\circ,$$

$$\alpha = \tan^{-1}\left(\frac{1.51^2}{1.47^2}\tan 60^\circ\right) - 60^\circ = 1.31^\circ, \qquad \text{for } \theta = 60^\circ.$$

From these numerical examples, we find that the walk-off angle does not vary monotonically with θ (see Problem 1.6.15).

A polarizing beam splitter can be made by cutting a KDP crystal at an angle, such as 45°, with respect to its optical axis for a parallel plate of thickness l. A beam at 1 μm wavelength that consists of a mix of extraordinary and ordinary polarizations is normally incident on the plate for $\theta = 45^\circ$ and $\alpha = 1.54^\circ$. Because the ordinary wave does not have walk-off, the Poynting vectors of the extraordinary and ordinary components of the beam separate at an angle of $\alpha = 1.54^\circ$. If a minimum spatial separation of $d = 100$ μm between the extraordinary and ordinary components is desired on the exit surface of the KDP plate, the minimum thickness of the plate has to be $l > d/\tan\alpha = 3.7$ mm.

Optical anisotropy and crystal symmetry

The optical anisotropy of a crystal depends on its structural symmetry. Crystals are classified into seven systems according to their symmetry. The linear optical properties of these seven systems are summarized in Table 1.2. Some important remarks regarding the relation between the optical properties and the structural symmetry of a crystal are made:

1. A cubic crystal need not have an isotropic structure although its linear optical properties are isotropic. For example, most III–V semiconductors, such as GaAs, InP, InAs, AlAs, etc., are cubic crystals with isotropic linear optical properties. Nevertheless, they have well-defined crystal axes, $\hat{a}$, $\hat{b}$, and $\hat{c}$. They are also polar semiconductors, which have anisotropic nonlinear optical properties.

Table 1.2 *Linear optical properties of crystals*

Crystal symmetry	Optical property
Cubic	Isotropic: $n_x = n_y = n_z$
Trigonal, tetragonal, hexagonal	Uniaxial: $n_x = n_y \neq n_z$
Orthorhombic, monoclinic, triclinic	Biaxial: $n_x \neq n_y \neq n_z$

2. Although the principal axes may coincide with the crystal axes in certain crystals, they are not the same concept. The crystal axes, denoted by $\hat{a}$, $\hat{b}$, and $\hat{c}$, are defined by the structural symmetry of a crystal, whereas the principal axes, denoted by $\hat{x}$, $\hat{y}$, and $\hat{z}$, are determined by the symmetry of $\boldsymbol{\epsilon}$. The principal axes of a crystal are orthogonal to one another, but the crystal axes are not necessarily so.

1.7 Gaussian beam

Because the wave equation governs optical propagation, the transverse field distribution pattern and its variation along the longitudinal propagation direction have to satisfy this equation in order for the wave to exist and to propagate. A well-defined field pattern that can remain unchanged as the wave propagates is called a *mode* of wave propagation. Such a transverse field pattern is known as a *transverse mode*. The optical modes that exist in a given medium are determined by the optical properties of the medium together with any boundary conditions imposed on the wave equation by the optical structures in the medium. Here we consider the optical modes in a homogeneous medium. Modes in waveguides and optical fibers are discussed in Chapters 2 and 3.

A monochromatic optical wave propagating in an isotropic, homogeneous medium is governed by the wave equation given in (1.82). Clearly, the monochromatic plane wave expressed in (1.58) is a solution of this wave equation. Therefore, plane waves are normal modes in an isotropic, homogeneous medium. They are not the only normal modes, however, as the wave equation governing wave propagation in such a medium has other normal-mode solutions. One such important set of modes is the *Gaussian modes*. Like plane waves, Gaussian modes are normal modes of wave propagation in an isotropic, homogeneous medium. Different from a plane wave, however, a Gaussian mode has a finite cross-sectional field distribution defined by its *spot size*. Being an unguided field with a finite spot size, a Gaussian mode differs from a waveguide mode, discussed in Chapters 2 and 3, in that its spot size varies along its longitudinal axis, taken to be the z axis, of propagation though its pattern remains unchanged. Therefore, its transverse field distribution also changes with z though the field pattern does not change. A Gaussian mode field at a frequency ω can thus be expressed as

$$\mathbf{E}_{mn}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_{mn}(x, y, z) \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t) = \hat{e}\mathcal{E}_{mn}(x, y, z) \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t), \quad (1.132)$$

with a corresponding field distribution for its magnetic field component, where m and n are mode indices associated with the two transverse dimensions x and y, respectively. A Gaussian mode field has neither longitudinal electric nor longitudinal magnetic field components. It is a *TEM mode* that has only transverse electric and magnetic field components. Normal modes are orthonormal to each other and can be normalized, as

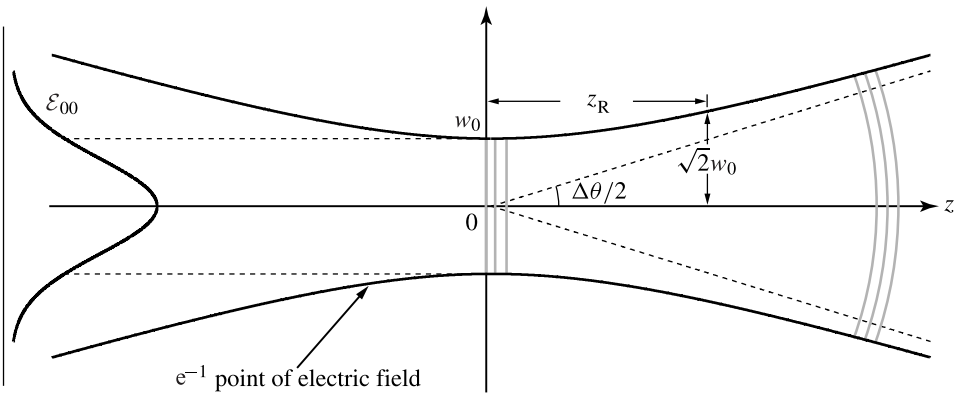


Figure 1.14 Gaussian beam characteristics.

is discussed in detail in Section 2.4. Gaussian modes are normalized by the following condition:

$$\frac{2k}{\omega\mu_0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\mathcal{E}_{mn}(x,y,z)|^2\mathrm{d}x\mathrm{d}y=1. \tag{1.133}$$

The location, taken to be $z=0$ for a beam propagating along the z axis, where the smallest spot size of the beam occurs, is known as the *waist* of a Gaussian beam. The minimum *Gaussian beam spot size*, w_0, is defined as the e^{-2} radius of the Gaussian beam intensity profile at the beam waist. The diameter of the beam waist is $d_0=2w_0$. As illustrated in Fig. 1.14, a Gaussian beam has a plane wavefront at its beam waist. The beam remains well collimated within a distance of

$$z_R=\frac{kw_0^2}{2}=\frac{\pi n w_0^2}{\lambda}, \tag{1.134}$$

known as the *Rayleigh range*, on either side of the beam waist. In (1.134), $k=2\pi n/\lambda$ is the propagation constant of the optical beam in a medium of refractive index n. The parameter $b=2z_R$ is called the *confocal parameter* of the Gaussian beam. Because of diffraction, a Gaussian beam diverges away from its waist and acquires a spherical wavefront. As a result, both its spot size, $w(z)$, and the radius of curvature, $\mathcal{R}(z)$, of its wavefront are functions of distance z from its beam waist:

$$w(z)=w_0\left(1+\frac{z^2}{z_R^2}\right)^{1/2}=w_0\left[1+\left(\frac{2z}{kw_0^2}\right)^2\right]^{1/2} \tag{1.135}$$

and

$$\mathcal{R}(z)=z\left(1+\frac{z_R^2}{z^2}\right)=z\left[1+\left(\frac{kw_0^2}{2z}\right)^2\right]. \tag{1.136}$$

We see from (1.135) that at $z = \pm z_R$, $w = \sqrt{2}w_0$. At $|z| \gg z_R$, far away from the beam waist, $\mathcal{R}(z) \approx z$ and $w(z) \approx 2|z|/kw_0$. Therefore, the far-field beam *divergence angle* is

$$\Delta\theta = 2\frac{w(z)}{|z|} = \frac{4}{kw_0} = \frac{2\lambda}{\pi n w_0}. \tag{1.137}$$

For the far field at $|z| \gg z_R$, we find that the beam spot size $w(z)$ is inversely proportional to the beam waist spot size w_0 but is linearly proportional to the distance $|z|$ from the beam waist. This characteristic does not exist for the near field at $|z| \leq z_R$.

A complete set of Gaussian modes includes the fundamental TEM_{00} mode and high-order TEM_{mn} modes. The specific forms of the mode fields depend on the transverse coordinates of symmetry: the mode fields are described by a set of Hermite–Gaussian functions in rectangular coordinates, whereas they are described by the Laguerre–Gaussian functions in cylindrical coordinates. Because there is no structurally determined symmetry in free space, either set is equally valid. Usually the Hermite–Gaussian functions in the rectangular coordinates are used. In a transversely isotropic and homogeneous medium, a normalized TEM_{mn} Hermite–Gaussian mode field propagating along the z axis can be expressed as

$$\begin{aligned}\hat{\mathcal{E}}_{mn}(x, y, z) &= \frac{A_{mn}}{w(z)} H_m\left[\frac{\sqrt{2}x}{w(z)}\right] H_n\left[\frac{\sqrt{2}y}{w(z)}\right] \exp\left[\mathrm{i}\frac{k}{2}\frac{x^2+y^2}{q(z)}\right] \exp\left[\mathrm{i}\zeta_{mn}(z)\right] \\ &= \frac{A_{mn}}{w(z)} H_m\left[\frac{\sqrt{2}x}{w(z)}\right] H_n\left[\frac{\sqrt{2}y}{w(z)}\right] \exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \exp\left[\mathrm{i}\frac{k}{2}\frac{x^2+y^2}{\mathcal{R}(z)}\right] \\ &\quad \times \exp\left[\mathrm{i}\zeta_{mn}(z)\right], \end{aligned} \tag{1.138}$$

where $A_{mn} = (\omega\mu_0/\pi k)^{1/2}(2^{m+n}\, m!\, n!)^{-1/2}$ is the normalization constant, H_m is the Hermite polynomial of order m, $q(z)$ is the complex radius of curvature of the Gaussian wave,

$$q(z) = z - \mathrm{i}z_R \quad \text{or} \quad \frac{1}{q(z)} = \frac{1}{\mathcal{R}(z)} + \mathrm{i}\frac{2}{kw^2(z)}, \tag{1.139}$$

and $\zeta_{mn}(z)$ is a mode-dependent on-axis phase variation along the z axis given by

$$\zeta_{mn}(z) = -(m+n+1)\tan^{-1}\frac{z}{z_R} = -(m+n+1)\tan^{-1}\left(\frac{2z}{kw_0^2}\right). \tag{1.140}$$

The Hermite polynomials can be obtained using the following relation:

$$H_m(\xi) = (-1)^m \mathrm{e}^{\xi^2}\frac{\mathrm{d}^m \mathrm{e}^{-\xi^2}}{\mathrm{d}\xi^m}. \tag{1.141}$$

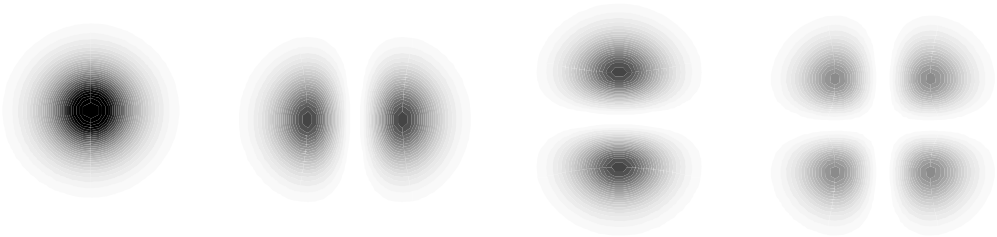


Figure 1.15 Intensity patterns of Hermite–Gaussian modes.

Some low-order Hermite polynomials are

$$H_0(\xi) = 1, \qquad H_1(\xi) = 2\xi, \qquad H_2(\xi) = 4\xi^2 - 2, \qquad H_3(\xi) = 8\xi^3 - 12\xi. \quad (1.142)$$

We see from (1.138) and (1.142) that the transverse field distribution $|\hat{\mathcal{E}}_{00}(x, y)|$ of the fundamental Gaussian mode, TEM_{00}, at any fixed longitudinal location z is simply a Gaussian function of the transverse radial distance $r = (x^2 + y^2)^{1/2}$ and that the spot size $w(z)$ is the e^{-1} radius of this Gaussian field distribution at z. The transverse field distribution of a high-order mode, TEM_{mn}, is the same Gaussian distribution spatially modulated by the Hermite polynomials H_m in x and H_n in y. As a result, its field distribution is more spread out radially than that of the fundamental TEM_{00} mode. In general, the higher the order of a mode, the farther its transverse field distribution spreads out. The intensity patterns of some Hermite–Gaussian modes are shown in Fig. 1.15.

EXAMPLE 1.7 A fundamental Gaussian beam in free space at the He–Ne laser wavelength of 632.8 nm has a spot size of $w_0 = 500$ μm at its beam waist. This beam has a Rayleigh range $z_R = \pi w_0^2/\lambda = 1.24$ m and a confocal parameter $b = 2z_R = 2.48$ m. Using (1.135) and (1.136), we find the following spot sizes and radii of curvature at a few different locations:

$w = 502$ μm, $\mathcal{R} = \pm 15.5$ m at $z = \pm 10$ cm,
$w = 642$ μm, $\mathcal{R} = \pm 2.54$ m at $z = \pm 1$ m,
$w \approx 40$ cm, $\mathcal{R} \approx \pm 1$ km at $z = \pm 1$ km.

From these numerical examples, we see that a Gaussian beam diverges very slowly, much like a plane wave, within the Rayleigh range on both sides of its beam waist. At the beam waist, a Gaussian beam has a plane wavefront with $\mathcal{R} = \infty$. At a distance much larger than the Rayleigh range on either side of the beam waist, a Gaussian beam approaches the characteristics of a spherical wave with $\mathcal{R} \approx z$. The Gaussian beam in this example has a far-field divergence angle of $\Delta\theta = 2\lambda/\pi w_0 = 0.8$ mrad.

1.8 Reflection and refraction

The characteristics of reflection and refraction of an optical wave at the interface of two different media depend on the properties of the media. We first consider the simple case of reflection and refraction at the planar interface of two dielectric media that are linear, lossless, and isotropic. In this situation, the permittivities ϵ_1 and ϵ_2 of the two media are constant real scalars, while the permeabilities are simply equal to μ_0 at optical frequencies. We assume that the optical wave is incident from medium 1 with a wavevector $\mathbf{k}_\mathrm{i}$, while the reflected wave has a wavevector $\mathbf{k}_\mathrm{r}$ and the transmitted wave has a wavevector $\mathbf{k}_\mathrm{t}$.

Because an optical wave varies with $\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r} - \mathrm{i}\omega t)$, the condition that

$$\mathbf{k}_\mathrm{i}\cdot\mathbf{r} = \mathbf{k}_\mathrm{r}\cdot\mathbf{r} = \mathbf{k}_\mathrm{t}\cdot\mathbf{r} \tag{1.143}$$

is required at the interface for the boundary conditions described by (1.17)–(1.20) to be satisfied at all points along the interface at all times. This implies that the three vectors $\mathbf{k}_\mathrm{i}$, $\mathbf{k}_\mathrm{r}$, and $\mathbf{k}_\mathrm{t}$ lie in the same plane known as the *plane of incidence*, as shown in Figs. 1.16 and 1.17. The projections of these three wavevectors on the interface are all equal so that

$$k_\mathrm{i}\sin\theta_\mathrm{i} = k_\mathrm{r}\sin\theta_\mathrm{r} = k_\mathrm{t}\sin\theta_\mathrm{t}, \tag{1.144}$$

where θ_i is the *angle of incidence*, and θ_r and θ_t are the angle of reflection and the angle of refraction, respectively, for the reflected and transmitted waves. All three angles are measured with respect to the normal $\hat{n}$ of the interface, as is shown in Figs. 1.16 and 1.17. Because $k_\mathrm{i} = k_\mathrm{r}$ and $k_\mathrm{i}/k_\mathrm{t} = n_1/n_2$, (1.144) yields the relation

$$\theta_\mathrm{i} = \theta_\mathrm{r} \tag{1.145}$$

and the following familiar *Snell law* for refraction:

$$n_1\sin\theta_\mathrm{i} = n_2\sin\theta_\mathrm{t}. \tag{1.146}$$

By expressing $\mathbf{H}$ in terms of $\mathbf{k}\times\mathbf{E}$ in the form of (1.86) with appropriate values of $\mathbf{k}$ for the incident, reflected, and refracted fields, the amplitudes of the reflected and transmitted fields can be obtained from the boundary conditions in (1.17) and (1.18). There are two different modes of field polarization.

TE polarization (s wave, σ wave)

The electric field is linearly polarized in a direction *perpendicular* to the plane of incidence while the magnetic field is polarized parallel to the plane of incidence, as

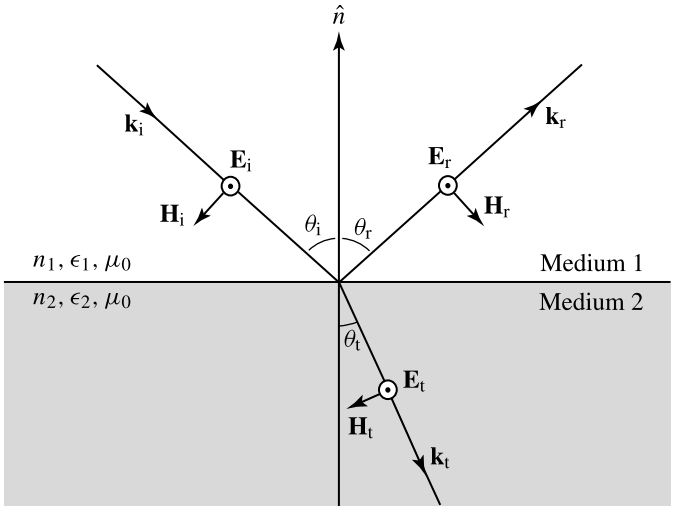


Figure 1.16 Reflection and refraction of a TE-polarized wave at the interface of two isotropic dielectric media. The three vectors $\mathbf{k}_i$, $\mathbf{k}_r$, and $\mathbf{k}_t$ lie in the plane of incidence. The relationship between θ_i and θ_t shown here is for the case $n_1 < n_2$.

shown in Fig. 1.16. This is called *transverse electric (TE) polarization* or *perpendicular polarization*. This wave is also called *s polarized*, or *σ polarized*. In this case, the *reflection coefficient*, r, and the *transmission coefficient*, t, of the electric field are given by the following *Fresnel equations*:

$$r_s \equiv \frac{\mathcal{E}_r}{\mathcal{E}_i} = \frac{n_1 \cos\theta_i - n_2 \cos\theta_t}{n_1 \cos\theta_i + n_2 \cos\theta_t} = \frac{n_1 \cos\theta_i - \sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}{n_1 \cos\theta_i + \sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}, \tag{1.147}$$

$$t_s \equiv \frac{\mathcal{E}_t}{\mathcal{E}_i} = \frac{2n_1 \cos\theta_i}{n_1 \cos\theta_i + n_2 \cos\theta_t} = \frac{2n_1 \cos\theta_i}{n_1 \cos\theta_i + \sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}, \tag{1.148}$$

respectively. The intensity *reflectance* and *transmittance*, R and T, which are also known as *reflectivity* and *transmissivity*, respectively, are given by

$$R_s \equiv \frac{I_r}{I_i} = \frac{|\overline{\mathbf{S}}_r \cdot \hat{n}|}{|\overline{\mathbf{S}}_i \cdot \hat{n}|} = \left| \frac{n_1 \cos\theta_i - n_2 \cos\theta_t}{n_1 \cos\theta_i + n_2 \cos\theta_t} \right|^2, \tag{1.149}$$

$$T_s \equiv \frac{I_t}{I_i} = \frac{|\overline{\mathbf{S}}_t \cdot \hat{n}|}{|\overline{\mathbf{S}}_i \cdot \hat{n}|} = 1 - R_s. \tag{1.150}$$

TM polarization (p wave, π wave)

The electric field is linearly polarized in a direction *parallel* to the plane of incidence while the magnetic field is polarized perpendicular to the plane of incidence, as shown

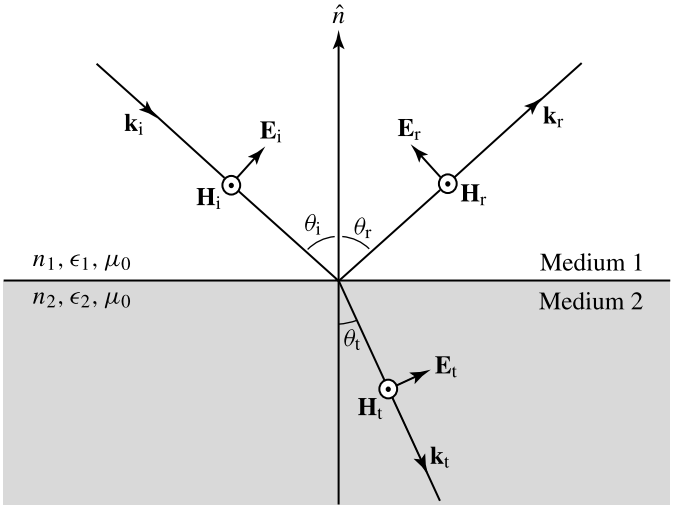


Figure 1.17 Reflection and refraction of a TM-polarized wave at the interface of two isotropic dielectric media. The three vectors $\mathbf{k}_i$, $\mathbf{k}_r$, and $\mathbf{k}_t$ lie in the plane of incidence. The relationship between θ_i and θ_t shown here is for the case $n_1 < n_2$.

in Fig. 1.17. This is called *transverse magnetic (TM) polarization* or *parallel polarization*. This wave is also called *p polarized*, or π *polarized*. In this case, the reflection and transmission coefficients of the electric field are given by the following Fresnel equations:

$$r_p \equiv \frac{\mathcal{E}_r}{\mathcal{E}_i} = \frac{n_2 \cos\theta_i - n_1 \cos\theta_t}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{n_2^2 \cos\theta_i - n_1\sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}{n_2^2 \cos\theta_i + n_1\sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}, \tag{1.151}$$

$$t_p \equiv \frac{\mathcal{E}_t}{\mathcal{E}_i} = \frac{2n_1 \cos\theta_i}{n_2 \cos\theta_i + n_1 \cos\theta_t} = \frac{2n_1 n_2 \cos\theta_i}{n_2^2 \cos\theta_i + n_1\sqrt{n_2^2 - n_1^2 \sin^2\theta_i}}, \tag{1.152}$$

respectively. The intensity reflectance and transmittance for TM polarization are given, respectively, by

$$R_p \equiv \frac{I_r}{I_i} = \left|\frac{n_2 \cos\theta_i - n_1 \cos\theta_t}{n_2 \cos\theta_i + n_1 \cos\theta_t}\right|^2, \tag{1.153}$$

$$T_p \equiv \frac{I_t}{I_i} = 1 - R_p. \tag{1.154}$$

Several important characteristics of the reflection and refraction of an optical wave at an interface between two media are summarized.

1. For both TE and TM polarizations, $R = |r|^2$ and $R + T = 1$, but $T \neq |t|^2$.
2. If $n_1 < n_2$, light is incident from a rare medium upon a dense medium. In this case, the reflection is called *external reflection*. If $n_1 > n_2$, light is incident from a dense medium on a rare medium, and the reflection is called *internal reflection*.
3. **Normal incidence.** In the case of normal incidence, $\theta_i = \theta_t = 0$. There is no difference between TE and TM polarizations, and

$$R = \left|\frac{n_1 - n_2}{n_1 + n_2}\right|^2, \quad T = 1 - R = \frac{4n_1 n_2}{(n_1 + n_2)^2}. \tag{1.155}$$

   For the case of external reflection at normal incidence, there is a 180° phase reversal for the reflected electric field with respect to the incident field. For internal reflection, the phase of the reflected field is not reversed at normal incidence. However, the values of R and T do not depend on which side of the interface the wave comes from.
4. **Brewster angle.** For a TE wave, R_s increases monotonically with the angle of incidence. For a TM wave, R_p first decreases then increases as the angle of incidence increases. For the interface between two lossless media, $R_p = 0$ at an angle of incidence $\theta_i = \theta_B$, where

$$\theta_B = \tan^{-1}\frac{n_2}{n_1} \tag{1.156}$$

   is known as the *Brewster angle*. When $\theta_i = \theta_B$, the angle of refraction for the transmitted wave is

$$\theta_t = \frac{\pi}{2} - \theta_B. \tag{1.157}$$

   It can be shown that this angle is the Brewster angle for the same wave incident from the other side of the interface. Figure 1.18(*a*) shows the reflectance of TE and TM waves as a function of the angle of incidence for external reflection at the interface between two media of refractive indices of 1 and 3.5. These characteristics are very useful in practical device applications: (a) at $\theta_i = \theta_B$, a TM-polarized wave is totally transmitted, resulting in a perfect lossless window for TM polarization – such windows are called *Brewster windows* and are useful as laser windows; (b) at $\theta_i = \theta_B$, the reflected wave is completely TE polarized – linearly polarized light can be produced by a *reflection-type polarizer* based on this principle.
5. **Critical angle.** In the case of internal reflection with $n_1 > n_2$, *total internal reflection* occurs if the angle of incidence θ_i is larger than the angle

$$\theta_c = \sin^{-1}\frac{n_2}{n_1}, \tag{1.158}$$

   which is called the *critical angle*. The reflectance of TE and TM waves as a function of angle of incidence for internal reflection at the interface between two media of

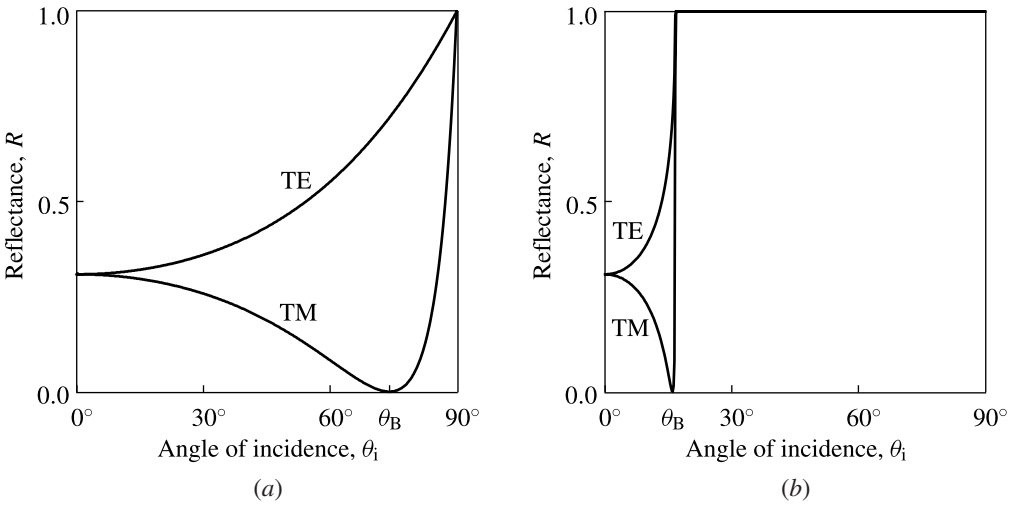


Figure 1.18 Reflectance of TE and TM waves at an interface of lossless media as a function of the angle of incidence for (*a*) external reflection and (*b*) internal reflection.

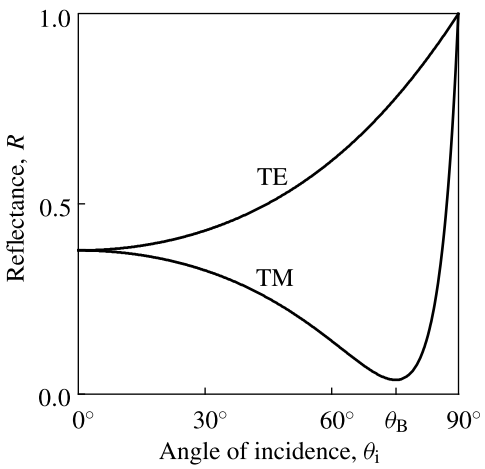


Figure 1.19 Reflectance of TE and TM waves at an interface of lossy or amplifying media as a function of the angle of incidence for external reflection.

refractive indices of 1 and 3.5 is shown in Fig. 1.18(*b*). Note that the Brewster angle for internal reflection is always smaller than the critical angle.

6. If one or both media have a loss or gain, the indices of refraction become complex. In this situation, the reflectance of the TM wave has a minimum that does not reach zero, as shown in Fig. 1.19 for external reflection.
7. For wave propagation in a general direction in an anisotropic medium, there are two normal modes that have different indices of refraction. The refracted fields of these two normal modes can propagate in different directions, resulting in the

phenomenon of *double refraction*. Meanwhile, the Poynting vector does not have to be in the plane of incidence.

8. Optical media are generally dispersive. Therefore, reflectance and transmittance, as well as the direction of the refracted wave, are generally frequency dependent.

EXAMPLE 1.8 Water has an index of refraction $n = 1.33$. The index of refraction of ordinary glass is approximately $n = 1.5$. For most semiconductors, such as Si, GaAs, and InP, the index of refraction is often in the range between 3 and 4, depending on the optical wavelength and the material. Here we take a nominal value of $n = 3.5$ for a semiconductor. Find the reflectivities at normal incidence, the Brewster angles, and the critical angles for these media at their interfaces with air.

Solution Using the formula given in (1.155) for the reflectivity at normal incidence, we find that $R = 0.02$ for water, $R = 0.04$ for ordinary glass, and R typically falls in the range of 0.3 and 0.32 for a semiconductor. Using (1.156) for the Brewster angle, we find that $\theta_B \approx 54^\circ$ for water, $\theta_B \approx 56^\circ$ for ordinary glass, and θ_B is typically around 74° for a semiconductor. Using (1.158) for the critical angle, we find that $\theta_c \approx 49^\circ$ for water, $\theta_c \approx 42^\circ$ for ordinary glass, and θ_c is around 17° for a semiconductor.

1.9 Phase velocity, group velocity, and dispersion

For a monochromatic plane optical wave traveling in the z direction, the electric field can be written as

$$\mathbf{E} = \boldsymbol{\mathcal{E}} \exp(\mathrm{i}kz - \mathrm{i}\omega t), \tag{1.159}$$

where $\boldsymbol{\mathcal{E}}$ is a constant vector independent of space and time. This represents a sinusoidal wave whose phase varies with z and t as

$$\varphi = kz - \omega t. \tag{1.160}$$

For a point of constant phase on the space- and time-varying field, $\varphi =$ constant and thus $k\mathrm{d}z - \omega \mathrm{d}t = 0$. If we track this point of constant phase, we find that it is moving with a velocity of

$$v_p = \frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\omega}{k}. \tag{1.161}$$

This is called the *phase velocity* of the wave. Note that the phase velocity is a function of optical frequency because the refractive index of a medium is a function of frequency. There is *phase-velocity dispersion* due to the fact that $\mathrm{d}n/\mathrm{d}\omega \neq 0$. In the case of *normal dispersion*, $\mathrm{d}n/\mathrm{d}\omega > 0$ and $\mathrm{d}n/\mathrm{d}\lambda < 0$; in the case of *anomalous dispersion*, $\mathrm{d}n/\mathrm{d}\omega < 0$ and $\mathrm{d}n/\mathrm{d}\lambda > 0$.

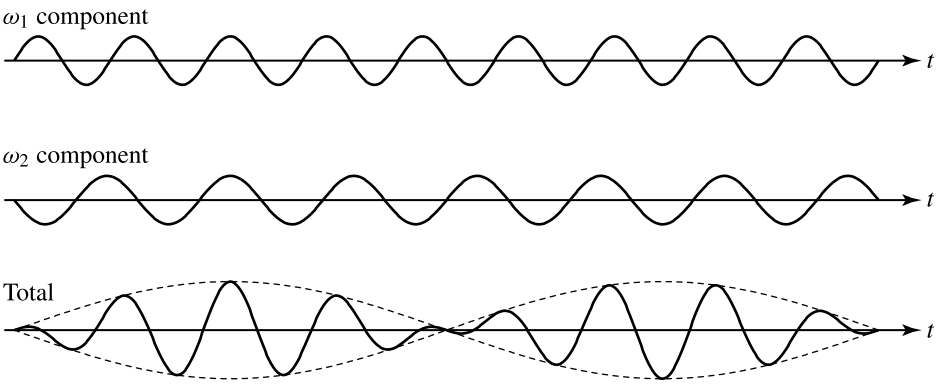


Figure 1.20 Wave packet composed of two frequency components showing the carrier and the envelope. The carrier travels at the phase velocity, whereas the envelope travels at the group velocity.

In real circumstances, a propagating optical wave rarely contains only one frequency. It usually consists of many frequency components grouped around some center frequency ω_0. For the simplicity of illustration, we consider a wave packet traveling in the z direction that is composed of two plane waves of equal real amplitude $\mathcal{E}$. The frequencies and propagation constants of the two component plane waves are

$$\begin{aligned} &\omega_1 = \omega_0 + \mathrm{d}\omega,\ k_1 = k_0 + \mathrm{d}k, \\ &\omega_2 = \omega_0 - \mathrm{d}\omega,\ k_2 = k_0 - \mathrm{d}k. \end{aligned} \tag{1.162}$$

The space- and time-dependent total real field of the wave packet is then given by

$$\begin{aligned} \boldsymbol{E} &= \mathcal{E}\exp(\mathrm{i}k_1 z - \mathrm{i}\omega_1 t) + \text{c.c.} + \mathcal{E}\exp(\mathrm{i}k_2 z - \mathrm{i}\omega_2 t) + \text{c.c.} \\ &= 2\mathcal{E}\{\cos[(k_0 + \mathrm{d}k)z - (\omega_0 + \mathrm{d}\omega)t] + \cos[(k_0 - \mathrm{d}k)z - (\omega_0 - \mathrm{d}\omega)t]\} \\ &= 4\mathcal{E}\cos(\mathrm{d}kz - \mathrm{d}\omega t)\cos(k_0 z - \omega_0 t). \end{aligned} \tag{1.163}$$

We find that the resultant wave packet has a *carrier*, which has a frequency ω_0 and a propagation constant k_0, and an *envelope*, which varies in space and time as $\cos(\mathrm{d}kz - \mathrm{d}\omega t)$. This is illustrated in Fig. 1.20. Therefore, a fixed point on the envelope is defined by $\mathrm{d}kz - \mathrm{d}\omega t = \text{constant}$, and it travels with a velocity of

$$v_\mathrm{g} = \frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}\omega}{\mathrm{d}k}. \tag{1.164}$$

This is the velocity of the wave packet and is called the *group velocity*. Because the energy of a harmonic wave is proportional to the square of its field amplitude, the energy carried by a wave packet that is composed of many frequency components is concentrated in regions where the amplitude of the envelope is large. Therefore, the energy in a wave packet is transported at group velocity v_g. *The constant-phase wavefront travels at the phase velocity, but the group velocity is the velocity at which energy and information travel.*

In reality, group velocity is usually a function of optical frequency. Then,

$$\frac{\mathrm{d}^2k}{\mathrm{d}\omega^2} = \frac{\mathrm{d}}{\mathrm{d}\omega}v_{\mathrm{g}}^{-1} \neq 0. \tag{1.165}$$

Therefore, $\mathrm{d}^2k/\mathrm{d}\omega^2$ represents *group-velocity dispersion*. A *dimensionless* coefficient for group-velocity dispersion can be defined as

$$D = c\omega\frac{\mathrm{d}^2k}{\mathrm{d}\omega^2} = \frac{2\pi c^2}{\lambda}\frac{\mathrm{d}^2k}{\mathrm{d}\omega^2}. \tag{1.166}$$

Group-velocity dispersion is an important consideration in the propagation of optical pulses. It can cause broadening of an individual pulse, as well as changes in the time delay between pulses of different frequencies. The sign of the group-velocity dispersion can be either positive or negative. In the case of *positive group-velocity dispersion*, $\mathrm{d}^2k/\mathrm{d}\omega^2 > 0$ and $D > 0$, a long-wavelength, or low-frequency, pulse travels faster than a short-wavelength, or high-frequency, pulse. In contrast, a short-wavelength pulse travels faster than a long-wavelength pulse in the case of *negative group-velocity dispersion*, $\mathrm{d}^2k/\mathrm{d}\omega^2 < 0$ and $D < 0$. In a given material, the sign of D generally depends on the spectral region of concern. Group-velocity dispersion and phase-velocity dispersion discussed earlier have different meanings. They should not be confused with each other.

When measuring the transmission delay or the broadening of optical pulses due to dispersion in optical fibers, another dispersion coefficient defined as

$$D_\lambda = -\frac{2\pi c}{\lambda^2}\frac{\mathrm{d}^2k}{\mathrm{d}\omega^2} = -\frac{D}{c\lambda} \tag{1.167}$$

is usually used. This coefficient is generally expressed as a function of wavelength in units of picoseconds per kilometer per nanometer. It is a direct measure of the chromatic pulse transmission delay over a unit transmission length.

In general, both $\epsilon(\omega)$ and $n(\omega)$ in an optical medium are frequency dependent, and the propagation constant is

$$k = \frac{\omega}{c}n(\omega). \tag{1.168}$$

Therefore, we have

$$v_{\mathrm{p}} = \frac{c}{n} \tag{1.169}$$

and

$$v_{\mathrm{g}} = \frac{c}{N}, \tag{1.170}$$

where

$$N = n + \omega\frac{\mathrm{d}n}{\mathrm{d}\omega} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \tag{1.171}$$

is called the *group index*. Using (1.166) and (1.167), we also have

$$D(\lambda) = \lambda^2 \frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \tag{1.172}$$

and

$$D_\lambda(\lambda) = -\frac{\lambda}{c} \frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2}, \tag{1.173}$$

respectively.

EXAMPLE 1.9 The index of refraction of a certain type of glass as a function of optical wavelength around $\lambda = 1.3$ μm can be approximated as $n = 1.465 - 0.0114(\lambda - 1.3) - 0.004(\lambda - 1.3)^3$, where λ is measured in micrometers. Therefore,

$$\frac{\mathrm{d}n}{\mathrm{d}\lambda} = -0.0114 - 0.012(\lambda - 1.3)^2,$$

$$N = n - \lambda \frac{\mathrm{d}n}{\mathrm{d}\lambda} = 1.48 + 0.0156(\lambda - 1.3)^2 + 0.008(\lambda - 1.3)^3,$$

$$D = \lambda^2 \frac{\mathrm{d}n^2}{\mathrm{d}\lambda^2} = -0.024\lambda^2(\lambda - 1.3).$$

We find that, in this spectral region, $\mathrm{d}n/\mathrm{d}\lambda < 0$ for any wavelength but $D > 0$ for $\lambda < 1.3$ μm and $D < 0$ for $\lambda > 1.3$ μm. Clearly, this glass has normal phase-velocity dispersion in the entire spectral region around $\lambda = 1.3$ μm, but it has positive group-velocity dispersion for $\lambda < 1.3$ μm and negative group-velocity dispersion for $\lambda > 1.3$ μm. As an example, we find that $n \approx 1.469$, $N \approx 1.481$, and $D \approx 0.0072$ at $\lambda = 1$ μm. We also find that $n \approx 1.463$, $N \approx 1.481$, and $D \approx -0.0108$ at $\lambda = 1.5$ μm. Because of normal phase-velocity dispersion, the group index is always larger than the refractive index, $N > n$, in this spectral region.

1.10 Material dispersion

As discussed in Sections 1.1 and 1.3, dispersion in the susceptibility of a medium is caused by the fact that the response of the medium to excitation by an optical field does not decay instantaneously. The general characteristics of the medium can be understood from its impulse response. In general, the impulse response of a medium decays exponentially while oscillating at some resonance frequencies. There may exist several exponential relaxation constants and several oscillation frequencies for a given material across the electromagnetic spectrum. This is true even within the optical spectral region. However, at a given optical frequency ω, the characteristics of the material response are dominated by the resonance frequency closest to ω and the relaxation constant associated with the oscillation at this particular resonance frequency. We therefore consider, for simplicity, a medium of a single resonance frequency at ω_0 with a relaxation

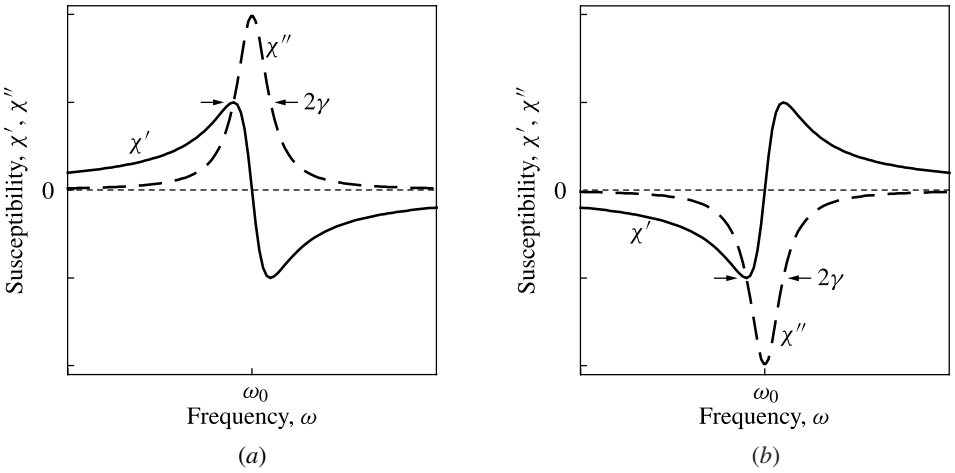


Figure 1.21 Real and imaginary parts, χ' and χ'', respectively, of susceptibility for a medium with (a) a loss and (b) a gain near a resonance frequency, ω_0.

constant γ. The susceptibility in the time domain is simply the impulse response of the medium, which is real and has the following general form:

$$\chi(t) \propto \begin{cases} e^{-\gamma t} \sin \omega_0 t, & t > 0, \\ 0, & t < 0. \end{cases} \tag{1.174}$$

Note that $\chi(t) = 0$ for $t < 0$ because a medium can respond only after, but not before, an excitation. This is the *causality* condition, which applies to all physical systems.

The Fourier transform of (1.174) yields

$$\chi(\omega) = \int_{-\infty}^{\infty} \chi(t) e^{i\omega t} dt \approx -\chi_b \frac{\omega_0}{\omega - \omega_0 + i\gamma} \tag{1.175}$$

in the frequency domain, where $\chi_b = \chi(\omega \ll \omega_0)$ is a constant equal to the background value of $\chi(\omega)$ at low frequencies far away from resonance. In (1.175), we have taken the so-called *rotating-wave approximation* by dropping a term that contains $\omega + \omega_0$ in its denominator because $\omega + \omega_0 \gg |\omega - \omega_0|$ in the optical spectral region (see Problem 1.10.1). This susceptibility has the following real and imaginary parts:

$$\chi'(\omega) = -\chi_b \frac{\omega_0(\omega - \omega_0)}{(\omega - \omega_0)^2 + \gamma^2}, \qquad \chi''(\omega) = \chi_b \frac{\omega_0 \gamma}{(\omega - \omega_0)^2 + \gamma^2}, \tag{1.176}$$

which are plotted in Fig. 1.21. Note that $\chi''(\omega)$ has a *Lorentzian lineshape*, which has a FWHM $\Delta\omega = 2\gamma$. The sign of χ'' depends on that of χ_b. In the normal state, $\chi_b > 0$, and the medium has an optical loss near resonance. This characteristic results in the

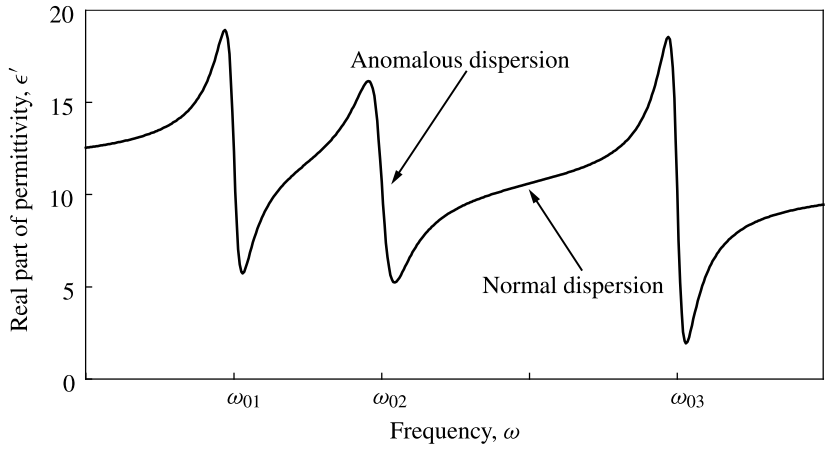
absorption of light at frequency $\omega = \omega_0$. When $\chi_b < 0$, the medium has optical gain, resulting in the amplification of light at $\omega = \omega_0$ such as in the case of a laser. Note that both χ' and χ'' are proportional to χ_b. Therefore, *when χ'' changes sign, χ' also changes sign.* When $\chi'' < 0$, χ' is negative for $\omega < \omega_0$ and positive for $\omega > \omega_0$, as is shown in Fig. 1.21(*b*).

EXAMPLE 1.10 For an atomic transition associated with absorption or emission of optical radiation at 1 μm wavelength, the resonance frequency is $\nu_0 = c/\lambda = 300$ THz, thus $\omega_0 = 2\pi\nu_0 = 1.885 \times 10^{15}$ s^{-1}. If the polarization associated with this resonant transition relaxes with a time constant of $\tau = 1$ ps, then $\gamma = 1/\tau = 10^{12}$ s^{-1} and $\Delta\omega = 2\gamma = 2 \times 10^{12}$ s^{-1}. Thus the Lorentzian spectral line has a FWHM linewidth of $\Delta\nu = \Delta\omega/2\pi \approx 318$ GHz, which is considered quite broad but is approximately only 0.1% of the center frequency ν_0 of the spectral line. If the relaxation time constant is $\tau = 1$ ns, we find a spectral linewidth of $\Delta\nu \approx 318$ MHz. For a relaxation time constant of $\tau = 1$ μs, we have a narrow linewidth of $\Delta\nu \approx 318$ kHz.

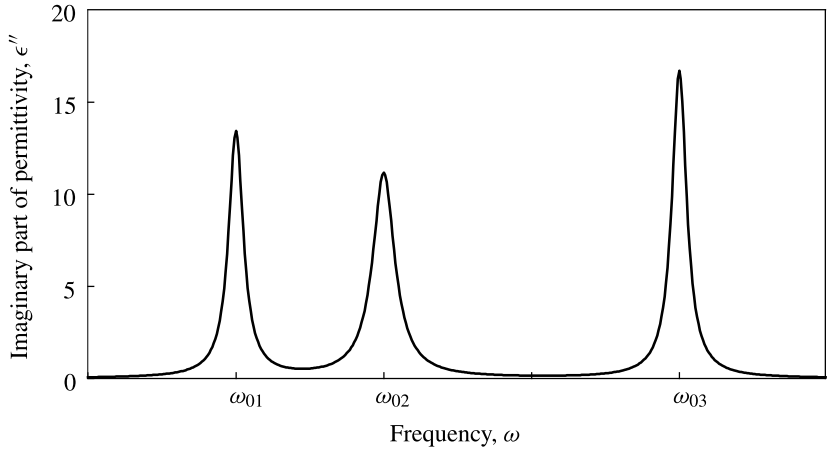
Note that the spectral linewidth is determined by the *polarization relaxation time* rather than by the *population relaxation time* of a material. The polarization relaxation time constant is generally much smaller than the population relaxation time constant for a given transition. Therefore, the spectral linewidth of a given transition can be quite broad even when the energy levels involved have long population relaxation times. One good example is the optical transitions in Nd : YAG discussed in Section 10.1.

A medium generally has many resonance frequencies, each corresponding to an absorption frequency in the normal state of the medium. Because $\epsilon(\omega) = \epsilon_0(1 + \chi(\omega))$, the dispersion characteristics of $\epsilon(\omega)$ depend directly on those of $\chi(\omega)$ given by (1.176). Its real and imaginary parts in the normal state as a function of ω over a spectral range covering a few resonances are shown in Fig. 1.22. Some important dispersion characteristics of $\chi(\omega)$ and $\epsilon(\omega)$ are summarized below.

1. It can be seen from Fig. 1.21(*a*) that $\chi'(\omega \ll \omega_0)$ is larger than $\chi'(\omega \gg \omega_0)$ in the normal state. Therefore, around any single resonance frequency, ϵ' at any frequency on the low-frequency side has a value larger than that at any frequency on the high-frequency side.
2. A medium is said to have *normal dispersion* in a spectral region where ϵ' increases with frequency so that $d\epsilon'/d\omega > 0$. It is said to have *anomalous dispersion* in a spectral region where ϵ' decreases with increasing frequency so that $d\epsilon'/d\omega < 0$. Because $dn/d\omega$ and $d\epsilon'/d\omega$ have the same sign, the index of refraction also increases with frequency in a spectral region of normal dispersion and decreases with frequency in a spectral region of anomalous dispersion.



(*a*)



(*b*)

Figure 1.22 Real and imaginary parts of ϵ as a function of ω for a medium in its normal state over a spectral range covering a few resonance frequencies.

3. It can be seen from Fig. 1.22 that when a material is in its normal state, normal dispersion appears everywhere except in the immediate neighborhood within the FWHM of a resonance frequency where anomalous dispersion occurs. This characteristic can be reversed near a resonance frequency where resonant amplification, rather than absorption, exists.
4. Note the distinction between the definition of normal and anomalous dispersion in terms of the sign of $d\epsilon'/d\omega$ or $dn/d\omega$ and that of positive and negative group-velocity dispersion in terms of the sign of D. Both positive and negative

group-velocity dispersion can appear in a spectral region where the dispersion defined in terms of $dn/d\omega$ is normal.[3]

5. In most transparent materials, such as glass and water, normal dispersion appears in the visible spectral region and may extend to the near infrared and near ultraviolet regions.

Kramers–Kronig relations

It can be seen from the discussions above that the real and imaginary parts of $\chi(\omega)$, or those of $\epsilon(\omega)$, are not independent of each other. The susceptibility of any physical system has to satisfy the causality requirement in the time domain. This requirement leads to a general relationship between χ' and χ'' in the frequency domain:

$$\chi'(\omega) = \frac{2}{\pi}\mathrm{P}\int_0^\infty \frac{\omega'\chi''(\omega')}{\omega'^2 - \omega^2}\mathrm{d}\omega', \qquad \chi''(\omega) = -\frac{2}{\pi}\mathrm{P}\int_0^\infty \frac{\omega\chi'(\omega')}{\omega'^2 - \omega^2}\mathrm{d}\omega', \tag{1.177}$$

where the principal values are taken for the integrals. These relations are known as the *Kramers–Kronig relations*. Therefore, once the real part of $\chi(\omega)$ is known over the entire spectrum, its imaginary part can be found, and vice versa. Note that the relations in (1.177) are consistent with the fact that $\chi'(\omega)$ is an even function, while $\chi''(\omega)$ is an odd function, of ω, as discussed in Section 1.3. The contradiction to this statement seen in (1.176) is only apparent but not real. It is caused by the rotating-wave approximation taken in (1.175). There is no contradiction when the approximation is removed and exact expressions are used for $\chi'(\omega)$ and $\chi''(\omega)$ (see Problem 1.10.1).

1.11 Photon nature of light

When considering the function of a device that involves the emission or absorption of light, a purely electromagnetic wave description of light is not adequate. In this situation, the photon nature of light cannot be ignored. Meanwhile, the material involved in this process also undergoes quantum mechanical transitions between its energy levels.

The energy of a photon is determined by its frequency ν or, equivalently, its angular frequency ω. Associated with the particle nature of a photon, there is a momentum

[3] In the literature, positive group-velocity dispersion is sometimes referred to as normal dispersion while negative group-velocity dispersion is referred to as anomalous dispersion. This is confusing and is, strictly speaking, not correct.

determined by its wavelength λ or, equivalently, its wavevector $\mathbf{k}$. These characteristics are summarized below for a photon in free space:

$$\begin{aligned} &\text{speed} && c = \lambda\nu, \\ &\text{energy} && h\nu = \hbar\omega = pc, \\ &\text{momentum} && p = \frac{h\nu}{c} = \frac{h}{\lambda}, \qquad \mathbf{p} = \hbar\mathbf{k}. \end{aligned}$$

The energy of a photon that has a wavelength λ in free space can be calculated using the following formula:

$$h\nu = \frac{1.2398}{\lambda}\ \mu\text{m eV} = \frac{1239.8}{\lambda}\ \text{nm eV}. \tag{1.178}$$

For example, at an optical wavelength of 1 μm, the photon energy is 1.2398 eV.

The energy of a photon is determined only by the frequency, or wavelength, of light, but not by its intensity. The intensity of light is related to the flux density, or number per unit time per unit area, of photons by

$$\text{photon flux density} = \frac{I}{h\nu} = \frac{I}{\hbar\omega}. \tag{1.179}$$

EXAMPLE 1.11 It is found that a piece of crystal transmits light at $\lambda = 500$ nm but absorbs light at $\lambda = 400$ nm. Make an intelligent guess of its bandgap from this limited information.

Solution Because a crystal transmits photons with energies below its bandgap but absorbs those with energies above its bandgap, we can reasonably guess that the bandgap of this crystal falls between the photon energies corresponding to 500 and 400 nm wavelengths. Using (1.178) for the photon energy, we find that

$$2.48\ \text{eV} < E_g < 3.10\ \text{eV}.$$

PROBLEMS

1.1.1 Verify that Maxwell's equations and the continuity equation are invariant under (a) space inversion, (b) time reversal, and (c) space inversion and time reversal simultaneously.

1.3.1 Verify the reality condition for electric susceptibility and electric permittivity given in (1.56) and (1.57), respectively.

1.4.1 Two polarizers placed in tandem along the line of propagation of an optical beam are called *cross polarizers* if their axes are arranged to be orthogonal to each

other. For the purpose of answering the following questions, consider polarizers of transmission type.

a. Show that no light of any polarization can pass through a set of cross polarizers.
b. A third polarizer is inserted in between the two cross polarizers. The transmission of this three-polarizer combination is not zero any more if the axis of the inserted polarizer is not parallel to either of the original two. Find the transmittance of this combination as a function of the angle between the axis of this polarizer and that of the polarizer at the input end.
c. Since each polarizer acts only as a polarization-sensitive filter to transmit the field component of a particular polarization, the phenomenon described in (b) may not seem possible. Can you give a physically intuitive explanation for it?

1.5.1 Express the wavenumber β and the attenuation coefficient α defined in (1.100) for propagation of an optical wave in an absorptive medium in terms of the real part, χ', and the imaginary part, χ'', of the electric susceptibility of the medium. Show that when $\chi'' \ll \chi'$, we have

$$\alpha \approx \beta \frac{\chi''}{n^2}. \tag{1.180}$$

1.5.2 The electric susceptibility of pure crystalline silicon at the optical wavelength of $\lambda = 532$ nm is $\chi = 15.48 + \mathrm{i}0.284$. An optical beam of 1 W power at 532 nm wavelength is normally incident from the air on the surface of a crystalline silicon wafer, which is polished to mirror finish. The surface on the other side of the silicon wafer is antireflection coated so that no reflection of light takes place on that surface.

a. How much light (in milliwatts) is reflected from the surface from which the light enters the silicon wafer? How much enters the silicon wafer?
b. How much of the light entering the wafer is transmitted from the other side if the thickness of the silicon wafer is 100 μm?
c. What is the thickness of the wafer if 1 mW of light is transmitted from the other side?

1.6.1 An optical isolator transmits light traveling in one direction and blocks its reflection traveling in the opposite direction. Show that isolation of light reflected from a plane mirror can be accomplished by using a combination of a polarizer and a quarter-wave plate with the axis of the quarter-wave plate set at 45° with respect to the transmission axis of the polarizer.

1.6.2 A polarizer and a half-wave plate can be used to make an attenuator of linearly polarized light. Sketch a diagram of how this can be achieved and then plot the output intensity of the system as a function of the angle between the axis of the wave plate and that of the polarizer.

1.6.3 A crystal has the following electric permittivity tensor in the (x, y, z) coordinate system:

$$\epsilon = \epsilon_0 \begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 2.13 & 0 \\ 0 & 0 & 2.02 \end{bmatrix}.$$

A linearly polarized optical wave that has a free-space wavelength $\lambda = 600$ nm is sent into the crystal. Find the wavelength of the wave in the crystal in each of the following arrangements.

a. The wave is polarized along $\hat{x}$ and propagates along $\hat{z}$.
b. The wave is polarized along $\hat{y}$ and propagates along $\hat{z}$.
c. The wave is polarized along $\hat{x}$ and propagates along $\hat{y}$.
d. The wave is polarized along $\hat{z}$ and propagates along $\hat{y}$.

1.6.4 When the electric permittivity of a crystal is measured at $\lambda = 1$ μm with respect to an arbitrary Cartesian coordinate system defined by $\hat{x}_1$, $\hat{x}_2$, and $\hat{x}_3$, it is found to be given by the following tensor:

$$\epsilon = \epsilon_0 \begin{bmatrix} 4.786 & 0 & 0.168 \\ 0 & 5.01 & 0 \\ 0.168 & 0 & 4.884 \end{bmatrix}.$$

a. Find the principal dielectric axes $\hat{x}$, $\hat{y}$, and $\hat{z}$ of the crystal and their corresponding principal indices of refraction.
b. Write down the equation that describes the index ellipsoid of the crystal in the original coordinate system. What is the equation for the index ellipsoid in the coordinate system defined by the principal axes?
c. Is the crystal uniaxial or biaxial? Find its optical axis if it is uniaxial or its optical axes if biaxial.
d. How do you arrange an optical wave to propagate in such a crystal so that the polarization of the wave remains unchanged throughout the entire path if the wave is linearly polarized? How about if the wave is circularly polarized?
e. Make a quarter-wave plate for the optical wave at $\lambda = 1$ μm. What is the thickness of the plate?

1.6.5 Under what condition can the polarization of an optical wave propagating in a birefringent crystal remain unchanged for any initial state of polarization and any distance of propagation?

1.6.6 Show that a linearly polarized wave can be converted into a circularly polarized wave by passing it through a quarter-wave plate, and vice versa. In converting a circularly polarized wave into a linearly polarized wave, how do you control

the direction of the linear polarization at the output? Design the arrangement in conducting such an experiment properly in terms of the orientation of the relevant axes and the direction of polarization.

1.6.7 How far must a linearly polarized wave at $\lambda = 1$ μm travel through a crystal that has $n_x = 1.55$ and $n_y = 1.52$ before its polarization is changed into each of the following states. In answering these questions, explain by showing the arrangements with sketches.

a. It is made circularly polarized.
b. It remains linearly polarized but with its polarization rotated by 90°.
c. It remains linearly polarized but with its polarization rotated by 60°.

1.6.8 Quartz is a positive uniaxial crystal, which has n_o=1.544 23 and n_e=1.553 32 at $\lambda = 600$ nm. A quartz plate is cut in such a way that its optical axis is parallel to the surfaces of the plate. A linearly polarized optical beam at 600 nm is sent to pass through such a quartz plate.

a. What is the thickness of a piece of quartz needed to change a linearly polarized beam into a circularly polarized beam at 600 nm wavelength? How should the quartz plate be arranged with respect to the polarization direction of the linearly polarized beam in order for this to happen?
b. What should the thickness of the quartz plate be to enable rotation of the linear polarization of the beam by 50°? How do you arrange the polarization direction with respect to the crystal axes in this case?
c. If instead we want to make sure that the linearly polarized beam stays linearly polarized in the same direction upon passing through the quartz plate irrespective of the polarization direction with respect to the optical axis of the quartz plate, what should the thickness of the plate be?

1.6.9 At what wavelength does a quarter-wave plate for $\lambda = 1$ μm function as a half-wave plate if the dispersion in the refractive indices of the plate is neglected? At what wavelength does light traveling through the plate always return to its input polarization state?

1.6.10 Quartz is a positive uniaxial crystal, which has $n_o = 1.544\,23$ and $n_e = 1.553\,32$ at $\lambda = 600$ nm.

a. Design a quartz waveplate to be used for rotating the polarization direction of a linearly polarized beam at 600 nm wavelength by 60°. Give the thickness of the plate and the arrangement of your setup.
b. If dispersion of the quartz plate can be neglected, at what other wavelengths can this plate be used as a polarization rotator for linearly polarized light?
c. Again, if dispersion can be neglected, at what optical wavelengths can this plate be used to convert a linearly polarized beam into a circularly polarized one?

d. Find the thickness of a plate that has the same function as the one found in (a) if it has to be thicker than 1 mm but thinner than 1.5 mm.

1.6.11 Rutile (TiO_2) is a uniaxial crystal. Its ordinary and extraordinary indices of refraction as a function of wavelength are given by

$$n_o^2 = 5.913 + \frac{0.2441}{\lambda^2 - 0.083}, \tag{1.181}$$

$$n_e^2 = 7.197 + \frac{0.3322}{\lambda^2 - 0.0843}, \tag{1.182}$$

where λ is in micrometers. A rutile plate of thickness l is cut in such a way that its surface normal is perpendicular to its optical axis.

a. If the plate is to be used as a first-order half-wave plate at an optical wavelength of 1 μm, what should its thickness l be? How do you arrange the plate with respect to the polarization of the incident beam if the polarization of a linearly polarized input beam is to be rotated 60° after passing through the plate?

b. With the thickness of the plate obtained in (a), find another wavelength at which the plate also functions as a half-wave plate. Find a wavelength at which it functions as a quarter-wave plate.

1.6.12 Consider wave propagation in a uniaxial crystal whose optical axis is $\hat{z}$.

a. By using the relationships among $\hat{k}$, $\hat{e}_o$, and $\hat{e}_e$ given in (1.121), verify that the unit vectors $\hat{e}_o$ and $\hat{e}_e$ are given by the expressions in (1.123) and (1.124), respectively.

b. By examining the index ellipsoid, show that $n_e(\theta)$ for the extraordinary wave is given by (1.125).

1.6.13 Explain why (1.118) is written in **E** whereas (1.126) is written in **D**. How would **D** be expressed for the wave that is described by (1.118)? Does it have the same form as (1.118)? Why? How would **E** be expressed for the wave that is described by (1.126)? Does it have the same form as (1.126)? Why?

1.6.14 Show that the walk-off angle as defined in Fig. 1.13(a) is given by (1.131). Given n_e and n_o for a uniaxial crystal, find the angle θ for the propagation direction $\hat{k}$ that results in the largest walk-off for an extraordinary wave.

1.6.15 An extraordinary optical wave propagates in a uniaxial crystal with its wavevector **k** making an angle θ with respect to the optical axis, $\hat{z}$, of the crystal. In the case when $\theta \neq 90°$, the Poynting vector, **S**, of the wave is not parallel to **k**. The angle α between **S** and **k** is the same as that between **E** and **D**.

a. Show that the vector **S** lies between **k** and the optical axis if the crystal is positive uniaxial, while **k** lies between **S** and $\hat{z}$ if it is negative uniaxial. What is the relationship among **E**, **D**, and $\hat{z}$ in either case?

b. Show that the walk-off angle given by (1.131) can also be expressed as

$$\alpha = \tan^{-1}\left[\frac{n_e^2(\theta)}{2}\left(\frac{1}{n_e^2} - \frac{1}{n_o^2}\right)\sin 2\theta\right], \tag{1.183}$$

where $n_e(\theta)$ is that given by (1.125).

c. Show that the maximum walk-off between **S** and **k** occurs at

$$\theta = \tan^{-1}\frac{n_e}{n_o} \tag{1.184}$$

for

$$\alpha = \tan^{-1}\frac{n_o}{n_e} - \tan^{-1}\frac{n_e}{n_o}. \tag{1.185}$$

1.6.16 Rutile (TiO_2) is a uniaxial crystal. Its ordinary and extraordinary indices of refraction as a function of wavelength are given by (1.181) and (1.182), respectively. A rutile plate of thickness l is cut in such a way that its surface normal is at an angle $\theta = 30^\circ$ with respect to its optical axis. If this plate is used as a polarizing beam splitter for normally incident light at $\lambda = 0.8$ μm, what is the separation between the two orthogonally polarized beams leaving the plate? If the spot size of a collimated incident beam is 100 μm in diameter, what is the minimum value of l for the two orthogonally polarized beams at the exit to be completely separated without spatial overlap?

1.7.1 The intensity profile of a fundamental Gaussian beam, whose field profile is given by (1.138) with $m = n = 0$, at any spatial location is a function of the transverse radial distance, $r = (x^2 + y^2)^{1/2}$, from the beam center and the longitudinal distance z from the beam waist.

a. Show that the intensity profile can be expressed as the following Gaussian function:

$$I(r, z) = I_0(z)\exp\left[-\frac{2(x^2 + y^2)}{w^2(z)}\right] = I_0(z)\exp\left[-\frac{2r^2}{w^2(z)}\right], \tag{1.186}$$

where $I_0(z)$ is the peak intensity of the beam at the longitudinal location z.

b. Express the power, P, of this Gaussian beam as a function of its peak intensity $I_0(z)$ and its spot size $w(z)$ at any location z.

c. Find the variation of the peak intensity $I_0(z)$ along the longitudinal axis of the beam by expressing it as a function of peak intensity I_0 at the beam waist and distance z from the beam waist.

1.7.2 A fundamental Gaussian laser beam of power $P = 1$ W at a wavelength of $\lambda = 532$ nm is focused to a small spot radius of $w_0 = 10$ μm at its beam waist. What is the peak intensity I_0 at the beam waist? What is the divergence angle

of the beam? What are its spot size and peak intensity at a distance of 1 m from the beam waist? If the spot size is reduced by half to $w_0 = 5$ μm at the beam waist, what are the changes of the peak intensities at the beam waist and at 1 m from the waist?

1.7.3 A circular aperture of radius a is placed in the path of a fundamental Gaussian beam with the center of the aperture located at the center of the beam. The Gaussian beam has a spot size w at the location of the aperture.

a. Find and plot the fraction of beam power transmitted through the aperture as a function of a and w.
b. What percentage of power is transmitted if the aperture has a radius equal to the beam spot size, $a = w$?
c. What is the minimum aperture diameter for the aperture to transmit at least 99% of the beam power?

1.7.4 A laser retroreflector was first placed on the lunar surface by the astronauts of the *Apollo 11* lunar landing mission in 1969. Similar retroreflectors were placed on different parts of the lunar surface by astronauts in later missions, including *Apollos 14* and *15*. These retroreflectors have since been used for precision lunar laser ranging to measure the distance between Earth and the Moon using nanosecond and picosecond laser pulses down to a precision of the order of 1 cm. The *Apollo 11* retroreflector consists of an array of 100 silica corner cubes in a 46 cm × 46 cm panel. Each corner cube has a diameter of 3.8 cm. The function of a corner cube is to reflect the light intercepted by it right back to the original direction from which the light comes without the need for critical alignment. The distance between the centers of Earth and the Moon is about 385 000 km, but the direct distance between their surfaces is shorter and is about the distance for light to travel in 1.25 s. In this problem, we consider a lunar ranging experiment using a telescope of 1.5 m diameter to collimate laser pulses of 350 ps duration at a wavelength of 532 nm from the second harmonic of a Nd : YAG laser. We assume that the laser beam has a fundamental Gaussian profile of waist spot size $w_0 = 0.5$ m at the aperture of the 1.5-m-diameter telescope. We also assume that each corner cube in the retroreflector reflects about 80% of the laser light it intercepts but adds a divergence of 14 μrad to the reflected beam due to diffraction. In answering the following questions, we first ignore the scattering, absorption, diffraction, and dispersion caused by the atmosphere.

a. What is the divergence angle of the out-going beam? What is the spot size of the beam on the Moon's surface?
b. If the laser beam is incident on the retroreflector with the beam center located at the center of the panel, what fraction of the laser energy is intercepted and reflected back by the retroreflector?

c. What is the spot size of the reflected beam on Earth? What fraction of the beam reflected by the retroreflector back to Earth is intercepted by the 1.5-m receiving aperture of the telescope?
d. What fraction of the energy in each laser pulse is finally received after making the round trip to the Moon and back? If we hope to detect at least one photon in each pulse, what is the minimum energy required of the original out-going pulse?
e. In reality, the effects of the atmosphere are significant and certainly cannot be ignored unless the entire station is located in space. In each pass, the atmosphere adds a divergence of about 18 μrad to the beam mainly due to dispersion and transmits only about 2% due to scattering and absorption. Answer questions (a)–(d) with the atmospheric effects accounted for.

1.7.5 The effect of sending a Gassian beam through a thin lens of focal length f can be described by the following relation:

$$\frac{1}{q'} = \frac{1}{q} - \frac{1}{f}, \tag{1.187}$$

where q and q' are the complex radii of curvature of the Gaussian beam immediately before and after the thin lens. The value of f can be either positive or negative for a positive or negative lense, respectively. A Gaussian beam of waist radius w_0 located at $z = 0$ is sent through a thin lens of focal length f located at $z = z_0$.

a. Show that the waist radius for the beam after passing through the lens is

$$w_0' = \frac{|f|}{[(z_0 - f)^2 + z_R^2]^{1/2}} w_0, \tag{1.188}$$

where z_R is the Rayleigh range of the incoming beam.
b. Show that the waist of the beam passing through the lens is located at

$$z = \frac{z_0^2(z_0 - f) + z_R^2(z_0 + f)}{(z_0 - f)^2 + z_R^2}. \tag{1.189}$$

c. How can the beam be best collimated? What are the waist radius and Rayleigh range of this optimally collimated beam?
d. If the lens is placed at the waist location of the incoming beam, what is the waist radius of the outgoing beam? Where is the waist located? What is the effect of the lens on the divergence of the beam?

1.8.1 Under what condition is an optical wave that is reflected from a dielectric surface completely polarized no matter whether the incident wave is polarized or not? What is its state of polarization?

1.8.2 A beam of circularly polarized light is incident from the air on the surface of an isotropic lossless dielectric material. The index of refraction of the dielectric material is unknown. However, it is found experimentally, by varying the angle of incidence, that the reflected light is linearly polarized when the angle of incidence is 60°. What is the index of refraction of the dielectric material? Explain what happens.

1.8.3 A reflection-type polarizer can be made simply with a glass plate. If the glass plate available has an index of refraction $n = 1.5$ at the wavelength of interest, what should the incident angle of the light be in order for the plate to function as a polarizer? Illustrate how this device should be used if the incident light is arbitrarily polarized.

1.8.4 During a sunny day on the equator when the sun rises at 6 a.m. and sets at 6 p.m., at what times is the sunlight reflected from the ocean surface most polarized?

1.8.5 When sunlight reflected from the surface of a body of water is viewed through linearly polarizing glass, the apparent glare from the water is reduced.

a. Upon which concept discussed in this chapter is this glare reduction based?

b. Suppose you have a beachfront house, and you want to use polarizing glass to reduce the glare from the sunlight reflected from the ocean. How should you orient the polarizing glass? (Should the glass block horizontally or vertically polarized light?)

c. For what angle of reflected sunlight will your polarizing glass be most effective? (Assume that the index of refraction of water is 1.33.)

1.8.6 The index of refraction of ordinary glass is $n = 1.5$.

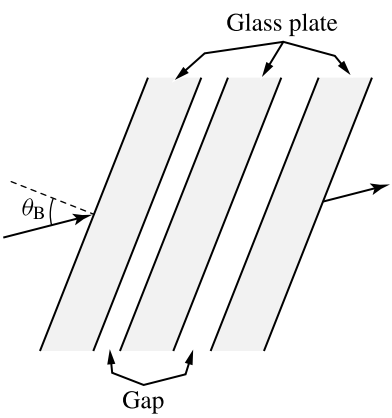


Figure 1.23 Stack of parallel flat glass plates.

a. Find the Brewster angles for the incidence of light from air to glass and from glass to air, respectively. What is the angle for total internal reflection?

b. For a stack of parallel flat glass plates separated by air gaps as shown in Fig. 1.23, show that if TM-polarized light is incident on the surface of the first plate at the Brewster angle, it is transmitted through the whole stack without reflection at any interface. Sketch the path of light through the stack. What are the effects of the thickness of the plates and that of the air gaps?

c. What happens if one air gap is filled with water whose index of refraction is 1.33? Illustrate by sketching the path of light.

1.8.7 The indices of refraction for water and diamond are 1.33 and 2.42, respectively.

a. For a piece of diamond exposed to the air, what are the critical angle for internal reflection, the Brewster angle for external reflection, and the reflectivities for TE and TM waves at an incident angle of 45°?

b. Answer the same question for a piece of diamond that is immersed in water.

1.8.8 A 90° symmetric prism with antireflection coating at the input surface as shown in Fig 1.24 can be used as a retroreflector. This kind of prism can losslessly reflect light with an adjustable lateral displacement between the paths of the incident and reflected beams.

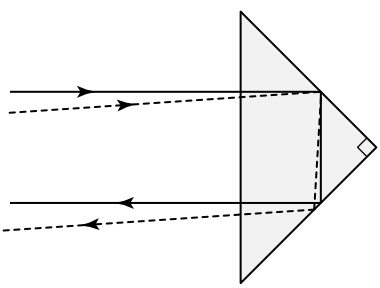


Figure 1.24 Prism retroreflector.

a. Show that the path of the reflected beam is parallel to that of the input beam for both normal and oblique incidence, thus requiring no critical alignment.

b. However, if the angle of incidence is too large, the reflected beam will suffer losses. What is the condition for a retroreflecting prism to have an angular tolerance of $\pm 5^\circ$ with respect to normal without substantial loss?

1.8.9 At the optical wavelength of 500 nm, GaAs is measured to have a reflectivity of 40% at normal incidence and an absorption coefficient of $\alpha = 10^7\ \mathrm{m}^{-1}$.

a. What is the complex refractive index of GaAs at 500 nm? What is the complex susceptibility?

b. Plot the reflectivity of GaAs at 500 nm as a function of incident angle for both TE and TM polarizations. What is the lowest reflectivity for the TM polarization? At what incident angle does it occur?

1.9.1 Explain how the primary rainbow is formed and describe the sequence of the rainbow colors from top to bottom. Answer the same questions for the secondary rainbow and explain the differences between the primary and secondary rainbows. Explain also why a rainbow has the shape of an arc. Find the arc angles for the primary and secondary rainbows.

1.9.2 The $LiNbO_3$ crystal is negative uniaxial. Its indices of refraction for the ordinary and extraordinary waves at room temperature as a function of optical wavelength are given by the following Sellmeier equations:

$$n_o^2 = 4.9130 + \frac{0.1188}{\lambda^2 - 0.045\,97} - 0.0278\lambda^2, \tag{1.190}$$

$$n_e^2 = 4.5798 + \frac{0.0994}{\lambda^2 - 0.042\,35} - 0.0224\lambda^2, \tag{1.191}$$

where λ is in micrometers. For both ordinary and extraordinary waves at an optical wavelength of 1.3 μm, find (a) the phase velocities, (b) the group velocities, and (c) the group-velocity dispersion parameters.

1.9.3 The BBO crystal is negative uniaxial. Its indices of refraction for the ordinary and extraordinary waves at room temperature as a function of optical wavelength are given by the following Sellmeier equations:

$$n_o^2 = 2.7359 + \frac{0.018\,78}{\lambda^2 - 0.018\,22} - 0.013\,54\lambda^2, \tag{1.192}$$

$$n_e^2 = 2.3753 + \frac{0.012\,24}{\lambda^2 - 0.016\,67} - 0.015\,16\lambda^2, \tag{1.193}$$

where λ is in micrometers. For both ordinary and extraordinary waves in the optical wavelength range between 0.5 and 2.0 μm, plot (a) phase velocity, (b) group velocity, and (c) group-velocity dispersion, as a function of wavelength.

1.10.1 Find the exact $\chi(\omega)$ corresponding to $\chi(t)$ given in (1.174) without making the rotating-wave approximation used in (1.175). Show that the real and imaginary parts of this exact $\chi(\omega)$ are even and odd functions of ω, respectively. Compare them with their respective approximate expressions in (1.176) to justify the applicability of the latter. Show that the exact expression for $\chi(\omega)$ satisfies the reality condition, as expected.

1.10.2 A material has two closely spaced resonance frequencies at ω_{01} and ω_{02} with respective response constants χ_{b1} and χ_{b2} and relaxation constants γ_1 and γ_2. The condition $0 \ll \omega_{02} - \omega_{01} \ll \omega_{01}$ is always valid in this problem.

a. Consider the case when $\chi_{b1} = \chi_{b2}$ and $\gamma_1 = \gamma_2 = \omega_{02} - \omega_{01}$. Sketch the real and imaginary parts of $\chi(\omega)$ as a function of ω near the two closely spaced

resonance frequencies. Also indicate the regions of normal and anomalous dispersion.

b. What changes to the dispersion of a laser material do you expect when its resonance at ω_{01} is pumped to population inversion but not that at ω_{02}, meaning that χ_{b1} changes sign but χ_{b2} does not? Sketch the real and imaginary parts of $\chi(\omega)$ as a function of ω near the two closely spaced resonance frequencies in this situation.

c. Sketch the real and imaginary parts of $\chi(\omega)$ as a function of ω near the two closely spaced resonance frequencies in the situation when population inversion occurs at both resonances so that both χ_{b1} and χ_{b2} change sign. Indicate the regions of normal and anomalous dispersion.

d. Answer questions (a)–(c) for the case when $\chi_{b1} = 3\chi_{b2}$ but $\gamma_1 = \gamma_2/3 = \omega_{02} - \omega_{01}$.

1.11.1 What is the separation in energy between the two energy levels that are responsible for emission at $\lambda = 1.064$ μm of a Nd : YAG laser?

1.11.2 A ruby is basically crystalline Al_2O_3 doped with Cr^{3+} impurities. Its red color is caused by the fact that the Cr^{3+} ions emit light at 694.3 nm when making the transition from an excited state to the ground state. What is the energy level of this excited state?

1.11.3 Silicon has a bandgap of 1.12 eV at room temperature. What wavelengths in the optical spectrum are transmitted through a pure silicon wafer?

1.11.4 GaAs has an energy bandgap of 1.424 eV at room temperature and absorbs any photon that has an energy higher than this value. For what optical wavelengths is GaAs transparent?

1.11.5 Consider monochromatic light illuminating a photographic film. The incident photons will be recorded if they have enough energy to dissociate the AgBr molecules in the film. The minimum energy required to do this is about 0.6 eV. Find the cutoff wavelength longer than which the incident light will not be recorded. In what spectral region does this wavelength fall?

1.11.6 A photon of 10.6 μm wavelength is combined with a photon of 1.06 μm wavelength to create a photon that combines the energies of both. What is the wavelength of the resultant photon?

SELECT BIBLIOGRAPHY

Born, M. and Wolf, E., *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th edn. Cambridge: Cambridge University Press, 1999.

Fowler, G. R., *Introduction to Modern Optics*, 2nd edn. New York: Dover, 1975.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Iizuka, K., *Elements of Photonics in Free Space and Special Media*, Vol. I. New York: Wiley, 2002.

Jackson, J. D., *Classical Electrodynamics*, 3rd edn. New York: Wiley, 1999.
Landau, L. D. and Lifshitz, E. M., *Electrodynamics of Continuous Media*. Oxford: Pergamon, 1960.
Nye, J. F., *Physical Properties of Crystals*. London: Oxford University Press, 1957.
Parker, S. P., *Optical Source Book*. New York: McGraw-Hill, 1987.
Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.
Sirotin, Yu. I. and Shaskolskaya, M. P., *Fundamentals of Crystal Physics*. Moscow: Mir Publishers, 1982.
Yariv, A. and Yeh, P., *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: Wiley, 1984.

Part II

Waveguides and couplers

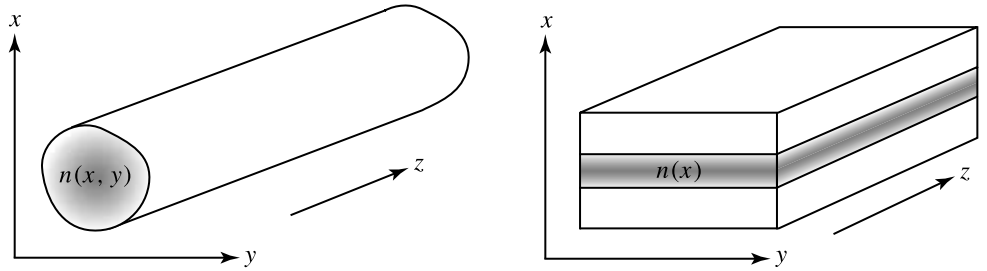
2 Optical waveguides

Optical waveguides are the basic elements for confinement and transmission of light over various distances, ranging from tens or hundreds of micrometers in integrated photonics to hundreds or thousands of kilometers in long-distance fiber-optic transmission. They are used to connect various photonic devices. In many devices, they form important parts or key structures, such as the waveguides providing optical confinement in semiconductor lasers. Furthermore, they form important active or passive photonic devices by themselves, such as waveguide couplers and modulators. In this chapter, we consider the basic characteristics of linear, lossless dielectric waveguides. Optical fibers are discussed in Chapter 3. Other waveguide devices are discussed in later chapters.

2.1 Waveguide modes

The basic structure of a dielectric optical waveguide consists of a longitudinally extended high-index optical medium, called the *core*, which is transversely surrounded by low-index media, called the *cladding*. A guided optical wave propagates in the waveguide along its longitudinal direction. We consider a straight waveguide whose longitudinal direction is taken to be the z direction, as shown in Fig. 2.1(a). The characteristics of a waveguide are determined by the transverse profile of its dielectric constant $\epsilon(x, y)/\epsilon_0$, which is independent of the z coordinate. For a waveguide made of optically isotropic media, we can simply characterize the waveguide with a single spatially dependent transverse profile of the index of refraction, $n(x, y)$.

In a *nonplanar waveguide* of two-dimensional transverse optical confinement, the core is surrounded by cladding in all transverse directions, and $n(x, y)$ is a function of both x and y coordinates. The *channel waveguides*, discussed in Section 2.8, and the circular optical fibers, discussed in Chapter 3, are such waveguides. In a *planar waveguide* that has optical confinement in only one transverse direction, the core is sandwiched between cladding layers in only one direction, say the x direction, with an index profile $n(x)$, as shown in Fig. 2.1(b). The core of a planar waveguide is also called the *film*, while the upper and lower cladding layers are called the *cover* and the



(*a*) (*b*)

Figure 2.1 (*a*) Nonplanar waveguide of two-dimensional transverse optical confinement. (*b*) Planar waveguide of one-dimensional transverse optical confinement.

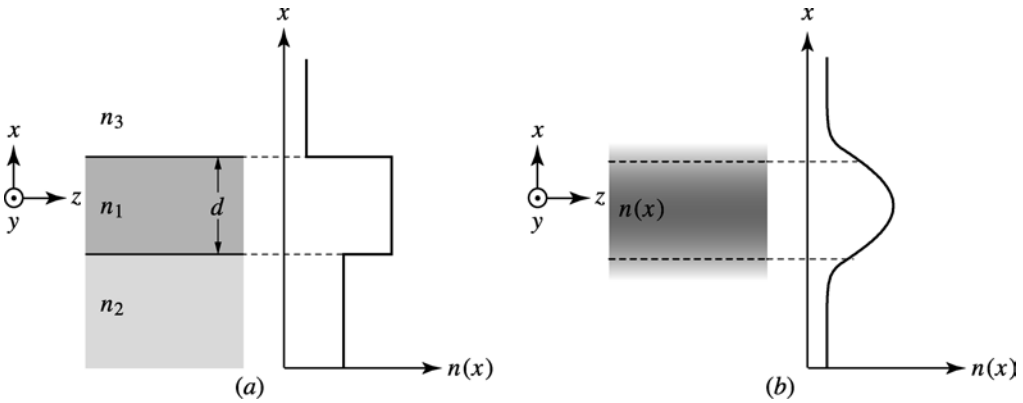


Figure 2.2 Index profiles of (*a*) a step-index planar waveguide and (*b*) a graded-index planar waveguide.

substrate, respectively. Optical confinement is provided only in the x direction by the planar waveguide shown in Fig. 2.1(*b*).

A waveguide in which the index profile has abrupt changes between the core and the cladding is called a *step-index waveguide*, while one in which the index profile varies gradually is called a *graded-index waveguide*. Figure 2.2 shows examples of step-index and graded-index planar waveguides.

Waveguide *modes* exist that are characteristic of a particular waveguide structure. A waveguide mode is a transverse field pattern whose amplitude and polarization profiles remain constant along the longitudinal z coordinate. Therefore, the electric and magnetic fields of a mode can be written in the following form:

$$\mathbf{E}_\nu(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_\nu(x, y) \exp(\mathrm{i}\beta_\nu z - \mathrm{i}\omega t), \tag{2.1}$$

$$\mathbf{H}_\nu(\mathbf{r}, t) = \boldsymbol{\mathcal{H}}_\nu(x, y) \exp(\mathrm{i}\beta_\nu z - \mathrm{i}\omega t), \tag{2.2}$$

where ν is the *mode index*, $\boldsymbol{\mathcal{E}}_\nu(x, y)$ and $\boldsymbol{\mathcal{H}}_\nu(x, y)$ are the mode field profiles, and β_ν is the propagation constant of the mode. For a waveguide of two-dimensional transverse

optical confinement, there are two degrees of freedom in the transverse xy plane, and the mode index ν consists of two parameters for characterizing the variations of the mode fields in these two transverse dimensions. For example, ν represents two mode numbers, $\nu = mn$ with integral m and n, for discrete guided modes. For the planar waveguide shown in Fig. 2.1(b), the mode fields do not depend on the y coordinate. Thus, (2.1) and (2.2) are reduced to

$$\mathbf{E}_\nu(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_\nu(x) \exp(\mathrm{i}\beta_\nu z - \mathrm{i}\omega t), \tag{2.3}$$

$$\mathbf{H}_\nu(\mathbf{r}, t) = \boldsymbol{\mathcal{H}}_\nu(x) \exp(\mathrm{i}\beta_\nu z - \mathrm{i}\omega t), \tag{2.4}$$

respectively. In this case, ν consists of only one parameter characterizing the field variation in the x dimension.

To get a general idea of the modes of a dielectric waveguide, it is instructive to consider the qualitative behavior of an optical wave in the asymmetric planar step-index waveguide shown in Fig. 2.2(a), where $n_1 > n_2 > n_3$. For an optical wave of angular frequency ω and free-space wavelength λ, the media in the three different regions of the waveguide define the following propagation constants:

$$k_1 = n_1 \frac{\omega}{c}, \qquad k_2 = n_2 \frac{\omega}{c}, \qquad \text{and} \qquad k_3 = n_3 \frac{\omega}{c}, \tag{2.5}$$

where $k_1 > k_2 > k_3$.

An intuitive picture can be obtained from studying ray optics by considering the path of an optical ray, or a plane optical wave, in the waveguide, as shown in the central column of Fig. 2.3. There are two critical angles associated with the internal reflections at the lower and upper interfaces:

$$\theta_{c2} = \sin^{-1} \frac{n_2}{n_1} \qquad \text{and} \qquad \theta_{c3} = \sin^{-1} \frac{n_3}{n_1}, \tag{2.6}$$

respectively. We see that $\theta_{c2} > \theta_{c3}$ because $n_2 > n_3$. The characteristics of the reflection and refraction of the ray at the interfaces depend on the angle of incidence θ and the polarization of the wave.

1. **Guided modes.** If $\theta > \theta_{c2} > \theta_{c3}$, the wave inside the core is totally reflected at both interfaces and is trapped by the core, resulting in *guided modes*. As the wave is reflected back and forth between the two interfaces, it interferes with itself. A guided mode can exist only when a transverse resonance condition is satisfied so that the repeatedly reflected wave has constructive interference with itself. In the core region, the x component of the wavevector is $k_1 \cos\theta$ for a ray with an angle of incidence θ, while the z component is $\beta = k_1 \sin\theta$. The phase shift in the optical field caused by a round-trip transverse passage in the core of thickness d is $2k_1 d \cos\theta$. In addition, there are phase shifts φ_2 and φ_3 associated with the internal reflections at the lower and upper interfaces, respectively. These phase shifts can be obtained from the phase angle of r_s in (1.147) for a TE wave and that of r_p in (1.151) for a TM wave

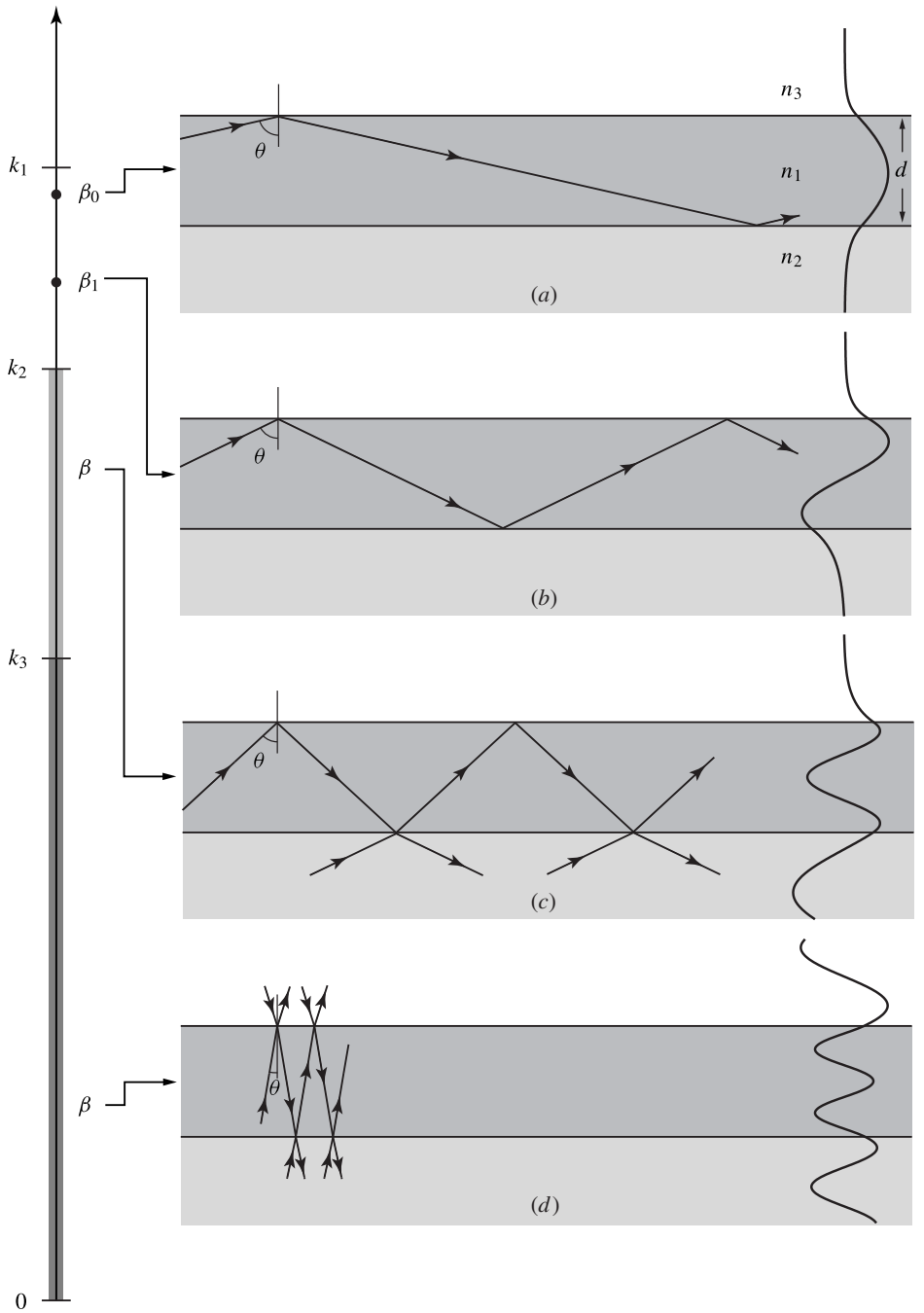


Figure 2.3 Modes of an asymmetric planar step-index waveguide where $n_1 > n_2 > n_3$. The range of the propagation constants, the zig-zag ray pictures, and the field patterns are shown correspondingly for (*a*) the guided fundamental mode, (*b*) the guided first high-order mode, (*c*) a substrate radiation mode for $\beta = 1.3k_3$, and (*d*) a substrate–cover radiation mode for $\beta = 0.3k_3$. The waveguide structure is chosen so that it supports only two guided modes. The mode field profiles are calculated mode field distributions that are normalized to their respective peak values.

for a given $\theta_i = \theta > \theta_{c2}, \theta_{c3}$. Because φ_2 and φ_3 are functions of θ, the transverse resonance condition for constructive interference in a round-trip transverse passage is

$$2k_1 d \cos\theta + \varphi_2(\theta) + \varphi_3(\theta) = 2m\pi, \tag{2.7}$$

where m is an integer. Because m can assume only integral values, only certain discrete values of θ can satisfy (2.7). This results in discrete values of the propagation constant β_m for guided modes identified by the mode number m. The guided mode with $m = 0$ is called the *fundamental mode* and those with $m \neq 0$ are *high-order modes*. Although the critical angles, θ_{c2} and θ_{c3}, do not depend on the polarization of the wave, the phase shifts, $\varphi_2(\theta)$ and $\varphi_3(\theta)$, caused by internal reflection at a given angle θ depend on the polarization. Therefore, TE and TM waves have different solutions for (2.7), resulting in different β_m and different mode characteristics for a given mode number m. For a given polarization, solution of (2.7) yields a smaller value of θ and a correspondingly smaller value of β for a larger value of m. Therefore, β_0 for the fundamental mode has the largest value among the allowed values for β, and $\beta_0 > \beta_1 > \ldots$, as shown in Figs. 2.3($a$) and ($b$).

2. **Substrate radiation modes.** When $\theta_{c2} > \theta > \theta_{c3}$, total reflection occurs only at the upper interface but not at the lower interface. As a result, an optical wave incident from either the core or the substrate can be refracted at the lower interface. This wave is not confined to the core, but is transversely extended to infinity in the substrate. It is called a *substrate radiation mode*. In this case, the angle θ is not dictated by a resonance condition like (2.7) but can assume any value in the range of $\theta_{c2} > \theta > \theta_{c3}$. As a result, the allowed values of β form a continuum between k_2 and k_3, and the modes are not discrete. These characteristics of a substrate radiation mode are illustrated in Fig. 2.3(c).
3. **Substrate–cover radiation modes.** When $\theta_{c2} > \theta_{c3} > \theta$, there is no total reflection at either interface. In this case, an optical wave incident from either side is refracted at both interfaces, and it can transversely extend to infinity on both sides of the waveguide, resulting in *substrate–cover radiation modes*. These modes are not discrete, and the allowed values of β for them form a continuum between k_3 and 0. These characteristics of a substrate–cover radiation mode are illustrated in Fig. 2.3(d).

In addition to the three types of modes discussed above, there are also *evanescent radiation modes*, which have purely imaginary values of β that are not discrete. Their fields decay exponentially along the z direction. Because the waveguide is lossless and does not absorb energy, the energy of an evanescent mode radiates away from the waveguide transversely. A lossless waveguide cannot generate energy, either. Therefore, the evanescent modes do not exist in perfect, longitudinally infinite waveguides. They exist at the longitudinal junctions or imperfections of a waveguide, as well as at the terminations of a realistic waveguide of finite length. In comparison, a substrate

radiation mode or a substrate–cover radiation mode has a real β; hence, its energy does not decay as it propagates. In such a radiation mode, the power flowing away from the center of the waveguide in the transverse direction is equal to that flowing toward the center.

The approach of ray optics gives a very intuitive picture of the waveguide modes and their key characteristics. Nevertheless, this approach has many limitations. In more complicated waveguide geometries such as that of a circular fiber, the idea of using the resonance condition based on the total internal reflection to find the allowed values of β for the guided modes does not necessarily yield correct results.[1] For a complete description of the waveguide fields, rigorous analyses using electromagnetic wave equations are required.

2.2 Field equations

For a linear, isotropic dielectric waveguide characterized by a spatial permittivity distribution of $\epsilon(x, y)$, Maxwell's equations in (1.40) and (1.41) can be written as

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}, \tag{2.8}$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}. \tag{2.9}$$

Because the optical fields in the waveguide have the form of (2.1) and (2.2), these two Maxwell's equations can be written in the following form:

$$\frac{\partial \mathcal{E}_z}{\partial y} - \mathrm{i}\beta \mathcal{E}_y = \mathrm{i}\omega\mu_0 \mathcal{H}_x, \tag{2.10}$$

$$\mathrm{i}\beta \mathcal{E}_x - \frac{\partial \mathcal{E}_z}{\partial x} = \mathrm{i}\omega\mu_0 \mathcal{H}_y, \tag{2.11}$$

$$\frac{\partial \mathcal{E}_y}{\partial x} - \frac{\partial \mathcal{E}_x}{\partial y} = \mathrm{i}\omega\mu_0 \mathcal{H}_z, \tag{2.12}$$

and

$$\frac{\partial \mathcal{H}_z}{\partial y} - \mathrm{i}\beta \mathcal{H}_y = -\mathrm{i}\omega\epsilon \mathcal{E}_x, \tag{2.13}$$

$$\mathrm{i}\beta \mathcal{H}_x - \frac{\partial \mathcal{H}_z}{\partial x} = -\mathrm{i}\omega\epsilon \mathcal{E}_y, \tag{2.14}$$

$$\frac{\partial \mathcal{H}_y}{\partial x} - \frac{\partial \mathcal{H}_x}{\partial y} = -\mathrm{i}\omega\epsilon \mathcal{E}_z. \tag{2.15}$$

[1] For an excellent detailed discussion on this point, see Marcuse, D., *Theory of Dielectric Optical Waveguides*. New York: Academic Press, 1974, p. 89.

From these equations, the transverse components of the electric and magnetic fields can be expressed in terms of the longitudinal components:

$$(k^2 - \beta^2)\mathcal{E}_x = \mathrm{i}\beta\frac{\partial \mathcal{E}_z}{\partial x} + \mathrm{i}\omega\mu_0\frac{\partial \mathcal{H}_z}{\partial y}, \tag{2.16}$$

$$(k^2 - \beta^2)\mathcal{E}_y = \mathrm{i}\beta\frac{\partial \mathcal{E}_z}{\partial y} - \mathrm{i}\omega\mu_0\frac{\partial \mathcal{H}_z}{\partial x}, \tag{2.17}$$

$$(k^2 - \beta^2)\mathcal{H}_x = \mathrm{i}\beta\frac{\partial \mathcal{H}_z}{\partial x} - \mathrm{i}\omega\epsilon\frac{\partial \mathcal{E}_z}{\partial y}, \tag{2.18}$$

$$(k^2 - \beta^2)\mathcal{H}_y = \mathrm{i}\beta\frac{\partial \mathcal{H}_z}{\partial y} + \mathrm{i}\omega\epsilon\frac{\partial \mathcal{E}_z}{\partial x}, \tag{2.19}$$

where

$$k^2 = \omega^2\mu_0\epsilon(x, y) \tag{2.20}$$

is a function of x and y to account for the transverse spatial inhomogeneity of the waveguide structure.

The relations in (2.16)–(2.19) are generally true for a longitudinally homogeneous waveguide of any transverse geometry and any transverse index profile where $\epsilon(x, y)$ is not a function of z. They are equally true for step-index and graded-index waveguides. In waveguides that have circular cross sections, such as optical fibers, the x and y coordinates of the rectangular system can be transformed to the r and ϕ coordinates of the cylindrical system for similar relations. Therefore, in a waveguide, once the longitudinal field components, $\mathcal{E}_z$ and $\mathcal{H}_z$, are known, all field components can be obtained. The fields in a waveguide can have various vectorial characteristics. They can be classified based on the characteristics of the longitudinal field components.

1. A *transverse electric and magnetic mode*, or TEM mode, has $\mathcal{E}_z = 0$ and $\mathcal{H}_z = 0$. Dielectric waveguides do not support TEM modes, as can be seen from (2.16)–(2.19).
2. A *transverse electric mode*, or TE mode, has $\mathcal{E}_z = 0$ and $\mathcal{H}_z \neq 0$.
3. A *transverse magnetic mode*, or TM mode, has $\mathcal{H}_z = 0$ and $\mathcal{E}_z \neq 0$.
4. A *hybrid mode* has both $\mathcal{E}_z \neq 0$ and $\mathcal{H}_z \neq 0$. Hybrid modes do not appear in planar waveguides but exist in nonplanar waveguides of two-dimensional transverse optical confinement. The HE and EH modes of optical fibers are hybrid modes.

2.3 Wave equations

The field equations obtained in the preceding section establish the relations among the field components. In general, it is only necessary to find $\mathcal{E}_z$ and $\mathcal{H}_z$. Then all other components can be calculated by simply using (2.16)–(2.19). The common approach

to finding $\mathcal{E}_z$ and $\mathcal{H}_z$ is to solve the wave equations together with boundary conditions. In this section, we examine the form of the wave equations for waveguides.

To obtain the wave equations, we need the other two Maxwell's equations in addition to (2.8) and (2.9). For the case of a linear, isotropic waveguide with a spatially dependent $\epsilon(x, y)$ discussed here, they can be written as

$$\boldsymbol{\nabla} \cdot (\epsilon \mathbf{E}) = 0, \tag{2.21}$$

$$\boldsymbol{\nabla} \cdot \mathbf{H} = 0. \tag{2.22}$$

Note that (2.21) implies that

$$\boldsymbol{\nabla} \cdot \mathbf{E} = -\frac{\boldsymbol{\nabla}\epsilon}{\epsilon} \cdot \mathbf{E}, \tag{2.23}$$

which does not vanish in general because $\epsilon(x, y)$ is spatially dependent. Using the four Maxwell's equations in (2.8), (2.9), (2.21), and (2.22), together with (2.23) and the vector identity $\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times = \boldsymbol{\nabla}\boldsymbol{\nabla} \cdot - \nabla^2$, we have

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = -\boldsymbol{\nabla}\left(\frac{\boldsymbol{\nabla}\epsilon}{\epsilon} \cdot \mathbf{E}\right), \tag{2.24}$$

$$\nabla^2 \mathbf{H} + k^2 \mathbf{H} = -\frac{\boldsymbol{\nabla}\epsilon}{\epsilon} \times \boldsymbol{\nabla} \times \mathbf{H}. \tag{2.25}$$

It can be seen that the three components E_x, E_y, and E_z for the electric field are generally coupled together because $\boldsymbol{\nabla}\epsilon \neq 0$ in a waveguide. For the same reason, H_x, H_y, and H_z are also coupled. This fact indicates that the vectorial characteristics of a mode field in a waveguide are strongly dependent on the geometry and index profile of the waveguide.

If the terms on the right-hand sides of (2.24) and (2.25) vanish, then the field components are decoupled. This condition exists only in certain special cases. For example, in the case of a TE mode,

$$\boldsymbol{\nabla}\epsilon \perp \mathbf{E} \quad \text{so that} \quad \boldsymbol{\nabla}\epsilon \cdot \mathbf{E} = 0. \tag{2.26}$$

As a consequence, each component of the electric field of a TE mode satisfies a homogeneous scalar differential equation. The magnetic field components of a TE mode are still coupled because the right-hand term of (2.25) does not vanish.

The index profile of a step-index waveguide is piecewise constant. We can write a homogeneous wave equation separately for each region of constant ϵ because $\boldsymbol{\nabla}\epsilon = 0$ within each region. By taking $\mathbf{E}$ and $\mathbf{H}$ in the forms of (2.1) and (2.2), respectively, and using (2.24) and (2.25) with $\boldsymbol{\nabla}\epsilon = 0$ for each region of constant ϵ, we obtain

$$\frac{\partial^2 \mathcal{E}_z}{\partial x^2} + \frac{\partial^2 \mathcal{E}_z}{\partial y^2} + (k_i^2 - \beta^2)\mathcal{E}_z = 0, \tag{2.27}$$

$$\frac{\partial^2 \mathcal{H}_z}{\partial x^2} + \frac{\partial^2 \mathcal{H}_z}{\partial y^2} + (k_i^2 - \beta^2)\mathcal{H}_z = 0, \tag{2.28}$$

where

$$k_i^2 = \omega^2 \mu_0 \epsilon_i = n_i^2 \frac{\omega^2}{c^2} \tag{2.29}$$

is a constant for region i, which has a constant index of refraction n_i. A homogeneous equation in the same form can be written for each of the other four field components, $\mathcal{E}_x$, $\mathcal{E}_y$, $\mathcal{H}_x$, and $\mathcal{H}_y$. However, it is not necessary to solve the wave equations for all field components because the transverse field components can be found from $\mathcal{E}_z$ and $\mathcal{H}_z$ using the relations in (2.16)–(2.19) once $\mathcal{E}_z$ and $\mathcal{H}_z$ are found. Therefore, the mode field pattern can be obtained by solving only (2.27) and (2.28) for each region of constant index and by requiring the fields to satisfy the boundary conditions at the interfaces between neighboring regions. Clearly, this approach does not work for graded-index waveguides because (2.27) and (2.28) are not valid for such waveguides.

Wave equations for planar waveguides

Homogeneous wave equations exist for planar waveguides of any index profile $n(x)$. For a planar waveguide, the modes are either TE or TM. Furthermore, $\partial/\partial y = 0$ because the index profile is independent of the y coordinate. The wave equations are substantially simplified.

1. **TE modes.** For any TE mode of a planar waveguide, $\mathcal{E}_z = 0$. It can be seen from (2.16)–(2.19) that $\mathcal{E}_x = \mathcal{H}_y = 0$ as well because $\partial \mathcal{H}_z/\partial y = 0$. The only nonvanishing field components are $\mathcal{H}_x$, $\mathcal{E}_y$, and $\mathcal{H}_z$. Because there is only one nonvanishing electric field component $\mathcal{E}_y$, the wave equation for $\mathcal{E}_y$ is naturally decoupled from the other field components. Therefore, we have

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} + (k^2 - \beta^2)\mathcal{E}_y = 0, \tag{2.30}$$

where

$$k^2 = \omega^2 \mu_0 \epsilon(x) = \frac{\omega^2}{c^2} n^2(x). \tag{2.31}$$

Using (2.10) and (2.12), the other two nonvanishing field components can be obtained from $\mathcal{E}_y$:

$$\mathcal{H}_x = -\frac{\beta}{\omega \mu_0} \mathcal{E}_y, \tag{2.32}$$

$$\mathcal{H}_z = \frac{1}{\mathrm{i}\omega \mu_0} \frac{\partial \mathcal{E}_y}{\partial x}. \tag{2.33}$$

2. **TM mode.** For any TM mode of a planar waveguide, $\mathcal{H}_z = 0$. Then, $\mathcal{H}_x = \mathcal{E}_y = 0$ because $\partial \mathcal{E}_z/\partial y = 0$. The only nonvanishing field components are $\mathcal{E}_x$, $\mathcal{H}_y$, and $\mathcal{E}_z$.

In this case, there is only one nonvanishing magnetic field component $\mathcal{H}_y$, and the wave equation for $\mathcal{H}_y$ is naturally decoupled from the other field components. From (2.25), we have

$$\frac{\partial^2 \mathcal{H}_y}{\partial x^2} + (k^2 - \beta^2)\mathcal{H}_y = \frac{1}{\epsilon}\frac{\mathrm{d}\epsilon}{\mathrm{d}x}\frac{\partial \mathcal{H}_y}{\partial x}, \tag{2.34}$$

where $k^2 = k^2(x)$ is the same as that given by (2.31). The other two nonvanishing field components can be obtained from $\mathcal{H}_y$:

$$\mathcal{E}_x = \frac{\beta}{\omega\epsilon}\mathcal{H}_y, \tag{2.35}$$

$$\mathcal{E}_z = -\frac{1}{\mathrm{i}\omega\epsilon}\frac{\partial \mathcal{H}_y}{\partial x}. \tag{2.36}$$

In the case of a planar waveguide, it is convenient to solve for the unique transverse field component first: $\mathcal{E}_y$ for a TE mode and $\mathcal{H}_y$ for a TM mode. The other field components, including the longitudinal component, then follow directly. Although there is only one nonvanishing longitudinal field component for each type of mode in a planar waveguide, it is coupled to a transverse field component. For example, $\mathcal{H}_z$ of a TE mode is coupled to $\mathcal{H}_x$ and is not described by a simple equation of the form of (2.30).

2.4 Power and orthogonality

Except for evanescent fields, the energy of the fields in a waveguide flows only in the longitudinal direction, as discussed in Section 2.1. The intensity of a waveguide mode ν is thus given by

$$I_\nu = \overline{\mathbf{S}}_\nu \cdot \hat{z} = (\mathbf{S}_\nu + \mathbf{S}_\nu^*) \cdot \hat{z} = (\boldsymbol{\mathcal{E}}_\nu \times \boldsymbol{\mathcal{H}}_\nu^* + \boldsymbol{\mathcal{E}}_\nu^* \times \boldsymbol{\mathcal{H}}_\nu) \cdot \hat{z}, \tag{2.37}$$

which is a function of x and y. The power, P_ν, of the mode is obtained by integrating $I_\nu(x, y)$ over the entire transverse cross section of the waveguide. It can be seen that the longitudinal components, $\mathcal{E}_z$ and $\mathcal{H}_z$, of the mode fields do not contribute to the mode intensity or the mode power.

For TE and TM modes, the power obtained by integrating $I_\nu(x, y)$ given in (2.37) can be transformed into other forms. It can be shown, using (2.10) and (2.11), that the power of a TE mode is simply given by

$$P_{\mathrm{TE}} = \frac{2\beta}{\omega\mu_0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\boldsymbol{\mathcal{E}}|^2 \mathrm{d}x\mathrm{d}y. \tag{2.38}$$

By using (2.13) and (2.14), the power of a TM mode can be expressed as

$$P_{\mathrm{TM}} = \frac{2\beta}{\omega} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\epsilon(x, y)} |\boldsymbol{\mathcal{H}}|^2 \mathrm{d}x\mathrm{d}y. \tag{2.39}$$

In a lossless isotropic waveguide, the mode fields have the following *orthogonality relation*:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\boldsymbol{\mathcal{E}}_\nu \times \boldsymbol{\mathcal{H}}_\mu^* + \boldsymbol{\mathcal{E}}_\mu^* \times \boldsymbol{\mathcal{H}}_\nu\right) \cdot \hat{z} \mathrm{d}x\mathrm{d}y = \pm P_\nu \delta_{\nu\mu}. \tag{2.40}$$

The mode fields can be normalized to have the following *orthonormality relation*:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\hat{\boldsymbol{\mathcal{E}}}_\nu \times \hat{\boldsymbol{\mathcal{H}}}_\mu^* + \hat{\boldsymbol{\mathcal{E}}}_\mu^* \times \hat{\boldsymbol{\mathcal{H}}}_\nu\right) \cdot \hat{z} \mathrm{d}x\mathrm{d}y = \pm \delta_{\nu\mu}, \tag{2.41}$$

where the plus sign is for forward-propagating modes while the minus sign is for backward-propagating modes. The electric and magnetic field patterns of a particular mode ν are represented by the normalized mode field distributions $\hat{\boldsymbol{\mathcal{E}}}_\nu(x, y)$ and $\hat{\boldsymbol{\mathcal{H}}}_\nu(x, y)$, respectively. Here $\delta_{\nu\mu}$ is the Kronecker delta function for discrete modes. For a nonplanar waveguide, $\nu = mn$ and $\mu = m'n'$; hence $\delta_{\nu\mu} = \delta_{mm'}\delta_{nn'}$. For a planar waveguide, $\nu = m$, $\mu = m'$, and $\delta_{\nu\mu} = \delta_{mm'}$. For continuous modes, $\delta_{\nu\mu}$ has to be replaced by the Dirac delta function $\delta(\nu - \mu)$. For a nonplanar waveguide, $\delta(\nu - \mu) = \delta(a - a')\delta(b - b')$ for $\nu = ab$ and $\mu = a'b'$. The Dirac delta function is defined as

$$\delta(\nu - \mu) = \begin{cases} 0, & \nu \neq \mu, \\ \infty, & \nu = \mu, \end{cases} \tag{2.42}$$

and

$$\int_{-\infty}^{\infty} \delta(\nu) \mathrm{d}\nu = 1. \tag{2.43}$$

For TE modes, the orthonormality relation in (2.41) can be transformed to

$$\frac{2\beta_\nu}{\omega\mu_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\nu \cdot \hat{\boldsymbol{\mathcal{E}}}_\mu^* \mathrm{d}x\mathrm{d}y = \delta_{\nu\mu}. \tag{2.44}$$

For TM modes, we have

$$\frac{2\beta_\nu}{\omega} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\epsilon(x, y)} \hat{\boldsymbol{\mathcal{H}}}_\nu \cdot \hat{\boldsymbol{\mathcal{H}}}_\mu^* \mathrm{d}x\mathrm{d}y = \delta_{\nu\mu}. \tag{2.45}$$

The orthogonality relation in (2.40), or the orthonormality relation in (2.41), indicates that power cannot be transferred between different modes in a linear, lossless waveguide.

For anisotropic or lossy waveguides, (2.40) and (2.41) do not apply, neither do (2.44) and (2.45). The orthogonality conditions for modes of such waveguides have other forms.

2.5 Step-index planar waveguides

A step-index planar waveguide is also called a *slab waveguide*. We have used it in Section 2.1 with the approach of ray optics to illustrate an intuitive picture and some basic characteristics of the wave behavior in a waveguide. In this section, the important characteristics of a slab waveguide are discussed, beginning with solution of the wave equations developed in Section 2.3. The structure and parameters of the three-layer slab waveguide under discussion are shown in Fig. 2.4.

Normalized waveguide parameters

The mode properties of a waveguide are commonly characterized in terms of a few dimensionless normalized waveguide parameters. The *normalized frequency and waveguide thickness*, also known as the *V number*, of a step-index planar waveguide is defined as

$$V = \frac{2\pi}{\lambda} d\sqrt{n_1^2 - n_2^2} = \frac{\omega}{c} d\sqrt{n_1^2 - n_2^2}, \tag{2.46}$$

where d is the thickness of the waveguide core. The propagation constant β can be represented by the following *normalized guide index*:

$$b = \frac{\beta^2 - k_2^2}{k_1^2 - k_2^2} = \frac{n_\beta^2 - n_2^2}{n_1^2 - n_2^2}, \tag{2.47}$$

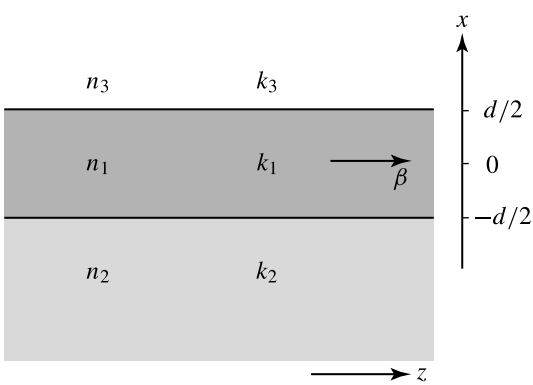


Figure 2.4 Three-layer planar slab waveguide.

where $n_\beta = c\beta/\omega = \beta\lambda/2\pi$ is the effective refractive index of the waveguide mode that has a propagation constant β. The measure of the asymmetry of the waveguide is represented by an *asymmetry factor* a, which depends on the polarization of the mode under consideration. For TE modes, we have

$$a_{\mathrm{E}} = \frac{n_2^2 - n_3^2}{n_1^2 - n_2^2}. \tag{2.48}$$

For TM modes, we have

$$a_{\mathrm{M}} = \frac{n_1^4}{n_3^4} \cdot \frac{n_2^2 - n_3^2}{n_1^2 - n_2^2}. \tag{2.49}$$

Note that for a given asymmetric structure, $a_{\mathrm{M}} > a_{\mathrm{E}}$. For symmetric waveguides, $n_3 = n_2$ and $a_{\mathrm{E}} = a_{\mathrm{M}} = 0$.

Mode parameters

For a guided mode, $k_1 > \beta > k_2 > k_3$. Therefore, positive real parameters h_1, γ_2, and γ_3 exist such that

$$k_1^2 - \beta^2 = h_1^2, \tag{2.50}$$

$$\beta^2 - k_2^2 = \gamma_2^2, \tag{2.51}$$

$$\beta^2 - k_3^2 = \gamma_3^2. \tag{2.52}$$

In correlation with the discussions in Section 2.1 leading to (2.7), it can be seen from (2.50) that $h_1 = k_1 \cos\theta$, which has the meaning of the transverse component of the wavevector in the core region of a refractive index n_1. For a guided mode, the transverse components of the wavevectors in the substrate and cover regions given by $h_2 = (k_2^2 - \beta^2)^{1/2}$ and $h_3 = (k_3^2 - \beta^2)^{1/2}$, respectively, are purely imaginary because $\beta > k_2 > k_3$. The field of the guided mode has to decay exponentially in the transverse direction in the substrate and cover regions with $\gamma_2 = |h_2|$ and $\gamma_3 = |h_3|$ being the decay constants in these regions.

For a substrate radiation mode, $k_1 > k_2 > \beta > k_3$. Then h_2 can be chosen to be real and positive, and (2.51) is replaced by

$$k_2^2 - \beta^2 = h_2^2, \tag{2.53}$$

while (2.52) is still valid in this case. For a substrate–cover radiation mode, $k_1 > k_2 > k_3 > \beta$. Then both h_2 and h_3 are real and positive. In this case, in addition to replacing (2.51) with (2.53), (2.52) is replaced by

$$k_3^2 - \beta^2 = h_3^2. \tag{2.54}$$

The transverse field pattern of a mode is characterized by the transverse parameters h_1, γ_2 (or h_2), and γ_3 (or h_3). Because k_1, k_2, and k_3 are well-defined parameters of a

given waveguide, the only parameter that has to be determined for a particular waveguide mode is the longitudinal propagation constant β. Once the value of β is found, the parameters associated with the transverse field pattern are completely determined. Therefore, a waveguide mode is completely specified by its β. Alternatively, if any one of its transverse parameters, such as h_1 for most instances, is determined, the value of its β is also determined, by (2.50), and the mode is completely specified also. As will be seen in the following, this approach is commonly taken for solving the normal modes of a waveguide.

Guided TE modes

The fields of a TE mode are obtained by solving (2.30) for $\mathcal{E}_y$ and by using (2.32) and (2.33) for $\mathcal{H}_x$ and $\mathcal{H}_z$, respectively. The boundary conditions require that $\mathcal{E}_y$, $\mathcal{H}_x$, and $\mathcal{H}_z$ be continuous at the interfaces at $x = \pm d/2$ between layers of different refractive indices. From (2.32) and (2.33), it can be seen that this is equivalent to requiring $\mathcal{E}_y$ and $\partial \mathcal{E}_y / \partial x$ be continuous at these interfaces.

For a guided mode, we have to use h_1, γ_2, and γ_3 defined above for the transverse field parameters in the core, substrate, and cover regions, respectively. The solutions of (2.30) and the requirement of the boundary conditions yield the following mode field distribution:

$$\hat{\mathcal{E}}_y = C_{\mathrm{TE}} \begin{cases} \cos(h_1 d/2 - \psi) \exp[\gamma_3 (d/2 - x)], & x > d/2, \\ \cos(h_1 x - \psi), & -d/2 < x < d/2, \\ \cos(h_1 d/2 + \psi) \exp[\gamma_2 (d/2 + x)], & x < -d/2, \end{cases} \tag{2.55}$$

and the following eigenvalue equations:

$$\tan h_1 d = \frac{h_1(\gamma_2 + \gamma_3)}{h_1^2 - \gamma_2 \gamma_3} \tag{2.56}$$

and

$$\tan 2\psi = \frac{h_1(\gamma_2 - \gamma_3)}{h_1^2 + \gamma_2 \gamma_3}. \tag{2.57}$$

To normalize the mode field, we apply the normalization relation of (2.44) to the field in (2.55). This procedure yields

$$C_{\mathrm{TE}} = \sqrt{\frac{\omega \mu_0}{\beta d_{\mathrm{E}}}}, \tag{2.58}$$

where

$$d_{\mathrm{E}} = d + \frac{1}{\gamma_2} + \frac{1}{\gamma_3} \tag{2.59}$$

is the *effective waveguide thickness* for a guided TE mode.

Guided TM modes

The fields of a TM mode are obtained by solving (2.34) for $\mathcal{H}_y$ and by using (2.35) and (2.36) for $\mathcal{E}_x$ and $\mathcal{E}_z$, respectively. Note that for the step-index waveguide considered here, $\mathrm{d}\epsilon/\mathrm{d}x = 0$ in each waveguide layer except at the boundaries. The boundary conditions require that $\mathcal{H}_y$, $\epsilon\mathcal{E}_x$, and $\mathcal{E}_z$ be continuous at the interfaces at $x = \pm d/2$ between layers of different refractive indices. Note that $\mathcal{E}_x$ is not continuous because it is the electric field component normal to the interfaces where discontinuities in ϵ occur. Similarly, $\partial\mathcal{H}_y/\partial x$ is not continuous at the interfaces. Rather, it is $\epsilon^{-1}\partial\mathcal{H}_y/\partial x$, or $n^{-2}\partial\mathcal{H}_y/\partial x$, that is continuous at the interfaces. Therefore, the boundary conditions are simply that $\mathcal{H}_y$ and $n^{-2}\partial\mathcal{H}_y/\partial x$ are continuous at the interfaces.

For a guided TM mode, the solutions of (2.34) and the requirement of the boundary conditions yield the following mode field distribution:

$$\hat{\mathcal{H}}_y = C_{\mathrm{TM}} \begin{cases} \cos(h_1 d/2 - \psi)\exp[\gamma_3(d/2 - x)], & x > d/2, \\ \cos(h_1 x - \psi), & -d/2 < x < d/2, \\ \cos(h_1 d/2 + \psi)\exp[\gamma_2(d/2 + x)], & x < -d/2, \end{cases} \tag{2.60}$$

and the following eigenvalue equations:

$$\tan h_1 d = \frac{(h_1/n_1^2)(\gamma_2/n_2^2 + \gamma_3/n_3^2)}{(h_1/n_1^2)^2 - \gamma_2\gamma_3/n_2^2 n_3^2} \tag{2.61}$$

and

$$\tan 2\psi = \frac{(h_1/n_1^2)(\gamma_2/n_2^2 - \gamma_3/n_3^2)}{(h_1/n_1^2)^2 + \gamma_2\gamma_3/n_2^2 n_3^2}. \tag{2.62}$$

To normalize the mode field, we apply the normalization relation of (2.45) to the field in (2.60). This procedure yields

$$C_{\mathrm{TM}} = \sqrt{\frac{\omega\epsilon_0 n_1^2}{\beta d_{\mathrm{M}}}}, \tag{2.63}$$

where the effective waveguide thickness for a guided TM mode is

$$d_{\mathrm{M}} = d + \frac{1}{\gamma_2 q_2} + \frac{1}{\gamma_3 q_3} \tag{2.64}$$

and

$$q_2 = \frac{\beta^2}{k_1^2} + \frac{\beta^2}{k_2^2} - 1, \tag{2.65}$$

$$q_3 = \frac{\beta^2}{k_1^2} + \frac{\beta^2}{k_3^2} - 1. \tag{2.66}$$

Modal dispersion

Guided modes have discrete allowed values of β. They are determined by the allowed values of h_1 because β and h_1 are directly related to each other through (2.50). Because γ_2 and γ_3 are uniquely determined by β through (2.51) and (2.52), respectively, they are also uniquely determined by h_1. In terms of the normalized waveguide parameters, we have

$$\gamma_2^2 d^2 = \beta^2 d^2 - k_2^2 d^2 = V^2 - h_1^2 d^2, \tag{2.67}$$

$$\gamma_3^2 d^2 = \beta^2 d^2 - k_3^2 d^2 = (1 + a_{\mathrm{E}}) V^2 - h_1^2 d^2. \tag{2.68}$$

Therefore, there is only one independent variable h_1 in the eigenvalue equations, (2.56) for TE modes and (2.61) for TM modes. The solutions of (2.56) yield the allowed parameters for guided TE modes, while those of (2.61) yield the parameters for guided TM modes. A transcendental equation such as (2.56) or (2.61) is usually solved graphically by plotting its left- and right-hand sides as a function of $h_1 d$ while using (2.67) and (2.68) to replace γ_2 and γ_3 by expressions in terms of $h_1 d$. The solutions yield the allowed values of β, or the normalized guide index b, as a function of the parameters a and V. The results for the first few guided TE modes are shown in Fig. 2.5. For a given waveguide, a guided TE mode has a larger propagation constant than the corresponding TM mode of the same order:

$$\beta_m^{\mathrm{TE}} > \beta_m^{\mathrm{TM}}. \tag{2.69}$$

However, for ordinary dielectric waveguides where $n_1 - n_2 \ll n_1$, the difference is very small. Then Fig. 2.5 can be used approximately for TM modes with $a = a_{\mathrm{M}}$.

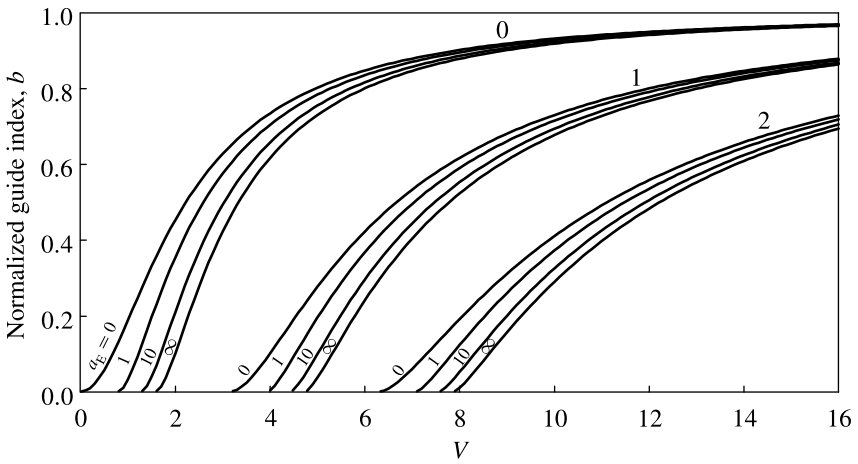


Figure 2.5 Allowed values of normalized guide index b as a function of the V number and the asymmetry factor a_{E} for the first three guided TE modes. The cutoff value V_{c} for a mode is the value of V at the intersection of its dispersion curve with the horizontal axis.

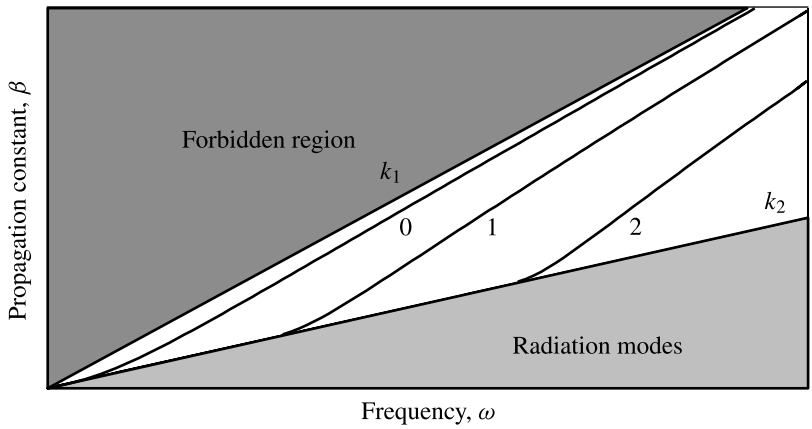


Figure 2.6 Mode propagation constant β as a function of optical frequency ω for a given step-index dielectric waveguide.

For a given waveguide, the values of a_E and a_M, as well as those of d and $n_1^2 - n_2^2$, are fixed. Then, β is a function of optical frequency ω because V depends on ω. Figure 2.6 illustrates a typical relation between β and ω for guided modes of different orders.

Comparing β, k_1, and k_2 in Fig. 2.6, it is seen that the propagation constant of a waveguide mode has a frequency dependence contributed by the structure of the waveguide in addition to that due to material dispersion. This extra contribution also causes different modes to have different dispersion properties, resulting in the phenomenon of *modal dispersion*. *Polarization dispersion* also exists because TE and TM modes generally have different propagation constants. Polarization dispersion is very small in *weakly guiding waveguides* where $n_1 - n_2 \ll n_1$.

EXAMPLE 2.1 An asymmetric slab waveguide is made of a polymer layer of thickness $d = 1$ μm deposited on a silica substrate. At 1 μm optical wavelength, $n_1 = 1.77$ for the polymer guiding layer, $n_2 = 1.45$ for the silica substrate, and $n_3 = 1$ for the air cover. Find the propagation constants of the guided TE and TM modes of this waveguide. Plot the mode field distributions.

Solution With the given parameters of the waveguide, we find that $V = 6.378$, $a_E = 1.07$, and $a_M = 10.5$ by using (2.46), (2.48), and (2.49). We also find that $k_1 = 2\pi n_1/\lambda = 11.12\ \mu\text{m}^{-1}$, $k_2 = 2\pi n_2/\lambda = 9.11\ \mu\text{m}^{-1}$, and $k_3 = 2\pi n_3/\lambda = 6.28\ \mu\text{m}^{-1}$. To find the propagation constant, the parameter h_1 has to be found by solving (2.56) for a TE mode or (2.61) for a TM mode. To solve the eigenvalue equations, we take the variable $\xi = h_1 d$ and express $\gamma_2 d$ and $\gamma_3 d$ in terms of ξ by using the relations in (2.67) and (2.68):

$$\gamma_2 d = (V^2 - \xi^2)^{1/2} \quad \text{and} \quad \gamma_3 d = [(1 + a_E)V^2 - \xi^2]^{1/2}.$$

Then the eigenvalue equation in (2.56) for the TE mode can be expressed in terms of a single variable ξ as

$$\tan\xi = \xi \frac{(V^2-\xi^2)^{1/2} + [(1+a_E)V^2-\xi^2]^{1/2}}{\xi^2 - (V^2-\xi^2)^{1/2}[(1+a_E)V^2-\xi^2]^{1/2}},$$

and the eigenvalue equation in (2.61) for the TM mode can be expressed as

$$\tan\xi = \xi \frac{n_1^2 n_3^2 (V^2-\xi^2)^{1/2} + n_1^2 n_2^2 [(1+a_E)V^2-\xi^2]^{1/2}}{n_2^2 n_3^2 \xi^2 - n_1^4 (V^2-\xi^2)^{1/2}[(1+a_E)V^2-\xi^2]^{1/2}}.$$

These equations yield only discrete eigenvalues for given values of waveguide parameters n_1, n_2, n_3, V, and a_E. They are transcendental equations that have to be solved graphically or numerically. With given waveguide parameters, numerical solution yields two eigenvalues for each of the two equations, indicating two guided TE modes and two guided TM modes. Once the eigenvalues for ξ are found, h_1, γ_2, and γ_3 are found. They can be used to find the phase ψ from (2.57) for a TE mode and from (2.62) for a TM mode. The propagation constant can be found using (2.50) as $\beta = (k_1^2 - h_1^2)^{1/2}$. The effective waveguide thickness can be calculated directly from (2.59) for a TE mode and from (2.64) for a TM mode. The numerical results, as well as the confinement factors Γ_{TE} and Γ_{TM} discussed later, are summarized below.

	h_1 (μm^{-1})	γ_2 (μm^{-1})	γ_3 (μm^{-1})	ψ (rad)	β (μm^{-1})	d_E, d_M (μm)	Γ_{TE}, Γ_{TM}
TE_0	2.47	5.88	8.84	-0.06	10.8432	1.28	0.974
TM_0	2.73	5.76	8.76	-0.10	10.7800	1.17	0.977
TE_1	4.86	4.13	7.78	$\pi/2 - 0.15$	10.0036	1.37	0.871
TM_1	5.28	3.58	7.50	$\pi/2 - 0.28$	9.7873	1.36	0.834

We see from the listed values that a TE mode has a larger propagation constant than a TM mode of the same order, confirming the relation stated in (2.69). Among all of the modes found for this waveguide, β has the largest value for the TE_0 mode and the smallest value for the TM_1 mode. Using the mode parameters listed above, the distributions of $\hat{\mathcal{E}}_y(x)$ given in (2.55) for the TE modes and $\hat{\mathcal{H}}_y(x)$ given in (2.60) for the TM modes are plotted in Fig. 2.7.

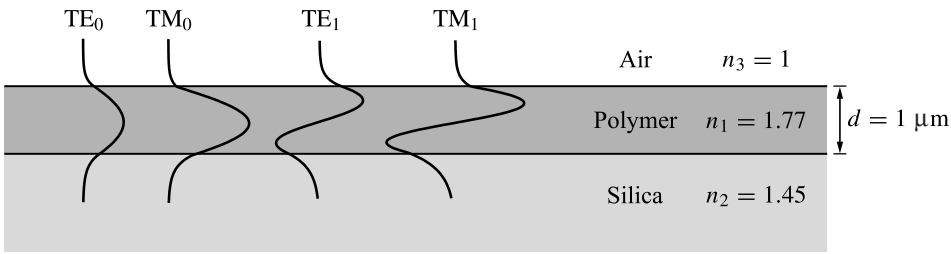


Figure 2.7 Transverse mode field distributions. These field profiles are plotted from real data.

Cutoff conditions

As discussed above, γ_2 and γ_3 are real and positive for a guided mode, so that the fields of the mode decay exponentially in the transverse direction outside the core region and remain bound to the core. This is equivalent to the condition that $\theta > \theta_{c2} > \theta_{c3}$ in the ray optics picture illustrated in Fig. 2.3 so that the ray in the core is totally reflected by both interfaces. Because $\theta_{c2} > \theta_{c3}$, the transition from a guided mode to an unguided radiation mode occurs when $\theta = \theta_{c2}$. This corresponds to the condition that $\beta = k_2$ and $\gamma_2 = 0$. As can be seen from the mode field solutions in (2.55) and (2.60), the fields extend to infinity on the substrate side for $\gamma_2 = 0$. This defines the *cutoff condition* for guided modes. The cutoff condition is determined by $\gamma_2 = 0$, rather than by $\gamma_3 = 0$, because $\gamma_3 > \gamma_2$ and γ_2 reaches zero first as their values are reduced.

At cutoff, $V = V_c$. The cutoff value V_c for a particular guided mode is the value of V at the point where the curve of its b versus V dispersion relation, shown in Fig. 2.5, intersects with the horizontal axis $b = 0$. From (2.67) and (2.68), we find by setting $\gamma_2 = 0$ that

$$h_1 d = V_c \quad \text{and} \quad \gamma_3 d = \sqrt{a_E} V_c \tag{2.70}$$

at cutoff. Substitution of (2.70) and $\gamma_2 = 0$ in (2.56) for a guided TE mode yields

$$\tan V_c = \sqrt{a_E}. \tag{2.71}$$

Therefore, the cutoff condition for the mth guided TE mode is

$$V_m^c = \tan^{-1}\sqrt{a_E} + m\pi, \qquad m = 0, 1, 2, \ldots \tag{2.72}$$

A similar mathematical procedure yields the following cutoff condition for the mth guided TM mode:

$$V_m^c = \tan^{-1}\sqrt{a_M} + m\pi, \qquad m = 0, 1, 2, \ldots \tag{2.73}$$

Because $a_M > a_E$ for a given asymmetric waveguide, the value of V_c for a TM mode is larger than that for a TE mode of the same order.

Using (2.46), we can write

$$V_m^c = \frac{2\pi}{\lambda_m^c} d\sqrt{n_1^2 - n_2^2} = \frac{\omega_m^c}{c} d\sqrt{n_1^2 - n_2^2}, \tag{2.74}$$

where λ_m^c is the *cutoff wavelength*, λ_c, and ω_m^c is the *cutoff frequency*, ω_c, of the mth mode. The mth mode is not guided at a wavelength longer than λ_m^c, or a frequency lower than ω_m^c. For given waveguide parameters, (2.72) and (2.73) can be used to determine the cutoff wavelengths of TE and TM modes, respectively, from (2.74). For a given optical wavelength, they can be used to determine the waveguide parameters that allow the existence of a particular guided mode. For given waveguide parameters and optical wavelength, they can be used to determine the number of guided modes for the waveguide. Therefore, for a given optical wavelength and a waveguide with a given V

number, the total number of guided TE modes is simply

$$M_{\mathrm{TE}} = \left[\frac{V}{\pi} - \frac{1}{\pi}\tan^{-1}\sqrt{a_{\mathrm{E}}}\right]_{\mathrm{int}}, \tag{2.75}$$

while that of the guided TM modes is

$$M_{\mathrm{TM}} = \left[\frac{V}{\pi} - \frac{1}{\pi}\tan^{-1}\sqrt{a_{\mathrm{M}}}\right]_{\mathrm{int}}, \tag{2.76}$$

where $[]_{\mathrm{int}}$ means the nearest integer larger than the value in the bracket.

A waveguide with $M = 1$ that supports only the fundamental TE_0 and/or TM_0 mode is called a *single-mode waveguide*. A waveguide that also supports any number of high-order modes is a *multimode waveguide*.

EXAMPLE 2.2 For the waveguide given in Example 2.1, verify that there are exactly two guided TE modes and two guided TM modes, as is found in the solution of Example 2.1. If the thickness d of the polymer layer is reduced without changing the index profile of the structure, which among these four modes gets cut off first? At what value of d is it cut off?

Solution We already find that $V = 6.378$, $a_{\mathrm{E}} = 1.07$, and $a_{\mathrm{M}} = 10.5$ for the waveguide with $d = 1$ μm and other parameters given in Example 2.1. Therefore, by applying (2.75) and (2.76), we find that

$$M_{\mathrm{TE}} = \left[\frac{6.378}{\pi} - \frac{1}{\pi}\tan^{-1}\sqrt{1.07}\right]_{\mathrm{int}} = [1.77]_{\mathrm{int}} = 2$$

and

$$M_{\mathrm{TM}} = \left[\frac{6.378}{\pi} - \frac{1}{\pi}\tan^{-1}\sqrt{10.5}\right]_{\mathrm{int}} = [1.63]_{\mathrm{int}} = 2,$$

verifying that there are exactly two guided TE modes and two guided TM modes.

From (2.72) and (2.73), we learn that (1) for TE and TM modes of the same mode number, the TM mode has a larger value of V_{c} because $a_{\mathrm{M}} > a_{\mathrm{E}}$; and (2) among modes of the same polarization, a higher-order mode has a larger value of V_{c}. Therefore, among all guided modes found in a waveguide, the highest-order TM mode gets cut off first when the V value is reduced. For the problem under consideration, the TM_1 mode is cut off first as d is reduced so that the value of V is reduced. Using (2.73) we find that the waveguide does not support the TM_1 mode when

$$V < V_1^{\mathrm{c}} = \tan^{-1}\sqrt{10.5} + \pi = 4.413.$$

This condition yields $d < 0.69$ μm, by using (2.46), for the TM_1 mode to be cut off from the waveguide. For $V = 4.413$ corresponding to $d = 0.69$ μm, $M_{\mathrm{TE}} = [1.15]_{\mathrm{int}} = 2$. Therefore, the TE_1 mode is still supported when TM_1 reaches its cutoff point.

Mode confinement

The mode confinement factor, Γ_{mode}, of a guided mode is defined as the fraction of its power in the core region. In an active waveguide, such as that in a semiconductor laser or in a waveguide amplifier, the core guiding region is where the optical gain is, whereas the substrate and cover regions are usually passive media without an optical gain. Only the fraction of power in the core region sees a gain, and the effective gain of a given mode is proportionally reduced. Therefore, the confinement factor is very important in assessing the effective gain of an active optical waveguide for a particular guided mode.

Because the power of a TE mode can be calculated using (2.38), the confinement factor for a TE mode in a slab waveguide is given by

$$\Gamma_{\text{TE}} = \frac{\int_{-d/2}^{d/2} |\mathcal{E}_y(x)|^2 \mathrm{d}x}{\int_{-\infty}^{\infty} |\mathcal{E}_y(x)|^2 \mathrm{d}x} = \frac{2\beta}{\omega\mu_0} \int_{-d/2}^{d/2} |\hat{\mathcal{E}}_y(x)|^2 \mathrm{d}x. \tag{2.77}$$

For a TM mode, the power can be calculated using (2.39), and the confinement factor is given by

$$\Gamma_{\text{TM}} = \frac{\int_{-d/2}^{d/2} n_1^{-2} |\mathcal{H}_y(x)|^2 \mathrm{d}x}{\int_{-\infty}^{\infty} n^{-2}(x) |\mathcal{H}_y(x)|^2 \mathrm{d}x} = \frac{2\beta}{\omega\epsilon_1} \int_{-d/2}^{d/2} |\hat{\mathcal{H}}_y(x)|^2 \mathrm{d}x = \frac{2\omega\epsilon_1}{\beta} \int_{-d/2}^{d/2} |\hat{\mathcal{E}}_x(x)|^2 \mathrm{d}x. \tag{2.78}$$

Using (2.55) to carry out the integration in (2.77) together with (2.56) and (2.57) to simplify the expression, it can be shown that

$$\Gamma_{\text{TE}} = \frac{1}{d_{\text{E}}} \left(d + \frac{1}{\gamma_2} \cdot \frac{1}{1 + h_1^2/\gamma_2^2} + \frac{1}{\gamma_3} \cdot \frac{1}{1 + h_1^2/\gamma_3^2} \right). \tag{2.79}$$

A similar procedure using (2.60) in (2.78) together with (2.61) and (2.62) yields

$$\Gamma_{\text{TM}} = \frac{1}{d_{\text{M}}} \left(d + \frac{1}{\gamma_2 q_2} \cdot \frac{1}{1 + h_1^2/\gamma_2^2} + \frac{1}{\gamma_3 q_3} \cdot \frac{1}{1 + h_1^2/\gamma_3^2} \right). \tag{2.80}$$

As discussed earlier and displayed in (2.69), for guided modes of the same order, the TE mode has a larger propagation constant than the corresponding TM mode, $\beta_{\text{TE}} > \beta_{\text{TM}}$. Therefore, from (2.50)–(2.52), we also have

$$h_1^{\text{TE}} < h_1^{\text{TM}}, \quad \gamma_2^{\text{TE}} > \gamma_2^{\text{TM}}, \quad \text{and} \quad \gamma_3^{\text{TE}} > \gamma_3^{\text{TM}} \tag{2.81}$$

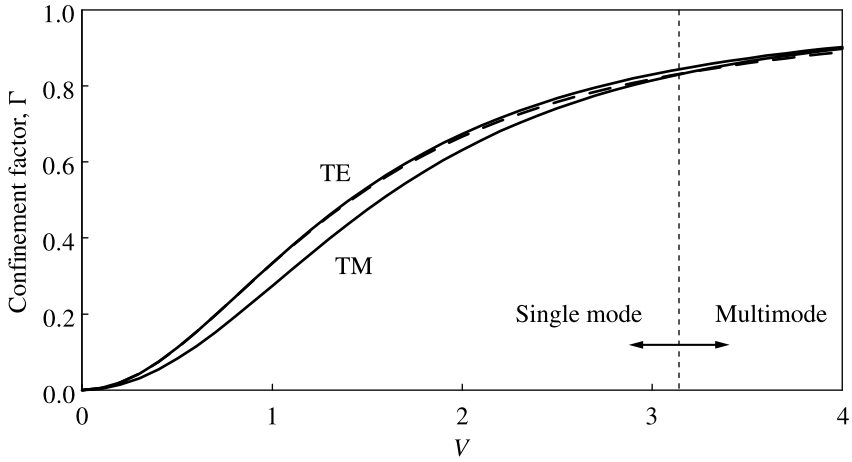


Figure 2.8 Confinement factors of the fundamental TE and TM modes of a symmetric slab waveguide as a function of the waveguide V number. Γ_{TE} is a function of V only, but Γ_{TM} is a function of both V and n_1/n_2. The solid curves are obtained from the exact relations of (2.91) for Γ_{TE} and (2.92) for Γ_{TM} with $n_1/n_2 = 1.09$. For comparison, the dashed curve shows the values of Γ_{TE} obtained from the approximate formula of (2.93).

for TE and TM modes of the same order. Because q_2 and q_3 defined in (2.65) and (2.66) for a TM mode can be either larger or smaller than unity, the relationship between d_E and d_M and that between Γ_{TE} and Γ_{TM} for modes of the same order are less straightforward. Indeed, for a given mode order, Γ_{TE} can be either larger or smaller than Γ_{TM}, but the difference between them is small. For modes of the same polarization, however, a low-order mode is more confined than a high-order mode. Therefore, we can only state that

$$\Gamma_{TE} \approx \Gamma_{TM}, \quad \text{but} \quad \Gamma_m > \Gamma_{m+1}. \tag{2.82}$$

The fundamental TE mode has the largest propagation constant but it may or may not have the largest confinement factor. Either the fundamental TE or the fundamental TM mode has the largest confinement factor among all guided modes. The confinement factors for the fundamental TE and TM modes of a symmetric waveguide where $n_2 = n_3$ is shown in Fig. 2.8.

EXAMPLE 2.3 Find the confinement factors for the guided TE and TM modes determined in Example 2.1 and compare them among modes of different polarizations and modes of different orders.

Solution With the values of h_1, γ_2, and γ_3, as well as those of d_E and d_M, found and listed in Example 2.1, the confinement factors can be calculated using (2.79) and (2.80) for the TE and TM modes, respectively. The results are listed in the last column of the table in Example 2.1. By examining the values of the confinement factors for different modes, we find the following characteristics. (1) The confinement factors for

TE and TM modes of the same order are about the same. There is no clear pattern that indicates whether a TE mode or a TM mode has a larger confinement factor. For example, the TM_0 mode has a slightly larger confinement factor than the TE_0 mode, but the relationship is reversed between TE_1 and TM_1 modes. (2) Among modes of the same polarization but different orders, it is clear that the confinement factor decreases as the mode order increases. For example, the TE_0 mode has a larger confinement factor than the TE_1 mode. The same statement can be made for TM modes.

2.6 Symmetric slab waveguides

In a symmetric slab waveguide, $n_3 = n_2$ and $a_E = a_M = 0$. In addition, we also have $\gamma_3 = \gamma_2$. Then, it can be seen from (2.57) and (2.62) that $\tan 2\psi = 0$ and

$$\psi = \frac{m\pi}{2}, \qquad m = 0, 1, 2, \ldots, \tag{2.83}$$

for both TE and TM modes. Therefore, the mode field patterns of a symmetric waveguide given by (2.55) and (2.60) are either even functions of x with $\cos h_1 x$ in the region $-d/2 < x < d/2$ for even values of m or odd functions of x with $\sin h_1 x$ in the region $-d/2 < x < d/2$ for odd values of m. This characteristic is expected because the mode field pattern in a symmetric structure is either symmetric or antisymmetric. Figure 2.9 shows the field patterns and the corresponding intensity distributions of the first few guided modes of a symmetric slab waveguide.

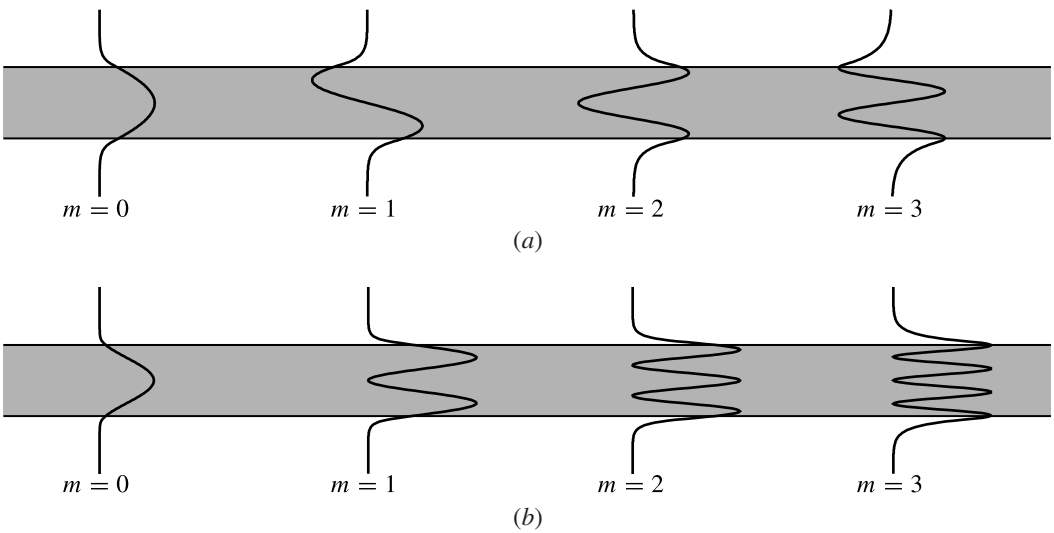


Figure 2.9 (*a*) Field patterns and (*b*) intensity distributions of the first few guided modes of a symmetric slab waveguide.

By equating γ_3 to γ_2 and using the following identity

$$\tan 2\theta = \frac{2\tan\theta}{1-\tan^2\theta} = \frac{2\cot\theta}{\cot^2\theta - 1}, \tag{2.84}$$

the eigenvalue equation (2.56) for TE modes can be transformed to the following two equations:

$$\tan\frac{h_1 d}{2} = \frac{\gamma_2}{h_1}, \qquad \text{for even modes,} \tag{2.85}$$

$$-\cot\frac{h_1 d}{2} = \frac{\gamma_2}{h_1}, \qquad \text{for odd modes,} \tag{2.86}$$

which yield the allowed parameters of guided TE modes. These two equations can be combined in one eigenvalue equation for all guided TE modes:

$$\tan\left(\frac{h_1 d}{2} - \frac{m\pi}{2}\right) = \frac{\gamma_2}{h_1} = \frac{\sqrt{V^2 - h_1^2 d^2}}{h_1 d}, \qquad m = 0, 1, 2, \ldots, \tag{2.87}$$

where m is the same mode number as the one in (2.83). Using (2.61), a similar procedure yields

$$\tan\left(\frac{h_1 d}{2} - \frac{m\pi}{2}\right) = \frac{n_1^2}{n_2^2}\frac{\gamma_2}{h_1} = \frac{n_1^2}{n_2^2}\frac{\sqrt{V^2 - h_1^2 d^2}}{h_1 d}, \qquad m = 0, 1, 2, \ldots, \tag{2.88}$$

for guided TM modes. The solutions of (2.87) yield the allowed values of $h_1 d$ for a given value of the waveguide parameter V for both even and odd TE modes. Those of (2.88) yield the allowed values of $h_1 d$ for both even and odd TM modes. Figure 2.10 shows an example with $V = 5\pi$. Because $n_1 > n_2$, it can be seen from comparison of (2.87) and (2.88) and from the graphic solution shown in Fig. 2.10 that for modes of the same order, $h_1^{\mathrm{TE}} < h_1^{\mathrm{TM}}$. This is consistent with the conclusion obtained from the general discussions in the preceding section.

Because $a_{\mathrm{E}} = a_{\mathrm{M}} = 0$, TE and TM modes of a symmetric waveguide have the same cutoff condition:

$$V_m^{\mathrm{c}} = m\pi \tag{2.89}$$

for the mth TE and TM modes alike. This can also be seen in Fig. 2.10. Because $m = 0$ for the fundamental modes, *neither fundamental TE nor fundamental TM mode in a symmetric waveguide has cutoff.* Any symmetric dielectric waveguide supports at least one TE and one TM mode. The number of TE modes supported by a given symmetric waveguide is the same as that of the TM modes and is simply

$$M_{\mathrm{TE}} = M_{\mathrm{TM}} = \left[\frac{V}{\pi}\right]_{\mathrm{int}}. \tag{2.90}$$

These conclusions are unique to symmetric waveguides. They are not true for an asymmetric waveguide. For example, a guided mode for an asymmetric slab waveguide at a

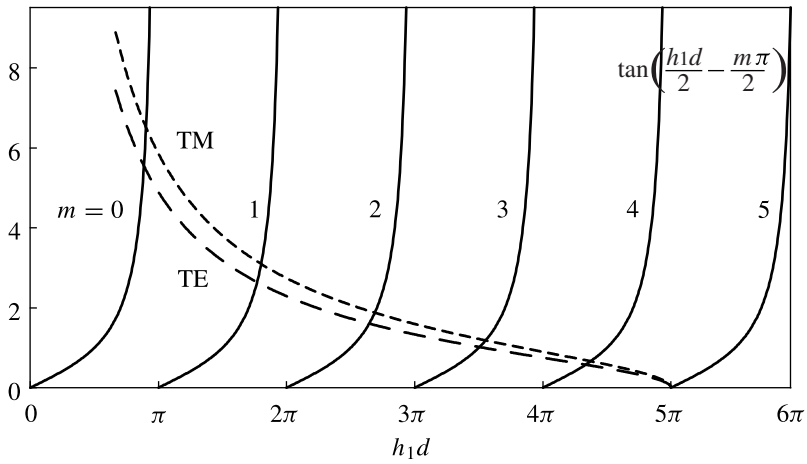


Figure 2.10 Graphic solutions for the eigenvalues of guided TE and TM modes of a symmetric waveguide of $V = 5\pi$. The intersections of dashed and solid curves yield the values of $h_1 d$ for eigenmodes.

given optical frequency may not exist because both its fundamental TE and TM modes have a nonzero cutoff.

The confinement factors of the modes of a symmetric waveguide have the following simple forms:

$$\Gamma_{\text{TE}} = \frac{\gamma_2 d(V^2 + 2\gamma_2 d)}{V^2(2 + \gamma_2 d)} \tag{2.91}$$

and

$$\Gamma_{\text{TM}} = \frac{\gamma_2 d(q_2 V^2 + 2\gamma_2 d)}{V^2(2 + q_2\gamma_2 d)}. \tag{2.92}$$

In general, the confinement factor has to be calculated by first solving the eigenvalue equation of a particular mode to find the mode parameters and then by using (2.91) or (2.92), or, in the case of an asymmetric waveguide, (2.79) or (2.80). However, for the fundamental TE mode of a symmetric slab waveguide, the following approximate formula can be used:

$$\Gamma_0^{\text{TE}} = \frac{V^2}{2 + V^2}. \tag{2.93}$$

Note that (2.93) is a function of V only. Thus, it can be used to calculate the confinement factor of the fundamental TE mode without knowing the mode parameters. It has an error of less than 1.5%.

EXAMPLE 2.4 A symmetric slab waveguide, shown in Fig. 2.11, is made by covering the structure in Example 2.1 with silica, thus sandwiching a polymer layer of thickness $d = 1$ μm between the silica substrate and cover as shown. At 1 μm optical wavelength,

$n_1 = 1.77$ for the polymer guiding layer, and $n_2 = 1.45$ for the silica substrate and cover. How many guided TE and TM modes are supported by this waveguide? Find the propagation constants and the confinement factors for the guided TE and TM modes of this waveguide.

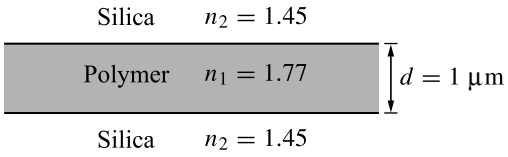


Figure 2.11 Symmetric slab waveguide.

Solution With the given parameters of the waveguide, we have $V = 6.378$, $k_1 = 11.12\ \mu\text{m}^{-1}$, and $k_2 = 9.11\ \mu\text{m}^{-1}$. Using (2.90), we find that the waveguide now supports three TE modes and three TM modes because

$$M_{\text{TE}} = M_{\text{TM}} = \left[\frac{6.378}{\pi}\right]_{\text{int}} = [2.03]_{\text{int}} = 3.$$

Compared with the asymmetric waveguide in Example 2.1, two additional modes, TE_2 and TM_2, are supported by this symmetric waveguide. Because $V/\pi = 2.03$ is very close to the integer 2, however, these two modes are only slightly above cutoff.

The mode characteristics of this asymmetric waveguide can be found in a procedure similar to, but much simpler than, that used to find those of the asymmetric waveguide in Example 2.1. By taking $\xi = h_1 d$, we see that the eigenvalue equations (2.87) and (2.88) are already expressed in terms of the variable ξ. They can be solved numerically to find the values of h_1 for the guided TE and TM modes. Then, $\beta = (k_1^2 - h_1^2)^{1/2}$ and $\gamma_2 = (\beta^2 - k_2^2)^{1/2}$ are found. The phase shift ψ need not be solved because it is simply $\psi = m\pi/2$ for the mth mode. By setting $\gamma_3 = \gamma_2$ for the symmetric waveguide, the effective waveguide thickness can be calculated from (2.59) for a TE mode and from (2.64) for a TM mode. The confinement factors are found using (2.91) and (2.92) for TE and TM modes, respectively. The numerical characteristics of the guided modes of this symmetric waveguide are summarized below.

	h_1 (μm^{-1})	γ_2 (μm^{-1})	ψ (rad)	β (μm^{-1})	$d_{\text{E}}, d_{\text{M}}$ (μm)	$\Gamma_{\text{TE}}, \Gamma_{\text{TM}}$
TE_0	2.38	5.92	0	10.8641	1.34	0.965
TM_0	2.57	5.84	0	10.8208	1.25	0.967
TE_1	4.65	4.37	$\pi/2$	10.1027	1.46	0.833
TM_1	4.92	4.06	$\pi/2$	9.9746	1.49	0.804
TE_2	6.37	0.28	π	9.1150	8.08	0.125
TM_2	6.37	0.20	π	9.1127	16.2	0.063

We still see the same characteristics as those in Example 2.1 that a TE mode has a larger propagation constant than a TM mode of the same order and that the TE_0 mode has the largest propagation constant. The characteristics discussed in Example 2.2 for the confinement factors are also seen here for the symmetric waveguide. Compared with the asymmetric waveguide in Example 2.1, it is interesting to see that the symmetric waveguide now supports TE_2 and TM_2 modes because of the increased index in the cover layer from 1, of the air, to 1.45, of silica. Because these two modes are very close to the cutoff point, they have very small values of γ_2. As a consequence, their fields penetrate deeply into the cover and the substrate, resulting in their large effective thicknesses and small confinement factors. Meanwhile, their propagation constants are only slightly larger than k_2 of the substrate and cover. The field and intensity distributions of these modes have the characteristics of the symmetric waveguide modes shown in Fig. 2.9 and are not repeated here.

We can use (2.93) to obtain an estimate of $\Gamma_0^{\mathrm{TE}} \approx 0.953$ for $V = 6.378$. Compared with the actual confinement factor of 0.965 for the TE_0 mode listed above, the error of (2.93) is only 1.2% in this case.

2.7 Graded-index planar waveguides

In the preceding two sections, we have considered slab waveguides that have step-index profiles. In this section, we consider graded-index planar waveguides, which do not have the piecewise-constant index profiles of step-index waveguides. Two types of graded-index planar waveguides, shown in Fig. 2.12, are of practical interest. One is the *smooth*

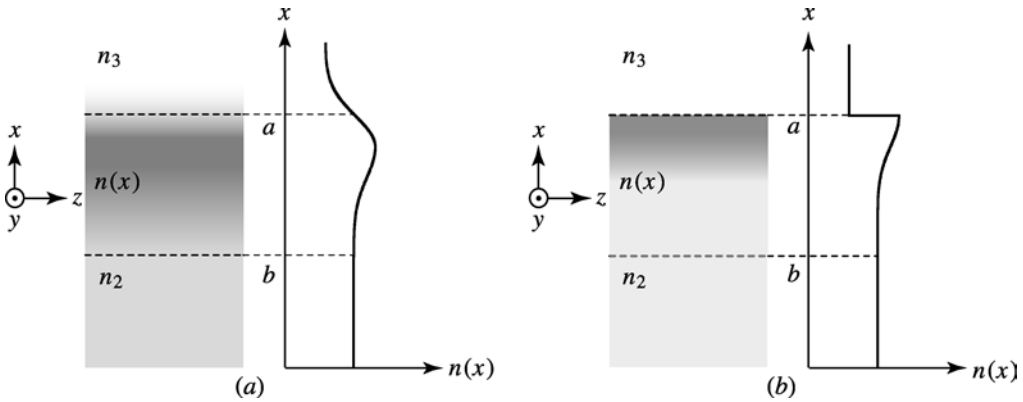


Figure 2.12 Two types of graded-index planar waveguides: (*a*) smooth graded-index waveguide with a completely smooth graded-index profile and (*b*) step-bounded graded-index waveguide with a graded-index profile bounded by an index step on one side. The index values of $n_1 = 2.232$, $n_2 = 1.700$, and $n_3 = 1.000$ are used for these profiles. At $x = a$ and $x = b$, $n(x) = n_2$.

graded-index waveguide, which has a smooth index profile across its entire structure in the x direction, as shown in Fig. 2.12(a). The other is the *step-bounded graded-index waveguide*, which has a graded-index profile on one side but is bounded by a step-index boundary on the other side of its core, shown in Fig. 2.12(b). A graded-index planar waveguide can be either asymmetric or symmetric. For the purpose of general discussion, we consider the index profiles shown in Fig. 2.12 with the index at the peak of the profile in the waveguide core being n_1 and the indices of the substrate and the cover far away from the core of the waveguide being n_2 and n_3, respectively, where $n_1 > n_2 > n_3$ in consistency with the designation of these three indices throughout this chapter. The waveguide core is the region between the two points $x = a$ and $x = b$ defined by $n(a) = n(b) = n_2$ within which an index $n(x)$ larger than both n_2 and n_3 can be found.

The discussions in Section 2.3 regarding planar waveguides apply to graded-index planar waveguides as well. Many of the general qualitative conclusions obtained in Section 2.5 for step-index planar waveguides are also valid for graded-index planar waveguides. The modes of graded-index planar waveguides are still either TE or TM. It is also true that the fundamental mode is TE_0 and that $\beta_m^{\mathrm{TE}} > \beta_m^{\mathrm{TM}}$.

The guided TE modes can be found by solving (2.30),

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} + \left[k^2(x) - \beta^2\right] \mathcal{E}_y = 0, \tag{2.94}$$

for $\mathcal{E}_y(x)$, followed by using (2.32) and (2.33) to obtain $\mathcal{H}_x(x)$ and $\mathcal{H}_z(x)$, respectively. Similarly, the guided TM modes can be found by solving (2.34),

$$\frac{\partial^2 \mathcal{H}_y}{\partial x^2} + \left[k^2(x) - \beta^2\right] \mathcal{H}_y = \frac{1}{\epsilon} \frac{\mathrm{d}\epsilon}{\mathrm{d}x} \frac{\partial \mathcal{H}_y}{\partial x}, \tag{2.95}$$

for $\mathcal{H}_y(x)$, followed by using (2.35) and (2.36) to obtain $\mathcal{E}_x(x)$ and $\mathcal{E}_z(x)$, respectively. For a graded-index waveguide, the propagation constant

$$k(x) = \frac{\omega}{c} n(x) = \frac{2\pi}{\lambda} n(x) \tag{2.96}$$

is a spatially varying function of x. Therefore, (2.94) and (2.95) cannot be readily solved analytically. Though such second-order ordinary differential equations can be solved numerically, here we are interested in gaining physical insight into the waveguide characteristics without resorting to complete numerical solutions. For this purpose, we consider the common situation where the term on the right-hand side of (2.95) is very small compared with $\partial^2 \mathcal{H}_y / \partial x^2$ so that it can be neglected for (2.95) to take the form of (2.94) approximately. We then find that these equations have the form of the Schrödinger equation in quantum mechanics. Approximate solutions can be obtained using the Wentzel–Kramers–Brillouin (WKB) approximation developed in quantum mechanics for solving the Schrödinger equation of a general graded potential.

The central concept of the WKB method is to realize the fact that a guided mode can be established if it forms a standing wave pattern in the transverse x direction

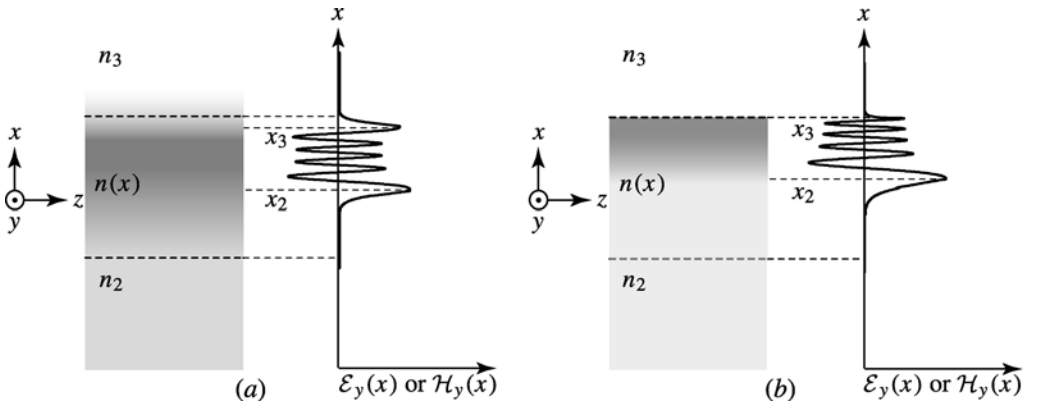


Figure 2.13 Standing wave patterns for guided modes of $m = 8$ of (a) the smooth graded-index planar waveguide and (b) the step-bounded graded-index planar waveguide, which are shown in Figs. 2.12(a) and (b), respectively. Calculated using the index profiles given in Fig. 2.12, these mode patterns show oscillatory field variations in the region between $x = x_2$ and $x = x_3$ and decaying fields outside.

that has oscillatory spatial variations in a certain region within the waveguide core but has decaying fields away from the core in the substrate and cover, as illustrated in Fig. 2.13. As discussed in Section 2.1, the condition for such a standing wave pattern to be formed is that the total phase shift in a round-trip transverse passage be an integral multiple of 2π. This condition is given in (2.7) for a step-index planar waveguide. A similar, but more complicated, condition can be obtained for a graded-index planar waveguide without going through the detailed WKB analysis by simply modifying (2.7). The key point here is to realize that the first term, $2k_1 d \cos\theta$, in (2.7) is the round-trip phase shift through the oscillatory region in the x direction. We notice that $k_1 \cos\theta = h_1 = (k_1^2 - \beta^2)^{1/2}$ for a step-index waveguide and that d is the range of the oscillatory region in the waveguide. For a graded-index waveguide, k_1 has to be replaced by $k(x)$, and the oscillatory region is where $k(x) > \beta$ so that the following function

$$p^2(x) = k^2(x) - \beta^2 \tag{2.97}$$

has positive values. Therefore, as seen in Fig. 2.13, the oscillatory region for a given mode of propagation constant β is the range bounded by the two *turning points* $x = x_2$ and $x = x_3$, where $p(x_2) = p(x_3) = 0$ for $k(x_2) = k(x_3) = \beta$. In the oscillatory region, where $x_2 < x < x_3$, we have $p^2(x) > 0$ so that a real, positive square root $p(x) = \left(k^2(x) - \beta^2\right)^{1/2}$ exists. In the evanescent regions, where $x < x_2$ and $x_3 < x$, $p^2(x) < 0$ and its square roots are purely imaginary.

From these discussions, it is clear that the following condition can be obtained for the guided modes of a graded-index planar waveguide:

$$2\int_{x_2}^{x_3} \left[k^2(x) - \beta^2\right]^{1/2} \mathrm{d}x = 2m\pi - \varphi_2 - \varphi_3, \quad m = 0, 1, 2, \ldots, \tag{2.98}$$

where x_2 and x_3 are the roots of $p^2(x) = k^2(x) - \beta^2 = 0$. Because x_2, x_3, φ_2, and φ_3 are all functions of β, this equation has to be solved self-consistently for β with a given integer m for the mth-order mode. The phase shifts, φ_2 and φ_3, are polarization dependent as well as mode dependent; therefore, TE and TM modes have different solutions, thus slightly different values of β, for the same value of m in (2.98). They are (see Problem 2.1.1)

$$\varphi_2 = -2\tan^{-1}\left[\frac{\beta^2 - k^2(x_2^-)}{k^2(x_2^+) - \beta^2}\right]^{1/2} \quad \text{and} \quad \varphi_3 = -2\tan^{-1}\left[\frac{\beta^2 - k^2(x_3^+)}{k^2(x_3^-) - \beta^2}\right]^{1/2} \tag{2.99}$$

for a TE mode of propagation constant β, and

$$\varphi_2 = -2\tan^{-1}\frac{k^2(x_2^+)}{k^2(x_2^-)}\left[\frac{\beta^2 - k^2(x_2^-)}{k^2(x_2^+) - \beta^2}\right]^{1/2} \text{and} \quad \varphi_3 = -2\tan^{-1}\frac{k^2(x_3^-)}{k^2(x_3^+)}\left[\frac{\beta^2 - k^2(x_3^+)}{k^2(x_3^-) - \beta^2}\right]^{1/2} \tag{2.100}$$

for a TM mode of propagation constant β. The values of the phase shifts are in the range of $-\pi < \varphi_2, \varphi_3 < 0$ for all guided modes. At $x = x_2$, where we have assumed that the index grading is smooth for both types of graded-index waveguides shown in Fig. 2.12, $\beta^2 - k^2(x_2^-) = k^2(x_2^+) - \beta^2$ as x_2^- and x_2^+ approach the turning point x_2 infinitesimally from the left and right, respectively. Therefore, we find that $\varphi_2 = -\pi/2$ from (2.99) and (2.100) for any guided TE or TM mode. Following the same reasoning, we also find that $\varphi_3 = -\pi/2$ for any guided TE or TM mode for the smooth graded-index waveguide shown in Fig. 2.12(a), which has smooth index grading at x_3. For the step-bounded graded-index waveguide shown in Fig. 2.12(b), however, the value of φ_3 cannot be so generalized but is a function of $k_1 = k(x_3^-)$ and $k_3 = k(x_3^+)$, with $k_1 > k_3$, because the turning point x_3 is located at the abrupt index step.

To summarize, for a smooth graded-index waveguide, the eigenvalue equation for the propagation constants of its guided modes can be simply expressed as

$$\int_{x_2}^{x_3} \left[k^2(x) - \beta^2\right]^{1/2} \mathrm{d}x = \left(m + \frac{1}{2}\right)\pi, \quad m = 0, 1, 2, \ldots \tag{2.101}$$

For a step-bounded graded-index waveguide, however, the eigenvalue equation can only be simplified to

$$\int_{x_2}^{x_3} \left[k^2(x) - \beta^2\right]^{1/2} \mathrm{d}x = \left(m + \frac{1}{4}\right)\pi - \frac{\varphi_3}{2}, \quad m = 0, 1, 2, \ldots, \tag{2.102}$$

where φ_3 takes the form in (2.99) for a TE mode and that in (2.100) for a TM mode with $k(x_3^-) = k_1$ and $k(x_3^+) = k_3$. For both types of graded-index waveguides at locations away from the immediate vicinity of the turning points, the unnormalized mode fields

for a guided mode have the following asymptotic form:

$$\begin{matrix}\text{TE } \mathcal{E}_y(x) \\ \text{TM } \mathcal{H}_y(x)\end{matrix} \sim \begin{cases} \dfrac{1}{\sqrt{|p(x)|}} \exp\left[-\displaystyle\int_x^{x_2} |p(x')|\mathrm{d}x'\right], & x < x_2, \\ \dfrac{2}{\sqrt{p(x)}} \cos\left[\displaystyle\int_{x_2}^{x} p(x')\mathrm{d}x' - \dfrac{\pi}{4}\right], & x_2 < x < x_3, \\ \dfrac{(-1)^m}{\sqrt{|p(x)|}} \exp\left[-\displaystyle\int_{x_3}^{x} |p(x')|\mathrm{d}x'\right], & x > x_3, \end{cases} \tag{2.103}$$

where a factor of $(-1)^m$ is used in the field pattern for $x > x_3$ to account for the correct phase of the field at $x = x_3$.

Number of modes

Graded-index waveguides are often used as multimode waveguides because of their low modal dispersion compared with step-index waveguides. This feature is discussed in great detail for optical fibers in Section 3.5, but the concept applies generally to planar waveguides as well. In contrast, single-mode waveguides are preferably step-index waveguides because the step-index geometry allows precise control of the waveguide parameters. In addition, the step-index geometry also conforms with the various junction structures discussed in Section 13.5 for high-performance optoelectronic devices, which normally requires single-mode characteristics.

The number of guided modes supported by a graded-index planar waveguide can be found by finding the largest integral value of m for a solution of β from its eigenvalue equation. Because $\beta > k_2 > k_3$ for any guided mode, the minimum possible value of β is $\beta = k_2$. The turning points for $\beta = k_2$, which are $x_2 = a$ and $x_3 = b$ shown in Fig. 2.12, are where $k(a) = k(b) = k_2$ and $n(a) = n(b) = n_2$. The number of guided modes in a given polarization that are supported by a waveguide is found by adding 1 to the mode number of the highest guided mode because the fundamental mode has a mode number of $m = 0$. For a smooth graded-index waveguide, we find that for $\beta = k_2$, the phase shifts given in (2.99) and (2.100) are not $-\pi/4$, obtained above for (2.101) where $\beta > k_2$, but are $\varphi_2 = \varphi_3 = 0$ for both TE and TM modes. Then, from (2.98), the numbers of guided TE and TM modes supported by a smooth graded-index waveguide as shown in Fig. 2.12(a) are

$$M_{\text{TE}} = M_{\text{TM}} = \left[\frac{2}{\lambda}\int_a^b \left[n^2(x) - n_2^2\right]^{1/2} \mathrm{d}x\right]_{\text{int}}, \tag{2.104}$$

where $[]_{\text{int}}$ means taking the nearest integer larger than the value in the brackets.

For a step-bounded graded-index waveguide, we find that $\varphi_2 = 0$ but $\varphi_3 = -2\tan^{-1}\sqrt{a_E}$ for a TE mode and $\varphi_3 = -2\tan^{-1}\sqrt{a_M}$ for a TM mode, where a_E and a_M are the asymmetric factors defined in (2.48) and (2.49), respectively. Therefore, the numbers of guided TE and TM modes supported by a step-bounded graded-index waveguide as shown in Fig. 2.12(b) are

$$M_{TE} = \left[\frac{2}{\lambda}\int_a^b \left[n^2(x) - n_2^2\right]^{1/2} dx - \frac{1}{\pi}\tan^{-1}\sqrt{a_E}\right]_{int} \tag{2.105}$$

and

$$M_{TM} = \left[\frac{2}{\lambda}\int_a^b \left[n^2(x) - n_2^2\right]^{1/2} dx - \frac{1}{\pi}\tan^{-1}\sqrt{a_M}\right]_{int}. \tag{2.106}$$

EXAMPLE 2.5 A planar $LiNbO_3$ waveguide made by Ti diffusion is a step-bounded graded-index waveguide that has an index profile like the one shown in Fig. 2.12(b). The optical axis of the $LiNbO_3$ crystal, which is negative uniaxial, is lined up with the z axis of this waveguide so that both TE and TM modes see only the ordinary index n_o. At $\lambda = 1.3$ μm, $n_o = 2.222$. We take the index step to be located at $x = b = 0$. Then $n(x) = n_3 = 1$ for $x > 0$. The graded-index profile created by Ti diffusion has the following Gaussian profile:

$$n(x) = n_o + \Delta n e^{-x^2/d^2}, \quad \text{for } x < 0,$$

where d is the diffusion depth of Ti in $LiNbO_3$ required to define the waveguiding region. The diffusion depth is determined by the Ti diffusion coefficient D and the total time duration Δt for the diffusion process as $d = \sqrt{D\Delta t}$. At a temperature of 1020 °C, $D = 1.4 \times 10^{-12}$ cm^2 s^{-1}. Design a single-mode Ti : $LiNbO_3$ waveguide that supports exactly one TE mode and one TM mode at $\lambda = 1.3$ μm.

Solution With the given index profile, the condition for the waveguide to support exactly one TE mode can be found from (2.105) for $M_{TE} = 1$ as

$$1 + \frac{1}{\pi}\tan^{-1}\sqrt{a_E} > \frac{2\sqrt{2n_o\Delta n}}{\lambda}\int_{-\infty}^{0} e^{-x^2/2d^2} dx > \frac{1}{\pi}\tan^{-1}\sqrt{a_E}$$

for $\Delta n \ll 1$. Using the identity

$$\int_{-\infty}^{0} e^{-x^2} dx = \int_0^{\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2},$$

this condition is reduced to

$$1 + \frac{1}{\pi}\tan^{-1}\sqrt{a_E} > 2\sqrt{\pi n_o \Delta n}\frac{d}{\lambda} > \frac{1}{\pi}\tan^{-1}\sqrt{a_E}.$$

The condition for the waveguide to support exactly one TM mode can be obtained by replacing a_E with a_M in the above relation. A practical index step, which is controlled by the thickness of Ti deposited on the surface of $LiNbO_3$ during the diffusion process, is chosen to be $\Delta n = 0.01$. For this waveguide, we then have $n_1 = n_o + \Delta n = 2.232$, $n_2 = n_o = 2.222$, and $n_3 = 1$. We also find that $a_E = 88$ and $a_M = 2188$. With these parameters and with $\lambda = 1.3$ μm, we find

$3.607\ \mu m > d > 1.147\ \mu m$ for a single TE mode

and

$3.674\ \mu m > d > 1.213\ \mu m$ for a single TM mode.

To satisfy both conditions so that the waveguide supports exactly one TE mode and one TM mode, a good choice for the diffusion depth is $d = 2$ μm, with the chosen index step of $\Delta n = 0.01$. Such a waveguide can be made by Ti diffusion at 1020 °C for 8 hours because $\Delta t = d^2/D \approx 8$ hours from the given value of D.

2.8 Channel waveguides

So far we have discussed the characteristics of planar waveguides. In practice, most waveguides used in device applications are nonplanar waveguides. For a nonplanar waveguide, the index profile $n(x, y)$ is a function of both transverse coordinates x and y. There are many different types of nonplanar waveguides that are differentiated by the distinctive features of their index profiles. One very unique group is the circular optical fibers discussed in Chapter 3. Another important group of nonplanar waveguides is the *channel waveguides*, which include the *buried channel waveguides*, the *strip-loaded waveguides*, the *ridge waveguides*, the *rib waveguides*, and the *diffused waveguides*, shown in Fig. 2.14.

A buried channel waveguide is formed with a high-index waveguiding core buried in a low-index surrounding medium. The waveguiding core can have any cross-sectional geometry though it is often intended to have a rectangular shape, as shown in Fig. 2.14(a). A strip-loaded waveguide is formed by loading a planar waveguide, which already provides optical confinement in the x direction, with a dielectric strip of index $n_3 < n_1$ or a metal strip to facilitate optical confinement in the y direction, as shown in Fig. 2.14(b). The waveguiding core of a strip waveguide is the n_1 region under the loading strip, with its thickness d determined by the thickness of the n_1 layer and its width w defined by the width of the loading strip. A ridge waveguide, shown in Fig. 2.14(c), has a structure that looks like a strip waveguide, but the strip, or the ridge, on top of its planar structure has a high index and is actually the waveguiding core. A ridge waveguide has strong optical confinement because it is surrounded on three sides by low-index air. A rib waveguide, shown in Fig. 2.14(d), has a structure similar to that of a strip or ridge waveguide, but the strip has the same index as the high-index planar layer

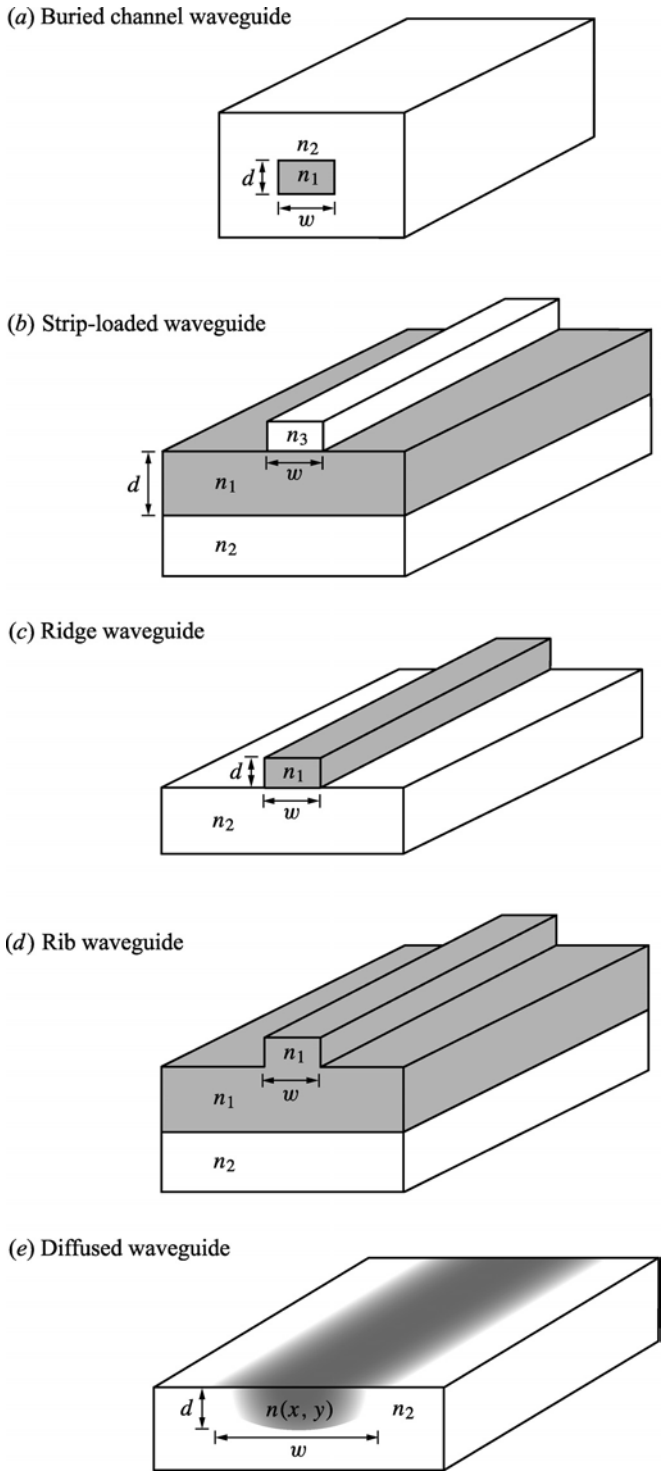


Figure 2.14 Representative channel waveguides.

beneath it and is part of the waveguiding core. These four types of waveguides are usually classified as *rectangular waveguides* with a thickness d in the x direction and a width w in the y direction, though their shapes are normally not exactly rectangular. A diffused waveguide, shown in Fig. 2.14(*e*), is formed by creating a high-index region in a substrate through diffusion of dopants, such as a $LiNbO_3$ waveguide with a core formed by Ti diffusion. Because of the diffusion process, the core boundaries in the substrate are not sharply defined. However, a diffused waveguide also has a thickness d defined by the diffusion depth of the dopant in the x direction and a width w defined by the distribution of the dopant in the y direction.

One distinctive property of nonplanar dielectric waveguides versus planar waveguides is that a nonplanar waveguide supports hybrid modes in addition to TE and TM modes, whereas a planar waveguide supports only TE and TM modes. Except for those few exhibiting special geometric structures, such as circular optical fibers, nonplanar dielectric waveguides generally do not have analytical solutions for their guided mode characteristics. Numerical methods, such as the *beam propagation method*, exist for analyzing such waveguides. Here we are interested in obtaining approximate solutions that give the mode characteristics without full-blown numerical analysis. One of the methods for this purpose is the *effective index method* discussed below.

Effective index method

The basic concept of the effective index method, illustrated in Fig. 2.15, is to convert the problem of a channel waveguide into that of two planar waveguides. The effective index method is a good approximation if the waveguide satisfies the following two conditions: (1) the waveguide width is larger than its thickness, $w > d$; and (2) waveguiding in the y direction across its width is not stronger than that in the x direction across its thickness. Many useful waveguides satisfy these conditions. The effective index method applies to both step-index and graded-index channel waveguides, including all of those shown in Fig. 2.14, as long as these two conditions are satisfied. When these two conditions are satisfied, the characteristics of the guided modes are primarily determined by the layered

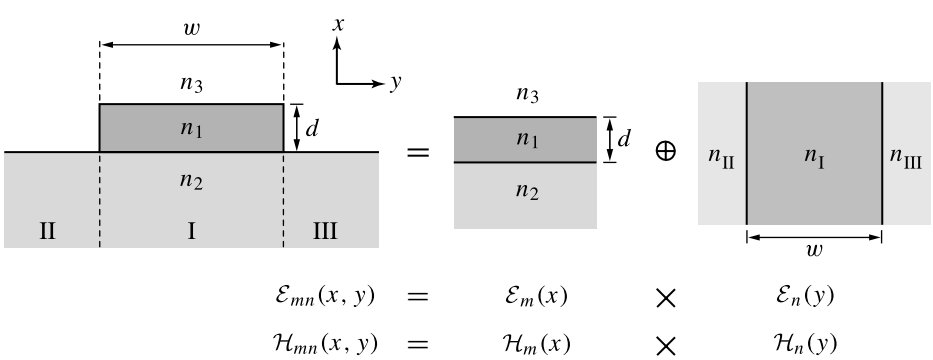


Figure 2.15 Basic concept of the effective index method.

structure perpendicular to the x direction, much like a planar waveguide of thickness d, but are modified by a lateral structure of width w. The planar structure defines TE and TM polarizations, but the lateral structure distorts them. Therefore, a mode with its electric field mostly in the y direction parallel to the planar layers is called a *TE-like mode*, and one with its magnetic field mostly in this direction is called a *TM-like mode*.

The procedure of applying the effective index method is straightforward. Because an effective index is mode dependent, we first decide on the specific mode, either TE_{mn} or TM_{mn}, with specific mode indices m and n, to be analyzed. The waveguide is then divided into three structures for the three vertical regions, I, II, and III, shown in Fig. 2.15. The structure associated with each region is then treated as a planar waveguide to find the propagation constant β_m for the mode m. The x dependence of the y component of the mode field, $\mathcal{E}_{m,y}(x)$ in the case of a TE-like mode or $\mathcal{H}_{m,y}(x)$ in the case of a TM-like mode, for central waveguide region I is also found through the same procedure. The propagation constants for the three regions are used to determine the effective indices, $n_{\text{I}} = c\beta_m^{\text{I}}/\omega = \lambda\beta_m^{\text{I}}/2\pi$, $n_{\text{II}} = c\beta_m^{\text{II}}/\omega = \lambda\beta_m^{\text{II}}/2\pi$, and $n_{\text{III}} = c\beta_m^{\text{III}}/\omega = \lambda\beta_m^{\text{III}}/2\pi$, for a vertical planar waveguide of core width w. This structure is then treated as a planar slab waveguide to solve for the propagation constant β_{mn} of the desired mode and for the y dependence of the y component of the mode field, $\mathcal{E}_{n,y}(y)$ in the case of a TE-like mode or $\mathcal{H}_{n,y}(y)$ in the case of a TM-like mode. Note that $\mathcal{E}_{n,y}(y)$ for a TE-like mode of the original channel waveguide is obtained from the $\mathcal{E}_y$ component of the TM_n field of the effective vertical planar waveguide, whereas $\mathcal{H}_{n,y}(y)$ for a TM-like mode of the original waveguide is obtained from the $\mathcal{H}_y$ component of the TE_n field of the effective vertical planar waveguide. Finally, the y component of the total mode field for the original channel waveguide is $\mathcal{E}_{mn,y}(x, y) = \mathcal{E}_{m,y}(x)\mathcal{E}_{n,y}(y)$, in the case when the TE_{mn} mode is considered, or $\mathcal{H}_{mn,y}(x, y) = \mathcal{H}_{m,y}(x)\mathcal{H}_{n,y}(y)$, in the case when the TM_{mn} mode is considered. Other significant field components are found by using (2.32) and (2.33) for a TE-like mode and by using (2.35) and (2.36) for a TM-like mode. The propagation constant is simply β_{mn} found from the effective vertical planar waveguide.

EXAMPLE 2.6 A strip-loaded waveguide can be made by loading the planar waveguide illustrated in Example 2.1 (Fig. 2.7) with a silica strip on top of the polymer layer, as shown in Fig. 2.16. The silica loading strip has a width of $w = 5$ μm and a thickness of $t = 2$ μm. We are interested in the TM-like modes that have fundamental-mode characteristics in the x direction. How many such modes exist at $\lambda = 1$ μm? Find their characteristics using the effective index method.

Solution To apply the effective index method to this problem, the waveguide is divided into three regions as shown in Fig. 2.16. The structure in region I can be treated as a symmetric waveguide if t is sufficiently large so that the evanescent wave in the x direction does not reach the air above the strip. From Example 2.4, we find that $\gamma_2 = 5.84\ \mu\text{m}^{-1}$ for the TM_0 mode of interest here. Because $\exp(-\gamma_2 t) \approx 8 \times 10^{-6}$

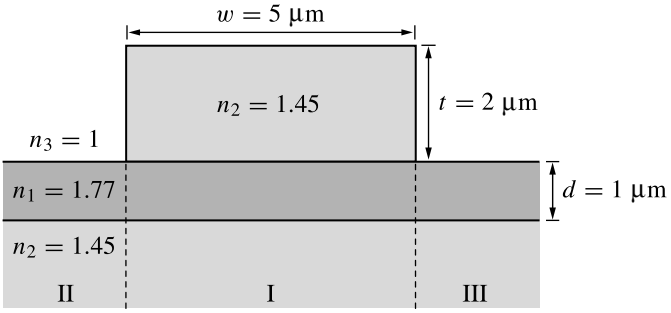


Figure 2.16 Strip-loaded waveguide for the effective index method.

for $t = 2$ μm, we can safely say that the evanescent field is completely confined within the strip. Therefore, the structure in region I is simply the symmetric waveguide solved in Example 2.4 with $\beta_0^{\mathrm{I}} = 10.8208$ μm^{-1} for the TM_0 mode. The structures in regions II and III are just the asymmetric waveguide solved in Example 2.1 with $\beta_0^{\mathrm{II}} = \beta_0^{\mathrm{III}} = 10.7800$ μm^{-1} for the TM_0 mode. The vertical planar waveguide is thus a symmetric waveguide that has a width of $w = 5$ μm and effective indices of $n_{\mathrm{I}} = 1.7222$ and $n_{\mathrm{II}} = n_{\mathrm{III}} = 1.7157$. The V number of this effective planar waveguide at $\lambda = 1$ μm is

$$V = \frac{2\pi}{\lambda} w \sqrt{n_{\mathrm{I}}^2 - n_{\mathrm{II}}^2} = 4.696.$$

It has two TE modes and two TM modes because $2\pi > V > \pi$.

Because the TM-like modes of the strip waveguide are polarized in the x direction, we have to consider the TE modes of the effective vertical planar waveguide. Because there are two such TE modes, we have two TM-like modes, TM_{00} and TM_{01}, which are associated with the TM_0 mode of the horizontal planar waveguide and the TE_0 and TE_1 modes, respectively, of the effective vertical planar waveguide. The characteristics of these TE_0 and TE_1 modes can be solved in the same manner as that described in Example 2.4 because the effective vertical waveguide is symmetric. The results are summarized below.

	h_{I} (μm^{-1})	γ_{II} (μm^{-1})	ψ_n (rad)	β_n (μm^{-1})	w_n (μm)	Γ_n	β_{mn} (μm^{-1})	Γ_{mn}
TM_{00}	0.435	0.832	0	10.8120	7.4	0.930	10.8120	0.899
TM_{01}	0.826	0.446	$\pi/2$	10.7892	9.5	0.634	10.7892	0.613

In this table, w_n is the effective width of the TM_{0n} mode in the y direction similar to the effective thickness d_{M} in the x direction for the TM_0 mode.

Except for those in the last two columns, the parameters listed above are for the y dependence of the modes. Both mode fields have the same x dependence described by $\mathcal{H}_{0,y}(x)$, which has the characteristics of the TM_0 mode listed in Example 2.4. The y dependence is found by using the parameters listed above to obtain $\mathcal{E}_{0,x}(y)$

and $\mathcal{E}_{1,x}(y)$ of the TE_0 and TE_1 modes of the effective planar waveguide, respectively. Using $\mathcal{H}_y = -\beta\mathcal{E}_x/\omega\mu_0$ after exchanging the x and y coordinates for (2.32), we then find $\mathcal{H}_{0,y}(y)$ and $\mathcal{H}_{1,y}(y)$. The y component of the total mode field for the TM_{00} mode is then $\mathcal{H}_{00,y}(x, y) = \mathcal{H}_{0,y}(x)\mathcal{H}_{0,y}(y)$ and that for the TM_{01} mode is $\mathcal{H}_{01,y}(x, y) = \mathcal{H}_{0,y}(x)\mathcal{H}_{1,y}(y)$. The propagation constants are simply those found from solving for the effective vertical waveguide: $\beta_{00} = \beta_0 = 10.8120\ \mu m^{-1}$ and $\beta_{01} = \beta_1 = 1.7892\ \mu m^{-1}$. The effective mode confinement factor Γ_{mn}, defined as its fractional power in the $d \times w$ two-dimensional guiding core, can be found by multiplying its two confinement factors in the x and y dimensions. Thus, we have $\Gamma_{00} = 0.967 \times 0.930 = 0.899$ for the TM_{00} mode and $\Gamma_{01} = 0.967 \times 0.634 = 0.613$ for the TM_{01} mode.

PROBLEMS

2.1.1 When total internal reflection occurs at the interface of two dielectric media under the condition that the incident angle θ_i is larger than the critical angle θ_c, as shown in Fig. 2.17, the reflected wave acquires a phase shift with respect to the incident wave. This phase shift depends on the polarization of the wave and is a function of the incident angle θ_i. Assume that $n_1 > n_2$ and that the plane of incidence is the zx plane as shown.

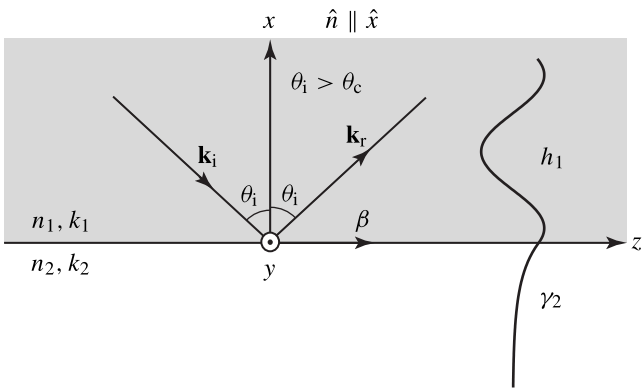


Figure 2.17 Total internal reflection.

a. Show that the incident and reflected waves in a medium of refractive index n_1 and the evanescent wave in a medium of refractive index n_2 all vary with z for a propagation constant β, which is a function of θ_i. Show also that $k_1 > \beta > k_2$, where $k_1 = n_1\omega/c$ and $k_2 = n_2\omega/c$, and ω is the frequency of the optical wave.

b. The positive real parameters h_1 and γ_2 are defined by the relations given in (2.50) and (2.51), respectively. Describe how these two parameters characterize the spatial variations of the incident, reflected, and evanescent waves.

c. How are h_1 and γ_2 related to n_1, n_2, and θ_i?

d. Show that for a TE-polarized wave, the phase shift upon total internal reflection is given by

$$\varphi_{TE} = -2\tan^{-1}\frac{\gamma_2}{h_1}. \tag{2.107}$$

e. Show that for a TM-polarized wave, the phase shift upon total internal reflection is

$$\varphi_{TM} = -2\tan^{-1}\frac{n_1^2\gamma_2}{n_2^2 h_1}. \tag{2.108}$$

f. Plot the phase shifts as a function of incident angle for TE-polarized and TM-polarized waves at the interface of air and ordinary glass, which has a refractive index of 1.5.

2.2.1 Derive the field equations given in (2.16)–(2.19) for the field components of a waveguide mode from the two Maxwell's equations in (2.8) and (2.9) by using the mode field definition given in (2.1) and (2.2).

2.2.2 Can a dielectric waveguide support a TEM mode? Why?

2.2.3 Show that hybrid modes do not exist in a planar waveguide, such as a slab waveguide.

2.2.4 What kinds of guided modes can exist in a planar dielectric waveguide? What kinds can exist in a nonplanar dielectric waveguide? What kinds can exist in a nonplanar metallic waveguide? Under what condition can each kind of guided mode exist in a particular dielectric or metallic waveguide?

2.3.1 Show that Maxwell's equations lead to the two inhomogeneous wave equations given in (2.24) and (2.25) in the case when $\nabla\epsilon \neq 0$. Show that (2.24) and (2.25) result in (2.30) and (2.34), respectively, in the case of a planar waveguide, even when the waveguide has an arbitrary index profile $n(x)$ such that $\nabla\epsilon \neq 0$.

2.4.1 In general, the orthonormality relation of dielectric waveguide modes takes the form of (2.41), and the power of a waveguide mode ν has to be obtained by using (2.40) with the mode indices $\mu = \nu$. Though TE and TM modes can be treated generally like any other types of modes, they are somewhat special.

a. Show that the orthonormality relation among TE modes of a waveguide can be expressed in the form of (2.44), and the power of a TE mode can be obtained by using (2.38).

b. Show that the orthonormality relation among TM modes of a waveguide can be expressed in the form of (2.45), and the power of a TM mode can be obtained by using (2.39).

c. Show that (2.44) does not apply to TM modes and that (2.45) does not apply to TE modes.

d. Show that neither (2.44) nor (2.45) applies to hybrid waveguide modes.

2.5.1 What is the mode that has the largest propagation constant in a symmetric slab waveguide? Sketch its field and intensity profiles.

2.5.2 How does the propagation constant β of a particular normal mode vary with optical frequency in a given slab waveguide? What is the allowed range of β for the mode to exist and remain guided?

2.5.3 Light can enter a semiconductor slab waveguide in different polarization states. Describe what happens when light in each of the following input polarization states is coupled into the waveguide: (a) linearly polarized in a direction parallel to the boundaries of the core layer, (b) linearly polarized in a direction perpendicular to the boundaries of the core layer, (c) linearly polarized in a direction that is neither parallel nor perpendicular to the boundaries of the core layer, and (d) circularly polarized.

2.5.4 Show that the cutoff conditions for the mth guided TE and TM modes of an asymmetric slab waveguide are those given in (2.72) and (2.73), respectively.

2.5.5 Show that the confinement factors for guided TE and TM modes of an asymmetric slab waveguide are those given in (2.79) and (2.80), respectively. Show that they reduce to those in (2.91) and (2.92), respectively, in the case of a symmetric slab waveguide.

2.5.6 It is found that an asymmetric slab waveguide supports two TE modes and two TM modes at each of the two wavelengths $\lambda = 1.3$ and 1.55 μm.

a. Among these modes, which one has the largest β and which one has the smallest β?

b. Which ones have the largest and the smallest confinement factors?

c. If we start reducing the core thickness of the waveguide while maintaining the index profile, which mode gets cut off first? Which one gets cut off last?

2.5.7 In an asymmetric slab waveguide that guides three TE modes and three TM modes, which mode has the largest propagation constant? As the core thickness of this waveguide is reduced while the index profile is maintained, which mode gets cut off first?

2.5.8 A multimode asymmetric slab waveguide has exactly five guided modes. What are they? Which one has the largest propagation constant? Which one has the smallest propagation constant? Within what range do their propagation constants fall? Which one has the largest confinement factor? Which one has the smallest confinement factor?

2.5.9 For the asymmetric waveguide discussed in Example 2.1, find the wavelength range within which it is a single-mode waveguide for TE polarization. What is the wavelength range for it to be single moded for TM polarization?

2.5.10 An asymmetric slab waveguide made of a polymer layer of thickness $d = 2$ μm deposited on a silica substrate has the same index profile as the one discussed in

Example 2.1. At 1 μm optical wavelength, $n_1 = 1.77$ for the polymer guiding layer, $n_2 = 1.45$ for the silica substrate, and $n_3 = 1$ for the air cover.

a. How many guided modes are supported by this waveguide?
b. Find the propagation constants and the confinement factors for the guided TE and TM modes of this waveguide.
c. Plot the mode field distributions of the guided modes.
d. If the thickness is reduced, some existing modes will be cut off. Which one is cut off first? At what thickness is it cut off?
e. If the thickness is increased, additional modes will be supported. At what thickness will one additional mode be supported? What is that mode?

2.5.11 Use the index profile, $n_1 = 1.77, n_2 = 1.45$, and $n_3 = 1$, as that of the waveguide in Example 2.1 to design a single-mode asymmetric waveguide for $\lambda = 1$ μm wavelength.

a. Design a waveguide that supports only one mode for each polarization by choosing a proper thickness d for the guiding layer. What are these two modes? Find the propagation constants and the confinement factors for these two modes.
b. It is possible to choose a thickness for the waveguide to support one and exactly one mode between both polarizations. Design such a waveguide. What is this mode? What are its propagation constant and confinement factor?

2.6.1 Eigenvalue equations similar to (2.87) and (2.88) can be obtained for asymmetric waveguides.

a. Show that for TE modes, we have

$$\tan\left(\frac{h_1 d}{2} - \frac{m\pi}{2}\right) = \frac{\sqrt{(h_1^2 + \gamma_2^2)(h_1^2 + \gamma_3^2)} - h_1^2 + \gamma_2\gamma_3}{h_1(\gamma_2 + \gamma_3)}. \tag{2.109}$$

Express this equation in terms of $h_1 d$, the waveguide parameters V, and a_E.

b. Show that for TM modes, the eigenvalue equation can be obtained by replacing the parameters h_1, γ_2, and γ_3 on the right-hand side of (2.109) with h_1/n_1^2, γ_2/n_2^2, and γ_3/n_3^2, respectively. Express this equation in terms of $h_1 d$, V, and a_M.

2.6.2 For TE and TM modes of the same order in the same symmetric slab waveguide, which one has a larger propagation constant? Which one has a higher cutoff frequency?

2.6.3 Sketch the electric field and intensity distribution patterns of the TE_5 and TM_5 modes of a symmetric slab waveguide.

2.6.4 Sketch the field distribution patterns of the first two TE modes and the first two TM modes of the multilayer symmetric waveguide shown in Fig. 2.18.

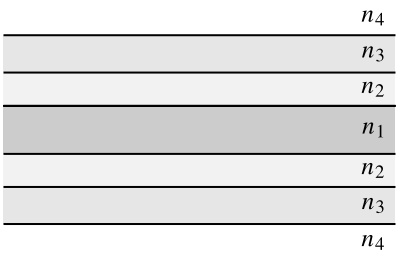


Figure 2.18 Seven-layer symmetric slab waveguide.

2.6.5 What is the largest thickness d of a planar symmetric dielectric waveguide with refractive indices of $n_1 = 1.50$ and $n_2 = 1.46$ for which there is only one TE mode at $\lambda = 1.3$ μm? What is the number of modes if a waveguide with this thickness is used at λ = 850 nm instead? What are those modes? What are the confinement factors of the fundamental TE mode for these two cases?

2.6.6 A symmetric slab waveguide has a confinement factor of $\Gamma_0^{\mathrm{TE}} = 0.7$ for the TE_0 mode at 1.55 μm optical wavelength. Neglect the dispersion of the waveguide material.

a. What is the confinement factor for the TE_0 mode at 1.3 μm wavelength?
b. Is the waveguide single moded or multimoded at each wavelength? Explain briefly.
c. If the thickness of the waveguide core is quadrupled while the same index profile is maintained, how many modes (including both TE and TM modes) can the waveguide support at each of the two wavelengths considered here?

2.6.7 A symmetric slab waveguide has a core thickness of 2 μm. Ignoring the dispersion of the waveguide material, we find the indices to be $n_1 = 1.50$ and $n_2 = 1.46$.

a. Is this waveguide single moded or multimoded at $\lambda = 1.5$ and 1.3 μm?
b. What is the range of wavelength in which this waveguide is single moded?
c. If we want to make the waveguide to be single moded at both $\lambda = 1.5$ and 1.3 μm, how should we change the waveguide parameter? (A qualitative answer is sufficient.)

2.6.8 A symmetric slab waveguide is found to support exactly five TE and five TM modes at an optical wavelength $\lambda = 500$ nm. Assume that dispersion of the waveguide material is negligible.

a. How many TE and TM modes does it support at $\lambda = 1$ μm?
b. Which mode among those at 500 nm and 1 μm wavelengths has the largest propagation constant?

2.6.9 A symmetric slab waveguide is used to guide signals at two wavelengths, 1.55 and 1.3 μm, simultaneously. Neglecting the dispersion in the material, the

refractive indices are $n_1 = 1.50$ and $n_2 = 1.46$ in the core and cladding layers, respectively. The core thickness is 1.5 μm.

a. How many TE modes does the waveguide support at each of the wavelengths?
b. Which mode among those TE modes in (a) has the largest confinement factor and how much is it?
c. If the core thickness is increased, additional modes will show up. At what core thickness will the first additional TE mode show up? What is this TE mode? Specify the order of the mode and its wavelength.

2.6.10 A symmetric waveguide is made by sandwiching a layer of $Al_xGa_{1-x}As$ between two layers of $Al_yGa_{1-y}As$. At an optical wavelength $\lambda = 900$ nm, the index of refraction of $Al_xGa_{1-x}As$ is $n = 3.59 - 0.710x + 0.091x^2$. With these parameters, design a symmetric single-mode waveguide for TE polarization by properly choosing the waveguide core thickness d and the material composition indices x and y. Calculate the confinement factor for the guided TE mode.

2.6.11 A semiconductor slab waveguide is formed by a double heterostructure, with the core of thickness d being GaAs and both cladding layers being $Al_xGa_{1-x}As$, as shown in Fig. 2.19. The index of refraction of the ternary semiconductor material $Al_xGa_{1-x}As$ depends on the fractional aluminum content x. It has a lower index of refraction than GaAs. At an optical wavelength $\lambda = 900$ nm, the refractive index of GaAs is $n_{GaAs} = 3.59$, while that of $Al_xGa_{1-x}As$ is $n_x = 3.59 - 0.710x + 0.091x^2$. With these parameters, we want to design symmetric single-mode waveguides for TE polarization by properly choosing the waveguide core thickness d and the material composition x.

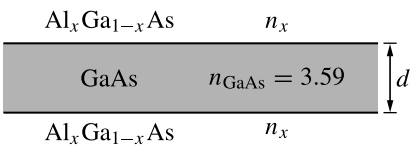


Figure 2.19 Symmetric GaAs/AlGaAs slab waveguide.

a. If x is fixed at 0.3, what is the range of d that allows only the fundamental mode?
b. If d is fixed at 2 μm, what is the range of x that allows only the fundamental mode?
c. Design a single-mode waveguide that has a confinement factor of $\Gamma_{TE} = 0.5$ for the TE_0 mode.

2.6.12 The index of refraction of $Al_xGa_{1-x}As$ at $\lambda = 1.3$ μm is $n = 3.41 - 0.52x$ and that at $\lambda = 1.5$ μm is $n = 3.38 - 0.52x$. You are asked to design a symmetric slab waveguide with $Al_xGa_{1-x}As$ cladding layers for $x < 0.3$ and a GaAs core. It is desired that the waveguide has the following characteristics: (a) it is single

moded at both wavelengths, (b) it has a core thickness as small as possible, (c) it has the largest confinement factor possible for both wavelengths.

2.6.13 In this problem, we would like to design a single-mode symmetric InGaAsP/InP semiconductor slab waveguide for the 1.3 μm optical wavelength. The substrate and cladding of such a waveguide are made of InP, which has an index of refraction of $n_2 = 3.205$ at 1.3 μm. The core is made of $In_{1-x}Ga_xAs_yP_{1-y}$, which has a higher index than InP. Both the bandgap and the refractive index of the core material are determined by the composition indices x and y. In order to keep the absorption at 1.3 μm low enough so that the waveguide has a low loss, we need to keep $y < 0.5$. This means that the core index is limited to $n_1 < 3.435$. Assume that we are able to control the material composition only to an accuracy that allows an accuracy in the refractive index no better than $\delta(\Delta n) = 0.01$. We want to design a single-mode waveguide that has a confinement factor of at least 0.8.

a. What is the allowed range of the value for the V parameter?
b. What are the maximum and minimum limits of the waveguide core thickness d set by the requirements?
c. Give one example of your design that satisfies the requirements.

2.7.1 Find the wavelength range within which the $Ti:LiNbO_3$ waveguide designed in Example 2.5 is single moded for both TE and TM polarizations. In what wavelength range does the waveguide support a single TE mode but not support any TM mode? If the index step is doubled to $\Delta n = 0.02$ but the Ti diffusion depth remains unchanged at $d = 2$ μm, within what wavelength range does the waveguide remain single moded for both polarizations? Ignore the dispersion of $LiNbO_3$ and that of the index profile in sloving this problem.

2.7.2 A $Ti:LiNbO_3$ waveguide similar to the one designed in Example 2.5 is made by Ti diffusion at 1020 °C for 6 hours. What index step Δn should be chosen so that the waveguide is single moded for both TE and TM polarizations in the wavelength range between 500 nm and 1.3 μm?

2.7.3 Show that the relations given in (2.105) and (2.106) for a step-bounded graded-index waveguide reduce to those given in (2.75) and (2.76) for a step-index waveguide if we transform the step-bounded graded-index waveguide into a step-index waveguide by setting $n(x) = n_1$ for $a < x < b$ and $n(x) = n_2$ for $x < a$.

2.8.1 How many TE_{0n} modes does the strip-loaded waveguide discussed in Example 2.6 support at $\lambda = 1$ μm? Find their characteristics by using the effective index method.

2.8.2 A rib waveguide, shown in Fig. 2.20, is formed out of the planar waveguide discussed in Example 2.1 by raising a strip of polymer that has a width of $w = 5$ μm and a thickness of $t = 1$ μm atop the polymer guiding layer. In this structure, the rib is part of the waveguiding core, which has a width of

$w = 5$ μm and a thickness of $d + t = 2$ μm. At $\lambda = 1$ μm, how many TE_{0n} modes and how many TM_{0n} modes are supported by this waveguide? What are they? Find the characteristics of the TE_{00} and TM_{00} modes of this rib waveguide at $\lambda = 1$ μm using the effective index method. Which mode has the largest propagation constant? Which one has the largest confinement factor?

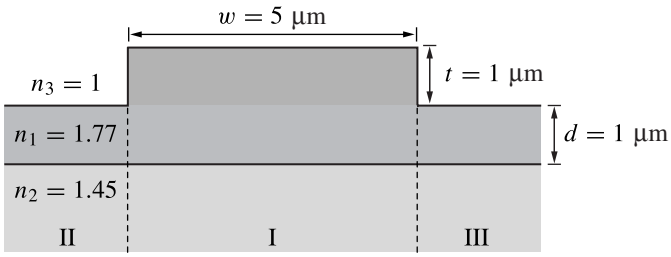


Figure 2.20 Rib waveguide.

SELECT BIBLIOGRAPHY

Buckman, A. B., *Guided-Wave Photonics*. Fort Worth, TX: Sauders College Publishing, 1992.

Ebeling, K. J., *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*. Berlin: Springer-Verlag, 1993.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Iizuka, K., *Elements of Photonics for Fiber and Integrated Optics*, Vol. II. New York: Wiley, 2002.

Kasap, S. O., *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice-Hall, 2001.

Kressel, H. and Butler, J. K., *Semiconductor Lasers and Heterojunction LEDs*. New York: Academic Press, 1977.

Marcuse, D., *Theory of Dielectric Optical Waveguides*, 2nd edn. Boston, MA: Academic Press, 1991.

Nishihara, H., Haruna, M., and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Syms, R. and Cozens, J., *Optical Guided Waves and Devices*, London: McGraw-Hill, 1992.

Tamir, T., ed., *Integrated Optics*, New York: Springer-Verlag, 1982.

Yariv, A. and Yeh, P., *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: Wiley, 1984.

ADVANCED READING LIST

Buus, J., "The effective index method and its application to semiconductor lasers," *IEEE Journal of Quantum Electronics* **QE-18**(7): 1083–1089, July 1982.

Chiang, K. S., "Review of numerical and approximate methods for the modal analysis of general optical dielectric waveguides," *Optical and Quantum Electronics* **26**(3): S113–S134, Mar. 1994.

Hocker, G. B. and Burns, W. K., "Mode dispersion in diffused channel waveguides by the effective index method," *Applied Optics* **16**(1): 113–118, Jan. 1977.

Marcatili, E. A. J., "Dielectric rectangular waveguide and directional coupler for integrated optics," *Bell System Technical Journal* **8**(7): 2071–2102, Sep. 1969.

Olshansky, R., "Propagation in glass optical waveguides," *Reviews of Modern Physics* **51**(2): 341–367, Apr. 1979.

Tien, P. K., "Integrated optics and new wave phenomena in optical waveguides," *Reviews of Modern Physics* **49**(2): 361–420, Apr. 1977.

3 Optical fibers

An optical fiber is basically a cylindrical dielectric waveguide with a circular cross section where a high-index waveguiding *core* is surrounded by a low-index *cladding*. Optical fibers are usually made of silica (SiO_2) glass. The index step and profile are controlled by the concentration and distribution of dopants. For example, the core can be doped with germania (GeO_2) or alumina (Al_2O_3) or other oxides, such as P_2O_5 or TiO_2, for a slightly higher index than that of a silica cladding. Alternatively, to take advantage of low-loss pure silica, the cladding can be doped with fluorine for a slightly lower index while the core contains undoped pure silica. Silica fibers are ideal for light transmission in the visible and near-infrared regions because of their low loss and low dispersion in these spectral regions. They are therefore suitable for optical communications and most laser applications in this range of the spectrum. Optical fibers made of other materials are also developed for special applications. For example, low-cost plastic fibers can be used for short-distance interconnections between personal computers and printers in offices. Fibers composed of ZrF_4, BaF_2, AlF_3, LiF_3, and other fluorides have a low loss in the range of 2–4 μm in the mid infrared. They can be used for mid-infrared optical communication or medical applications. Fibers for other spectral regions, such as the 10-μm region of CO_2 laser wavelengths, are also developed.

Optical fibers have a wide range of applications. Owing to their low losses and large bandwidths, their most important applications are fiber-optic communications and interconnections. Other important applications include fiber sensors, guided optical imaging, remote monitoring, and medical applications. With active dopants, such as neodymium or erbium, fibers with an optical gain under optical pumping are also used as optical amplifiers and fiber lasers, opening up many new applications. In addition, because optical fibers provide strong optical confinement over long distances, they also present unique conditions for many interesting nonlinear optical processes, which lead to such applications as optical soliton formation and propagation, optical pulse compression, and optical frequency conversion. Within the photonics community, fiber-optic components and systems form a major industry by themselves. In this chapter, we discuss the important characteristics of optical fibers.

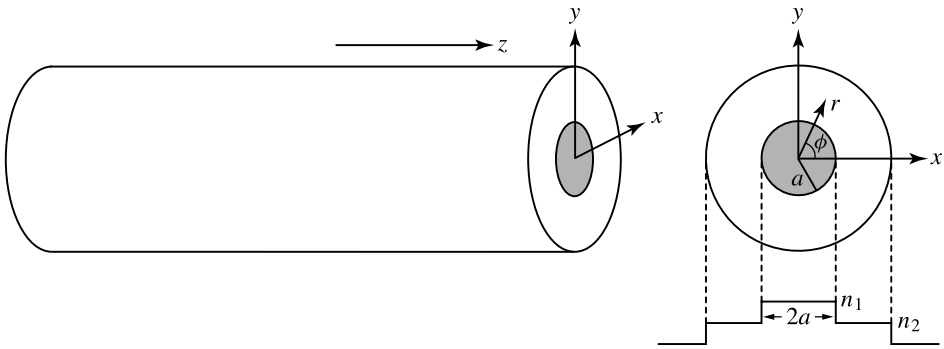


Figure 3.1 Step-index optical fiber with a core radius a.

3.1 Step-index fibers

A step-index fiber is a nonplanar step-index waveguide that has a circular cross section, as shown in Fig. 3.1. The core has a radius a. The *core diameter*, $2a$, typically ranges from a few micrometers for a *single-mode fiber* to less than 100 μm for a *multimode fiber*. It is designed for the fiber to support a desired number of guided modes. The *outer diameter* of a fiber is that of the outside boundary of its cladding, which is typically about 100 μm or somewhat larger. The outer diameter of a fiber is determined by the requirement that the cladding be thicker than the penetration depth of a guided-mode field to prevent the field from reaching the air–cladding boundary and by the consideration of easy handling. The standard outer diameter size for multimode silica fibers is 125 μm.

For a step-index fiber, the waveguide parameter V, also called the *V number* of the fiber, is defined as

$$V = \frac{2\pi}{\lambda} a\sqrt{n_1^2 - n_2^2} = \frac{\omega}{c} a\sqrt{n_1^2 - n_2^2}. \tag{3.1}$$

The *numerical aperture* of the fiber is

$$\mathrm{NA} = \sqrt{n_1^2 - n_2^2} = \sin\theta_\mathrm{a}, \tag{3.2}$$

which determines the *acceptance angle*, θ_a, of an optical fiber. Therefore, the acceptance angle of a circular fiber is simply $\theta_\mathrm{a} = \sin^{-1}(\mathrm{NA}) = \sin^{-1}\sqrt{n_1^2 - n_2^2}$. The acceptance angle is the largest incident angle, with respect to the normal of the end surface of a fiber, that allows an optical beam to be coupled into the fiber core. A wave entering the fiber at an incident angle smaller than the acceptance angle will be totally reflected at the core–cladding interface and thus will be guided in the fiber core. A wave entering at an incident angle larger than θ_a will be partially transmitted through the core–cladding interface after entering the fiber and will not be guided.

EXAMPLE 3.1 A step-index silica fiber has a core index of 1.452, a cladding index of 1.449, and a core diameter of 8 μm. What are its numerical aperture and acceptance angle? What is the value of its V number at 850 nm wavelength?

Solution For this fiber, $n_1 = 1.452$, $n_2 = 1.449$, and $a = 4$ μm. The numerical aperture is

$$\text{NA} = \sqrt{n_1^2 - n_2^2} = 0.093.$$

The acceptance angle is

$$\theta_\text{a} = \sin^{-1} 0.093 = 5.34^\circ.$$

The V number at $\lambda = 850$ nm is

$$V = \frac{2\pi}{\lambda} a \sqrt{n_1^2 - n_2^2} = 2.758.$$

The mode fields of a circular fiber are best described in cylindrical coordinates with

$$\mathbf{E}_{mn}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_{mn}(\phi, r) \exp(\mathrm{i}\beta_{mn} z - \mathrm{i}\omega t), \tag{3.3}$$

$$\mathbf{H}_{mn}(\mathbf{r}, t) = \boldsymbol{\mathcal{H}}_{mn}(\phi, r) \exp(\mathrm{i}\beta_{mn} z - \mathrm{i}\omega t). \tag{3.4}$$

Note that the first index, m, is associated with the coordinate ϕ, while the second index, n, is associated with the coordinate r. This designation of indices will become clear later. The field equations obtained in Section 2.2 are general equations for waveguides. They can be used for a circular fiber by transforming x and y coordinates to r and ϕ coordinates. For example, (2.16)–(2.19) become

$$(k^2 - \beta^2)\mathcal{E}_r = \mathrm{i}\beta \frac{\partial \mathcal{E}_z}{\partial r} + \mathrm{i}\omega\mu_0 \frac{1}{r} \frac{\partial \mathcal{H}_z}{\partial \phi}, \tag{3.5}$$

$$(k^2 - \beta^2)\mathcal{E}_\phi = \mathrm{i}\beta \frac{1}{r} \frac{\partial \mathcal{E}_z}{\partial \phi} - \mathrm{i}\omega\mu_0 \frac{\partial \mathcal{H}_z}{\partial r}, \tag{3.6}$$

$$(k^2 - \beta^2)\mathcal{H}_r = \mathrm{i}\beta \frac{\partial \mathcal{H}_z}{\partial r} - \mathrm{i}\omega\epsilon \frac{1}{r} \frac{\partial \mathcal{E}_z}{\partial \phi}, \tag{3.7}$$

$$(k^2 - \beta^2)\mathcal{H}_\phi = \mathrm{i}\beta \frac{1}{r} \frac{\partial \mathcal{H}_z}{\partial \phi} + \mathrm{i}\omega\epsilon \frac{\partial \mathcal{E}_z}{\partial r}, \tag{3.8}$$

where $k^2 = \omega^2 \mu_0 \epsilon(r) = \omega^2 n^2(r)/c^2$.

For a step-index fiber, (2.27) and (2.28) in Section 2.3 are also valid, but they take the following form in cylindrical coordinates:

$$\frac{\partial^2 \mathcal{E}_z}{\partial r^2} + \frac{1}{r} \frac{\partial \mathcal{E}_z}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \mathcal{E}_z}{\partial \phi^2} + (k_i^2 - \beta^2)\mathcal{E}_z = 0, \tag{3.9}$$

$$\frac{\partial^2 \mathcal{H}_z}{\partial r^2} + \frac{1}{r} \frac{\partial \mathcal{H}_z}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \mathcal{H}_z}{\partial \phi^2} + (k_i^2 - \beta^2)\mathcal{H}_z = 0, \tag{3.10}$$

where $k_1^2 = \omega^2 n_1^2/c^2$ for the core and $k_2^2 = \omega^2 n_2^2/c^2$ for the cladding for $i = 1$ and 2, respectively. For guided modes, we have $k_1 > \beta > k_2$ and

$$k_1^2 - \beta^2 = h^2, \tag{3.11}$$

$$\beta^2 - k_2^2 = \gamma^2. \tag{3.12}$$

In general, fiber modes can be hybrid modes with $\mathcal{E}_z \neq 0$ and $\mathcal{H}_z \neq 0$. Therefore, (3.9) and (3.10) have to be solved simultaneously. They can be solved by separation of variables. For example, for $\mathcal{E}_z$, the solution for ϕ dependence yields

$$\mathcal{E}_z(\phi, r) = R(r)\mathrm{e}^{\pm \mathrm{i}m\phi}, \qquad m = 0, 1, 2, \ldots, \tag{3.13}$$

where $R(r)$ satisfies

$$\frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}R}{\mathrm{d}r} + \left(h^2 - \frac{m^2}{r^2}\right) R = 0, \qquad \text{for } r < a, \tag{3.14}$$

$$\frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}R}{\mathrm{d}r} - \left(\gamma^2 + \frac{m^2}{r^2}\right) R = 0, \qquad \text{for } r > a. \tag{3.15}$$

Equations of the same form define the dependence of $\mathcal{H}_z$ on ϕ and r. The solution of (3.14) with the requirement that $R(r)$ be finite at $r = 0$ is $J_m(hr)$, the Bessel function of the first kind of order m. Meanwhile, the solution of (3.15) with the requirement that $rR^2(r) \to 0$ as $r \to \infty$ yields $K_m(\gamma r)$, the modified Bessel function of the second kind of order m. Thus the r dependence of $\mathcal{E}_z$ and $\mathcal{H}_z$ is found to be $J_m(hr)$ for $r < a$ and $K_m(\gamma r)$ for $r > a$. The leading orders of $J_m(x)$ and $K_m(x)$ are plotted in Figs. 3.2(a) and (b), respectively. These Bessel functions have the following properties:

$$J_0(0) = 1, \qquad J_{m\neq 0}(0) = 0, \tag{3.16}$$

$$K_m(0) = \infty, \tag{3.17}$$

and, for large values of x,

$$J_m(x) \approx \sqrt{\frac{2}{\pi x}}\left[\cos\left(x - \frac{m\pi}{2} - \frac{\pi}{4}\right) - \frac{4m^2 - 1}{8x}\sin\left(x - \frac{m\pi}{2} - \frac{\pi}{4}\right)\right], \tag{3.18}$$

$$K_m(x) \approx \sqrt{\frac{\pi}{2x}}\left(1 + \frac{4m^2 - 1}{8x}\right)\mathrm{e}^{-x}. \tag{3.19}$$

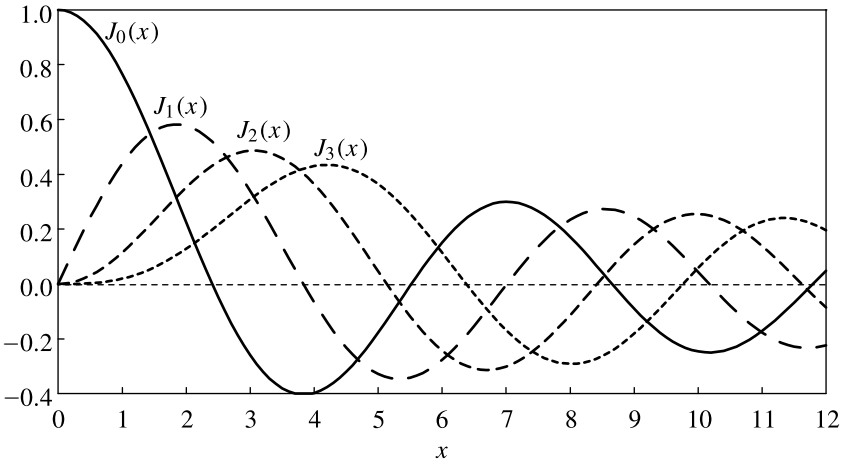
The following identities of the Bessel functions are also found to be useful:

$$J_{-m} = (-1)^m J_m, \qquad K_{-m} = K_m, \tag{3.20}$$

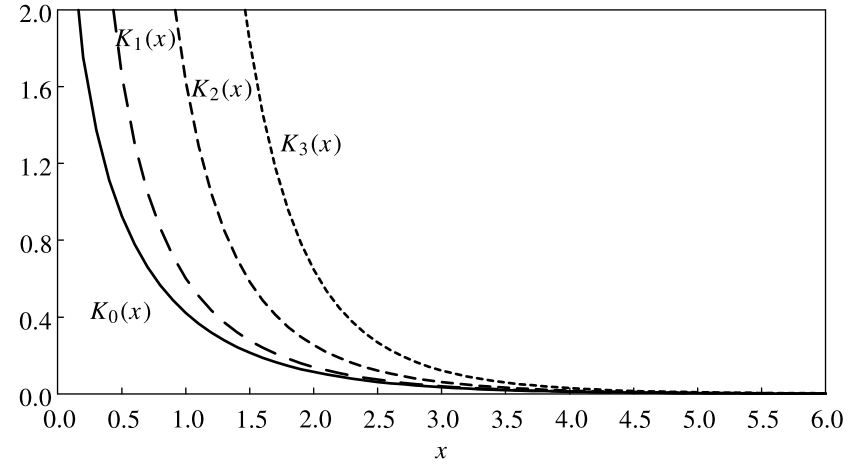
$$J_m' = \frac{1}{2}(J_{m-1} - J_{m+1}), \qquad K_m' = -\frac{1}{2}(K_{m-1} + K_{m+1}), \tag{3.21}$$

$$\frac{m}{x}J_m = \frac{1}{2}(J_{m-1} + J_{m+1}), \qquad \frac{m}{x}K_m = -\frac{1}{2}(K_{m-1} - K_{m+1}). \tag{3.22}$$

Once $\mathcal{E}_z$ and $\mathcal{H}_z$ are solved, the other field components can be found using (3.5)–(3.8). The boundary conditions require that the tangential field components, $\mathcal{E}_z$, $\mathcal{E}_\phi$, $\mathcal{H}_z$, and



(*a*)



(*b*)

Figure 3.2 Leading orders of (*a*) the Bessel functions $J_m(x)$ and (*b*) the modified Bessel functions $K_m(x)$.

$\mathcal{H}_\phi$, be continuous at the boundary, $r = a$, between the core and the cladding. These conditions result in the requirement that the ϕ dependence of $\mathcal{E}_z$ be 90° out-of-phase with respect to that of $\mathcal{H}_z$. Therefore, we can choose

$$\mathcal{E}_z(\phi, r) = \begin{cases} A_m J_m(hr) \cos m\phi, & r < a, \\ B_m K_m(\gamma r) \cos m\phi, & r > a, \end{cases} \tag{3.23}$$

$$\mathcal{H}_z(\phi, r) = \begin{cases} C_m J_m(hr) \sin m\phi, & r < a, \\ D_m K_m(\gamma r) \sin m\phi, & r > a, \end{cases} \tag{3.24}$$

where A_m, B_m, C_m, and D_m are constants to be found for a particular fiber mode. Alternatively, we can choose

$$\mathcal{E}_z(\phi, r)=\begin{cases} A'_m J_m(hr)\sin m\phi, & r < a, \\ B'_m K_m(\gamma r)\sin m\phi, & r > a, \end{cases} \tag{3.25}$$

$$\mathcal{H}_z(\phi, r)=\begin{cases} C'_m J_m(hr)\cos m\phi, & r < a, \\ D'_m K_m(\gamma r)\cos m\phi, & r > a, \end{cases} \tag{3.26}$$

where A'_m, B'_m, C'_m, and D'_m are also constants for a particular fiber mode. For $m \neq 0$, these two sets of choices are degenerate because one can be transformed into the other by a change of reference of the angle ϕ for one or the other, which has no physical significance in a circular fiber. However, for $m = 0$, they represent distinctly different sets of modes, as discussed below.

Application of the boundary conditions for a nontrivial solution of A_m, B_m, C_m, and D_m for (3.23) and (3.24) or that of A'_m, B'_m, C'_m, and D'_m for (3.25) and (3.26) yields the following eigenvalue equation for the allowed values of h and γ for the guided modes:

$$\left[\frac{J'_m(ha)}{haJ_m(ha)}+\frac{K'_m(\gamma a)}{\gamma aK_m(\gamma a)}\right]\left[\frac{n_1^2J'_m(ha)}{haJ_m(ha)}+\frac{n_2^2K'_m(\gamma a)}{\gamma aK_m(\gamma a)}\right] = m^2\frac{c^2\beta^2}{\omega^2}\left(\frac{1}{h^2a^2}+\frac{1}{\gamma^2a^2}\right)^2, \tag{3.27}$$

where J'_m and K'_m are the derivatives of the Bessel functions. Recall that each mode in a circular fiber is characterized by two mode indices m and n. As seen above, the first index m refers to the angular dependence $\cos m\phi$ or $\sin m\phi$. The second index n refers to the order of the allowed solutions for eigenvalues h or, equivalently, γ. Therefore, m is called the *azimuthal mode index*, or the *angular mode index*, while n is called the *radial mode index*. In general, (3.27) has to be solved numerically.

Fiber modes

It can be seen that when $m = 0$, the first set of solutions for the longitudinal components of the mode fields given in (3.23) and (3.24) results in the TM fields with $\mathcal{H}_z = 0$, while the second set of solutions given in (3.25) and (3.26) results in the TE fields with $\mathcal{E}_z = 0$. Therefore, *for $m = 0$, the guided modes are either TE or TM modes*, and (3.27) becomes two separate eigenvalue equations:

$$\frac{J_1(ha)}{haJ_0(ha)}+\frac{K_1(\gamma a)}{\gamma aK_0(\gamma a)} = 0, \qquad \text{for TE modes,} \tag{3.28}$$

and

$$\frac{n_1^2J_1(ha)}{haJ_0(ha)}+\frac{n_2^2K_1(\gamma a)}{\gamma aK_0(\gamma a)} = 0, \qquad \text{for TM modes,} \tag{3.29}$$

where the relations $J'_0 = -J_1$ and $K'_0 = -K_1$ are used.

For $m \geq 1$, the guided modes in a circular fiber are hybrid modes. Both $\mathcal{E}_z$ and $\mathcal{H}_z$ exist in these modes. As a result, all six field components exist. In this case, the solution given in (3.23) and (3.24) is degenerate with that given in (3.25) and (3.26) with

$$\frac{A_m}{C_m} = -\frac{A'_m}{C'_m} \qquad \text{and} \qquad \frac{B_m}{D_m} = -\frac{B'_m}{D'_m}. \tag{3.30}$$

Therefore, for a hybrid mode, we only have to consider the solutions given by, say, (3.23) and (3.24). The hybrid modes can be classified into two groups. Those with A_m and C_m having the same sign are called *HE modes*, while those with A_m and C_m having opposite signs are called *EH modes*. For each given $m \geq 1$, the eigenvalue equation in (3.27) yields two sets of solutions, one for HE modes and another for EH modes.

Using (3.5)–(3.8), all field components can be found from $\mathcal{E}_z$ and $\mathcal{H}_z$. For the fields in the core region, the resulting field expression can be simplified by using the identities in (3.21) and (3.22) for $J_m(x)$.

1. For TE_{0n} modes, $\mathcal{E}_z = \mathcal{E}_r = \mathcal{H}_\phi = 0$ and

$$\mathcal{H}_z = J_0(hr), \qquad \mathcal{E}_\phi = \frac{\mathrm{i}\omega\mu_0}{h} J_1(hr), \qquad \mathcal{H}_r = -\frac{\mathrm{i}\beta}{h} J_1(hr). \tag{3.31}$$

2. For TM_{0n} modes, $\mathcal{H}_z = \mathcal{E}_\phi = \mathcal{H}_r = 0$ and

$$\mathcal{E}_z = J_0(hr), \qquad \mathcal{E}_r = -\frac{\mathrm{i}\beta}{h} J_1(hr), \qquad \mathcal{H}_\phi = -\frac{\mathrm{i}\omega\epsilon_1}{h} J_1(hr). \tag{3.32}$$

3. For HE_{mn} and EH_{mn} modes, all six field components exist and are given by

$$\mathcal{E}_z = J_m(hr)\cos m\phi, \tag{3.33}$$

$$\mathcal{H}_z = \frac{\beta}{\omega\mu_0}\eta J_m(hr)\sin m\phi, \tag{3.34}$$

$$\mathcal{E}_r = \frac{\mathrm{i}\beta}{h}\left[\frac{1+\eta}{2}J_{m-1}(hr) - \frac{1-\eta}{2}J_{m+1}(hr)\right]\cos m\phi, \tag{3.35}$$

$$\mathcal{E}_\phi = -\frac{\mathrm{i}\beta}{h}\left[\frac{1+\eta}{2}J_{m-1}(hr) + \frac{1-\eta}{2}J_{m+1}(hr)\right]\sin m\phi, \tag{3.36}$$

$$\mathcal{H}_r = \frac{\mathrm{i}\omega\epsilon_1}{h}\left[\frac{1+\eta\beta^2/k_1^2}{2}J_{m-1}(hr) + \frac{1-\eta\beta^2/k_1^2}{2}J_{m+1}(hr)\right]\sin m\phi, \tag{3.37}$$

$$\mathcal{H}_\phi = \frac{\mathrm{i}\omega\epsilon_1}{h}\left[\frac{1+\eta\beta^2/k_1^2}{2}J_{m-1}(hr) - \frac{1-\eta\beta^2/k_1^2}{2}J_{m+1}(hr)\right]\cos m\phi, \tag{3.38}$$

where

$$\eta = \frac{\omega\mu_0 C_m}{\beta A_m}. \tag{3.39}$$

The value of the constant η is a characteristic of a particular HE or EH mode and is determined by the boundary conditions through solution of (3.27). For $\eta > 0$, (3.33)–(3.38) represent the field components of the HE_{mn} mode. For $\eta < 0$, they represent the field components of the EH_{mn} mode.

Note that a multiplicative constant common to all of the field components in a mode is omitted in the above representation. Thus, these mode fields are not normalized.

The intensity of a mode has to be calculated using (2.37). For the modes of a circular fiber, it is reduced to

$$I = 2(\mathcal{E}_r \mathcal{H}_\phi^* - \mathcal{E}_\phi \mathcal{H}_r^*). \tag{3.40}$$

The power in a mode is obtained by integrating the intensity over the fiber cross section:

$$P = 2 \int_0^\infty \int_0^{2\pi} (\mathcal{E}_r \mathcal{H}_\phi^* - \mathcal{E}_\phi \mathcal{H}_r^*) r \, \mathrm{d}\phi \, \mathrm{d}r. \tag{3.41}$$

In accordance with the discussions in Section 2.4, it can be shown that (3.41) is equivalent to (2.38) for a TE mode and is equivalent to (2.39) for a TM mode. For HE and EH hybrid modes, (3.40) and (3.41) cannot be reduced to the form of only an electric field or that of only a magnetic field.

Cutoff conditions

The cutoff for a particular guided mode of an optical fiber is determined by the condition $\gamma = 0$, at which instant the guided mode ceases to be guided. This is the same condition as that for a guided mode of a planar waveguide discussed in Section 2.5. At cutoff, we have

$$V_\mathrm{c} = ha, \tag{3.42}$$

which has a form similar to that of (2.70). The equation for finding the cutoff value V_c depends on the type of mode:

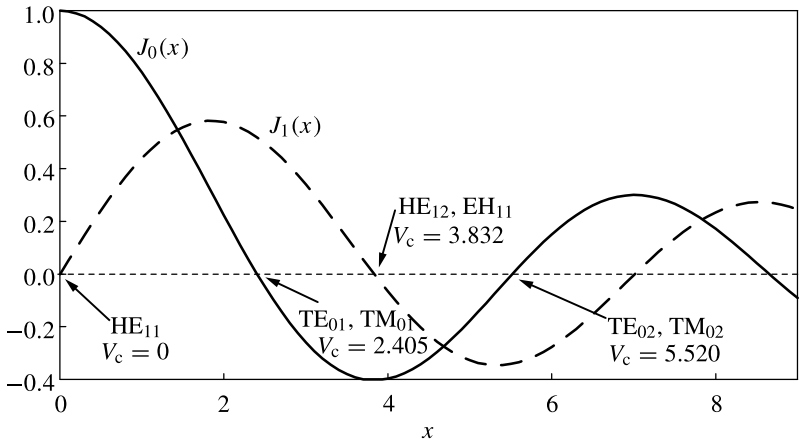
1. For TE_{0n} and TM_{0n} modes, V_c is the nth root of the equation

   $$J_0(x) = 0. \tag{3.43}$$

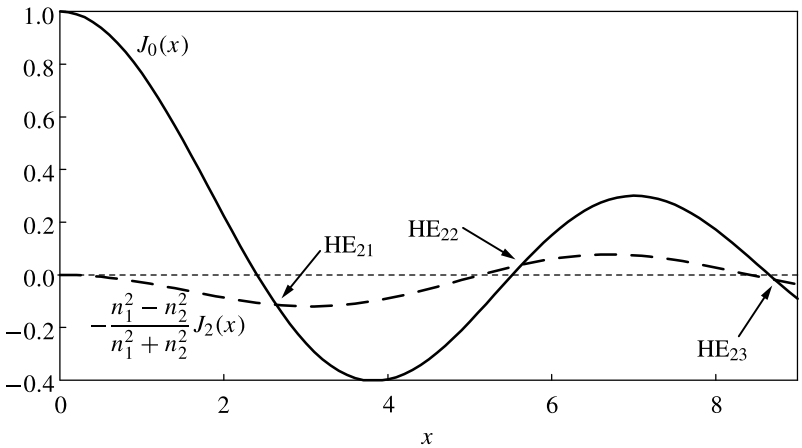
2. For HE_{1n} modes, V_c is the nth root of the equation

   $$J_1(x) = 0, \tag{3.44}$$

   the first of which being $x = 0$. Therefore, $V_\mathrm{c} = 0$ for the HE_{11} mode. For HE_{mn}



(*a*)



(*b*)

Figure 3.3 Graphic solutions of V_c for (*a*) TE_{0n}, TM_{0n}, HE_{1n}, and EH_{1n} modes and (*b*) HE_{2n} modes.

modes with $m \geq 2$, V_c is the nth *nonzero* root of the equation

$$J_{m-2}(x) + \frac{n_1^2 - n_2^2}{n_1^2 + n_2^2} J_m(x) = 0. \tag{3.45}$$

Because $J_{-1}(x) = -J_1(x)$, (3.45) reduces to (3.44) for the HE_{1n} modes when $m = 1$. Note that the values of V_c for HE_{mn} modes with $m \geq 2$ depend on the specific values of the refractive indices n_1 and n_2.

3. For all EH_{mn} modes, $m \geq 1$, and V_c is the nth *nonzero* root of the equation

$$J_m(x) = 0. \tag{3.46}$$

Figure 3.3 shows the graphic solution of V_c for some leading modes.

As can be seen in Fig. 3.3(*a*), *the fundamental mode of a circular fiber is the* HE_{11} *mode, which has no cutoff*. The first high-order modes are the TE_{01} and TM_{01} modes, which have the same cutoff value of $V_c = 2.405$. Note that although the TE_{01} and TM_{01} modes have the same cutoff V_c, they are not degenerate because they have different β defined by different eigenvalue equations in (3.28) and (3.29), respectively, when they are above cutoff. This is also true for other modes that have the same cutoff V_c, such as the HE_{12} and EH_{11} modes.

A fiber that has a waveguide parameter

$$V = \frac{2\pi}{\lambda} a\sqrt{n_1^2 - n_2^2} < 2.405 \tag{3.47}$$

supports only the fundamental HE_{11} mode and is called a *single-mode fiber*. A fiber with $V > 2.405$ can support more than just the HE_{11} mode and is called a *multimode fiber*. Clearly, whether a fiber is single moded or multimoded depends not only on its index step and core radius, but also on the optical wavelength being considered. For a given fiber, $V = 2.405$ determines its cutoff wavelength, λ_c, for its single-mode characteristics. The fiber is single moded for $\lambda > \lambda_c$, but is multimoded for $\lambda < \lambda_c$.

EXAMPLE 3.2 Is the silica fiber described in Example 3.1 single moded at 850 nm wavelength? What is the cutoff wavelength for its single-mode operation?

Solution As found in Example 3.1, $V = 2.758 > 2.405$ for the fiber at $\lambda = 850$ nm. Therefore, this fiber is not a single-mode fiber at 850 nm wavelength. The cutoff wavelength corresponds to $V_c = 2.405$. It is found as

$$\lambda_c = \frac{2\pi}{V_c} a\sqrt{n_1^2 - n_2^2} = 975 \text{ nm}.$$

This fiber is single moded at wavelengths longer than 975 nm but is multimoded at shorter wavelengths. For example, it is a single-mode fiber at 1.3 μm wavelength.

3.2 Weakly guiding fibers

Most optical fibers for practical applications are *weakly guiding fibers* that have a small index step, Δn, between the core and the cladding:

$$\Delta = \frac{\Delta n}{n_1} = \frac{n_1 - n_2}{n_1} \ll 1. \tag{3.48}$$

The mathematics for the modes of a weakly guiding fiber can be greatly simplified by taking proper approximations. For example, the cutoff V_c for the HE_{mn} modes with $m \geq 2$ of a weakly guiding fiber can be approximated by the nth nonzero root of

the equation

$$J_{m-2}(x) = 0, \tag{3.49}$$

which is obtained from (3.45) under the condition of (3.48). Meanwhile, for the modes of a weakly guiding fiber, $\beta^2/k_2^2 \approx 1$, and the parameter η defined in (3.39) has a value of $\eta \approx 1$ for HE modes and a value of $\eta \approx -1$ for EH modes. Therefore, (3.35)–(3.38) are reduced to a simple form that is useful for obtaining a visualization of the field patterns and intensity distributions of the modes. The resulting approximate transverse electric field components, $\mathcal{E}_r$ and $\mathcal{E}_\phi$, and intensity distribution, I, for the four types of fiber modes are

$$\left.\begin{array}{llll} \mathrm{TE}_{0n}: & \mathcal{E}_r = 0, & \mathcal{E}_\phi \sim J_1(hr), & I \sim J_1^2(hr), \\ \mathrm{TM}_{0n}: & \mathcal{E}_r \sim J_1(hr), & \mathcal{E}_\phi = 0, & I \sim J_1^2(hr), \\ \mathrm{HE}_{mn}: & \mathcal{E}_r \sim J_{m-1}(hr)\cos m\phi, & \mathcal{E}_\phi \sim -J_{m-1}(hr)\sin m\phi, & I \sim J_{m-1}^2(hr), \\ \mathrm{EH}_{mn}: & \mathcal{E}_r \sim -J_{m+1}(hr)\cos m\phi, & \mathcal{E}_\phi \sim -J_{m+1}(hr)\sin m\phi, & I \sim J_{m+1}^2(hr). \end{array}\right\} \tag{3.50}$$

Transverse magnetic field components also have a simple form similar to that of transverse electric field components. Because transverse magnetic field lines are simply orthogonal to transverse electric field lines, the magnetic field components are not spelled out explicitly in (3.50). The patterns of the field lines and intensity distributions of several leading modes are shown in Fig. 3.4. Note that the intensity distributions for all four types of modes do not depend on ϕ and have only radial variations.

Linearly polarized modes

It can be seen that except for the HE_{11} mode, the fields of the fiber modes shown in Fig. 3.4 are not plane polarized because the field lines are not straight parallel lines. However, in the weakly guiding approximation, it is possible to represent the fields in a fiber in terms of *linearly polarized modes*, called *LP modes*. Indeed, all of the HE_{1n} modes are very much plane polarized, particularly in weakly guiding fibers. For other modes, many are nearly degenerate, and plane polarized fields can be formed by linear combinations of these nearly degenerate modes if the weakly guiding approximation leading to (3.50) is valid. For example, in the weakly guiding limit, the cutoff V_c determined by (3.49) for the HE_{21} mode is the same as that of TE_{01} and TM_{01} modes. These three modes are nearly degenerate. Combinations of these nearly degenerate modes result in LP modes.

The discussions above can be demonstrated by considering the x and y components of the transverse electric field:

$$\mathcal{E}_x = \mathcal{E}_r \cos\phi - \mathcal{E}_\phi \sin\phi, \tag{3.51}$$

$$\mathcal{E}_y = \mathcal{E}_r \sin\phi + \mathcal{E}_\phi \cos\phi. \tag{3.52}$$

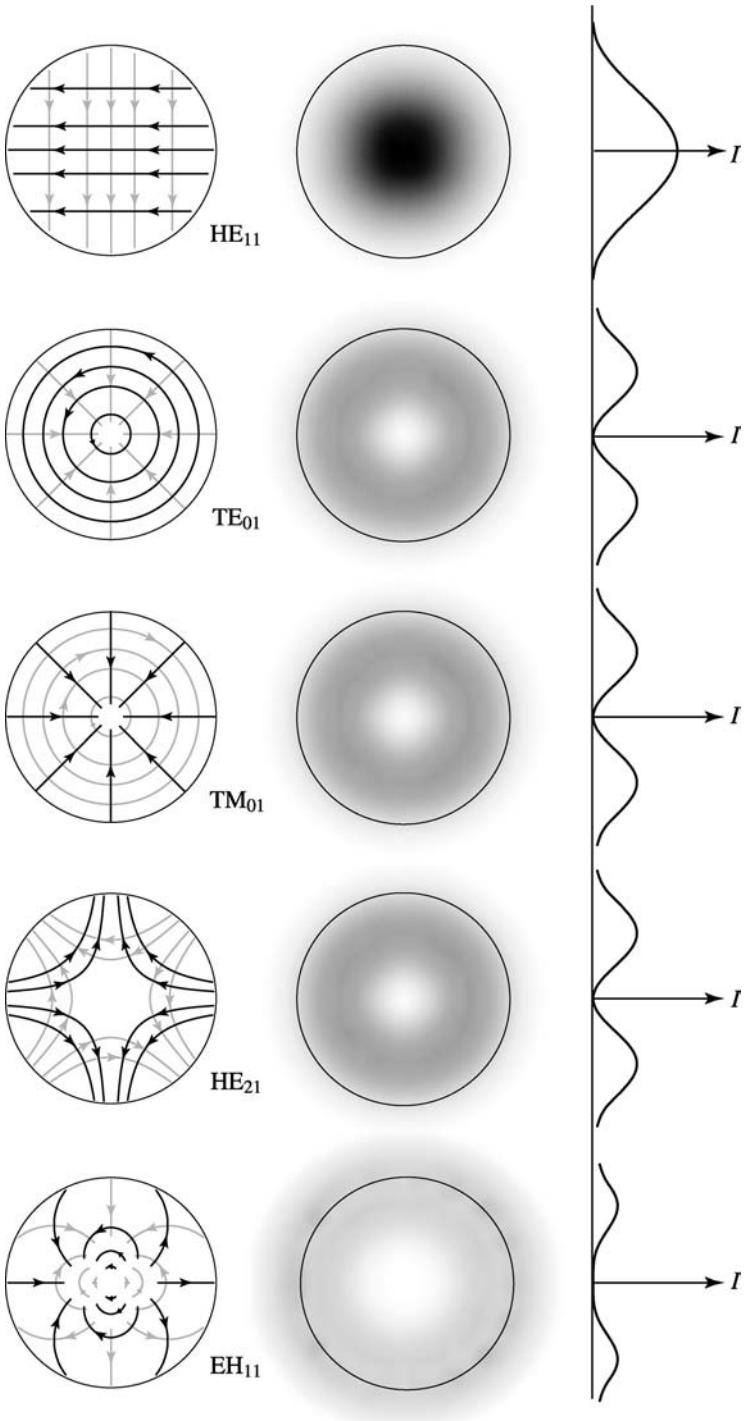


Figure 3.4 Field line patterns and intensity distributions for several leading modes of a circular fiber. The dark curves in the field patterns are the electric field lines, and the gray curves are the magnetic field lines. The thin circle in each profile locates the core boundary of a step-index fiber.

For any HE_{1n} mode, we have

$$\mathcal{E}_r \sim J_0(hr)\cos\phi \qquad \text{and} \qquad \mathcal{E}_\phi \sim -J_0(hr)\sin\phi \tag{3.53}$$

from (3.50). Using (3.51) and (3.52), this results in

$$\mathcal{E}_x \sim J_0(hr) \qquad \text{and} \qquad \mathcal{E}_y = 0. \tag{3.54}$$

Therefore, the transverse electric fields of all of the HE_{1n} modes given in the form of (3.53) are plane polarized in the x direction. They are designated as LP_{0n} modes. The LP_{01} mode is simply the HE_{11} mode and is the fundamental LP mode. There is two-fold degeneracy in LP_{0n} modes because all HE_{1n} modes are two-fold degenerate.

Before we proceed further, we have to note that each of the HE and EH modes has two-fold degeneracy, whereas TE and TM modes have no degeneracy. This is because the field patterns of the HE and EH modes are functions of ϕ, but those of the TE and TM modes are independent of ϕ. An orthogonal field pattern can be generated by rotating the field pattern of any HE_{mn} or EH_{mn} mode by an angle of $\pi/2m$ in ϕ. For example, an HE_{1n} mode given by the form in (3.50), such as the HE_{11} mode shown in Fig. 3.4, has its field lines parallel to the x direction, as is demonstrated above. Its degenerate orthogonal mode pattern is one with the field lines parallel to the y direction. For the HE_{21} mode given by (3.50) and shown in Fig. 3.4, its degenerate orthogonal mode pattern HE'_{21} can be obtained by substituting ϕ in (3.50) with $\phi + \pi/4$ for $m = 2$. Thus we have

$$\begin{aligned} &HE_{21}: \quad \mathcal{E}_r \sim J_1(hr)\cos 2\phi, \quad \mathcal{E}_\phi \sim -J_1(hr)\sin 2\phi, \\ &HE'_{21}: \quad \mathcal{E}_r \sim -J_1(hr)\sin 2\phi, \quad \mathcal{E}_\phi \sim -J_1(hr)\cos 2\phi. \end{aligned} \tag{3.55}$$

The TE_{01} and TM_{01} modes have no degeneracy. Their $\mathcal{E}_r$ and $\mathcal{E}_\phi$ field components are simply those given by (3.50). Using (3.51) and (3.52), it can be shown that

$$\begin{aligned} &TE_{01} + HE'_{21}: \quad \mathcal{E}_x \sim -2J_1(hr)\sin\phi, \quad \mathcal{E}_y = 0, \\ &TE_{01} - HE'_{21}: \quad \mathcal{E}_x = 0, \quad \mathcal{E}_y \sim 2J_1(hr)\cos\phi, \\ &TM_{01} + HE_{21}: \quad \mathcal{E}_x \sim 2J_1(hr)\cos\phi, \quad \mathcal{E}_y = 0, \\ &TM_{01} - HE_{21}: \quad \mathcal{E}_x = 0, \quad \mathcal{E}_y \sim 2J_1(hr)\sin\phi. \end{aligned} \tag{3.56}$$

These are plane polarized fields. They are designated as the LP_{11} mode. There is four-fold degeneracy in the LP_{11} mode because it contains four nearly degenerate modes, TE_{01}, TM_{01}, HE_{21}, and HE'_{21}. The LP_{11} mode is the first high-order LP mode above the fundamental mode.

The discussions above can be extended to other LP modes. Except for LP_{0n} modes, which are just HE_{1n} modes, all other LP modes can be constructed from linear combinations of different basic fiber modes. Their relationships are summarized in Table 3.1.

The eigenvalue equation and the equation defining the cutoff conditions of the LP modes, as well as their field and intensity patterns, are much simplified. These

Table 3.1 *Fiber modes*

LP modes	Degeneracy	Core intensity pattern	Basic modes	Degeneracy
LP_{0n}	2	$J_0^2(hr)$	HE_{1n}	2
LP_{1n}	4	$J_1^2(hr)\cos^2\phi$	TE_{0n}	1
			TM_{0n}	1
			HE_{2n}	2
LP_{mn} $(m \geq 2)$	4	$J_m^2(hr)\cos^2 m\phi$	$HE_{m+1,n}$	2
			$EH_{m-1,n}$	2

characteristics are summarized below.

1. **Eigenvalue equation.** The eigenvalue equation for all LP_{mn} modes can be written as

$$\frac{haJ_{m-1}(ha)}{J_m(ha)} = -\frac{\gamma aK_{m-1}(\gamma a)}{K_m(\gamma a)}. \tag{3.57}$$

   For $m = 0$, the relations $J_{-1}(x) = -J_1(x)$ and $K_{-1}(x) = K_1(x)$ from (3.20) can be used. Note that (3.57) reduces to (3.28) for $m = 1$ because the eigenvalue of the LP_{1n} mode is approximately that of the TE_{0n} mode.
2. **Cutoff conditions.** Except for the LP_{0n} mode, the cutoff V_c value for the LP_{mn} mode is the nth *nonzero* root of the equation

$$J_{m-1}(x) = 0. \tag{3.58}$$

   This condition can be obtained by considering the cutoff conditions for the TE, TM, HE, and EH modes discussed in the preceding section in the weakly guiding limit of (3.48). It can also be obtained by directly applying the cutoff condition of $\gamma = 0$ to the eigenvalue equation in (3.57) for the LP modes. For the LP_{0n} mode, $m = 0$ and (3.58) becomes

$$J_1(x) = 0. \tag{3.59}$$

   The first root, $x = 0$, counts even though it is a trivial root. The LP_{01} mode, which is simply the HE_{11} mode, has no cutoff, as discussed earlier. Therefore, the cutoff V_c for the LP_{0n} mode is the nth root of (3.59), counting $x = 0$ as the first one.
3. **Number of modes.** For a multimode fiber with a large V number, the number of modes supported by the fiber can be estimated. Since the cutoff V_{mn}^c for the LP_{mn} mode is the nth nonzero root of (3.58), we have

$$V_{mn}^c = \left(m + 2n - \frac{3}{2}\right)\frac{\pi}{2} \approx (m + 2n)\frac{\pi}{2} \tag{3.60}$$

   from (3.18). This means that for a given large value of V, the maximum value of m is $m_{\max} \approx 2V/\pi$, while the maximum value of n for a given m is $n_{\max} = V/\pi - m/2$.

Since there is a four-fold degeneracy for each LP_{mn} mode with $m \neq 0$, the total number of modes is approximately

$$M \approx 4 \sum_{m=0}^{2V/\pi} \sum_{n=1}^{V/\pi - m/2} 1 = \frac{4V^2}{\pi^2} + \frac{2V}{\pi} \approx \frac{4V^2}{\pi^2}. \tag{3.61}$$

4. **Field patterns.** The fields of the LP modes are plane polarized. Because of the degeneracy in each LP mode, there are two possible polarizations for an LP_{0n} mode and four possible combinations of polarizations and angular distributions for an LP_{mn} mode with $m \geq 1$. This characteristic is discussed above for the LP_{01} and LP_{11} modes and can be seen in (3.56) for the LP_{11} mode. For simplicity, we consider the field to be polarized in the y direction and the azimuthal angular distribution to be such that $\mathcal{E}_y$ has a maximum at $\phi = 0$. Then, for any LP_{mn} mode, the field pattern is simply

$$\mathcal{E}_y \sim \begin{cases} \dfrac{1}{J_m(ha)} J_m(hr) \cos m\phi, & r < a, \\ \dfrac{1}{K_m(\gamma a)} K_m(\gamma r) \cos m\phi, & r > a, \end{cases} \tag{3.62}$$

and $\mathcal{E}_x = 0$. Note that the boundary conditions for a circular fiber do not require $\mathcal{E}_y$ to be continuous at $r = a$. Rather, they require $\mathcal{E}_\phi$ and $\mathcal{H}_\phi$ to be continuous at $r = a$. Because (3.62) does not satisfy the boundary conditions exactly, it is only an approximation under the weakly guiding condition of (3.48).

5. **Intensity distributions.** The intensity distribution of the LP_{mn} mode has the following pattern:

$$I(\phi, r) \sim \begin{cases} \dfrac{1}{J_m^2(ha)} J_m^2(hr) \cos^2 m\phi, & r < a, \\ \dfrac{1}{K_m^2(\gamma a)} K_m^2(\gamma r) \cos^2 m\phi, & r > a. \end{cases} \tag{3.63}$$

This characteristic is also summarized in Table 3.1. Figure 3.5 shows the intensity profiles of a few LP modes.

6. **Confinement factor.** The confinement factor for a mode is the fractional power in the core region and is given by

$$\Gamma_{\mathrm{mode}} = \frac{P_{\mathrm{core}}}{P_{\mathrm{mode}}} = \frac{\displaystyle\int_0^a \int_0^{2\pi} I(\phi, r) r \, \mathrm{d}r \, \mathrm{d}\phi}{\displaystyle\int_0^\infty \int_0^{2\pi} I(\phi, r) r \, \mathrm{d}r \, \mathrm{d}\phi}. \tag{3.64}$$

For the LP_{mn} mode, the integrals in (3.64) can be calculated using the intensity

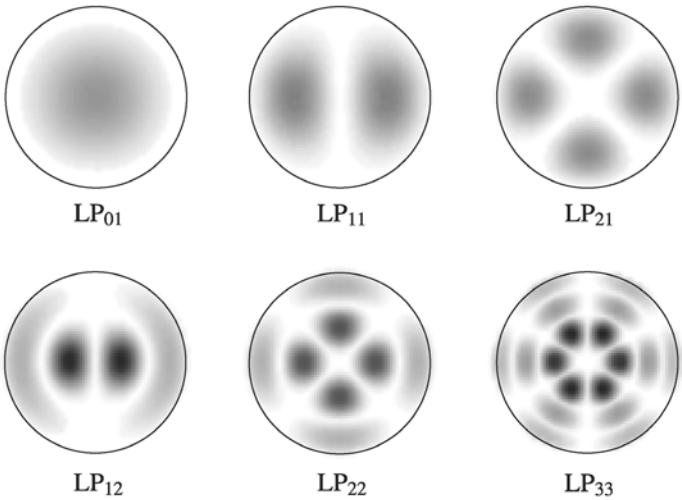


Figure 3.5 Intensity profiles of a few LP modes. The intensity pattern of the LP_{mn} mode consists of m node lines intersecting at the center and n intensity peaks counted radially out from the center. The thin circle in each profile locates the core boundary of a step-index fiber.

distribution given in (3.63), resulting in

$$\Gamma_{mn} = 1 - \frac{h^2a^2}{V^2}\left[1 - \frac{K_m^2(\gamma a)}{K_{m-1}(\gamma a)K_{m+1}(\gamma a)}\right]. \tag{3.65}$$

This expression has to be evaluated numerically. An approximate expression is

$$\Gamma_{mn} = 1 - \frac{h^2a^2}{V^2}\frac{1}{\sqrt{\gamma^2a^2 + m^2 + 1}}. \tag{3.66}$$

The confinement factors for some leading LP modes are shown as a function of the fiber V number in Fig. 3.6. We see that the fundamental LP_{01} mode has a confinement factor $\Gamma_{01} \approx 0.84$ at the cutoff point of $V = 2.405$ for the LP_{11} mode. Note that as cutoff is approached, the power for a mode with $m = 0$ or $m = 1$ moves away from the core to the cladding so that $\Gamma_{mn} \to 0$. However, for LP modes with $m \geq 2$, a large fraction of power remains in the core at cutoff. For a mode with large m, the power remains primarily in the core.

EXAMPLE 3.3 A multimode silica fiber has a core index of 1.48 and a core diameter of 50 μm. Find the index step needed for it to support at least 1000 guided modes at 850 nm wavelength. How many modes does this fiber support at 1.3 μm wavelength if dispersion can be ignored?

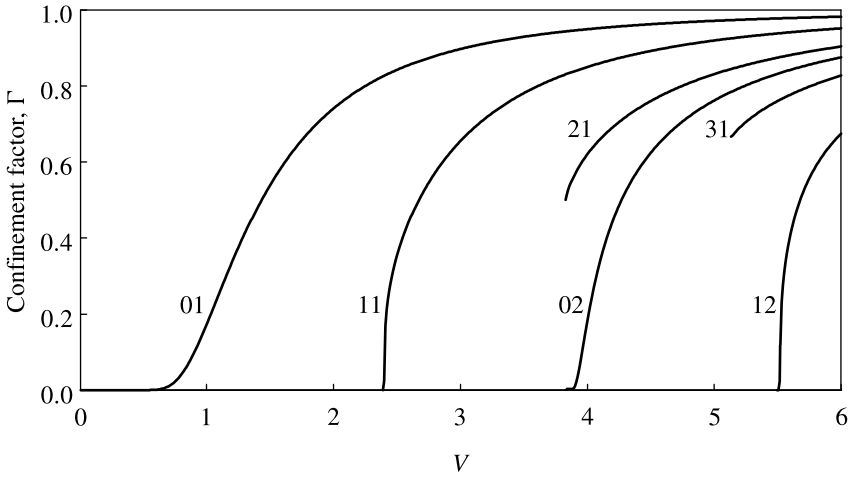


Figure 3.6 Confinement factors of leading LP modes as a function of the fiber V number.

Solution According to (3.61), the V number needs to be $V = \pi\sqrt{M}/2 > 49.67$ so that $M > 1000$ for the fiber to support at least 1000 modes. Using $n_1 = 1.48$, $a = 50\ \mu\text{m}/2 = 25\ \mu\text{m}$, and $\lambda = 850$ nm as given, we find that

$$n_2 \approx \sqrt{n_1^2 - \left(\frac{V\lambda}{2\pi a}\right)^2} < 1.4554.$$

Therefore, we can choose an index step $\Delta n = 0.025$ for $n_2 = 1.455$, which corresponds to $\Delta = 1.69\%$. With $n_2 = 1.455$, we find that $V = 50.05 > 49.67$ and $M = 1015 > 1000$, as required.

Because $M \propto V^2 \propto \lambda^{-2}$, we can find the number of modes at 1.3 μm directly from that at 850 nm if dispersion is ignored. Therefore, the number of modes at 1.3 μm is

$$M = \frac{0.85^2}{1.3^2} \times 1015 \approx 434.$$

It has to be noted that although eigenvalue equations and cutoff conditions are written for the LP modes, they are approximations valid only in the weakly guiding limit. Except for LP_{0n} modes, which are simply HE_{1n} normal modes, *the LP modes are not the exact solutions of Maxwell's equations for a fiber and thus are not true normal modes of a fiber.* This concept can be understood from the fact that an LP_{mn} mode with $m \geq 1$ is a linear combination of some *nearly, but not exactly, degenerate* modes. Consider the combination $LP_{11} = TM_{01} + HE_{21}$ given in (3.56). Because the TM_{01} and HE_{21} modes are not exactly degenerate, there is a slight difference, $\Delta\beta$, in their propagation constants. As the LP_{11} field propagates over a long enough distance, this small $\Delta\beta$ eventually causes the phase relation between the TM_{01} and HE_{21} fields, which together

make up the LP_{11} field, to change. As a result, the combined field will not always be plane polarized in the same direction. Therefore, the LP_{11} mode is not a true normal mode because it is not truly invariant in propagation. However, as can be expected, $\Delta\beta$ decreases with Δn and becomes insignificant for most practical applications, except for very-long-distance propagation of the mode. For practical applications, because the true modes that make up an LP mode are very nearly degenerate, they can be excited simultaneously if they are above cutoff. Consequently, if a plane polarized optical wave in free space is coupled into a fiber, it usually results in the excitation of an LP mode. The mode patterns shown in Fig. 3.5 are those usually seen at the output of a fiber.

3.3 Graded-index fibers

In a graded-index fiber, the index profile, $n(r)$, in the core of the fiber is a function of the radial distance, r, from the center of the fiber, as shown in Fig. 3.7. It starts at a value of $n(0) = n_1$ at the center of the fiber and gradually decreases to a value of $n(a) = n_2$ at the boundary between the core and the cladding. The fiber V number defined in (3.1) can still be used, but the properties of a graded-index fiber are also determined by the specific functional dependence of $n(r)$ on r. The numerical aperture, for example, is a function of radial position:

$$\text{NA}(r) = \sqrt{n^2(r) - n_2^2}, \tag{3.67}$$

which decreases from $\text{NA}(0) = \sqrt{n_1^2 - n_2^2}$ at the core center to $\text{NA}(a) = 0$ at the core–cladding boundary.

Because the index profile is no longer piecewise constant, the electric susceptibility $\epsilon(r) = \epsilon_0 n^2(r)$ is also a function of radial position. As a result, piecewise homogeneous wave equations for $\mathcal{E}_z$ and $\mathcal{H}_z$, such as those in (3.9) and (3.10), cannot be used, as discussed in Section 2.3. For a graded-index glass fiber with $\epsilon(r)$, it can be shown by

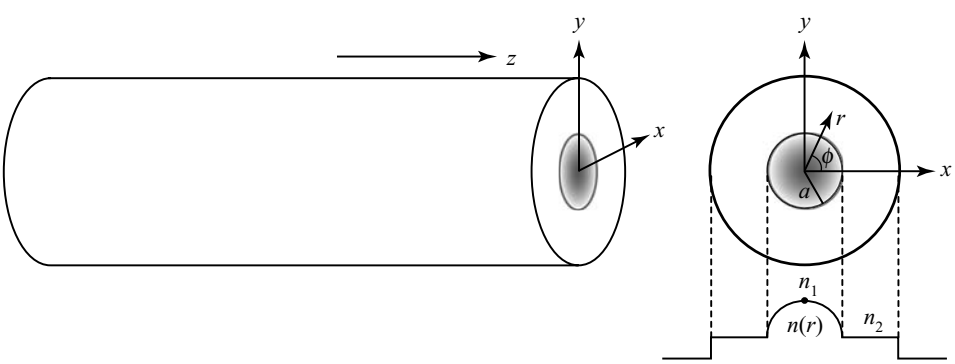


Figure 3.7 Graded-index fiber with a core radius a.

substitution of (3.3) and (3.4) in (2.24) and (2.25), respectively, that

$$\frac{\partial^2 \mathcal{E}_z}{\partial r^2} + \frac{1}{r}\frac{\partial \mathcal{E}_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathcal{E}_z}{\partial \phi^2} + (k^2 - \beta^2)\mathcal{E}_z = -\mathrm{i}\beta \frac{\mathrm{d}\ln\epsilon}{\mathrm{d}r}\mathcal{E}_r, \tag{3.68}$$

$$\frac{\partial^2 \mathcal{H}_z}{\partial r^2} + \frac{1}{r}\frac{\partial \mathcal{H}_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathcal{H}_z}{\partial \phi^2} + (k^2 - \beta^2)\mathcal{H}_z = \frac{\mathrm{d}\ln\epsilon}{\mathrm{d}r}\left(-\mathrm{i}\beta\mathcal{H}_r + \frac{\partial \mathcal{H}_z}{\partial r}\right), \tag{3.69}$$

where $k^2 = \omega^2\mu_0\epsilon(r) = \omega^2 n^2(r)/c^2$ is a function of r, and the relation

$$\frac{\boldsymbol{\nabla}\epsilon}{\epsilon} = \frac{\mathrm{d}\ln\epsilon}{\mathrm{d}r}\hat{r} \tag{3.70}$$

is used. The longitudinal field components $\mathcal{E}_z$ and $\mathcal{H}_z$ cannot be solved without solving the transverse field components simultaneously. This is a manifestation of the complicated vectorial nature of the mode fields in a fiber, which is caused by the geometry and structure of the dielectric fiber waveguide. In general, the problem has to be solved numerically with vectorial wave equations. However, some approximate analytic approaches exist that allow us to gain much understanding of the key characteristics of a graded-index fiber without numerical solutions.

All graded-index optical fibers in practical applications are weakly guiding fibers with very small index changes satisfying the condition in (3.48). In addition, the index profiles are usually quite smooth so that the index gradients are small. Under these assumptions, $\boldsymbol{\nabla}\epsilon/\epsilon$ is very small, and an approximation can be made to neglect the terms on the right-hand sides of (3.68) and (3.69), resulting in the following approximate homogeneous equations for $\mathcal{E}_z$ and $\mathcal{H}_z$:

$$\frac{\partial^2 \mathcal{E}_z}{\partial r^2} + \frac{1}{r}\frac{\partial \mathcal{E}_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathcal{E}_z}{\partial \phi^2} + (k^2 - \beta^2)\mathcal{E}_z \approx 0, \tag{3.71}$$

$$\frac{\partial^2 \mathcal{H}_z}{\partial r^2} + \frac{1}{r}\frac{\partial \mathcal{H}_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \mathcal{H}_z}{\partial \phi^2} + (k^2 - \beta^2)\mathcal{H}_z \approx 0. \tag{3.72}$$

By separation of variables, these equations yield the same ϕ dependence as that of the mode fields of a step-index fiber discussed in Section 3.1 and expressed in (3.23)–(3.26). The r dependence is given by solution of

$$\frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}R}{\mathrm{d}r} + \left(k^2 - \beta^2 - \frac{m^2}{r^2}\right)R = 0. \tag{3.73}$$

Because the ϕ dependence of the mode fields of a graded-index fiber is exactly the same as that of the mode fields of a step-index fiber and is independent of its index profile, the classification of the normal modes into the basic TE, TM, HE, and EH types discussed in Section 3.1 and the concept of the LP modes as appropriate linear combinations of basic normal modes for a weakly guiding fiber discussed in Section 3.2 are still valid. However, the r dependence of the mode fields is no longer simply

described by the Bessel functions. It depends on the specific functional form of the index profile $n(r)$.

Approximate solutions of (3.73) can be obtained for the r dependence of the guided mode fields of a graded-index fiber using the WKB approximation in a manner similar to, but somewhat more complicated than, that outlined in Section 2.7 for obtaining the approximate solutions for a graded-index planar waveguide. Here we summarize the results without going through the details.

The existence and the characteristics of guided modes in a graded-index fiber depend on the sign of the following function:

$$p^2(r) = k^2(r) - \frac{m^2 - 1/4}{r^2} - \beta^2. \tag{3.74}$$

For any given guided mode, there exist two *turning points* $r = r_1$ and $r = r_2$, with $0 < r_1 < r_2 < a$, where $p(r) = 0$. For the mode field to exist and be guided, it is necessary that $p^2(r) > 0$ within the radial range $r_1 < r < r_2$, but $p^2(r) < 0$ for $r < r_1$ and $r > r_2$. The eigenvalue equation for the guided modes of a graded-index fiber is given in terms of the function $p(r)$:

$$\int_{r_1}^{r_2} p(r)\mathrm{d}r = \int_{r_1}^{r_2} \left[k^2(r) - \frac{m^2 - 1/4}{r^2} - \beta_{mn}^2\right]^{1/2} \mathrm{d}r = \left(n - \frac{1}{2}\right)\pi,$$
$$n = 1, 2, 3, \ldots \tag{3.75}$$

The allowed values of β can be obtained by solving this equation with integral values of m and n. Evidently, the solution depends on the precise form of $n(r)$. For a given azimuthal index m, there exists only a finite number of discrete values of β_{mn} that are allowed for the guided modes.

While the ϕ dependence of the mode fields remains the same as that of the mode fields discussed in Section 3.1, the radial variations can be approximated with the following asymptotic form at radial locations away from the immediate vicinity of the turning points where $p(r) = 0$:

$$R(r) \sim \begin{cases} \dfrac{1}{\sqrt{r|p(r)|}} \exp\left[-\displaystyle\int_r^{r_1} |p(r')|\mathrm{d}r'\right], r < r_1, \\ \dfrac{2}{\sqrt{rp(r)}} \cos\left[\displaystyle\int_{r_1}^{r} p(r')\mathrm{d}r' - \dfrac{\pi}{4}\right], r_1 < r < r_2, \\ \dfrac{(-1)^{n-1}}{\sqrt{r|p(r)|}} \exp\left[-\displaystyle\int_{r_2}^{r} |p(r')|\mathrm{d}r'\right], r > r_2, \end{cases} \tag{3.76}$$

where the factor of $(-1)^{n-1}$ is used for the correct phase at $r = r_2$. For a graded-index

fiber, this radial solution replaces the Bessel functions $J_m(hr)$ and $K_m(hr)$ in (3.23)–(3.26) of the mode fields of a step-index fiber. Note the similarity between the form of $R(r)$ in oscillatory and evanescent regions and that of the first terms in (3.18) and (3.19), respectively, for the asymptotic behavior of $J_m(x)$ and $K_m(x)$ at large values of x.

Number of modes

Graded-index fibers are primarily used as low-dispersion multimode fibers. The total number of modes supported by a graded-index multimode fiber can be estimated using the eigenvalue equation in (3.75). It can be seen from (3.75) that for a given azimuthal mode index m, the largest number, $n(m)$, for the radial mode index is obtained when β has a minimum value. Since $\beta > k_2$ for a guided mode, the minimum value of β can be approximated by k_2 for a fiber that has a large number of modes. Therefore, $n(m)$ is approximately given by

$$n(m) \approx \frac{1}{\pi} \int_{r_1}^{r_2} \left[k^2(r) - \frac{m^2 - 1/4}{r^2} - k_2^2 \right]^{1/2} \mathrm{d}r \approx \frac{1}{\pi} \int_{r_1}^{r_2} \left[k^2(r) - \frac{m^2}{r^2} - k_2^2 \right]^{1/2} \mathrm{d}r. \quad (3.77)$$

Meanwhile, for guided modes, it is necessary that $p(r) > 0$ for $r_1 < r < r_2$, as discussed above. Therefore, the largest number, $m_{\max}$, of the azimuthal mode index for guided modes is

$$m_{\max} = r\sqrt{k^2(r) - k_2^2}. \quad (3.78)$$

The total number of guided modes can then be estimated as

$$M = 4 \sum_{m=0}^{m_{\max}} n(m) \approx 4 \int_0^{m_{\max}} n(m) \mathrm{d}m, \quad (3.79)$$

where the factor 4 accounts for the four-fold degeneracy of most high-order guided modes as discussed in the preceding section, and the summation over m is replaced by an integral for a fiber of a large number of densely spaced modes. Substituting (3.77) and (3.78) in (3.79) and noting that the minimum value of r_1 is $r = 0$ while the maximum value of r_2 is $r = a$, we have

$$\begin{aligned} M &\approx \frac{4}{\pi} \int_0^a \int_0^{r\sqrt{k^2(r)-k_2^2}} \left[k^2(r) - k_2^2 - \frac{m^2}{r^2} \right]^{1/2} \mathrm{d}m \mathrm{d}r \\ &= \int_0^a \left[k^2(r) - k_2^2 \right] r \mathrm{d}r. \end{aligned} \quad (3.80)$$

In terms of the index profile of the fiber, the total number of modes is

$$M \approx \frac{\omega^2}{c^2} \int_0^a \left[n^2(r) - n_2^2\right] r \mathrm{d}r. \tag{3.81}$$

Following the line of argument leading to (3.80), we can find the number M_β of guided modes that have propagation constants larger than β to be

$$M_\beta = \int_0^{r_2(\beta)} \left[k^2(r) - \beta^2\right] r \mathrm{d}r, \tag{3.82}$$

where $r_2(\beta)$ is determined by $k(r_2) = \beta$.

Power-law index profiles

We consider here the following power-law index profile:

$$n(r) = \begin{cases} n_1 \left[1 - 2\Delta \left(\dfrac{r}{a}\right)^\alpha\right]^{1/2}, & 0 \le r \le a, \\ n_2, & r > a, \end{cases} \tag{3.83}$$

where

$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_1}. \tag{3.84}$$

With $\Delta \ll 1$, the fiber core has a linear index profile for $\alpha = 1$. It becomes a step-index fiber for $\alpha = \infty$. In terms of Δ, the V number of a fiber is

$$V = \frac{\omega}{c} a \sqrt{n_1^2 - n_2^2} = \frac{\omega}{c} a n_1 \sqrt{2\Delta}. \tag{3.85}$$

Substituting (3.83) in (3.81) and using (3.85), the total number of modes can be obtained:

$$M = \frac{\alpha}{\alpha + 2} \frac{V^2}{2}. \tag{3.86}$$

From (3.82), the number of modes with propagation constants larger than β is found to be

$$M_\beta = M \left(\frac{1 - \beta^2/k_1^2}{2\Delta}\right)^{(\alpha+2)/\alpha}. \tag{3.87}$$

Therefore, the propagation constant can be written

$$\beta = k_1 \left[1 - 2\Delta \left(\frac{M_\beta}{M}\right)^{\alpha/(\alpha+2)}\right]^{1/2}. \tag{3.88}$$

The relation in (3.86) between M and V for a graded-index fiber applies only when M is a large number. It fails for a fiber that supports only a few modes and clearly is not applicable to single-mode fibers. A single-mode graded-index fiber is also determined by a cutoff V number similar to (3.47) for a step-index fiber. However, the cutoff V number for a given graded-index fiber depends on the particular index profile of the fiber. Specifically, the condition for a fiber with a quadratic index profile of $\alpha = 2$ to be single moded is $V < 3.53$. For other power-law profiles, the condition is approximately $V < 2.405\sqrt{1 + 2/\alpha}$.

EXAMPLE 3.4 If a graded-index fiber has all of the parameters of the step-index fiber designed in Example 3.3, except for a quadratic index profile of $\alpha = 2$, how many guided modes does it support at 850 nm? What is the propagation constant of its 200th mode? What should its core diameter be for the graded-index fiber to support at least 1000 modes at 850 nm if its index parameters and profile remain unchanged?

Solution The fiber designed in Example 3.3 has $V = 50.05$ and supports 1015 modes at 850 nm wavelength. A graded-index fiber with the same parameters but with $\alpha = 2$ also has the same V number according to (3.85), but its mode number is given by (3.86). Therefore, the number of modes it supports is

$$M = \frac{2}{2+2} \times \frac{50.05^2}{2} = 626,$$

which is much smaller than that of the step-index fiber. With $n_1 = 1.48$, we have $k_1 = 10.94\ \mu\text{m}^{-1}$. From Example 3.3, we know that $\Delta = 0.0169$. The propagation constant for the 200th mode can then be found using (3.88) by taking $M_\beta = 200$ to be $\beta = 0.99k_1 = 10.83\ \mu\text{m}^{-1}$, which is only 1% below k_1 because Δ is only 1.69%.

For $M > 1000$, we find that $V > 63.25$ is required by using (3.86) with $\alpha = 2$. With $n_1 = 1.48$, $n_2 = 1.455$, and $\lambda = 850$ nm, we find from (3.85) that $a > 31.6\ \mu\text{m}$. Therefore, the minimum core diameter is 63.2 μm, which is clearly larger than the 50 μm core of the step-index fiber that supports the same number of modes.

3.4 Attenuation in fibers

Several factors contribute to attenuation of the power of an optical wave propagating in an optical fiber. As discussed in Section 1.5, when an optical wave propagates in a lossy medium with an attenuation coefficient α, its intensity decays exponentially with distance according to (1.103). Since the power of an optical wave in a fiber is simply the integration of its intensity over the cross section of the fiber, the attenuation of optical power over a propagation distance l in a fiber having an attenuation coefficient α is

given by

$$P_{\text{out}} = P_{\text{in}} e^{-\alpha l}, \tag{3.89}$$

where P_{in} and P_{out} are the input and output power, respectively. In (3.89), P_{in} and P_{out} are measured in watts or, for example, milliwatts or microwatts in low-power applications or kilowatts or megawatts in high-power applications, while α is given per meter. In practical applications, α is also measured per centimeter or per kilometer when l is measured in centimeters or kilometers.

In practical engineering applications, it is convenient to use *decibels* (dB) as a measure of relative changes of quantities. The attenuation coefficient α is then measured in decibels per meter. In the case of low-loss fibers, the propagation length in a fiber is usually measured in kilometers, and α is conventionally given in decibels per kilometer:

$$\alpha(\text{dB km}^{-1}) = -\frac{1}{l(\text{km})} 10 \log \frac{P_{\text{out}}}{P_{\text{in}}}, \tag{3.90}$$

where P_{in} and P_{out} are measured in watts, milliwatts, or microwatts. Comparing (3.90) with (3.89), we have

$$\alpha(\text{dB km}^{-1}) = 4.34\alpha(\text{km}^{-1}) \quad \text{and} \quad \alpha(\text{km}^{-1}) = 0.23\alpha(\text{dB km}^{-1}). \tag{3.91}$$

Power can also be measured in decibels and has units of decibel-watts (dBW), decibel-milliwatts (dBm), or decibel-microwatts (dBμ) defined as follows:

$$P(\text{dBW}) = 10 \log P(\text{W}), \quad P(\text{dBm}) = 10 \log P(\text{mW}), \quad P(\text{dB}\mu) = 10 \log P(\mu\text{W}). \tag{3.92}$$

When power is given in decibel-watts or decibel-milliwatts and the attenuation coefficient is in decibels per kilometer, (3.89) can be expressed as

$$P_{\text{out}}(\text{dBW}) = P_{\text{in}}(\text{dBW}) - \alpha(\text{dB km}^{-1}) l(\text{km}), \tag{3.93}$$

or, equivalently,

$$P_{\text{out}}(\text{dBm}) = P_{\text{in}}(\text{dBm}) - \alpha(\text{dB km}^{-1}) l(\text{km}). \tag{3.94}$$

A similar formula can be written for power measured in decibel-microwatts. These formulas are very convenient and useful in practical applications as they relate the input power, output power, and attenuation in a simple arithmetic relation.

EXAMPLE 3.5 A fiber of 40 km length has an attenuation coefficient of 0.6 dB km^{-1} at 1.3 μm and 0.3 dB km^{-1} at 1.55 μm. If 1 mW of optical power at each wavelength is launched into the fiber, what is the output power at each wavelength?

Solution We can convert the attenuation coefficient given in decibels per kilometer into that measured per kilometer and then use (3.89) to find the output power. Alternatively, we can convert the input power given in milliwatts into that in decibel-milliwatts or decibel-microwatts and then use (3.94) to find the output power. The results are the same. Here we use the second approach. Then, $P_{\text{in}} = 1$ mW is converted to $P_{\text{in}} = 0$ dBm $= 30$ dBμ using (3.92). The output power at 1.3 μm is

$$P_{\text{out}} = 0 \text{ dBm} - 0.6 \text{ dB km}^{-1} \times 40 \text{ km} = -24 \text{ dBm} = 6 \text{ dB}\mu,$$

which is $P_{\text{out}} \approx 4$ μW from (3.92). Similarly, the output power at 1.55 μm is

$$P_{\text{out}} = 0 \text{ dBm} - 0.3 \text{ dB km}^{-1} \times 40 \text{ km} = -12 \text{ dBm} = 18 \text{ dB}\mu,$$

which is $P_{\text{out}} \approx 63$ μW. Comparing the results at the two wavelengths, we see the importance of reducing the losses in a fiber: a reduction in the attenuation coefficient by a factor of 2 increases the output power by a factor of nearly 16 in this particular example. The effect is even more dramatic at high losses. A doubling of the attenuation coefficient from 0.6 to 1.2 dB km^{-1} results in an output power of only 15.8 nW, which is a reduction of more than 250 times from the 4 μW output for the 0.6 dB km^{-1} attenuation.

Attenuation of light in a fiber is primarily caused by absorption and scattering. In addition, there are mechanical losses and losses due to nonlinear optical effects. The effects of these loss mechanisms vary, but they all add up to the total loss in a fiber. Since the majority of optical fibers are silica fibers, we discuss the loss mechanisms and their effects in silica fibers below.

1. **Electronic absorption.** The bandgap of fused silica is about 8.9 eV, which corresponds to the photon energy of light at the ultraviolet wavelength of approximately 140 nm. This causes strong absorption of light in the ultraviolet spectral region due to electronic transitions across the bandgap. Light in the visible and infrared regions has photon energies less than the bandgap energy and is not expected to be absorbed through direct electronic transitions across the bandgap. However, in practice, the bandgap of a material is not sharply defined but usually has bandtails extending from the conduction and valence bands into the bandgap due to a variety of reasons, such as thermal vibrations of the lattice ions and microscopic imperfections of the material structure. In particular, an amorphous material like fused silica generally has very long bandtails. These bandtails lead to an absorption tail extending into the visible and infrared regions. Empirically, it is found that the absorption tail at photon energies below the bandgap falls off exponentially with photon energy.
2. **Molecular absorption.** In the infrared region, the absorption of photons is accompanied by transitions between different vibrational modes of silica molecules. The

fundamental vibrational transition of fused silica causes a very strong absorption peak at about 9 μm wavelength. Nonlinear effects contribute to important harmonics and combination frequencies corresponding to minor absorption peaks at 4.4, 3.8, and 3.2 μm wavelengths. The result is a long absorption tail extending into the near infrared, causing a sharp rise in absorption at optical wavelengths longer than 1.6 μm. Molecular absorption is the major cause of attenuation in the infrared spectral region for a silica fiber.

3. **Impurity absorption.** Impurity absorption could be very important in the near infrared region because most impurity ions such as OH^-, Fe^{2+}, and Cu^{2+} form absorption bands in this region where both electronic and molecular absorption losses of the host silica glass are very low. Near the peaks of the impurity absorption bands, an impurity concentration as low as one part per billion can contribute to an absorption loss as high as 1 dB km^{-1}. In fact, fiber-optic communications were not considered possible until it was realized in 1966 that most losses in earlier fibers were caused by impurity absorption and then ultra-pure fibers were produced in the early 1970s. Today, impurities in fibers have been reduced to levels where losses associated with their absorption are negligible, with the exception of the OH^- radical. The OH^- radical results from the presence of water, which can enter a fiber through the manufacturing process or as humidity in the environment. Therefore, fibers are manufactured in ultra-dry conditions and are protected by plastic coating from water in the environment to reduce the loss caused by OH^- absorption. The absorption peak due to the fundamental vibration of the OH^- ions appears at 2.73 μm wavelength where intrinsic molecular absorption of silica is strong. The most important absorption peaks are those at the harmonics and combination frequencies of 1.39, 1.25, and 0.95 μm wavelengths.
4. **Rayleigh scattering.** The intrinsic Rayleigh scattering in a fiber is caused by variations in density and composition that are built into the fiber during the manufacturing process. They are primarily a result of thermal fluctuations in the density of silica glass and variations in the concentration of dopants before silica passes its glass transition point to become a solid. These variations are a fundamental thermodynamic phenomenon and cannot be completely removed. They create microscopic fluctuations in the index of refraction, which scatter light in the same manner as microscopic fluctuations of the density of air scatter sunlight. This elastic Rayleigh scattering process creates a loss given by

$$\alpha_R = \frac{8\pi^2}{3\lambda^4}(n^2 - 1)\beta k_B T, \tag{3.95}$$

where n is the index of refraction, k_B is the Boltzmann constant, T is the glass transition temperature, and β is isothermal compressibility. Note that $\alpha_R \propto \lambda^{-4}$. The loss due to Rayleigh scattering is very important in the short-wavelength region but falls off rapidly as the wavelength increases.

5. **Waveguide scattering.** Imperfections in the waveguide structure of a fiber, such as nonuniformity in the size and shape of the core, perturbations in the core–cladding boundary, and defects in the core or cladding, can be generated in the manufacturing process. In addition, environmentally induced effects, such as stress and temperature variations, also cause imperfections. The imperfections in a fiber waveguide result in additional scattering losses. They sometimes also induce coupling between different guided modes. Losses caused by waveguide scattering due to imperfections can be measured experimentally.
6. **Nonlinear losses.** In an optical fiber, because light is confined over long distances, nonlinear optical effects can become important even at a relatively moderate optical power. Nonlinear optical processes such as stimulated Brillouin scattering and stimulated Raman scattering can cause significant attenuation in the power of an optical signal. Other nonlinear processes can induce mode mixing or frequency shift, all contributing to the loss of a particular guided mode at a particular frequency. Because nonlinear effects are intensity dependent, they can become very important at high optical powers.

Figure 3.8 summarizes the contributions of various loss mechanisms, except those of waveguide scattering and nonlinear losses, to the total attenuation in a fiber as a function of wavelength. The limiting effect at short wavelengths is the Rayleigh scattering, which dominates the electronic absorption of fused silica in this spectral region. In the infrared region beyond 1.6 μm, attenuation is completely dominated by intrinsic absorption due to molecular vibrations of silica. In the near-infrared region, attenuation strongly depends on the concentration of the OH^- impurity. In addition, in this

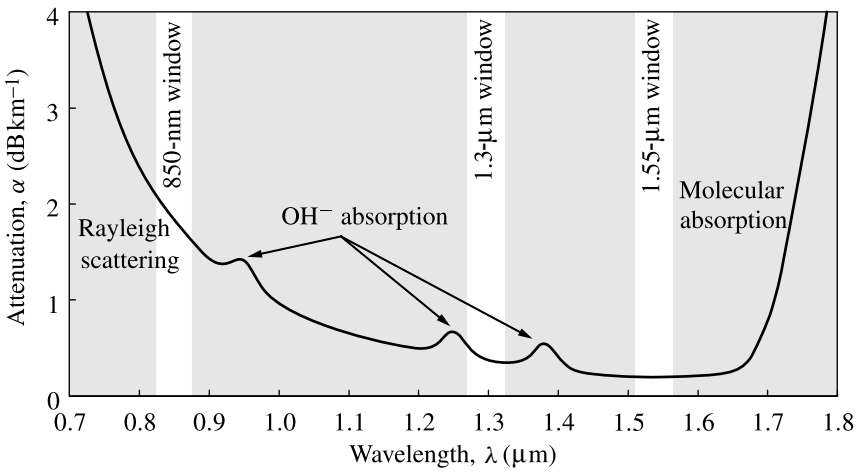


Figure 3.8 Spectral dependence of loss mechanisms and total attenuation in a fiber. Also shown are spectral ranges for three communication windows. (Based on data from assorted sources.)

low-loss region, any amount of loss caused by waveguide scattering would be relatively significant. Therefore, attenuation in this spectral region varies with the quality of the fiber.

The attenuation coefficient is also mode dependent. The fundamental mode generally has lower attenuation than high-order modes because its power is more confined to the core. Therefore, single-mode fibers usually have lower attenuation than multimode fibers. Among multimode fibers of a fixed outer diameter, such as the standard 125-μm size, the ones with larger cores, and simultaneously thinner claddings, typically have higher attenuation because the intensity distribution spreads farther out. A graded-index multimode fiber usually has lower attenuation than a comparable step-index multimode fiber because the intensity in a graded-index fiber is more concentrated at the center of the fiber.

There are three wavelength windows for applications in the transmission of light with fibers. They are the 850-nm window, corresponding to the wavelengths of GaAs/AlGaAs lasers, and the 1.3- and 1.55-μm windows, corresponding to the wavelengths of InGaAsP/InP lasers. It can be seen from Fig. 3.8 that the lowest attenuation in the entire spectral range occurs at 1.55 μm while the attenuation at 1.3 μm is slightly higher. At present, the best fibers have attenuation as low as 0.15 dB km^{-1} at 1.55 μm and 0.3 dB km^{-1} at 1.3 μm, while attenuation at 850 nm is typically 2 dB km^{-1}. This is the reason why the wavelength of 1.55 μm is chosen for long-distance optical communication systems and the wavelength of 1.3 μm is suitable for metropolitan and wide-area networks, while wavelengths in the 850-nm window are only useful for local optical links.

In addition to the losses discussed above, there are also *bending losses* caused by *macrobends* and *microbends* in a fiber and *connection losses* incurred at the junctions of fibers. Macrobends are bends visible from outside and are encountered in the looping or routing of fibers. Microbends are not visible from outside and are typically created by mechanical stresses associated with bundling, packaging, and handling of the fiber. Bending loss can be understood from the viewpoint of ray optics or that of wave optics. For simplicity, consider the fact that the evanescent field of a guided mode actually extends to infinity in all radial directions. When a fiber is bent, the evanescent field on the outside of the bent curve has to travel along a path that has a larger radius of curvature than that traveled by the field in the core of the fiber. Because different parts of a mode field have to stay in phase as a single entity, this evanescent field has to travel faster in order to keep up with the field in the core. The farther outside the field is, the faster it has to travel. At a critical radius, the required speed would exceed the speed of light. At this point, the field cannot keep up and radiates away, resulting in bending loss. Losses caused by controlled bending can be quantified. Fiber sensors based on bending-induced losses can be constructed for many useful applications.

3.5 Dispersion in fibers

Dispersion is the primary cause of limitation on the bandwidth of the transmission of optical signals through an optical fiber. As discussed in Chapter 2, there are *waveguide* and *modal dispersions* in an optical waveguide in addition to *material dispersion*, which is discussed in Chapter 1. Both material dispersion and waveguide dispersion are examples of *chromatic dispersion* because both are frequency dependent. Waveguide dispersion is caused by frequency dependence of the propagation constant β of a specific mode due to the waveguiding effect. The combined effect of material and waveguide dispersions for a particular mode alone is called *intramode dispersion*. Modal dispersion is caused by the variation in propagation constant between different modes: it is also called *intermode dispersion*. Modal dispersion appears only when more than one mode is excited in a multimode fiber. However, it exists even when chromatic dispersion disappears. In contrast, if only one mode is excited in a fiber, only intramode chromatic dispersion has to be considered even when the fiber is a multimode fiber.

Material dispersion

The physical mechanism responsible for material dispersion is discussed in Section 1.10. For optical fibers, the materials of interest are pure silica and doped silica. We first consider the characteristics of relevant parameters for these materials using the general mathematical definitions given in Section 1.9. The parameters of interest are the index of refraction, n, the group index, N, and the group-velocity dispersion, D. Although it is more natural to consider the propagation constant, k, or β in a waveguide, and its derivatives as a function of frequency, ω, in practice these parameters are commonly given as a function of the free-space wavelength, λ.

The index of refraction of pure silica in the wavelength range between 200 nm and 4 μm is given by the following empirically fitted Sellmeier equation:

$$n^2 = 1 + \frac{0.696\,166\,3\lambda^2}{\lambda^2 - (0.068\,404\,3)^2} + \frac{0.407\,942\,6\lambda^2}{\lambda^2 - (0.116\,241\,4)^2} + \frac{0.897\,479\,4\lambda^2}{\lambda^2 - (9.896\,161)^2}, \qquad (3.96)$$

where λ is in micrometers. As discussed at the beginning of this chapter, the index of refraction can be changed by adding dopants to silica, thus facilitating the means to control the index profile of a fiber. The amount of index change depends on the type and concentration of dopant or dopants. Specifically, doping with germania or alumina increases the index of refraction. Therefore, the coefficients in (3.96) actually depend on the composition of the glass. The indices of refraction as functions of wavelength for pure silica and a germania–silica glass that has 13.5 mol % GeO_2 and 86.5 mol % SiO_2 are shown in Fig. 3.9(*a*). The group index N and the group-velocity dispersion D

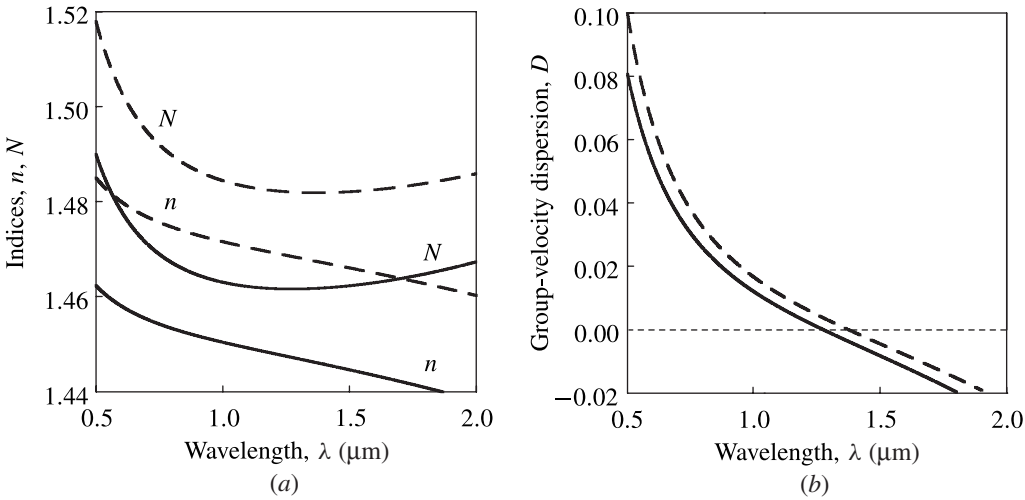


Figure 3.9 (*a*) Index of refraction n and group index N and (*b*) group-velocity dispersion D as functions of wavelength for pure silica (solid curves) and germania–silica containing 13.5 mol % GeO_2 (dashed curves). Zero group-velocity dispersion appears at 1.284 μm for pure silica.

can be calculated using (1.171) and (1.172), respectively. The group indices for both glasses are also shown in Fig. 3.9(*a*), while the group-velocity dispersion is shown in Fig. 3.9(*b*). It can be seen that the addition of GeO_2 to silica not only increases the index of refraction, but also increases material dispersion. As a result, the point of zero material group-velocity dispersion is shifted from 1.284 μm for pure silica to 1.383 μm for germania–silica glass. This increase in index of refraction and in dispersion is reduced if the percentage of GeO_2 is reduced. The effects of other dopants vary. For example, doping with 9.1 mol % P_2O_5 increases the index of refraction by more than 1% but only slightly shifts the dispersion curve, whereas doping with 13.3 mol % B_2O_3 results in a reduction of the index of refraction by less than 0.4% but shifts the point of zero dispersion to 1.231 μm. Clearly, it is possible to control the modification of material dispersion by dopants.

Waveguide dispersion

The propagation constant of a guided mode of a fiber is determined both by the parameters of the fiber, such as its index profile and core size, and by the material properties. Therefore, the frequency dependence of β of a particular mode has mixed contributions from material dispersion and waveguide dispersion. It is in fact more convenient to consider this combined effect directly. To do so, we only have to replace k in all of the formulas in Section 1.9 by β of the particular mode under consideration, thus defining the effective refractive index n_β, the effective group index N_β, and the effective

group-velocity dispersion D_β for the mode:

$$n_\beta = \frac{c\beta}{\omega}, \tag{3.97}$$

$$N_\beta = c\frac{\mathrm{d}\beta}{\mathrm{d}\omega} = n_\beta - \lambda\frac{\mathrm{d}n_\beta}{\mathrm{d}\lambda}, \tag{3.98}$$

and

$$D_\beta = c\omega\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2} = \lambda^2\frac{\mathrm{d}^2 n_\beta}{\mathrm{d}\lambda^2}. \tag{3.99}$$

The exact frequency dependence of these parameters depends on the parameters of the fiber, which are, specifically, the V number, the normalized index difference Δ, and, in the case of the power-law profiles, the parameter α. Since most optical fibers are weakly guiding, we consider only weakly guiding fibers in the following to simplify the mathematics.

In the case of a step-index fiber, it is convenient to use the normalized guide index b, which has the same form as that defined in (2.47), for the step-index planar waveguide:

$$b = \frac{n_\beta^2 - n_2^2}{n_1^2 - n_2^2}. \tag{3.100}$$

Taking the weakly guiding approximation of (3.48) and using (3.97), we have

$$n_\beta \approx n_2(1 + b\Delta). \tag{3.101}$$

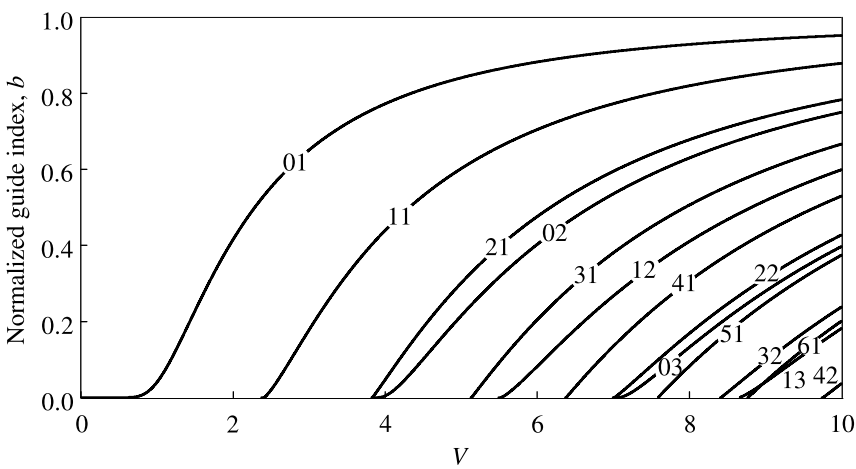


Figure 3.10 Normalized propagation constant b as a function of fiber V number for some LP modes of a weakly guiding step-index fiber.

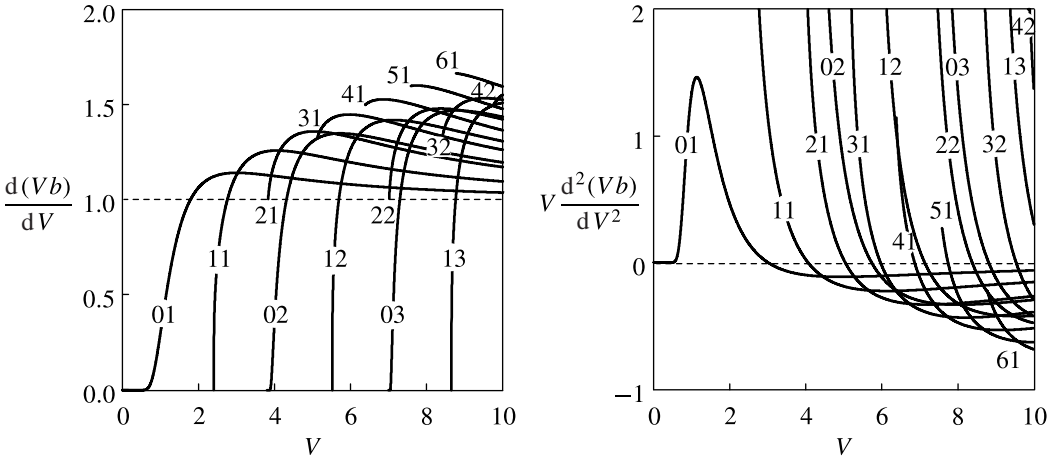


Figure 3.11 (*a*) Waveguide group delay parameter, $\mathrm{d}(Vb)/\mathrm{d}V$, and (*b*) waveguide dispersion parameter, $V\mathrm{d}^2(Vb)/\mathrm{d}V^2$, as a function of fiber V number for some LP modes of a weakly guiding step-index fiber.

The b parameter can be found by solving (3.57) together with (3.11) and (3.12) for β. It is plotted as a function of the fiber V number in Fig. 3.10 for some LP modes. The frequency or wavelength dependence of n_β can be found from the dependence of b on V. Using (3.98) and (3.99), we also find that

$$N_\beta \approx N_2 \left[1 + \frac{\mathrm{d}(Vb)}{\mathrm{d}V}\Delta\right], \tag{3.102}$$

where N_2 is the group index of the cladding material of the fiber, and

$$D_\beta \approx D_2 \left[1 + \frac{\mathrm{d}(Vb)}{\mathrm{d}V}\Delta\right] + \frac{N_2^2}{n_2}\frac{V\mathrm{d}^2(Vb)}{\mathrm{d}V^2}\Delta, \tag{3.103}$$

where D_2 is the group-velocity dispersion of the fiber cladding. In deriving (3.102) and (3.103), terms such as $\mathrm{d}\Delta/\mathrm{d}\omega$ and $\mathrm{d}^2\Delta/\mathrm{d}\omega^2$ that contain the differential material dispersion are dropped because they are usually very small compared with the terms we keep. For more accurate calculations, they should be included. In each of the relations in (3.101)–(3.103), the first term is the material contribution while the other terms are the waveguide contribution. The *waveguide group delay parameter*, $\mathrm{d}(Vb)/\mathrm{d}V$, and the *waveguide dispersion parameter*, $V\mathrm{d}^2(Vb)/\mathrm{d}V^2$, are plotted as a function of fiber V number in Figs. 3.11(*a*) and (*b*), respectively.

It can be seen from the discussion above and from the data plotted in Fig. 3.10 that n_β is bounded by n_1 and n_2, reaching n_2 near cutoff and approaching n_1 far away from cutoff. In contrast, Fig. 3.11(*a*) shows that only LP_{0n} and LP_{1n} modes have N_β reaching N_2 at cutoff because only LP_{0n} and LP_{1n} modes have their power moved completely away from the core into the cladding at cutoff. An LP_{mn} mode with $m \geq 2$

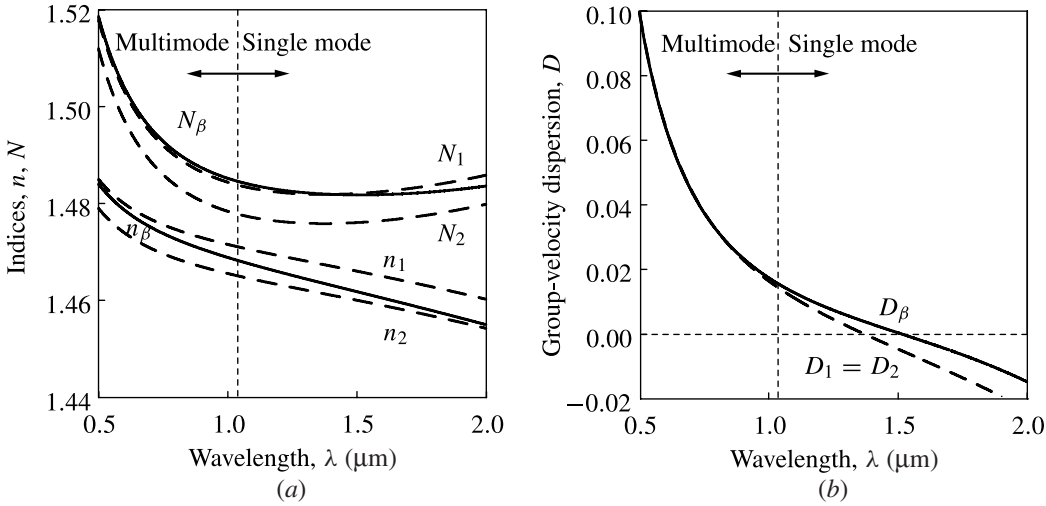


Figure 3.12 (*a*) Effective index of refraction and group index and (*b*) group-velocity dispersion of the fundamental mode as a function of wavelength. The solid curves show the effective parameters of the mode with both material and waveguide contributions. The dashed curves show only the material contribution to the core and cladding regions, labeled 1 and 2, respectively.

still has a large fraction of its power concentrated in the core at cutoff, as discussed in Section 3.2 and shown in Fig. 3.6. Figure 3.11(*b*) shows that the group-velocity dispersion can be modified by the waveguide contribution. As a practical example, Fig. 3.12 shows the combined material and waveguide contributions for the fundamental mode of a step-index germania–silica fiber with an index step of $n_1 - n_2 = 0.006$ and a core radius of $a = 3$ μm. The core is assumed to have 13.5 mol % GeO_2 for n_1, N_1, and D_1 to be consistent with those of the material properties of the germania–silica glass shown in Fig. 3.9. It is also assumed that the differential material dispersion, $d\Delta/d\omega$, is negligible so that $D_1 = D_2$ to show the effect of waveguide dispersion clearly. As shown in Fig. 3.12(*b*), the point of zero dispersion is shifted from that of the germania–silica material at 1.383 μm to 1.5 μm because of the waveguide contribution.

EXAMPLE 3.6 A step-index single-mode fiber for transmitting a signal at $\lambda = 1.35$ μm is 100 km long. At this wavelength, the fiber has the following parameters for its silica cladding: $n_2 = 1.446$, $N_2 = 1.466$, and $D_2 = -0.0027$. Its core has a radius of $a = 4$ μm and an index of $n_1 = 1.450$. What are the propagation constant, the group velocity, and the group-velocity dispersion of the signal propagating as the guided mode of the fiber? If a 10-ps pulse that has a spectral width of $\Delta\lambda_{ps} = 2$ nm is sent through the fiber, what is its transmission time through the fiber? What is the pulse duration when it arrives at the other end of the fiber?

Solution With the given parameters, we find that $V = 2$ at $\lambda = 1.35$ μm. Thus, this fiber is indeed a single-mode fiber for this wavelength. From Figs. 3.10 and 3.11, we find the following parameters for $V = 2$:

$$b = 0.416, \qquad \frac{\mathrm{d}(Vb)}{\mathrm{d}V} = 1.065, \qquad \frac{V\mathrm{d}^2(Vb)}{\mathrm{d}V^2} = 0.461.$$

Using these parameters, together with the given values of n_2, N_2, and D_2, we find from (3.101), (3.102), and (3.103) that

$$n_\beta = 1.448, \qquad N_\beta = 1.470, \qquad D_\beta = -0.000\,82.$$

From these results, we then find the following parameters for the mode:

$$\beta = \frac{2\pi n_\beta}{\lambda} = 6.739\ \mu\mathrm{m}^{-1}, \quad v_\mathrm{g} = \frac{c}{N_\beta} = 2.04 \times 10^8\ \mathrm{m\,s}^{-1},$$

$$D_\lambda = -\frac{D_\beta}{c\lambda} = 2.02\ \mathrm{ps\,km}^{-1}\ \mathrm{nm}^{-1}.$$

The transmission time of the pulse through the 100 km length of the fiber is

$$t_\mathrm{tr} = \frac{l}{v_\mathrm{g}} = 490\ \mu\mathrm{s}.$$

The spread of the pulse due to group-velocity dispersion is

$$\Delta t_\mathrm{g} = |D_\lambda| \Delta\lambda_\mathrm{ps} l = 404\ \mathrm{ps}.$$

Therefore, the pulse arrives at the end of the fiber after 490 μs with a substantially broadened pulse duration of $\Delta t_\mathrm{ps} = 10\ \mathrm{ps} + 404\ \mathrm{ps} = 414\ \mathrm{ps}$.

Although the data shown in Figs. 3.10–3.12 are specific for step-index fibers, the formulas obtained in (3.101)–(3.103) are equally applicable to graded-index fibers. However, in order to use (3.101)–(3.103) for a graded-index fiber, exact solution of the eigenvalue equation (3.75) has to be carried out to obtain the frequency dependence of the propagation constant β, and thus the dependence of b on V, for each guided mode of interest. This would be the required procedure if the fiber under consideration were a single-mode graded-index fiber or a multimode graded-index fiber that supported only a few modes.

For a multimode graded-index fiber that supports a very large number of modes, the approximate solution of β given by (3.88) can be used. Then, instead of expressing the index and dispersion parameters in terms of b and V, we can use (3.97)–(3.99) directly

to obtain

$$n_\beta = n_1(1 - 2\zeta\Delta)^{1/2}, \tag{3.104}$$

$$N_\beta \approx N_1\left(1 + \frac{\alpha - 2 - \delta}{\alpha + 2}\zeta\Delta + \frac{3\alpha - 2 - 2\delta}{\alpha + 2}\frac{\zeta^2\Delta^2}{2}\right), \tag{3.105}$$

and

$$D_\beta \approx D_1\left(1 + \frac{\alpha - 2 - \delta}{\alpha + 2}\zeta\Delta\right) - \frac{N_1^2}{n_1}\frac{2(\alpha - \delta/2)(\alpha - 2 - \delta)}{(\alpha + 2)^2}\zeta\Delta, \tag{3.106}$$

where

$$\zeta = \left(\frac{M_\beta}{M}\right)^{\alpha/(\alpha+2)} \tag{3.107}$$

and

$$\delta = \frac{2n_1}{N_1}\frac{\omega}{\Delta}\frac{\mathrm{d}\Delta}{\mathrm{d}\omega} = -\frac{2n_1}{N_1}\frac{\lambda}{\Delta}\frac{\mathrm{d}\Delta}{\mathrm{d}\lambda}. \tag{3.108}$$

Again, the first term in each of (3.104)–(3.106) represents the material contribution, while the other terms account for waveguide contributions.

Modal dispersion

Modal dispersion exists because different modes in a multimode waveguide propagate at different group velocities, as indicated by (3.105). Note that D_β given by (3.106) is the total *intramode* dispersion including material and waveguide contributions for a mode that has a propagation constant β in a multimode fiber. It has nothing to do with *intermode* dispersion. To find the modal dispersion, we have to consider the difference in N_β between modes of different β. This difference exists even when there is no intramode chromatic dispersion so that D_β vanishes.

In a multimode fiber, the modal dispersion between the fundamental mode and the highest-order mode supported by the fiber can be estimated. Because the fundamental mode HE_{11}, or LP_{01}, has two-fold degeneracy, it corresponds to $M_\beta = 2$. For the highest mode, $M_\beta = M$. Therefore, for a fiber with a very large number of modes, $\zeta_{\text{low}} = 2/M \approx 0$, while $\zeta_{\text{high}} = 1$. They determine the minimum and maximum values of N_β among the modes. The modal dispersion can then be expressed as

$$N_{\text{high}} - N_{\text{low}} = N_1\left(\frac{\alpha - 2 - \delta}{\alpha + 2}\Delta + \frac{3\alpha - 2 - 2\delta}{\alpha + 2}\frac{\Delta^2}{2}\right). \tag{3.109}$$

Note that $N_{\text{high}} > N_{\text{low}}$ when $\alpha > 2 + \delta$, but $N_{\text{high}} < N_{\text{low}}$ when $\alpha < 2 + \delta$. Because $v_{\text{g}} = c/N$, this dispersion represents the difference in the group velocity between different modes. *Although it is always true that a low-order mode has a larger β and thus*

a smaller phase velocity than a high-order mode, the relationship between their group velocities is less straightforward. It depends on many factors, including the waveguide structure, the index profile, the material properties, and how far away the modes are from cutoff. For example, it can be seen from (3.109) that a low-order mode travels faster than a high-order mode if $\alpha > 2 + \delta$, whereas the reverse is true if $\alpha < 2 + \delta$. In addition, it has to be kept in mind that even this statement is not always true for modes near cutoff, as can be seen from the discussions for step-index fibers and from Fig. 3.11(a) using (3.102). Therefore, modal dispersion can also be modified by choosing appropriate waveguide and material parameters.

Dispersion compensation

We have seen that all three types of dispersion in a fiber can be modified to a certain extent by various means. Therefore, it is possible to engineer a desired dispersion characteristic through careful choice of the type and concentration of dopants to control the material dispersion while designing the fiber parameters to adjust the waveguide dispersion and the modal dispersion. In some special applications, one might want a certain nonzero value of positive or negative dispersion at a particular wavelength. For example, one would need a finite amount of positive group-velocity dispersion in the application of fiber-grating compression of optical pulses, whereas one would need finite negative group-velocity dispersion for the generation and propagation of soliton pulses in a fiber. Nevertheless, in most applications using fibers to transmit optical signals, dispersion in a fiber causes undesirable spreading of the signal, limiting the bandwidth of transmission. It is desirable to reduce the dispersion to zero, if possible, for such applications.

For applications that require the largest bandwidths, single-mode fibers are the choice because modal dispersion does not exist in a single-mode fiber. Because the zero-dispersion point of pure silica is near the window of a local minimum of attenuation at 1.3 μm, transmission systems based on this wavelength have the combined advantage of low attenuation and low dispersion and are a choice for long-distance optical communications. As discussed in the preceding section, the real minimum of attenuation appears at 1.55 μm wavelength. Therefore, it is usually desirable to shift the point of zero dispersion to this wavelength. This task can be accomplished by a combination of choices of dopants and waveguide parameters, as demonstrated by the example shown in Fig. 3.12 where the point of zero dispersion is shifted to 1.5 μm already. Such fibers are known as *dispersion-shifted fibers*. Zero dispersion in both 1.3- and 1.55-μm windows can also be accomplished by special profiling of the fiber, resulting in so-called *dispersion-flattened fibers* that have low dispersion in the region between 1.3 and 1.55 μm with zero crossings at both wavelengths. Figure 3.13 shows an example of a dispersion-flattened fiber.

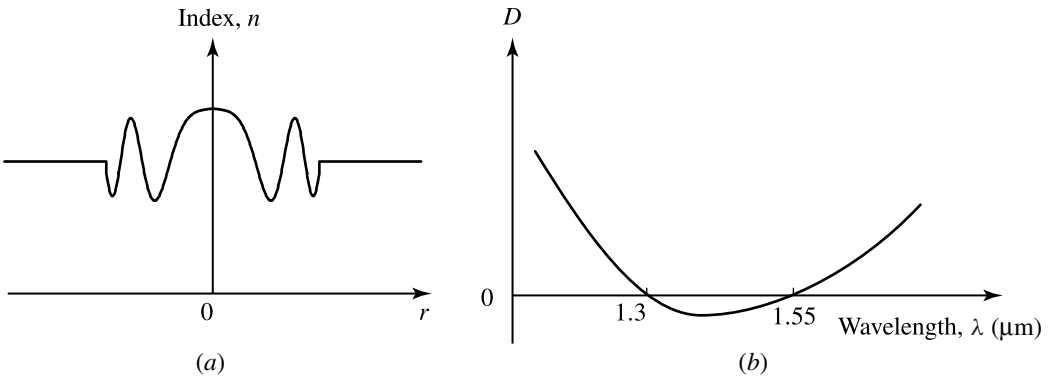


Figure 3.13 (*a*) Index profile and (*b*) dispersion characteristics of a dispersion flattened fiber.

For multimode fibers, modal dispersion dominates intramode waveguide dispersion. It is then important to minimize the modal dispersion. From (3.109), it can be seen that modal dispersion can be minimized if we choose

$$\alpha = 2 + \delta. \tag{3.110}$$

The value of δ depends on the dopants and the optical wavelength. In the near infrared spectral region, it is usually within the range of ± 0.3 for most dopants. Therefore, *the optimum profile for a low-dispersion multimode fiber is one close to a quadratic graded-index profile*. This results in a modal dispersion of

$$N_{\text{high}} - N_{\text{low}} = N_1 \frac{\Delta^2}{2}. \tag{3.111}$$

In comparison, a step-index multimode fiber has $\alpha = \infty$ and

$$N_{\text{high}} - N_{\text{low}} = N_1 \Delta. \tag{3.112}$$

Because Δ is a very small number, the modal dispersion in an optimized graded-index fiber is substantially lower than that in a step-index multimode fiber. It is interesting to see that the total intramode dispersion D_β given by (3.106) is also mode dependent. However, when α is chosen to be the optimum value given by (3.110), $D_\beta = D_1$, and the intramode dispersion becomes mode independent, indicating that waveguide dispersion is minimized. Therefore, *a graded-index fiber that has a minimum modal dispersion also has a minimum waveguide contribution to the intramode dispersion for each individual mode*.

EXAMPLE 3.7 A multimode fiber of 10 km length has a core group index of $N_1 = 1.5$ and an index step of $\Delta = 2\%$. If an optical signal sent through this fiber is carried by all of its guided modes, what is the transmission time of the signal? What is the temporal

broadening of the signal due to modal dispersion if the fiber is a dispersion-optimized graded-index fiber? What is the broadening if the fiber is a step-index fiber?

Solution The transmission time of the signal is

$$t_{\text{tr}} = \frac{l}{v_{\text{g}}} = \frac{l}{c} N_\beta \approx \frac{l}{c} N_1 = 50\ \mu\text{s}.$$

In the case of an optimized graded-index fiber, the temporal broadening due to modal dispersion is

$$\Delta t_{\text{mode}} = \frac{l}{c}(N_{\text{high}} - N_{\text{low}}) = \frac{l N_1}{2c}\Delta^2 = 10\ \text{ns}.$$

In the case of a step-index fiber, the temporal broadening due to modal dispersion is

$$\Delta t_{\text{mode}} = \frac{l}{c}(N_{\text{high}} - N_{\text{low}}) = \frac{l N_1}{c}\Delta = 1\ \mu\text{s}.$$

Clearly, temporal broadening of the signal due to modal dispersion is much worse in a step-index fiber than in an optimized graded-index fiber.

PROBLEMS

3.1.1 The only TE and TM modes that exist in a circular fiber are TE_{0n} and TM_{0n}. For simplicity, consider the case of a step-index optical fiber.
   a. Show that TE_{mn} modes with $m \neq 0$ cannot exist in a circular fiber by considering the continuity of the transverse magnetic field component $\mathcal{H}_\phi$ at the boundary between the core and the cladding.
   b. Show that TM_{mn} modes with $m \neq 0$ cannot exist in a circular fiber by considering the continuity of the transverse electric field component $\mathcal{E}_\phi$ at the boundary between the core and the cladding.
   c. What are the difference between the TE_{0n} mode and other TE modes and that between the TM_{0n} mode and other TM modes that allow the TE_{0n} and TM_{0n} modes to exist?

3.1.2 Which of the following are legitimate modes in a circular fiber: TE_{05}, TM_{32}, HE_{02}, EH_{22}, HE_{13}, EH_{04}, TEM_{00}, HE_{20}?

3.1.3 What is the fundamental mode of a circular fiber? Sketch its field and intensity profiles.

3.1.4 In this problem, we compare a circular fiber and a slab waveguide.
   a. Name the types of guided modes that exist in a slab waveguide and those of true modes that exist in a circular fiber.
   b. Name the modes that have the largest propagation constant in a circular fiber and in a slab waveguide, respectively.

c. Is it possible for a slab waveguide to support no guided mode at all? Is it possible for a circular fiber to support no guided mode?

3.1.5 Counting all possible degeneracies, a dielectric waveguide has exactly six guided modes. When its core dimension is reduced, some or all of these modes will be cut off.

a. If the waveguide is an asymmetric slab waveguide, which mode or modes will be cut off first? Which mode or modes will never be cut off unless the waveguide core disappears completely?

b. Answer the questions in (a) for a symmetric slab waveguide.

c. Answer the questions in (a) for a circular waveguide such as an optical fiber.

3.1.6 A step-index waveguide has an index of refraction of $n_1 = 1.45$ for its core and an index step of $\Delta n = 0.005$. Neglect the wavelength dependence of the index of refraction.

a. If the waveguide is a slab waveguide of $d = 3$ μm core thickness, what is the shortest wavelength for it to be a single-mode waveguide?

b. If it is an optical fiber of $a = 3$ μm core radius, what is the shortest wavelength for it to be a single-mode fiber?

3.1.7 A step-index optical fiber has a numerical aperture of NA $= 0.1$. Its cladding is pure silica and has a refractive index of $n_2 = 1.465$.

a. What is the largest core *diameter* for which the fiber remains single moded at an optical wavelength of 1.3 μm?

b. If the core diameter is found to be 8 μm, what is the wavelength range in which the fiber remains single moded?

c. What is the refractive index n_1 of the core?

3.1.8 The index of refraction of fused silica is $n = 1.452$, 1.447, and 1.444 at $\lambda =$ 850 nm and 1.30 and 1.55 μm, respectively. A step-index silica fiber is found to be single moded at a wavelength of 1.55 μm and multimoded at a wavelength of 1.30 μm.

a. Is the fiber single moded or multimoded at $\lambda = 850$ nm?

b. The core diameter is known to be $2a = 9$ μm. Estimate the index step, $\Delta n = n_1 - n_2$. What is the numerical aperture?

c. If a fiber of the same index step is single moded at $\lambda = 1$ μm, what is the limit of the acceptable values for its core diameter?

3.2.1 What are the LP modes of a circular fiber? Which LP modes can propagate like true normal modes? Which ones cannot?

3.2.2 Can the LP_{01} mode of a circular fiber propagate without changing its field distribution pattern? How about the LP_{11} mode?

3.2.3 Sketch the intensity distribution pattern of the LP_{21} mode in a weakly guiding step-index fiber.

3.2.4 Which of the following are legitimate fiber modes: HE_{32}, TE_{11}, EH_{03}, TM_{02}, TEM_{01}, LP_{01}, LP_{35}, LP_{00}, LP_{11}? Which are true normal modes? Which are approximate modes?

3.2.5 In this problem, we consider the behavior of a monochromatic laser beam that is coupled into an optical fiber.

a. The beam is coupled into a multimode fiber. After the optical wave travels for a certain distance in the fiber, the diameter of the fiber shrinks to become a single-mode fiber. What will happen to the optical field in the fiber? Are all of the different mode components of the optical field still there?

b. Now the beam is initially coupled into a single-mode fiber. After a certain distance, the fiber expands to become a multimode fiber. What happens to the optical field in the fiber?

3.2.6 The index step Δ of a practical single-mode silica fiber is typically in the range of 0.1–0.2%, whereas that of a practical multimode silica fiber is typically in the range of 1–2%. Discuss the reasons for such a difference. Use numerical examples to illustrate your arguments.

3.2.7 Determine the cladding index of refraction of a fiber that has a core index of 1.5, a core radius of 5 μm, and $V = 2.0$ at $\lambda = 1.5$ μm. What is the shortest wavelength λ_c at which the fiber is a single-mode fiber? What is the number of modes if this fiber is used at $\lambda_c/2$?

3.2.8 The refractive index of a glass fiber varies with optical wavelength relatively linearly in the neighborhood of $\lambda = 1.3$ μm. For the cladding, the index of refraction is approximately given by

$$n_2 = 1.465 - 0.0114(\lambda - 1.3)$$

in the wavelength range between 1 and 1.6 μm, where the optical wavelength λ is measured in micrometers. Assume that the index of refraction n_1 of the core of a step-index fiber has the same wavelength dependence as that of the cladding but is larger by a fixed amount Δn, so that $n_1 = n_2 + \Delta n$ for the entire spectral range of interest here. Use the wavelength dependence of n_2 given above to design a step-index optical fiber that supports only the fundamental LP_{01} mode for wavelengths at 1.3 and 1.55 μm but becomes a multimode fiber at 1 μm wavelength. For practical purposes, we want the diameter of the fiber core to be less than ten optical wavelengths.

3.2.9 A step-index silica fiber has a core index of 1.453, an index step, $\Delta = \Delta n/n_1$, of 0.2%, and a core diameter of 10 μm.

a. Is the fiber single moded or multimoded at $\lambda = 1.3$ μm? What is the cutoff wavelength for single-mode operation of the fiber?

b. If a fiber with the given index profile were to support 400 modes at $\lambda = 1.3$ μm, what should its core diameter be? Compare your answer with the standard outer diameter of 125 μm for multimode fibers.

c. If the core diameter of a multimode fiber with the given core index were to be fixed at the standard value of 50 μm, what should its index step be in order for it to support 400 modes at $\lambda = 1.3$ μm?

3.2.10 A multimode optical fiber is found to have 1000 guided modes at an optical wavelength $\lambda = 500$ nm. Neglect the dispersion of the refractive index of the fiber material when answering the following questions.

a. How many guided modes does it have at 1.3 μm wavelength?

b. If another fiber is found to have exactly the same index step and index profile but only half the diameter, how many modes does this fiber have at 500 nm wavelength?

3.2.11 The refractive index of pure silica is 1.444 at 1.55 μm wavelength. The cladding layers of optical fibers considered in this problem are made of pure silica.

a. Design a single-mode silica fiber for $\lambda = 1.55$ μm that has a numerical aperture in the range of $0.1 > \text{NA} > 0.07$ and is multimoded at $\lambda = 1.3$ μm.

b. Design a multimode fiber that supports 500 modes at $\lambda = 1.55$ μm and has a core diameter of 50 μm.

c. How many modes does the multimode fiber support at $\lambda = 1.3$ μm?

3.3.1 A multimode graded-index optical fiber for $\lambda = 1.3$ μm has a power-law index profile characterized by the following parameters: $n_1 = 1.466$, $n_2 = 1.451$, and $\alpha = 2$. The core radius of the fiber is 50 μm.

a. How many guided modes does this fiber support at $\lambda = 1.3$ μm?

b. Estimate the propagation constant of the LP_{01} mode and that of the highest-order guided mode.

c. Estimate the propagation constant of the 500th guided mode if it is supported.

3.3.2 A multimode silica fiber has a 50 μm core diameter. The refractive index of its cladding is 1.453 at 850 nm wavelength and 1.448 at 1.3 μm wavelength.

a. If the fiber has a step-index profile, what is the minimum refractive index of the core at 850 nm wavelength for it to support at least 500 modes at this wavelength? What is the numerical aperture?

b. Neglecting the dispersion of the index step, how many modes can the fiber determined in (a) support at 1.3 μm wavelength?

c. A graded-index fiber has a quadratic index profile with the same core diameter and the same numerical aperture as the step-index fiber found above. How many modes can it support at 850 nm and 1.3 μm wavelengths, respectively?

d. At 1.3 μm, the combined material and waveguide dispersion is approximately zero for these fibers so that modal dispersion completely dominates. The group index of the core at this wavelength is $N_1 = 1.462$. Find the distances of propagation for a 1-ns pulse to double its width in the step-index and graded-index fibers, respectively.

3.3.3 A single-mode fiber has a core index of $n_1 = 1.448$, an index step of $\Delta = 0.1\%$, and a core diameter of 12 μm. What is its cutoff wavelength for single-mode performance if it is a step-index fiber? What is its cutoff wavelength if it is a graded-index fiber with $\alpha = 2$? Ignore dispersion in solving this problem.

3.4.1 What limits the use of ordinary silica optical fibers in the infrared region? What is the physical mechanism that contributes to the limitation?

3.4.2 What are the two wavelength windows for long-distance fiber-optic communications? What advantages do they offer over other wavelengths?

3.4.3 A detector used in an optical communication system has the sensitivity to detect signals at a power level of 1 μW at the output end of a fiber transmission line. The attenuation coefficient of the fiber is 0.3 dB km^{-1} at 1.3 μm and is 0.2 dB km^{-1} at 1.55 μm. At an input power level of 1 mW, what is the maximum distance over which the signals can be sent through this system and be detected if the signal wavelength is 1.3 μm? What is the distance if the signal wavelength is 1.55 μm?

3.4.4 An optical fiber has an attenuation coefficient of 0.2 dB km^{-1} at 1.55 μm wavelength. If the output power is required to be at least 100 nW and the transmission distance is 200 km, what is the required input power in milliwatts and in decibel-milliwatts, respectively? If the fiber attenuation coefficient is only increased by 10% to 0.22 dB km^{-1}, how much increase in the input power is required to maintain the same transmission distance?

3.4.5 A detector used in an optical communication system has the sensitivity to detect signals at 1.3 μm optical wavelength at a power level as low as 1 μW. If such a detector is used to monitor the signals at the output end of a fiber transmission line, what is the maximum distance over which the signals can be sent through a low-loss silica fiber at an input power level of 1 mW? What could be done if the signals were to be transmitted over a distance twice as long? What if it is necessary to send the signal across the country for a distance of as long as 5000 km? Assume that the low-loss fiber has an attenuation coefficient of 0.3 dB km^{-1} at 1.3 μm wavelength. Note that the power that can be sent into a fiber for long-distance communications is limited to the order of 10–20 mW by nonlinear optical effects, while the power that can be detected by the most sensitive detectors is limited to the order of about 1 pW for practical purposes.

3.4.6 A fiber-optic link operating at 1.55 μm wavelength consists of many sections of low-loss fibers, which are connected with fiber connectors. The optical power coupled into the transmission link at the input end is −5 dBm. The detector at the output end of the link requires that the minimum optical power incident upon it be at least −50 dBm. The maximum length of fiber that can be used between connectors is 2 km. The loss of each connector is 2 dB.

a. If the fiber attenuation is 1 dB km^{-1}, what is the maximum length of the link?

b. What is the ultimate limitation of the link length if the fiber attenuation is reduced to the absolute minimum?

3.4.7 What are the three most important loss mechanisms that ultimately determine the attenuation characteristics of a silica fiber in the visible and infrared spectral regions? At what wavelength is the minimum attenuation typically found in a silica fiber?

3.5.1 Verify (3.103) and discuss the physical meaning of each term.

3.5.2 In the case of a regular step-index silica fiber, where does zero group-velocity dispersion normally occur? How does this zero dispersion point shift from a single-mode fiber of a large core diameter with $V = 2.4$ to one with a reduced core diameter if the two have exactly the same index profile and same material compositions?

3.5.3 Determine the core radius of a multimode step-index fiber with a numerical aperture of NA $= 0.1$ if the number of modes is $M = 5000$ when the wavelength is 870 nm. If the core refractive index is $n_1 = 1.445$, the group index is $N_1 = 1.456$, and Δ is approximately independent of wavelength, determine the modal dispersion for a 2-km fiber.

3.5.4 The total dispersion in a single-mode step-index fiber includes waveguide dispersion and material dispersion. Therefore, the point of zero dispersion can be shifted by choosing appropriate waveguide parameters for the fiber. Consider a step-index fiber of pure silica cladding.

a. Show that for a single-mode fiber of a simple index profile with $n = n_1$ for $r < a$ and $n = n_2$ for $r > a$, the waveguide dispersion always shifts the point of zero dispersion to a wavelength longer than that of the zero material dispersion.

b. Using the data in Figs. 3.9, 3.10, and 3.11, design a single-mode step-index fiber of the smallest possible index step that has zero dispersion at 1.55 μm wavelength due to compensation of material dispersion by waveguide dispersion.

3.5.5 What is a dispersion-shifted fiber? What is the major motivation for making such a fiber?

3.5.6 What is the primary consideration that may lead one to choose a graded-index fiber over a step-index fiber for application?

3.5.7 A step-index optical fiber is designed for single-mode applications in the transmission window covering the optical wavelength range from 800 to 900 nm. Its numerical aperture is NA $= 0.15$. Its core diameter is chosen to be the largest possible. It has an attenuation coefficient of 5 dB km^{-1} in this spectral range. Its group-velocity dispersion, including the material and waveguide effects, at the center wavelength of 850 nm in this range is $D_\lambda = -820$ ps nm^{-1} km^{-1}.

a. What is the core diameter of this fiber?

b. If 500 μW of input power is coupled into this fiber for transmission over a distance of 2 km, what is the output power?

c. What is the maximum transmission delay between pulses of different wavelengths in this spectral range?

3.5.8 A step-index optical fiber has a core diameter of 9 μm. Its numerical aperture is NA = 0.1. At a wavelength of 1.55 μm, the attenuation coefficient of the fiber is 0.15 dB km^{-1}, and the group-velocity dispersion D_λ is 17 ps nm^{-1} km^{-1}.

a. Is this fiber single moded or multimoded at the wavelength of 1.55 μm? Why?

b. Suppose this fiber is used for long-distance optical pulse transmission. Optical amplifiers with a gain of 30 dB and a minimum required input power of −30 dBm are used to compensate for the loss in the fiber. If 1 mW of input power is coupled into the fiber and the sensitivity of the receiver is −50 dBm, how many amplifiers are needed for a transmission distance of 1000 km?

c. If an optical pulse being transmitted has a temporal pulsewidth of 1 ns and a spectral width of 0.016 nm, how far can we extend the transmission distance by adding more amplifiers before hitting the limitation imposed by group-velocity dispersion? Assume that the dispersion limitation is reached when the temporal pulsewidth is doubled.

SELECT BIBLIOGRAPHY

Buckman, A. B., *Guided-Wave Photonics*. Fort Worth, TX: Sauders College Publishing, 1992.

Cheo, P. K., *Fiber Optics and Optoelectronics*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Cherin, A. H., *An Introduction to Optical Fibers*. New York: McGraw-Hill, 1983.

Gowar, J., *Optical Communication Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Iizuka, K., *Elements of Photonics for Fiber and Integrated Optics*, Vol. II. New York: Wiley, 2002.

Kao, C. K., *Optical Fibre*. London: Peter Peregrinus, 1988.

Kasap, S. O., *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice-Hall, 2001.

Krohn, D. A., *Fiber Optic Sensors*. Research Triangle Park, NC: Instrument Society of America, 1988.

Marcuse, D., *Theory of Dielectric Optical Waveguides*, 2nd edn. Boston, MA: Academic Press, 1991.

Murata, H., *Handbook of Optical Fibers and Cables*. New York: Marcel Dekker, 1988.

Neumann, E. G., *Single-Mode Fibers Fundamentals*. New York: Springer-Verlag, 1988.

Okoshi, T., *Optical Fibers*. New York: Academic Press, 1982.

Powers, J., *An Introduction to Fiber Optic Systems*, 2nd edn. Chicago, IL: Irwin, 1997.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Snyder, A. W. and Love, J. D., *Optical Waveguide Theory*. New York: Chapman and Hall, 1983.

Udd, E., ed., *Fiber Optic Sensors*. New York: Wiley, 1990.

ADVANCED READING LIST

Ainslie, J., and Day, C. R., "A review of single-mode fibers with modified dispersion characteristics," *Journal of Lightwave Technology* **LT-4**(8): 967–979, Aug. 1986.

Cancellieri, G. and Chiaraluce, F., "Recent progress in fibre optics," *Progress in Quantum Electronics* **18**(1): 39–95, 1994.

Gambling, W. A., "The rise and rise of optical fibers," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1084–1093, Nov.–Dec. 2000.

Gloge, D., "Weakly guiding fibers," *Applied Optics* **10**(10): 2252–2258, Oct. 1971.

Goyal, I. C., "Dispersion in telecommunication optical fibres: a tutorial review," *Journal of the Institution of Electronics and Telecommunication Engineers* **32**(4): 196–205, July–Aug. 1986.

Kogelnik, H., "High-speed lightwave transmission in optical fibers," *Science* **228**(4703): 1043–1048, May 1985.

Noda, J., Okamoto, K., and Sasaki, Y., "Polarization-maintaining fibers and their applications," *Journal of Lightwave Technology* **LT-4**(8): 1071–1089, Aug. 1986.

Tran, D. C., Sigel, G. H., Jr., and Bendow, B., "Heavy metal fluoride glasses and fibers: a review," *Journal of Lightwave Technology* **LT-2**(5): 566–585, Oct. 1984.

4 Coupling of waves and modes

The principles of many photonic devices are based on the coupling between optical fields of different frequencies or different spatial modes. In general, the coupling mechanism can be described by a polarization $\Delta\mathbf{P}$ on top of a background polarization representing the property of the medium in the absence of the coupling mechanism. In this chapter, we present the general coupled-wave and coupled-mode formalisms, which provide the foundation for understanding the functions of many devices in the following chapters. The coupled-wave formalism deals with the coupling of optical waves of different frequencies, whereas coupled-mode theory applies to the coupling of optical fields of different spatial modes.

4.1 Coupled-wave theory

In this section, the general formulation of the coupled-wave formalism for coupling of optical waves of different frequencies is presented. For simplicity, we consider only coupling among plane optical waves, but the formulation can be easily extended for nonplane waves, such as optical waves of Gaussian beam profiles.

As discussed in Section 1.3, coupling among optical waves of different frequencies is possible only if the optical property of the medium in which the optical waves propagate is either time varying or optically nonlinear. Time-varying optical properties can be induced by time-varying electric, magnetic, or acoustic fields through electro-optic, magneto-optic, or acousto-optic effects, which are discussed in Chapters 6, 7, and 8, respectively. In particular, an acoustic wave always induces time-varying changes in the optical property of a medium, whereas changes induced by electro-optic or magneto-optic effects can be static when they are caused by static electric or magnetic fields. Nonlinear optical properties are discussed in Chapter 9. Here we consider the general formulation without specifying the physical mechanism responsible for the coupling of optical waves. Applications of the couple-wave formalism to specific situations are seen in later chapters, particularly Chapters 8 and 9.

The time-varying or nonlinear optical property responsible for coupling of optical waves of different frequencies can be generally described by a polarization, $\Delta\mathbf{P}$, induced

by the underlying effect. In the absence of a coupling mechanism, an optical wave propagating in a medium is described by the linear wave equation

$$\nabla \times \nabla \times \mathbf{E} + \mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2} = 0, \tag{4.1}$$

where $\mathbf{D}$ accounts for only the linear, static property of the medium. For a monochromatic optical wave of constant amplitude at a frequency ω, this equation reduces to

$$\nabla \times \nabla \times \mathbf{E} - \omega^2 \mu_0 \boldsymbol{\epsilon}(\mathbf{k}, \omega) \cdot \mathbf{E} = 0, \tag{4.2}$$

where $\boldsymbol{\epsilon}(\mathbf{k}, \omega)$ describes the linear, time-independent optical property of the medium. Among the solutions of (4.2) are monochromatic plane waves and Gaussian waves. Here we consider only the plane waves, but the same concept applies to Gaussian waves as well.

Clearly, an optical wave of frequency ω that is governed by (4.2) propagates independently of waves of other frequencies. Therefore, optical waves of different frequencies do not couple if each of them is governed by (4.1) with $\mathbf{D}$ characterizing only the linear, static property of the medium. To describe the coupling, a certain polarization $\Delta\mathbf{P}$ that characterizes the coupling mechanism has to be included in the wave equation:

$$\nabla \times \nabla \times \mathbf{E} + \mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2} = -\mu_0 \frac{\partial^2 \Delta\mathbf{P}}{\partial t^2}. \tag{4.3}$$

Because $\Delta\mathbf{P}$ couples waves of different frequencies, an optical wave at a given frequency ω does not propagate independently of waves of other frequencies any more. A monochromatic wave that is coupled to other frequencies cannot propagate without changing its amplitude, which includes magnitude, phase, and polarization. Consequently, a monochromatic plane wave of constant amplitude is not a solution of (4.3). In most cases of interest, however, the condition

$$|\Delta\mathbf{P}| \ll |\mathbf{D}| \tag{4.4}$$

is valid; hence the wave-coupling mechanism can be considered as a perturbation on the linear, static property of the medium. Then, the total field of the waves being coupled can be expressed as a linear combination of plane waves of different frequencies, each of which has a spatially varying amplitude:

$$\mathbf{E}(\mathbf{r}, t) = \sum_q \mathbf{E}_q(\mathbf{r}) \exp(-\mathrm{i}\omega_q t) = \sum_q \boldsymbol{\mathcal{E}}_q(\mathbf{r}) \exp(\mathrm{i}\mathbf{k}_q \cdot \mathbf{r} - \mathrm{i}\omega_q t). \tag{4.5}$$

We can also expand $\Delta\mathbf{P}$ as a linear combination of its various frequency components:

$$\Delta\mathbf{P}(\mathbf{r}, t) = \sum_q \Delta\mathbf{P}_q(\mathbf{r}) \exp(-\mathrm{i}\omega_q t). \tag{4.6}$$

Substitution of (4.5) and (4.6) in (4.3) yields the following *coupled-wave equation* (see Problem 4.1.1):

$$\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times \mathbf{E}_q - \omega_q^2 \mu_0 \boldsymbol{\epsilon}(\mathbf{k}_q, \omega_q) \cdot \mathbf{E}_q = \omega_q^2 \mu_0 \Delta \mathbf{P}_q. \tag{4.7}$$

Note that $\Delta \mathbf{P}_q(\mathbf{r})$ is generally not proportional to $\mathbf{E}_q(\mathbf{r})$. It contains the fields of other frequencies to facilitate the coupling. Moreover, it does not necessarily contain a spatial phase factor of $\exp(\mathrm{i}\mathbf{k}_q \cdot \mathbf{r})$. In the special case when the spatial phase factor of $\Delta \mathbf{P}_q(\mathbf{r})$ is $\exp(\mathrm{i}\mathbf{k}_q \cdot \mathbf{r})$, the coupling interaction is most efficient and is called *phase matched*.

Slowly varying amplitude approximation

The coupled-wave equation expressed in (4.7) is a second-order differential equation. It can be reduced to a first-order differential equation by applying the *slowly varying amplitude approximation*, which assumes that variation of the wave amplitude $\boldsymbol{\mathcal{E}}_q(\mathbf{r})$ caused by coupling to other frequencies is negligibly small over the distance of an optical wavelength. This approximation is valid in almost all situations of practical interest.

We first consider the situation in an isotropic medium where $\boldsymbol{\epsilon}$ reduces to a scalar ϵ. From the discussions in Section 1.5, we find that $\boldsymbol{\nabla} \cdot \mathbf{E} = 0$ in this case. Then, (4.7) becomes

$$\nabla^2 \mathbf{E}_q + \omega_q^2 \mu_0 \epsilon(\mathbf{k}_q, \omega_q) \mathbf{E}_q = -\omega_q^2 \mu_0 \Delta \mathbf{P}_q. \tag{4.8}$$

Substitution of the relation $\mathbf{E}_q = \boldsymbol{\mathcal{E}}_q \exp(\mathrm{i}\mathbf{k}_q \cdot \mathbf{r})$ in (4.8), followed by application of the condition $k_q^2 = \omega_q^2 \mu_0 \epsilon(\mathbf{k}_q, \omega_q)$, yields

$$\nabla^2 \boldsymbol{\mathcal{E}}_q + 2\mathrm{i}(\mathbf{k}_q \cdot \boldsymbol{\nabla}) \boldsymbol{\mathcal{E}}_q = -\omega_q^2 \mu_0 \Delta \mathbf{P}_q \mathrm{e}^{-\mathrm{i}\mathbf{k}_q \cdot \mathbf{r}}. \tag{4.9}$$

Under the slowly varying amplitude approximation, we have

$$|\nabla^2 \boldsymbol{\mathcal{E}}_q| \ll |(\mathbf{k}_q \cdot \boldsymbol{\nabla}) \boldsymbol{\mathcal{E}}_q|. \tag{4.10}$$

Consequently, the coupled-wave equation in an isotropic medium can be written as

$$(\mathbf{k}_q \cdot \boldsymbol{\nabla}) \boldsymbol{\mathcal{E}}_q \approx \frac{\mathrm{i}\omega_q^2 \mu_0}{2} \Delta \mathbf{P}_q \mathrm{e}^{-\mathrm{i}\mathbf{k}_q \cdot \mathbf{r}}. \tag{4.11}$$

In the special situation when the amplitudes of all waves being coupled vary only in a particular direction, say the z direction, we can write $\boldsymbol{\mathcal{E}}_q(\mathbf{r}) = \boldsymbol{\mathcal{E}}_q(z)$ even though $\Delta \mathbf{P}_q(\mathbf{r})$ might have variations in other directions. Then, the coupled-wave equation can be written as

$$\frac{\mathrm{d}\boldsymbol{\mathcal{E}}_q(z)}{\mathrm{d}z} \approx \frac{\mathrm{i}\omega_q^2 \mu_0}{2k_{q,z}} \Delta \mathbf{P}_q(\mathbf{r}) \mathrm{e}^{-\mathrm{i}\mathbf{k}_q \cdot \mathbf{r}}. \tag{4.12}$$

If, furthermore, the interaction is collinear along the z direction, all participating waves have parallel or antiparallel wavevectors such that $\mathbf{k}_q = k_q \hat{z}$ for all q. In this situation,

$\Delta\mathbf{P}_q$ can have variations only along the z direction. Then (4.12) can be further simplified to

$$\frac{\mathrm{d}\boldsymbol{\mathcal{E}}_q(z)}{\mathrm{d}z} \approx \frac{\mathrm{i}\omega_q^2\mu_0}{2k_q}\Delta\mathbf{P}_q(z)\mathrm{e}^{-\mathrm{i}k_q z}. \tag{4.13}$$

For an optical wave propagating in an anisotropic medium, $\mathbf{E}$ is not necessarily perpendicular to $\mathbf{k}$ and, in general, $\boldsymbol{\nabla}\cdot\mathbf{E} \neq 0$, as discussed in Section 1.6. Consequently, (4.8) and the equations that follow are not valid in an anisotropic medium. In this situation, the field $\mathbf{E}_q$ propagating in the $\mathbf{k}_q = k_q\hat{k}_q$ direction can be divided into a transverse and a longitudinal component:

$$\mathbf{E}_q = \mathbf{E}_{q,\mathrm{T}} + \mathbf{E}_{q,\mathrm{L}}, \tag{4.14}$$

where the transverse component is given by

$$\mathbf{E}_{q,\mathrm{T}} = (\hat{k}_q \times \mathbf{E}_q) \times \hat{k}_q \tag{4.15}$$

and the longitudinal component is given by

$$\mathbf{E}_{q,\mathrm{L}} = (\hat{k}_q \cdot \mathbf{E}_q)\hat{k}_q. \tag{4.16}$$

Clearly, $\boldsymbol{\nabla}\cdot\mathbf{E}_{q,\mathrm{T}} = 0$ but $\boldsymbol{\nabla}\cdot\mathbf{E}_{q,\mathrm{L}} \neq 0$. Therefore, an equation similar to (4.8) can be written for the transverse component (see Problem 4.1.2):

$$\nabla^2\mathbf{E}_{q,\mathrm{T}} + \omega_q^2\mu_0[\boldsymbol{\epsilon}(\mathbf{k}_q, \omega_q)\cdot\mathbf{E}_q]_{\mathrm{T}} = -\omega_q^2\mu_0\Delta\mathbf{P}_{q,\mathrm{T}}, \tag{4.17}$$

where $\Delta\mathbf{P}_{q,\mathrm{T}} = (\hat{k}_q \times \Delta\mathbf{P}_q) \times \hat{k}_q$. Note that $\Delta\mathbf{P}_{q,\mathrm{T}}$ can have contributions from the longitudinal field components of the interacting waves. Following the same procedure as leads to (4.11), the coupled-wave equation in an anisotropic medium under the slowing varying amplitude approximation can be written as (see Problem 4.1.3)

$$(\mathbf{k}_q\cdot\boldsymbol{\nabla})\boldsymbol{\mathcal{E}}_{q,\mathrm{T}} \approx \frac{\mathrm{i}\omega_q^2\mu_0}{2}\Delta\mathbf{P}_{q,\mathrm{T}}\mathrm{e}^{-\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}}. \tag{4.18}$$

In the special situation when (4.11) can be reduced to (4.12) or (4.13), an equation similar to (4.12) or (4.13), but expressed in terms of the transverse field components, can be obtained from (4.18) for wave coupling in an anisotropic medium.

4.2 Coupled-mode theory

Coupled-mode theory deals with the coupling of spatial modes of different spatial distributions or different polarizations, or both. Although the theory described in this section is formulated specifically in terms of the coupling of waveguide modes, it can be easily extended to other kind of spatial modes, such as Gaussian spatial modes.

The mode fields in a lossless waveguide can be expressed in the forms of (2.1) and (2.2), which satisfy Maxwell's equations in (2.8) and (2.9). For fields at a single

frequency ω, we can write

$$\mathbf{E}_\nu(\mathbf{r}) = \boldsymbol{\mathcal{E}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.19}$$

$$\mathbf{H}_\nu(\mathbf{r}) = \boldsymbol{\mathcal{H}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.20}$$

for the spatial dependence of the mode fields, while the two Maxwell's equations in (2.8) and (2.9) become

$$\boldsymbol{\nabla} \times \mathbf{E} = \mathrm{i}\omega\mu_0\mathbf{H}, \tag{4.21}$$

$$\boldsymbol{\nabla} \times \mathbf{H} = -\mathrm{i}\omega\epsilon\mathbf{E}. \tag{4.22}$$

The normal modes with fields given by (4.19) and (4.20) are characteristic solutions of Maxwell's equations in (4.21) and (4.22).

Mode expansion

The normal modes are orthogonal and can be normalized to have the orthonormality relation given by (2.41). They form a basis for linear expansion of any optical field at a given frequency ω in the waveguide:

$$\mathbf{E}(\mathbf{r}) = \sum_\nu A_\nu \hat{\boldsymbol{\mathcal{E}}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.23}$$

$$\mathbf{H}(\mathbf{r}) = \sum_\nu A_\nu \hat{\boldsymbol{\mathcal{H}}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.24}$$

where $\hat{\boldsymbol{\mathcal{E}}}_\nu$ and $\hat{\boldsymbol{\mathcal{H}}}_\nu$ are normalized mode fields satisfying (2.41), and the summation sums over all discrete indices of the guided modes and integrates over all continuous indices of the radiation and evanescent modes. In an ideal waveguide where these modes are defined, the normal modes do not couple. Then, the expansion coefficients A_ν are constants that are independent of x, y, and z.

When there is a spatially dependent perturbation to a waveguide, the modes defined by the unperturbed ideal waveguide are no longer exact normal modes of the perturbed waveguide. They can now be coupled by the perturbation as they propagate along the waveguide. As a result, if the fields are still expanded in terms of the normal modes of the unperturbed waveguide, the expansion coefficients are no longer constants of propagation but vary with z as the fields propagate down the waveguide:

$$\mathbf{E}(\mathbf{r}) = \sum_\nu A_\nu(z)\hat{\boldsymbol{\mathcal{E}}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.25}$$

$$\mathbf{H}(\mathbf{r}) = \sum_\nu A_\nu(z)\hat{\boldsymbol{\mathcal{H}}}_\nu(x, y)\exp(\mathrm{i}\beta_\nu z), \tag{4.26}$$

where again $\hat{\boldsymbol{\mathcal{E}}}_\nu$ and $\hat{\boldsymbol{\mathcal{H}}}_\nu$ are normalized mode fields, and the summation is taken over all guided, radiation, and evanescent modes.

Single-waveguide mode coupling

We first consider the coupling between normal modes in a single waveguide that is subject to some perturbation. The spatially dependent perturbation to the waveguide can be represented by a perturbing polarization $\Delta\mathbf{P}(\mathbf{r})$ also at frequency ω. The following Maxwell's equations then replace (4.21) and (4.22):

$$\boldsymbol{\nabla} \times \mathbf{E} = \mathrm{i}\omega\mu_0\mathbf{H}, \tag{4.27}$$

$$\boldsymbol{\nabla} \times \mathbf{H} = -\mathrm{i}\omega\epsilon\mathbf{E} - \mathrm{i}\omega\Delta\mathbf{P}. \tag{4.28}$$

The fields in the perturbed waveguide, which can be expanded as (4.25) and (4.26), are governed by these two equations with $\Delta\mathbf{P} \neq 0$. Meanwhile, the normal mode fields of the unperturbed waveguide, which are defined by (4.21) and (4.22), also satisfy these two equations with $\Delta\mathbf{P} = 0$. Using (4.27) and (4.28), we have

$$\boldsymbol{\nabla} \cdot (\mathbf{E}_1 \times \mathbf{H}_2^* + \mathbf{E}_2^* \times \mathbf{H}_1) = -\mathrm{i}\omega(\mathbf{E}_1 \cdot \Delta\mathbf{P}_2^* - \mathbf{E}_2^* \cdot \Delta\mathbf{P}_1). \tag{4.29}$$

This is the *Lorentz reciprocity theorem*, which holds for any two arbitrary sets of fields $(\mathbf{E}_1, \mathbf{H}_1)$ and $(\mathbf{E}_2, \mathbf{H}_2)$.

If we take $(\mathbf{E}_1, \mathbf{H}_1)$ to be those of (4.25) and (4.26) and $(\mathbf{E}_2, \mathbf{H}_2)$ to be the normal mode fields given in (4.19) and (4.20), we have $\Delta\mathbf{P}_1 = \Delta\mathbf{P}$ and $\Delta\mathbf{P}_2 = 0$. Substituting these into (4.29) and integrating both sides of the resultant equation over the cross section of the waveguide, we have

$$\sum_\nu \frac{\mathrm{d}}{\mathrm{d}z} A_\nu(z)\mathrm{e}^{\mathrm{i}(\beta_\nu - \beta_\mu)z} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\hat{\boldsymbol{\mathcal{E}}}_\nu \times \hat{\boldsymbol{\mathcal{H}}}_\mu^* + \hat{\boldsymbol{\mathcal{E}}}_\mu^* \times \hat{\boldsymbol{\mathcal{H}}}_\nu) \cdot \hat{z}\,\mathrm{d}x\mathrm{d}y$$

$$= \mathrm{i}\omega\mathrm{e}^{-\mathrm{i}\beta_\mu z} \int_{-\infty}^{\infty}\int_{\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\mu^* \cdot \Delta\mathbf{P}\,\mathrm{d}x\mathrm{d}y. \tag{4.30}$$

By applying the orthonormality relation (2.41), we find from (4.30) the following *coupled-mode equation* (see Problem 4.2.1):

$$\pm\frac{\mathrm{d}A_\nu}{\mathrm{d}z} = \mathrm{i}\omega\mathrm{e}^{-\mathrm{i}\beta_\nu z} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\nu^* \cdot \Delta\mathbf{P}\,\mathrm{d}x\mathrm{d}y, \tag{4.31}$$

where the plus sign is used when $\beta_\nu > 0$ and mode ν is a forward-propagating mode, and the minus sign is used when $\beta_\nu < 0$ and mode ν is a backward-propagating mode.

The result in (4.31) can be used for mode coupling caused by any kind of spatially dependent perturbation on the characteristics of the waveguide. For example, $\Delta\mathbf{P}$ can be a perturbing polarization due to the effects of nonlinear optical interactions on the fields at frequency ω in the waveguide. For the simple case where the perturbation can

be represented by a change in linear polarization as

$$\Delta \mathbf{P} = \Delta \epsilon \mathbf{E} = \Delta \epsilon \sum_{\nu} A_{\nu} \hat{\boldsymbol{\mathcal{E}}}_{\nu} e^{i\beta_{\nu} z}, \tag{4.32}$$

we have

$$\pm \frac{dA_{\nu}}{dz} = \sum_{\mu} i\kappa_{\nu\mu} A_{\mu} e^{i(\beta_{\mu} - \beta_{\nu})z}, \tag{4.33}$$

where

$$\kappa_{\nu\mu} = \omega \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Delta \epsilon \hat{\boldsymbol{\mathcal{E}}}_{\nu}^{*} \cdot \hat{\boldsymbol{\mathcal{E}}}_{\mu} dx dy \tag{4.34}$$

is the *coupling coefficient* between mode ν and mode μ. This result can also be extended to anisotropic waveguides by simply considering $\Delta \mathbf{P}$ to be a polarization involving anisotropy as

$$\Delta \mathbf{P} = \Delta \boldsymbol{\epsilon} \cdot \mathbf{E}, \tag{4.35}$$

where $\Delta \boldsymbol{\epsilon}$ is a tensor. In this situation, the coupled-mode equation is still given by (4.33), but the coupling coefficient is given by

$$\kappa_{\nu\mu} = \omega \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_{\nu}^{*} \cdot \Delta \boldsymbol{\epsilon} \cdot \hat{\boldsymbol{\mathcal{E}}}_{\mu} dx dy. \tag{4.36}$$

In a lossless waveguide, the dielectric tensor is a Hermitian matrix, as discussed in Section 1.6. Therefore, $\Delta \epsilon_{ij} = \Delta \epsilon_{ji}^{*}$ and

$$\kappa_{\nu\mu} = \kappa_{\mu\nu}^{*} \tag{4.37}$$

in a lossless waveguide.

Multiple-waveguide mode coupling

In an optical structure that consists of more than one waveguide, we can certainly solve Maxwell's equations directly with the boundary conditions defined by the entire structure to find its normal modes. Alternatively, we can divide the structure into separate individual waveguides, expand the fields in terms of the normal modes of the individual waveguides, and treat the problem with a coupled-mode approach. The first approach can yield exact solutions and is sometimes desirable. However, it is not generally possible to obtain the exact solutions for complicated structures. The coupled-mode approach yields approximate solutions, but it can be applied to most structures without difficulty. In addition, it gives an intuitive picture of how optical waves interact in a multiple-waveguide structure. In the following, we consider the coupled-mode formulation for multiple parallel waveguides.

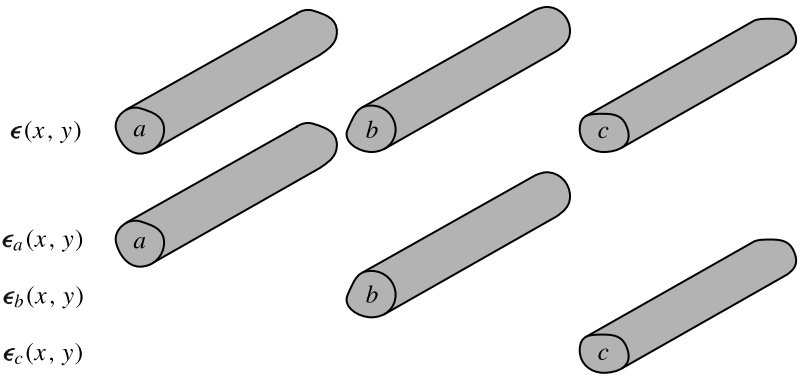


Figure 4.1 Representation of a multiple-waveguide structure in terms of a combination of individual single waveguides.

The concept of dividing a multiple-waveguide structure into a combination of individual single waveguides is illustrated in Fig. 4.1. The multiple-waveguide structure is described by $\epsilon(x, y)$ while the individual waveguides are described by $\epsilon_a(x, y)$, $\epsilon_b(x, y)$, $\epsilon_c(x, y)$, and so on. The normal modes are solved for each individual waveguide. The fields in the entire structure can be expanded in terms of these normal modes in the same form as that of (4.25) and (4.26) but with the index ν representing modes of different individual waveguides. From the standpoint of any individual waveguide ν, the entire structure looks like $\epsilon_\nu(x, y)$ plus a perturbation of

$$\Delta\epsilon_\nu(x, y) = \epsilon(x, y) - \epsilon_\nu(x, y). \tag{4.38}$$

This concept is schematically illustrated in Fig. 4.2. The coupled-mode equation for the multiple-waveguide structure can be obtained by using the reciprocity theorem of (4.29) and then following a procedure similar to that taken above to obtain the coupled-mode equation for the single waveguide. Because the mathematics is quite involved, we only give the results in the following without detailed derivation.

The coupled-mode equation for a multiple-waveguide structure can still be written in the same form as that of (4.33):

$$\pm\frac{\mathrm{d}A_\nu}{\mathrm{d}z} = \sum_\mu \mathrm{i}\kappa_{\nu\mu}A_\mu \mathrm{e}^{\mathrm{i}(\beta_\mu - \beta_\nu)z}, \tag{4.39}$$

where the plus sign is taken if mode ν is forward propagating, and the minus sign is used if it is backward propagating. It is noted that the summation over the index μ runs through the modes of every individual waveguide, not just the modes of one single waveguide. In addition, the coupling coefficients $\kappa_{\nu\mu}$ have a complicated form and are best expressed in terms of the matrix elements:

$$\kappa_{\nu\mu} = c_{\nu\nu}[\mathbf{c}^{-1} \cdot \tilde{\boldsymbol{\kappa}}]_{\nu\mu}, \tag{4.40}$$

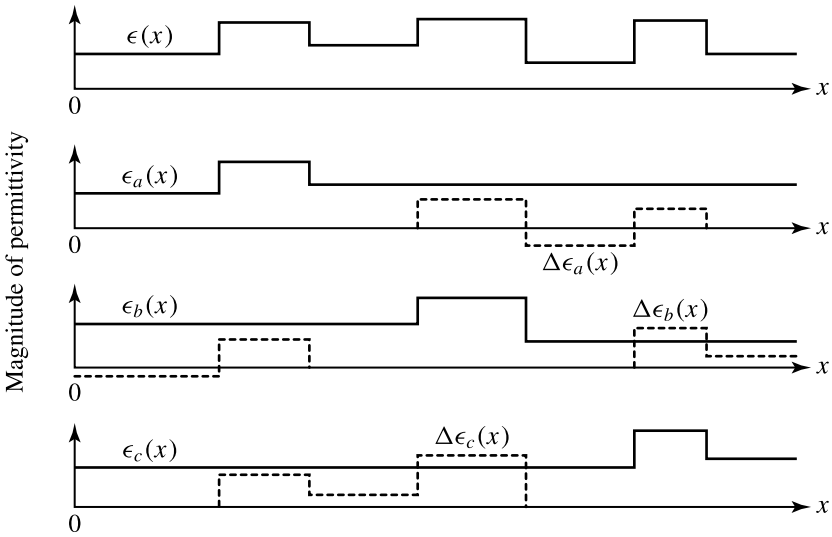


Figure 4.2 Schematic diagram of three coupled waveguides showing the decomposition into individual waveguides, in solid curves, plus the corresponding perturbation, in dashed curves, for each of them.

where $c_{\nu\nu} = 1$ if mode ν is forward propagating and $c_{\nu\nu} = -1$ if it is backward propagating, as can be seen from (4.41) below. The elements of the matrices $\mathbf{c} = [c_{\nu\mu}]$ and $\tilde{\boldsymbol{\kappa}} = [\tilde{\kappa}_{\nu\mu}]$ are given by

$$c_{\nu\mu} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\hat{\boldsymbol{\mathcal{E}}}_{\nu}^{*} \times \hat{\boldsymbol{\mathcal{H}}}_{\mu} + \hat{\boldsymbol{\mathcal{E}}}_{\mu} \times \hat{\boldsymbol{\mathcal{H}}}_{\nu}^{*}\right) \cdot \hat{z}\,\mathrm{d}x\mathrm{d}y = c_{\mu\nu}^{*} \tag{4.41}$$

and

$$\tilde{\kappa}_{\nu\mu} = \omega \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_{\nu}^{*} \cdot \Delta\boldsymbol{\epsilon}_{\mu} \cdot \hat{\boldsymbol{\mathcal{E}}}_{\mu}\mathrm{d}x\mathrm{d}y, \tag{4.42}$$

respectively. The coefficient $c_{\nu\mu}$ represents the *overlap coefficient* of $(\hat{\boldsymbol{\mathcal{E}}}_{\nu}, \hat{\boldsymbol{\mathcal{H}}}_{\nu})$ and $(\hat{\boldsymbol{\mathcal{E}}}_{\mu}, \hat{\boldsymbol{\mathcal{H}}}_{\mu})$, which are the mode fields of different individual waveguides. Note that $c_{\nu\mu} \neq 0$ in general because modes of different waveguides are not necessarily orthogonal to each other. Because the mode fields used in (4.41) are normalized, we have $c_{\nu\nu} = 1$ or -1, depending on whether the mode ν is forward or backward propagating as mentioned above, and $|c_{\nu\mu}| \leq 1$. Note also the difference between the form of $\tilde{\kappa}_{\nu\mu}$ and that of the coupling coefficients of modes in a single waveguide given by (4.36).

In general,

$$\tilde{\kappa}_{\nu\mu} \neq \tilde{\kappa}_{\mu\nu}^{*} \qquad \text{and} \qquad \kappa_{\nu\mu} \neq \kappa_{\mu\nu}^{*} \tag{4.43}$$

where ν and μ refer to modes of two different waveguides. Therefore, (4.37) is not always valid for coupling between waveguides. Indeed, it can be shown by using the reciprocity theorem that

$$\tilde{\kappa}_{\nu\mu} - \tilde{\kappa}^*_{\mu\nu} = \frac{c_{\nu\mu} + c^*_{\mu\nu}}{2}(\beta_\nu - \beta_\mu) = c_{\nu\mu}(\beta_\nu - \beta_\mu). \tag{4.44}$$

It can be seen that there is a direct relationship between the coupling coefficients and the propagation constants. This is an important relation. It has the following implications.

1. Coupling between two modes is not symmetric, $\kappa_{\nu\mu} \neq \kappa^*_{\mu\nu}$, unless $\beta_\nu = \beta_\mu$ or $c_{\nu\mu} = c^*_{\mu\nu} = 0$. This is because the normal modes in different individual waveguides are not necessarily orthogonal to each other.
2. Coupling of modes of the same order between two identical waveguides is always symmetric, resulting in $\tilde{\kappa}_{\nu\mu} = \tilde{\kappa}^*_{\mu\nu}$ and $\kappa_{\nu\mu} = \kappa^*_{\mu\nu}$.
3. The relation in (4.44) applies to modes in a single waveguide as well. In this situation, $c_{\nu\mu} = c^*_{\mu\nu} = 0$ if $\nu \neq \mu$, but $\tilde{\kappa}_{\nu\mu} = \kappa_{\nu\mu}$. Therefore, (4.37) holds in a single waveguide because the normal modes in the same waveguide are always orthogonal to each other.
4. It is not possible to change the coupling between two modes without simultaneously changing their overlap coefficient or their propagation constants.

4.3 Two-mode coupling

In most applications, we are interested in the coupling between two modes. This includes coupling between two modes in the same waveguide, such as that in a *periodic waveguide*, or coupling between two parallel waveguides, such as that in a *directional coupler*. For coupling between two modes, the coupled-mode equations can be written in a simple form that can be solved analytically. In this section, we consider the general formulation and general solutions for this important case of two-mode coupling. The characteristics of specific couplers are discussed in Chapter 5.

We have shown that both coupling among modes in the same waveguide and coupling among multiple waveguides can be described by coupled-mode equations of the same form as given in (4.33) and (4.39). The only difference is that the coupling coefficients in (4.39) for multiple-waveguide coupling are defined differently from those in (4.33) for single-waveguide mode coupling. This is convenient because general solutions of the coupled-mode equations can be applied to both cases. For a particular problem, we only have to calculate the coupling coefficients specific to the problem under consideration.

For two-mode coupling either in a single waveguide or between two separate waveguides, the field expansion in (4.25) and (4.26) consists of only two modes with amplitudes A and B. Thus, coupled-mode equations of the form given in (4.33) or (4.39)

simply reduce to the following two coupled equations:

$$\pm\frac{\mathrm{d}A}{\mathrm{d}z} = \mathrm{i}\kappa_{aa}A + \mathrm{i}\kappa_{ab}B\mathrm{e}^{\mathrm{i}(\beta_b-\beta_a)z}, \tag{4.45}$$

$$\pm\frac{\mathrm{d}B}{\mathrm{d}z} = \mathrm{i}\kappa_{bb}B + \mathrm{i}\kappa_{ba}A\mathrm{e}^{\mathrm{i}(\beta_a-\beta_b)z}. \tag{4.46}$$

For coupling in a single waveguide, the coupling coefficients in these equations are simply given by (4.34) in the case of an isotropic waveguide or by (4.36) in the case of an anisotropic waveguide. According to (4.37), we also have $\kappa_{ab} = \kappa_{ba}^*$ if the waveguide is lossless. For coupling between two waveguides, the coupling coefficients are given by (4.40), which can be expressed explicitly as

$$\begin{aligned}
\kappa_{aa} &= \frac{\tilde{\kappa}_{aa} - c_{ab}\tilde{\kappa}_{ba}/c_{bb}}{1 - c_{ab}c_{ba}/c_{aa}c_{bb}},\ \kappa_{ab} = \frac{\tilde{\kappa}_{ab} - c_{ab}\tilde{\kappa}_{bb}/c_{bb}}{1 - c_{ab}c_{ba}/c_{aa}c_{bb}},\\
\kappa_{ba} &= \frac{\tilde{\kappa}_{ba} - c_{ba}\tilde{\kappa}_{aa}/c_{aa}}{1 - c_{ab}c_{ba}/c_{aa}c_{bb}},\ \kappa_{bb} = \frac{\tilde{\kappa}_{bb} - c_{ba}\tilde{\kappa}_{ab}/c_{aa}}{1 - c_{ab}c_{ba}/c_{aa}c_{bb}}.
\end{aligned} \tag{4.47}$$

As discussed earlier and expressed in (4.43) and (4.44), in general, $\kappa_{ab} \neq \kappa_{ba}^*$ for coupling between two waveguides.

There is a self-coupling term in each of the coupled equations (4.45) and (4.46). These terms are caused by the fact that normal modes see an index profile in the perturbed waveguide different from that of the original waveguide where the modes are defined. They can be removed from these equations by expressing the normal-mode expansion coefficients as follows:

$$A(z) = \tilde{A}(z)\exp\left[\pm\mathrm{i}\int_0^z \kappa_{aa}(z)\mathrm{d}z\right], \tag{4.48}$$

$$B(z) = \tilde{B}(z)\exp\left[\pm\mathrm{i}\int_0^z \kappa_{bb}(z)\mathrm{d}z\right]. \tag{4.49}$$

where a plus or minus sign is chosen for a forward-propagating or backward-propagating mode, respectively.

Before transforming (4.45) and (4.46) into two coupled equations in terms of $\tilde{A}$ and $\tilde{B}$ to remove the self-coupling terms, we have to consider the fact that all of the coupling coefficients can be a function of z because $\Delta\epsilon$ can be a function of z but the integration in (4.36) and (4.42) is carried out only over x and y. In case $\kappa_{ab}(z)$ and $\kappa_{ba}(z)$ are arbitrary functions of z, the coupled-mode equations in (4.45) and (4.46) cannot be solved analytically. In this situation, there is no need to simplify the coupled-mode equations further because they can only be solved numerically. However, for waveguide structures of practical interest that are designed for two-mode coupling, $\Delta\epsilon$ is either independent of z or is a periodic function of z. Then, the coupling coefficients are either constant or periodic in z. In either case, (4.45) and (4.46) can be reduced to the

following general form:

$$\pm\frac{\mathrm{d}\tilde{A}}{\mathrm{d}z} = \mathrm{i}\kappa_{ab}\tilde{B}\mathrm{e}^{\mathrm{i}2\delta z}, \tag{4.50}$$

$$\pm\frac{\mathrm{d}\tilde{B}}{\mathrm{d}z} = \mathrm{i}\kappa_{ba}\tilde{A}\mathrm{e}^{-\mathrm{i}2\delta z}, \tag{4.51}$$

in terms of $\tilde{A}$ and $\tilde{B}$ with κ_{ab} and κ_{ba} in these two equations being constants independent of z. The parameter 2δ is the *phase mismatch* between the two modes being coupled. Phase-matched coupling with $\delta = 0$ between two modes is always symmetric with $\kappa_{ab} = \kappa_{ba}^*$ irrespective of whether these two modes belong to the same waveguide or two different waveguides (see Problem 4.3.1).

The general form of (4.50) and (4.51) applies to both cases of constant and periodic perturbations, but the details of the parameters in these two equations vary.

1. **Constant perturbation.** In this case, $\Delta\epsilon$ is not a function of z. Then all of the coupling coefficients κ_{aa}, κ_{bb}, κ_{ab}, and κ_{ba} are constants that are independent of z. We then find that

   $$A(z) = \tilde{A}(z)\mathrm{e}^{\pm\mathrm{i}\kappa_{aa}z} \quad \text{and} \quad B(z) = \tilde{B}(z)\mathrm{e}^{\pm\mathrm{i}\kappa_{bb}z} \tag{4.52}$$

   and

   $$2\delta = (\beta_b \pm \kappa_{bb}) - (\beta_a \pm \kappa_{aa}). \tag{4.53}$$

   The choice of sign in each $\pm$ here is consistent with that in (4.48) and (4.49) discussed above. The physical meaning of the self-coupling coefficients is a change in the propagation constant of each normal mode. While the propagation constants of the normal modes in the original waveguide are β_a and β_b, their values are changed because of the perturbation on the waveguide. These modes now propagate with the modified propagation constants $\beta_a \pm \kappa_{aa}$ and $\beta_b \pm \kappa_{bb}$, respectively, which take into account the effect of the perturbation. In addition, they couple to each other through κ_{ab} and κ_{ba}. Details of this type of coupling are discussed in Section 5.2.
2. **Periodic perturbation.** In this case, $\Delta\epsilon$ is a periodic function of z and so are the coupling coefficients $\kappa_{aa}(z)$, $\kappa_{bb}(z)$, $\kappa_{ab}(z)$, and $\kappa_{ba}(z)$. The periodic perturbation has a period Λ and a wavenumber

   $$K = \frac{2\pi}{\Lambda}. \tag{4.54}$$

   The coupling coefficients $\kappa_{ab}(z)$ and $\kappa_{ba}(z)$, being periodic in z with a periodicity Λ, can be expanded in a Fourier series with constant coefficients $\kappa_{ab}(q)$ and $\kappa_{ba}(q)$ and a phase factor qK, where q is an integer. Because $\kappa_{aa}(z)$ and $\kappa_{bb}(z)$ are periodic in z, we find that

$$\left|\int_0^z \kappa_{aa}(z)\mathrm{d}z\right| \ll Kz \quad \text{and} \quad \left|\int_0^z \kappa_{bb}(z)\mathrm{d}z\right| \ll Kz. \tag{4.55}$$

Therefore, the contribution to the phase-mismatch parameter 2δ by κ_{aa} and κ_{bb} is negligible compared to the contribution by qK. As a result, we find that the coupled-mode equations in the case of periodic perturbation can also be expressed in the form of (4.50) and (4.51) but with $\kappa_{ab} = \kappa_{ab}(q)$ and $\kappa_{ba} = \kappa_{ba}(q)$ being constants that are independent of z and

$$2\delta = \Delta\beta + qK = \beta_b - \beta_a + qK, \tag{4.56}$$

where $\Delta\beta = \beta_b - \beta_a$ and q is an integer that minimizes the value of δ. Details of this type of coupling are discussed in Section 5.1.

With these general considerations, (4.50) and (4.51) represent the most general coupled equations for two-mode coupling in waveguide structures of practical interest. They can be solved analytically and their solutions apply to many different two-mode coupling problems.

Codirectional coupling

First, we consider the coupling of two modes propagating in the same direction, say the forward direction in z, over a length l, as is shown in Fig. 4.3. In this case, $\beta_a > 0$ and $\beta_b > 0$. The coupled equations are

$$\frac{\mathrm{d}\tilde{A}}{\mathrm{d}z} = \mathrm{i}\kappa_{ab}\tilde{B}\mathrm{e}^{\mathrm{i}2\delta z}, \tag{4.57}$$

$$\frac{\mathrm{d}\tilde{B}}{\mathrm{d}z} = \mathrm{i}\kappa_{ba}\tilde{A}\mathrm{e}^{-\mathrm{i}2\delta z}. \tag{4.58}$$

These equations for codirectional coupling are generally solved as an initial-value problem with the initial values of $\tilde{A}(z_0)$ and $\tilde{B}(z_0)$ given at $z = z_0$ to find the values

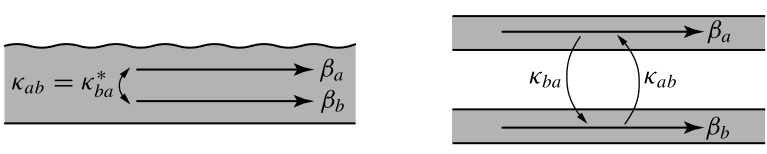


Figure 4.3 Codirectional coupling between two modes of propagation constants β_a and β_b (*a*) in the same waveguide and (*b*) in two parallel waveguides. A perturbation is required for codirectional coupling in the same waveguide but is not required for codirectional coupling between two waveguides.

of $\tilde{A}(z)$ and $\tilde{B}(z)$ at any other location z. The general solution can be expressed in the following matrix form:

$$\begin{bmatrix} \tilde{A}(z) \\ \tilde{B}(z) \end{bmatrix} = \mathbf{F}(z; z_0) \begin{bmatrix} \tilde{A}(z_0) \\ \tilde{B}(z_0) \end{bmatrix}, \tag{4.59}$$

where the *forward-coupling matrix* $\mathbf{F}(z; z_0)$ relates the field amplitudes at the location z_0 to those at the location z. It has the form (see Problem 4.3.2)

$$\mathbf{F}(z; z_0) = \begin{bmatrix} \dfrac{\beta_c \cos \beta_c(z - z_0) - i\delta \sin \beta_c(z - z_0)}{\beta_c} e^{i\delta(z-z_0)} & \dfrac{i\kappa_{ab}}{\beta_c} \sin \beta_c(z - z_0) e^{i\delta(z+z_0)} \\ \dfrac{i\kappa_{ba}}{\beta_c} \sin \beta_c(z - z_0) e^{-i\delta(z+z_0)} & \dfrac{\beta_c \cos \beta_c(z - z_0) + i\delta \sin \beta_c(z - z_0)}{\beta_c} e^{-i\delta(z-z_0)} \end{bmatrix}, \tag{4.60}$$

where

$$\beta_c = \left(\kappa_{ab}\kappa_{ba} + \delta^2\right)^{1/2}. \tag{4.61}$$

We consider a simple case when power is launched only into mode a at $z = 0$. Then the initial values are $\tilde{A}(0) \neq 0$ and $\tilde{B}(0) = 0$. By applying these conditions to (4.59) and taking $z_0 = 0$ in (4.60), we find that

$$\tilde{A}(z) = \tilde{A}(0) \left(\cos \beta_c z - \frac{i\delta}{\beta_c} \sin \beta_c z\right) e^{i\delta z}, \tag{4.62}$$

$$\tilde{B}(z) = \tilde{A}(0) \left(\frac{i\kappa_{ba}}{\beta_c} \sin \beta_c z\right) e^{-i\delta z}. \tag{4.63}$$

The power in the two modes varies with z as follows:

$$\frac{P_a(z)}{P_a(0)} = \left|\frac{A(z)}{A(0)}\right|^2 = \left|\frac{\tilde{A}(z)}{\tilde{A}(0)}\right|^2 = \frac{\kappa_{ab}\kappa_{ba}}{\beta_c^2} \cos^2 \beta_c z + \frac{\delta^2}{\beta_c^2}, \tag{4.64}$$

$$\frac{P_b(z)}{P_a(0)} = \left|\frac{B(z)}{A(0)}\right|^2 = \left|\frac{\tilde{B}(z)}{\tilde{A}(0)}\right|^2 = \frac{|\kappa_{ba}|^2}{\beta_c^2} \sin^2 \beta_c z. \tag{4.65}$$

The *coupling efficiency* for a length l is

$$\eta = \frac{P_b(l)}{P_a(0)} = \frac{|\kappa_{ba}|^2}{\beta_c^2} \sin^2 \beta_c l. \tag{4.66}$$

Thus, power is exchanged periodically between two modes with a *coupling length*

$$l_c = \frac{\pi}{2\beta_c}, \tag{4.67}$$

where maximum power transfer occurs. Figure 4.4 shows the periodic power exchange between the two coupled modes as a function of z. As can be seen from Fig. 4.4, complete power transfer can occur only in the *phase-matched condition* when $\delta = 0$.

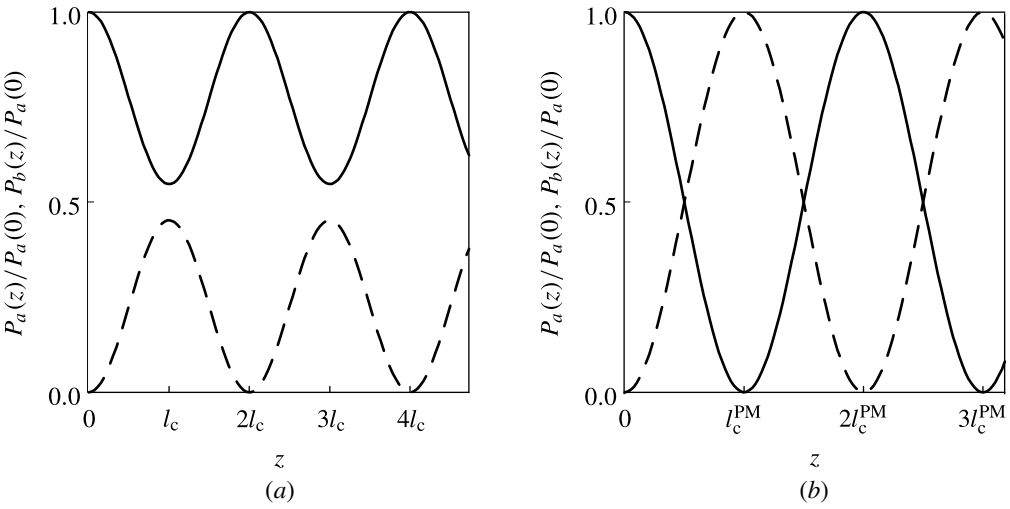


Figure 4.4 Periodic power exchange between two codirectionally coupled modes for (*a*) phase-mismatched condition $\delta \neq 0$ and (*b*) phase-matched condition $\delta = 0$. The solid curves represent $P_a(z)/P_a(0)$, and the dashed curves represent $P_b(z)/P_a(0)$.

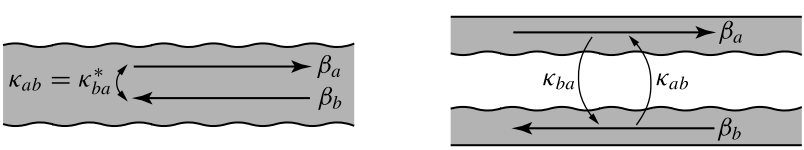
Contradirectional coupling

We now consider the coupling of two modes propagating in opposite directions over a length l, as is shown in Fig. 4.5 where mode a is forward propagating and mode b is backward propagating. In this case, $\beta_a > 0$ and $\beta_b < 0$. Thus, the coupled equations are

$$\frac{\mathrm{d}\tilde{A}}{\mathrm{d}z} = \mathrm{i}\kappa_{ab}\tilde{B}\mathrm{e}^{\mathrm{i}2\delta z}, \tag{4.68}$$

$$-\frac{\mathrm{d}\tilde{B}}{\mathrm{d}z} = \mathrm{i}\kappa_{ba}\tilde{A}\mathrm{e}^{-\mathrm{i}2\delta z}. \tag{4.69}$$

These equations for contradirectional coupling are generally solved as a boundary-value problem with the boundary values of $\tilde{A}(0)$ at one end and $\tilde{B}(l)$ at the other end to find the values of $\tilde{A}(z)$ and $\tilde{B}(z)$ at any location z between the two ends. The general



(*a*) (*b*)

Figure 4.5 Contradirectional coupling between two modes of propagation constants β_a and β_b (*a*) in the same waveguide and (*b*) in two parallel waveguides. A significant perturbation is required for contradirectional coupling in either case.

solution can be expressed in the following matrix form:

$$\begin{bmatrix} \tilde{A}(z) \\ \tilde{B}(z) \end{bmatrix} = \mathbf{R}(z;0,l) \begin{bmatrix} \tilde{A}(0) \\ \tilde{B}(l) \end{bmatrix}, \tag{4.70}$$

where the *reverse-coupling matrix* $\mathbf{R}(z;0,l)$ relates the field amplitudes $\tilde{A}(0)$ at $z=0$ and $\tilde{B}(l)$ at $z=l$ to those at location z. It has the following form (see Problem 4.3.4):

$$\mathbf{R}(z;0,l) = \begin{bmatrix} \dfrac{\alpha_c \cosh\alpha_c(l-z) + \mathrm{i}\delta \sinh\alpha_c(l-z)}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{\mathrm{i}\delta z} & \dfrac{\mathrm{i}\kappa_{ab} \sinh\alpha_c z}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{\mathrm{i}\delta(z+l)} \\ \dfrac{\mathrm{i}\kappa_{ba} \sinh\alpha_c(l-z)}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{-\mathrm{i}\delta z} & \dfrac{\alpha_c \cosh\alpha_c z + \mathrm{i}\delta \sinh\alpha_c z}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{-\mathrm{i}\delta(z-l)} \end{bmatrix} \tag{4.71}$$

where

$$\alpha_c = \left(\kappa_{ab}\kappa_{ba} - \delta^2\right)^{1/2}. \tag{4.72}$$

We consider a simple case when power is launched only into mode a at $z=0$. Then the boundary values are $\tilde{A}(0) \neq 0$ and $\tilde{B}(l) = 0$. By applying these conditions to (4.70), we find that

$$\tilde{A}(z) = \tilde{A}(0) \frac{\alpha_c \cosh\alpha_c(l-z) + \mathrm{i}\delta \sinh\alpha_c(l-z)}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{\mathrm{i}\delta z}, \tag{4.73}$$

$$\tilde{B}(z) = \tilde{A}(0) \frac{\mathrm{i}\kappa_{ba} \sinh\alpha_c(l-z)}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l} \mathrm{e}^{-\mathrm{i}\delta z}. \tag{4.74}$$

The power in the two contradirectionally coupled modes varies with z as follows:

$$\frac{P_a(z)}{P_a(0)} = \left|\frac{A(z)}{A(0)}\right|^2 = \left|\frac{\tilde{A}(z)}{\tilde{A}(0)}\right|^2 = \frac{\cosh^2\alpha_c(l-z) - \delta^2/\kappa_{ab}\kappa_{ba}}{\cosh^2\alpha_c l - \delta^2/\kappa_{ab}\kappa_{ba}}, \tag{4.75}$$

$$\frac{P_b(z)}{P_a(0)} = \left|\frac{B(z)}{A(0)}\right|^2 = \left|\frac{\tilde{B}(z)}{\tilde{A}(0)}\right|^2 = \frac{\kappa_{ba}^*}{\kappa_{ab}} \frac{\sinh^2\alpha_c(l-z)}{\cosh^2\alpha_c l - \delta^2/\kappa_{ab}\kappa_{ba}}. \tag{4.76}$$

Because mode b is propagating backward with no input at $z=l$ but an output at $z=0$, the coupling efficiency for a length l is

$$\eta = \frac{P_b(0)}{P_a(0)} = \frac{\kappa_{ba}^*}{\kappa_{ab}} \frac{\sinh^2\alpha_c l}{\cosh^2\alpha_c l - \delta^2/\kappa_{ab}\kappa_{ba}}. \tag{4.77}$$

Figure 4.6 shows the power exchange between the two contradirectionally coupled modes as a function of z. As can be seen from Fig. 4.6, complete power transfer occurs as $l \to \infty$ if $\delta^2 < \kappa_{ab}\kappa_{ba}$.

In the case when $\tilde{A}(0) \neq 0$ and $\tilde{B}(l) = 0$, as considered above, contradirectional coupling can be viewed as reflection of the field amplitude $\tilde{A}(0)$ at $z=0$ with a reflection coefficient

$$r = |r|\mathrm{e}^{\mathrm{i}\varphi_{\mathrm{DBR}}} = \frac{\tilde{B}(0)}{\tilde{A}(0)} = \frac{\mathrm{i}\kappa_{ba} \sinh\alpha_c l}{\alpha_c \cosh\alpha_c l + \mathrm{i}\delta \sinh\alpha_c l}. \tag{4.78}$$

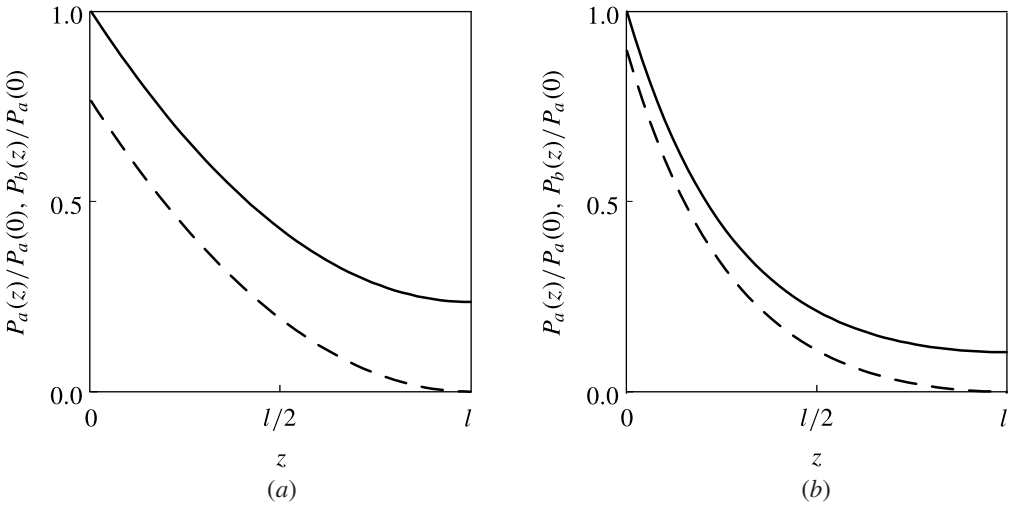


Figure 4.6 Power exchange between two contradirectionally coupled modes for (*a*) phase-mismatched condition $\delta \neq 0$ and (*b*) phase-matched condition $\delta = 0$. The solid curves represent $P_a(z)/P_a(0)$ and the dashed curves represent $P_b(z)/P_a(0)$.

The reflectivity is $R = |r|^2 = \eta$ as is given in (4.77). The phase shift is (see Problem 4.3.7)

$$\varphi_{\mathrm{DBR}} = \varphi_{\mathrm{B}} - \tan^{-1}\left(\frac{\delta}{\alpha_{\mathrm{c}}}\tanh\alpha_{\mathrm{c}} l\right). \tag{4.79}$$

Conservation of power

Conservation of power requires that in a lossless waveguide structure the net power flowing across any cross section of the waveguide be a constant independent of the longitudinal location of the cross section. For codirectional coupling with the power initially launched into only one mode so that $P_a(0) \neq 0$ but $P_b(0) = 0$, this requirement suggests that the sum of power in the two waveguides, $P_a(z) + P_b(z)$, is a constant because the power in the two modes flows in the same direction. For contradirectional coupling with the power launched into only one mode so that $P_a(0) \neq 0$ and $P_b(l) = 0$, this requirement suggests that $P_a(z) - P_b(z)$ is a constant because the power in mode b flows in the backward direction while that in mode a flows in the forward direction. These conclusions are correct for mode coupling in a single waveguide, but they do not generally hold for coupling between different waveguides.

It can be seen from (4.64) and (4.65) that $P_a(z) + P_b(z)$ is not a constant for codirectional coupling unless $\kappa_{ab} = \kappa_{ba}^*$. Similarly, from (4.75) and (4.76), it is also found that $P_a(z) - P_b(z)$ is not a constant for contradirectional coupling when $\kappa_{ab} \neq \kappa_{ba}^*$. It seems that the total power is not conserved in a lossless waveguide structure in the case of

asymmetric coupling with $\kappa_{ab} \neq \kappa_{ba}^*$. A close examination reveals that because $c_{ab} \neq 0$ in this case of asymmetric coupling, the two modes being coupled are not orthogonal to each other. Therefore, the total power flow cannot be fully accounted for by gathering the power in each individual mode as if the modes were orthogonal to each other. Indeed, by taking the total electric field and the total magnetic field expanded as (4.25) and (4.26), respectively, for two modes to calculate the power of the entire structure, we find that the total power flow as a function of space is (see Problem 4.3.8(a))

$$\begin{aligned} P(z) &= c_{aa}|A(z)|^2 + c_{bb}|B(z)|^2 + 2\mathrm{Re}\left[c_{ab}A^*(z)B(z)\mathrm{e}^{\mathrm{i}\Delta\beta z}\right] \\ &= c_{aa}P_a(z) + c_{bb}P_b(z) + P_{ab}(z), \end{aligned} \tag{4.80}$$

where $P_{ab}(z) = 2\mathrm{Re}\left[c_{ab}A^*(z)B(z)\mathrm{e}^{\mathrm{i}\Delta\beta z}\right]$ can be considered as the power residing between the two nonorthogonal modes of the two different waveguides. As defined in the preceding section, $c_{\nu\nu} = 1$ if mode ν is forward propagating and $c_{\nu\nu} = -1$ if mode ν is backward propagating. It can be shown, using (4.62) and (4.63) for the case of codirectional coupling and using (4.73) and (4.74) for the case of contradirectional coupling, that $P(z)$ is a constant independent of z no matter whether $\kappa_{ab} = \kappa_{ba}^*$ or $\kappa_{ab} \neq \kappa_{ba}^*$ (see Problem 4.3.8(b)). Therefore, conservation of power holds as expected.

When $P_{ab}(z) = 0$, it can be shown simply by applying conservation of power that $\kappa_{ab} = \kappa_{ba}^*$; hence the coupling is symmetric (see Problem 4.3.8(c)). Conversely, if the coupling is symmetric, $P_{ab}(z)$ always vanishes even when mode a and mode b are not orthogonal to each other. Two conclusions can thus be made:

1. When $c_{ab} = 0$, mode a and mode b are orthogonal to each other. Then $P_{ab}(z) = 0$ and $\kappa_{ab} = \kappa_{ba}^*$ even when $\delta \neq 0$ so that the two waveguide modes are not phase matched.
2. When the two modes are phase matched, $\delta = 0$. In this case, $P_{ab}(z) = 0$ and $\kappa_{ab} = \kappa_{ba}^*$ even when mode a and mode b are not orthogonal to each other with $c_{ab} \neq 0$ (see Problem 4.3.8(d)).

Consequently, *coupling between two modes a and b is symmetric with* $\kappa_{ab} = \kappa_{ba}^*$ *if these two modes are orthogonal to each other or if they are phase matched.*

Phase matching

As can be seen from Figs. 4.4 and 4.6, power transfer is most efficient when $\delta = 0$. The parameter δ is a measure of *phase mismatch* between the two modes being coupled. For the simple case when $2\delta = \Delta\beta = \beta_b - \beta_a$, the phase-matching condition $\delta = 0$ is achieved when $\beta_a = \beta_b$. Then, the two modes have the same phase velocity and are *synchronized*. In case δ includes a contribution from additional structure, such as a

periodic grating, phase matching of the two modes being coupled can be accomplished by matching the difference $\Delta\beta = \beta_b - \beta_a$ with a grating phase factor to make $\delta = 0$. When considering phase matching between two modes, it is important always to include all sources of contribution to the phase-mismatch parameter δ.

Phase-matched coupling is always symmetric, meaning that $\kappa_{ab} = \kappa_{ba}^*$ whenever $\delta = 0$. This statement is true even when $c_{ab} \neq 0$ and $\beta_a \neq \beta_b$ (see Problem 4.3.1). However, *symmetric coupling does not necessarily imply a phase-matched condition.* Therefore, it is also possible to have $\kappa_{ab} = \kappa_{ba}^*$ while $\delta \neq 0$. The most obvious example of this situation is the coupling between two phase-mismatched modes in the same waveguide.

When perfect phase matching is accomplished, we can take

$$\kappa = \kappa_{ab} = \kappa_{ba}^* \quad \text{with} \quad \kappa = |\kappa| e^{i\varphi}. \tag{4.81}$$

Because $\delta = 0$, we find that

$$\beta_c = \alpha_c = |\kappa|. \tag{4.82}$$

With these relations under the condition of perfect phase matching, the matrix $\mathbf{F}(z; z_0)$ for codirectional coupling is reduced to the simple form

$$\mathbf{F}_{PM}(z; z_0) = \begin{bmatrix} \cos|\kappa|(z - z_0) & ie^{i\varphi} \sin|\kappa|(z - z_0) \\ ie^{-i\varphi} \sin|\kappa|(z - z_0) & \cos|\kappa|(z - z_0) \end{bmatrix}, \tag{4.83}$$

and the matrix $\mathbf{R}(z; 0, l)$ for contradirectional coupling is reduced to

$$\mathbf{R}_{PM}(z; 0, l) = \begin{bmatrix} \dfrac{\cosh|\kappa|(l - z)}{\cosh|\kappa| l} & ie^{i\varphi} \dfrac{\sinh|\kappa| z}{\cosh|\kappa| l} \\ ie^{-i\varphi} \dfrac{\sinh|\kappa|(l - z)}{\cosh|\kappa| l} & \dfrac{\cosh|\kappa| z}{\cosh|\kappa| l} \end{bmatrix}. \tag{4.84}$$

For codirectional coupling with perfect phase matching, the coupling efficiency is

$$\eta_{PM} = \sin^2|\kappa| l, \tag{4.85}$$

and the coupling length is

$$l_c^{PM} = \frac{\pi}{2|\kappa|}. \tag{4.86}$$

By choosing the interaction length to be $l = l_c^{PM}$, or any odd multiple of l_c^{PM}, 100% power transfer from one mode to the other with $\eta_{PM} = 1$ can be accomplished.

EXAMPLE 4.1 A phase-matched codirectional coupler has a coupling length of $l_c^{PM} = 1$ mm for a 100% coupling efficiency. What is the coupling coefficient of the coupler? For the same coupling coefficient, what is the length of a 3-dB *codirectional coupler* that has a 50% coupling efficiency?

Solution From (4.86), we find that the coupling coefficient has a value

$$|\kappa| = \frac{\pi}{2l_c^{PM}} = 1.57 \text{ mm}^{-1}.$$

From (4.85), we find that $\eta_{PM} = 1/2$ when $|\kappa|l = \pi/4$. Therefore, the length of the 3-dB codirectional coupler is simply

$$l_{3dB} = \frac{\pi}{4|\kappa|} = \frac{l_c^{PM}}{2} = 0.5 \text{ mm}.$$

A 3-dB codirectional coupler can be made by cutting the length of a 100% codirectional coupler in half. This statement is true even if the 100% coupler has a length longer than l_c^{PM} at any odd integral multiple of l_c^{PM}.

For contradirectional coupling with perfect phase matching, the coupling efficiency is

$$\eta_{PM} = \tanh^2 |\kappa|l. \tag{4.87}$$

For an interaction length of $l = l_c^{PM}$ defined in (4.86), this gives a coupling efficiency of $\eta_{PM} \approx 84\%$. Although complete power transfer with 100% efficiency cannot be accomplished in the case of contradirectional coupling, most power is transferred in a length comparable to the coupling length of codirectional coupling if phase matching is accomplished.

EXAMPLE 4.2 A phase-matched contradirectional coupler has the same coupling coefficient as that of the codirectional coupler in Example 4.1. What is the length of the contradirectional coupler for a 99% coupling efficiency? What is the length of a 3-dB *contradirectional coupler* with a 50% coupling efficiency?

Solution A contradirectional coupler only approaches 100% efficiency asymptotically. From (4.87), we find that $\eta_{PM} = 99\%$ when $|\kappa|l = 3 \approx 0.96\pi$. Therefore, the length of the 99% contradirectional coupler with $|\kappa| = 1.57 \text{ mm}^{-1}$ as found in Example 4.1 is

$$l = \frac{3}{|\kappa|} = 1.91 \text{ mm},$$

which is almost twice the length of the 100% codirectional coupler of the same coupling coefficient. We also find from (4.87) that $\eta_{PM} = 0.5$ when $|\kappa|l = 0.88 \approx 0.28\pi$. The length of the 3-dB contradirectional coupler is thus

$$l_{3dB} = \frac{0.88}{|\kappa|} = 0.56 \text{ mm},$$

which again is longer than the 3-dB codirectional coupler of the same coupling coefficient found in Example 4.1. We also see that, unlike codirectional couplers, a 3-dB

contradirectional coupler cannot be made by cutting in half a contradirectional coupler of nearly complete coupling at 99% efficiency.

In the presence of phase mismatch, symmetric coupling with $\kappa_{ab} = \kappa_{ba}^*$ is also true for coupling between two modes in the same waveguide but is not necessarily true for coupling between two different waveguides. Nevertheless, to illustrate the effect of phase mismatch on the coupling efficiency between two modes, we consider the simple case that $\kappa = \kappa_{ab} = \kappa_{ba}^*$, as expressed in (4.81). Then the coupling efficiency obtained in (4.66) for codirectionally coupled modes can be written as

$$\eta = \frac{1}{1 + |\delta/\kappa|^2} \sin^2 \left(|\kappa| l \sqrt{1 + |\delta/\kappa|^2} \right). \tag{4.88}$$

The maximum efficiency is

$$\eta_{\max} = \frac{1}{1 + |\delta/\kappa|^2} \tag{4.89}$$

at a coupling length of

$$l_c = \frac{l_c^{\mathrm{PM}}}{\sqrt{1 + |\delta/\kappa|^2}}. \tag{4.90}$$

The maximum coupling efficiency is clearly less than unity when $\delta \neq 0$. As shown in Fig. 4.7(a), both l_c and $\eta_{\max}$ decrease as $|\delta/\kappa|$ increases. If the interaction length is fixed at $l = l_c^{\mathrm{PM}}$, the efficiency also drops quickly as $|\delta/\kappa|$ increases, as shown in Fig. 4.7(b).

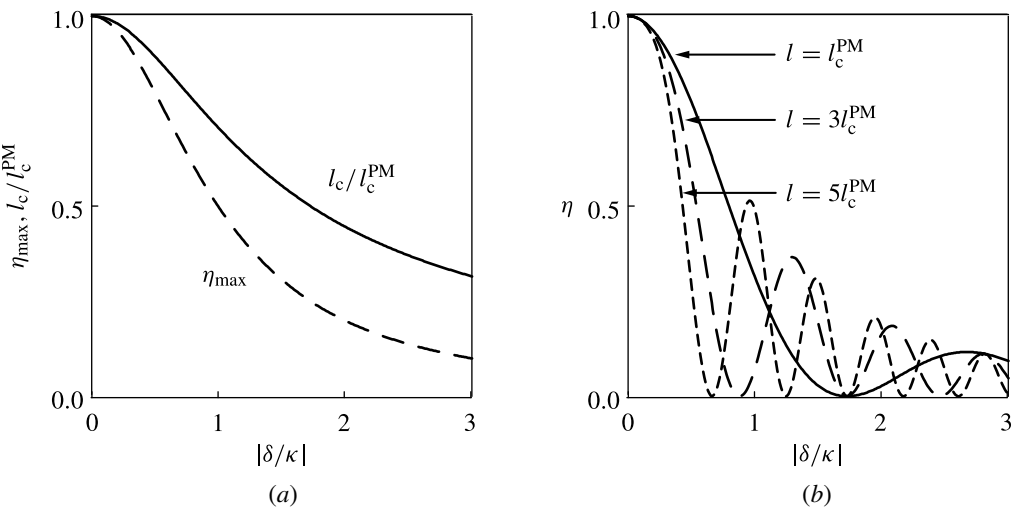


Figure 4.7 Effect of phase mismatch on codirectional coupling showing (a) the coupling length l_c, normalized as l_c/l_c^{PM} (solid curve) and the maximum coupling efficiency $\eta_{\max}$ (dashed curve) and (b) the coupling efficiency for fixed interaction lengths of $l = l_c^{\mathrm{PM}}$, $3l_c^{\mathrm{PM}}$, $5l_c^{\mathrm{PM}}$, both as a function of $|\delta/\kappa|$.

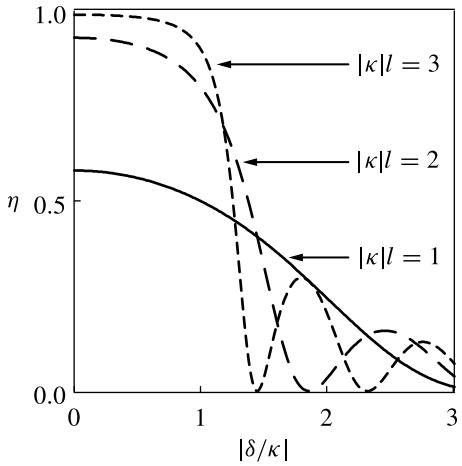


Figure 4.8 Effect of phase mismatch on contradirectional coupling showing the coupling efficiency for a few different values of $|\kappa|l$ as a function of $|\delta/\kappa|$.

For contradirectionally coupled modes, the coupling efficiency can also be expressed in terms of $|\kappa|l$ and $|\delta/\kappa|$:

$$\eta = \frac{\sinh^2\left(|\kappa|l\sqrt{1-|\delta/\kappa|^2}\right)}{\cosh^2\left(|\kappa|l\sqrt{1-|\delta/\kappa|^2}\right)-|\delta/\kappa|^2}. \tag{4.91}$$

The coupling efficiency also decreases as phase mismatch increases, as shown in Fig. 4.8.

EXAMPLE 4.3 Find the coupling efficiencies of codirectional and contradirectional couplers when the phase mismatch has the same magnitude as the coupling coefficient.

Solution For a codirectional coupler with $|\delta/\kappa| = 1$, we find from (4.88) that

$$\eta = \frac{1}{2}\sin^2\sqrt{2}|\kappa|l. \tag{4.92}$$

For a contradirectional coupler with $|\delta/\kappa| = 1$, we find from (4.91) that

$$\eta = \frac{|\kappa|^2l^2}{1+|\kappa|^2l^2}. \tag{4.93}$$

It is interesting to see that when the phase mismatch has the same magnitude as the coupling coefficient, a codirectional coupler can only have a maximum coupling efficiency of 50% but a contradirectional coupler can still have an efficiency higher than 50% if $|\kappa|l > 1$. However, the coupling efficiency of a contradirectional coupler varies with $|\kappa|l$ sinusoidally when $|\delta/\kappa| > 1$ rather than monotonically as it does when $|\delta/\kappa| < 1$.

In summary, to accomplish efficient coupling between two waveguide modes, the following three parameters have to be considered:

1. **Coupling coefficient.** The coupling coefficient κ has to exist and be large enough.
2. **Phase matching.** The phase mismatch has to be minimized so that $|\delta/\kappa|$ is made as small as possible. Ideally, perfect phase matching with $\delta = 0$ is desired.
3. **Interaction length.** For codirectional coupling, the length has to be properly chosen as the efficiency oscillates with interaction length. An overly long length is neither required nor beneficial. For contradirectional coupling, the length has to be sufficiently long but does not have to be critically chosen as the efficiency increases monotonically with interaction length. A very long length is not necessary, either.

PROBLEMS

4.1.1 Show, by expanding the field and the polarization into the linear combinations of their frequency components expressed in (4.5) and (4.6), respectively, that the general time-dependent wave equation given in (4.3) reduces to the coupled-wave equation given in (4.7).

4.1.2 Show that the coupled-wave equation given in (4.7), which is valid for both isotropic and anisotropic media, reduces to the form given in (4.8) for wave propagation in an isotropic medium but to the form given in (4.17) for wave propagation in an anisotropic medium.

4.1.3 Show that the coupled-wave equation given in (4.17) for wave propagation in an anisotropic medium can be reduced to that given in (4.18) under the slowly varying amplitude approximation. Show also that it can be further reduced to a form similar to that of (4.12) and (4.13) under proper conditions. What are the resulting expressions after such reduction? What are the conditions that allow such reduction?

4.2.1 Show that the coupled-mode equation given in (4.31) can be obtained by application of the Lorentz reciprocity theorem with mode expansion, followed by use of the orthonormality relation for waveguide modes.

4.2.2 Show that the coupled-mode equation given in (4.39) for multiple-waveguide mode coupling applies to mode coupling in a single waveguide as well by showing that $\kappa_{\nu\mu}$ given in (4.40) reduces to that given in (4.36) when all modes involved in the coupling belong to the same waveguide.

4.3.1 Coupling between two modes, a and b, is in general not symmetric if the two modes belong to two different waveguides and have different propagation constants such that $\beta_a \neq \beta_b$. Nevertheless, in lossless waveguides, if the coupling is phase matched in such a way that the total phase mismatch 2δ, which includes all the perturbations on the waveguides and appears in the coupled-mode equations

(4.50) and (4.51), is identically zero, the coupling is symmetric. This statement is true even when $\beta_a \neq \beta_b$ and $c_{ab} = c_{ba}^* \neq 0$ so that $\tilde{\kappa}_{ab} \neq \tilde{\kappa}_{ba}^*$. For simplicity, consider 2δ to have the form of $\Delta\beta$ given in (4.53).

a. Show that in lossless waveguides, $\tilde{\kappa}_{aa}$ and $\tilde{\kappa}_{bb}$ are both real quantities.
b. In the case of codirectional coupling, show that when $2\delta = 0$, $\kappa_{aa} - \kappa_{bb} = \tilde{\kappa}_{aa} - \tilde{\kappa}_{bb}$. In addition, verify that $\kappa_{ab} = \kappa_{ba}^*$; thus the coupling is symmetric.
c. In the case of contradirectional coupling, show that when $2\delta = 0$, $\kappa_{aa} + \kappa_{bb} = \tilde{\kappa}_{aa} + \tilde{\kappa}_{bb}$. In addition, verify that $\kappa_{ab} = \kappa_{ba}^*$; thus the coupling is also symmetric.

4.3.2 Solve (4.57) and (4.58) for two-mode codirectional coupling as an initial-value problem with given $\tilde{A}(z_0)$ and $\tilde{B}(z_0)$ to find the matrix $\mathbf{F}(z; z_0)$ expressed in (4.60). Show that $\mathbf{F}(z; z_1)\mathbf{F}(z_1; z_0) = \mathbf{F}(z; z_0)$ to demonstrate that codirectional coupling can be cascaded. Explain the physical meaning of this result.

4.3.3 In coupling light from one waveguide to another in a configuration of codirectional coupling, is it possible to increase the coupling efficiency continuously by increasing the length of the coupler? Why?

4.3.4 Solve (4.68) and (4.69) for two-mode contradirectional coupling as a boundary-value problem with given $\tilde{A}(0)$ and $\tilde{B}(l)$ to find the matrix $\mathbf{R}(z; 0, l)$ expressed in (4.71).

4.3.5 Instead of expressing contradirectional coupling in terms of the matrix $\mathbf{R}(z; 0, l)$ given in (4.71), the problem can be solved in terms of a matrix $\mathbf{S}(z; z_0)$ that relates the mode amplitudes $\tilde{A}(z_0)$ and $\tilde{B}(z_0)$ at z_0 to the mode amplitudes $\tilde{A}(z)$ and $\tilde{B}(z)$ at z for contradirectionally coupled modes as

$$\begin{bmatrix} \tilde{A}(z) \\ \tilde{B}(z) \end{bmatrix} = \mathbf{S}(z; z_0) \begin{bmatrix} \tilde{A}(z_0) \\ \tilde{B}(z_0) \end{bmatrix}, \tag{4.94}$$

in a manner similar to the matrix $\mathbf{F}(z; z_0)$ for codirectionally coupled modes. Show that this matrix for contradirectional coupling has the following form:

$$\mathbf{S}(z; z_0) = \begin{bmatrix} \dfrac{\alpha_c \cosh \alpha_c(z - z_0) - i\delta \sinh \alpha_c(z - z_0)}{\alpha_c} e^{i\delta(z - z_0)} & \dfrac{i\kappa_{ab}}{\alpha_c} \sinh \alpha_c(z - z_0) e^{i\delta(z + z_0)} \\ -\dfrac{i\kappa_{ba}}{\alpha_c} \sinh \alpha_c(z - z_0) e^{-i\delta(z + z_0)} & \dfrac{\alpha_c \cosh \alpha_c(z - z_0) + i\delta \sinh \alpha_c(z - z_0)}{\beta_c} e^{-i\delta(z - z_0)} \end{bmatrix}. \tag{4.95}$$

Show also that $\mathbf{S}(z; z_1)\mathbf{S}(z_1; z_0) = \mathbf{S}(z; z_0)$. What is the physical meaning of this result? Can contradirectional coupling be cascaded?

4.3.6 In coupling light from one waveguide to another in a configuration of contradirectional coupling, is it possible to increase the coupling efficiency continuously by increasing the length of the coupler? Why?

4.3.7 Verify the relation given in (4.79) for the phase shift φ_{DBR} that is defined in (4.78) for the reflection coefficient of contradirectional coupling.

4.3.8 The total power flow in a waveguide structure consisting of two coupled modes can be obtained by using the total field to calculate the time-averaged Poynting vector defined in (1.48) and then integrating it over the cross-sectional plane of the structure.

a. Show that the total power flow is that given in (4.80).
b. Verify the conservation of power for both codirectional coupling and contradirectional coupling in the general situation of asymmetric coupling.
c. Show that $\kappa_{ab} = \kappa_{ba}^*$ if and only if $P_{ab}(z) = 0$. What conclusion can be drawn regarding the condition for symmetric coupling from this fact?
d. Clearly, $P_{ab}(z) = 0$ if $c_{ab} = 0$. However, show that even when $c_{ab} \neq 0$, we have $P_{ab}(z) = 0$ if $\delta = 0$.

4.3.9 For codirectional coupling with a fixed value of coupling coefficient $|\kappa|$ and a fixed interaction length l, the coupling efficiency varies with the phase-mismatch parameter δ.

a. For highly efficient coupling with $l = l_c^{\mathrm{PM}}$ and $\eta_{\mathrm{PM}} = 1$ when perfect phase matching is achieved, find the values of δ for which the coupling efficiency is $\eta = \eta_{\mathrm{PM}}/2 = 1/2$.
b. In another limit of very low coupling efficiency with $\eta_{\mathrm{PM}} \ll 1$ because of a short interaction length of $l \ll l_c^{\mathrm{PM}}$, find the values of δ for which $\eta = \eta_{\mathrm{PM}}/2$.

4.3.10 For contradirectional coupling in the situation when $|\delta| > |\kappa|$ and when the phase-matched efficiency $\eta_{\mathrm{PM}} \ll 1$, find the values of δ for which $\eta = \eta_{\mathrm{PM}}/2$.

4.3.11 What are the first and second most likely causes if the coupling efficiency of a symmetric two-waveguide codirectional coupler is found to be $\eta = 0$? Also answer the same question for a symmetric two-waveguide contradirectional coupler.

4.3.12 A perfectly phase-matched coupler of high coupling efficiency is designed and fabricated. In this problem, we examine the tolerance on fabrication error.

a. If the coupler is a codirectional coupler designed to have a length of $l = l_c^{\mathrm{PM}} = \pi/2|\kappa|$ for a 100% efficiency, how much error in its length can be tolerated for a 10% variation in its efficiency? How much variation in its efficiency does a 10% error in its length cause?
b. If the coupler is a contradirectional coupler designed to have a length of $l = 3/|\kappa|$ for a 99% efficiency, how much error in its length can be tolerated for a 10% variation in its efficiency? How much variation in its efficiency does a 10% error in its length cause?

4.3.13 In designing a waveguide coupler of any geometry, what are the three major parameters that have to be considered in order to have a good efficiency? In what order of priority do they have to be considered? How are they optimized for best coupling efficiency in the case of (a) codirectional coupling and (b) contradirectional coupling?

SELECT BIBLIOGRAPHY

Buckman, A. B., *Guided-Wave Photonics*. Fort Worth, TX: Sauders College Publishing, 1992.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Marcuse, D., *Theory of Dielectric Optical Waveguides*, 2nd edn. Boston, MA: Academic Press, 1991.

Mentzer, M. A., *Principles of Optical Circuit Engineering*. New York: Marcel Dekker, 1990.

Nishihara, H., Haruna, M., and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Tamir, T., ed., *Integrated Optics*. New York: Springer-Verlag, 1982.

Yariv, A. and Yeh, P., *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: Wiley, 1984.

ADVANCED READING LIST

Chuang, S. L., "A coupled mode formulation by reciprocity and a variation principle," *Journal of Lightwave Technology* **LT-5**(1): 5–15, Jan. 1987.

"A coupled mode theory for multiwaveguide systems satisfying the reciprocity theorem and power conservation," *Journal of Lightwave Technology* **LT-5**(1): 174–183, Jan. 1987.

Hardy, A. and Streifer, W., "Coupled mode theory of parallel waveguides," *Journal of Lightwave Technology* **LT-3**(5): 1135–1146, Oct. 1985.

Haus, H. A. and Huang, W. P., "Coupled-mode theory," *Proceedings of the IEEE* **79**(10): 1505–1518, Oct. 1991.

Haus, H. A., Huang, W. P., Kawakami, S., and Whitaker, N.A., "Coupled-mode theory of optical waveguides," *Journal of Lightwave Technology* **LT-5**(1): 16–23, Jan. 1987.

Huang, W. P., "Coupled-mode theory for optical waveguides: an overview," *Journal of the Optical Society of America A* **11**(3): 963–983, Mar. 1994.

Streifer, W., Osinski, M., and Hardy, A., "Reformulation of the coupled-mode theory of multiwaveguide systems," *Journal of Lightwave Technology* **LT-5**(1): 1–4, Jan. 1987.

5 Optical couplers

Optical couplers are passive devices that couple light through waveguides or fibers. They play a very important role in the applications of photonic devices and systems. Optical couplers are used in many different ways. They can be the interface between devices in a system or can be important devices themselves. The most straightforward, yet important, application is to route optical waves around for coupling different devices. Sophisticated applications include devices such as polarization converters, mode converters, guided-wave beam splitters, beam combiners, directional couplers, branch couplers, wavelength filters, wavelength multiplexers, and so on. In this chapter, we discuss the waveguide couplers based on mode coupling. Input and output couplers, which couple light between free space and waveguides, are also discussed. Coupling due to active modulation, such as electro-optic switches, and coupling characteristics specific to a particular device are discussed in later chapters.

5.1 Grating waveguide couplers

Grating waveguide couplers have many useful applications and are one of the most important kinds of waveguide couplers. They consist of periodic fine structures that form gratings in waveguides. The grating in a waveguide can be either *periodic index modulation* or *periodic structural corrugation*. Periodic index modulation can be permanently written in a waveguide by periodically modulating the doping concentration in the waveguide medium, for example, or it can be created by an electro-optic, acousto-optic, or nonlinear optical effect. In the latter case, the grating can be time dependent if the modulation is time varying. It can also be a moving grating if the modulation signal is a traveling wave. In the case of periodic structural corrugation, the corrugation is a permanent structure of a waveguide. It is usually located at an interface between layers of different refractive indices, such as that between the guiding layer and the substrate or that between the guiding layer and the cover layer of a planar waveguide. It can also be placed away from the interfaces next to the guiding layer so long as the mode fields have sufficient penetration into the neighboring layers to see the corrugation. Figure 5.1 shows a few examples of grating structures in a planar waveguide.

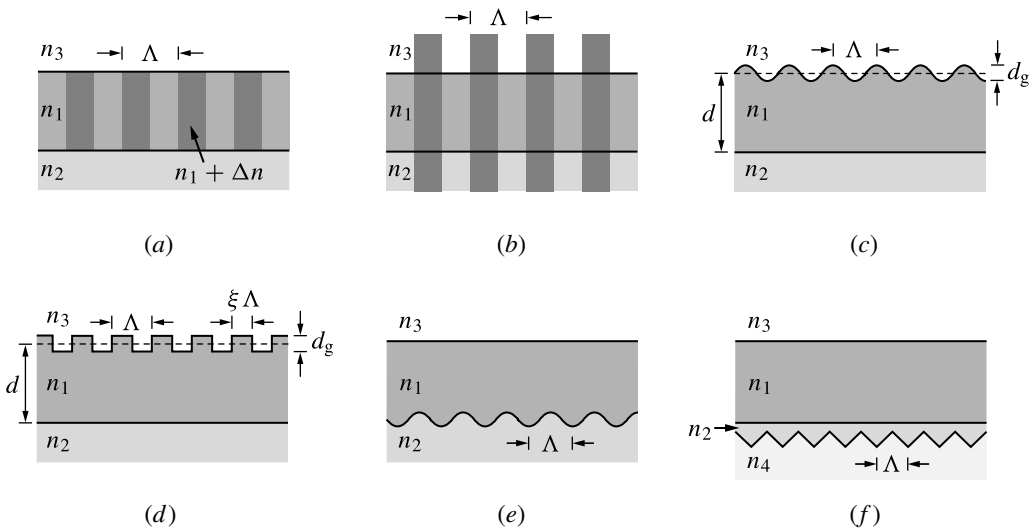


Figure 5.1 Structures of planar grating waveguide couplers with (*a*) and (*b*) periodic index modulation, (*c*), (*d*), (*e*), and (*f*) periodic structural corrugation.

In any event, the grating in a waveguide coupler can be considered as a periodic perturbation of $\Delta\epsilon$ that has a spatial periodicity characteristic of the grating. In a *coplanar coupler*, the grating can have a two-dimensional periodicity while the propagation vectors of the waves being coupled are in the same plane confined by the waveguide but not necessarily parallel to each other. In a *collinear coupler*, the waves being coupled are propagating either codirectionally or contradirectionally, and the grating is periodic only in the propagation direction of the guided waves. We consider here only the case of collinear coupling in a waveguide along the z direction. Then, $\Delta\epsilon$ is periodic only in z with a period Λ of the grating, as shown in Fig. 5.1.

With this periodically z-dependent perturbation, the coupling coefficients as defined in (4.36) and used in (4.45) and (4.46) are also periodic in z. In addition, for coupling in a single waveguide, we have $\kappa_{ab}(z) = \kappa_{ba}^*(z)$. They can be expressed in terms of the following Fourier series expansion:

$$\kappa_{ab}(z) = \omega \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_a^* \cdot \Delta\epsilon(x, y, z) \cdot \hat{\boldsymbol{\mathcal{E}}}_b \mathrm{d}x\mathrm{d}y = \sum_q \kappa_{ab}(q) \exp(\mathrm{i}qKz) \tag{5.1}$$

and

$$\kappa_{ba}(z) = \kappa_{ab}^*(z) = \sum_q \kappa_{ab}^*(q) \exp(-\mathrm{i}qKz), \tag{5.2}$$

where

$$K = \frac{2\pi}{\Lambda} \tag{5.3}$$

and

$$\kappa_{ab}(q) = \frac{1}{\Lambda}\int_0^{\Lambda} \kappa_{ab}(z)\mathrm{e}^{-\mathrm{i}qKz}\mathrm{d}z. \tag{5.4}$$

Considering the fact learned from Section 4.3 that efficient coupling exists only near phase matching, the coupled equations given in (4.45) and (4.46) can be transformed into (4.50) and (4.51) by the following approximations:

$$\pm\frac{\mathrm{d}\tilde{A}}{\mathrm{d}z} = \mathrm{i}\kappa_{ab}(z)\tilde{B}\mathrm{e}^{\mathrm{i}\Delta\beta z} = \mathrm{i}\tilde{B}\sum_q \kappa_{ab}(q)\mathrm{e}^{\mathrm{i}(\Delta\beta+qK)z} \approx \mathrm{i}\kappa\tilde{B}\mathrm{e}^{\mathrm{i}(\Delta\beta+qK)z}, \tag{5.5}$$

$$\pm\frac{\mathrm{d}\tilde{B}}{\mathrm{d}z} = \mathrm{i}\kappa_{ba}(z)\tilde{A}\mathrm{e}^{-\mathrm{i}\Delta\beta z} = \mathrm{i}\tilde{A}\sum_q \kappa^*_{ab}(q)\mathrm{e}^{-\mathrm{i}(\Delta\beta+qK)z} \approx \mathrm{i}\kappa^*\tilde{A}\mathrm{e}^{-\mathrm{i}(\Delta\beta+qK)z}, \tag{5.6}$$

where we can identify the phase mismatch as that given in (4.56):

$$2\delta = \Delta\beta + qK = \beta_b - \beta_a + qK. \tag{5.7}$$

Only one term in the Fourier series that yields a minimum value for $|\delta|$ is kept in each of the two coupled-mode equations because only this term will effectively couple the two waves. To be consistent with the notation used in the discussions following (4.81), we have also used

$$\kappa = \kappa_{ab}(q) \tag{5.8}$$

for the Fourier term that is kept in (5.5) and (5.6).

Note that though $\kappa_{ba}(z) = \kappa^*_{ab}(z)$, as is indicated in (5.2), $\kappa_{ba}(q)$ and $\kappa^*_{ab}(q)$ are not necessarily the same unless both happen to be real quantities. Instead, we have $\kappa_{ba}(q) = \kappa^*_{ab}(-q)$ among the Fourier components of $\kappa_{ba}(z)$ and $\kappa^*_{ab}(z)$. For this reason, the κs defined above have the following relations: $\kappa = \kappa_{ab}(q) = \kappa^*_{ba}(-q)$ and $\kappa^* = \kappa^*_{ab}(q) = \kappa_{ba}(-q)$.

We see from the above discussion that (5.5) and (5.6) are identical to (4.50) and (4.51), respectively, if we replace κ_{ab} and κ_{ba} in (4.50) and (4.51) respectively with κ and κ^*. Therefore, the general results obtained in Section 4.3 can be applied directly to the coupling of modes in a grating waveguide coupler with the coupling coefficients given by (5.8) and the phase mismatch given by (5.7).

EXAMPLE 5.1 Find the periods of the first- and second-order gratings for phase-matched coupling between contrapropagating (a) TE_0 and TE_0 mode fields, (b) TE_0 and TE_1 mode fields, and (c) TE_1 and TE_1 mode fields for the waveguide described in Example 2.1.

Solution For phase-matched coupling, it is required that $qK = -\Delta\beta$ because $\delta = 0$ in (5.7). Therefore, the grating period is

$$\Lambda = -q\frac{2\pi}{\Delta\beta}, \tag{5.9}$$

where $q = 1$ for the first-order grating and $q = 2$ for the second-order grating.

From Example 2.1, we find that $\beta_0 = 10.8432\ \mu\text{m}^{-1}$ for the TE_0 mode and $\beta_1 = 10.0036\ \mu\text{m}^{-1}$ for the TE_1 mode. (a) For the coupling from a forward-propagating TE_0 field to a backward-propagating TE_0 field, $\Delta\beta = -\beta_0 - \beta_0 = -21.6864\ \mu\text{m}^{-1}$. We then find by using (5.9) that $\Lambda_1 = 289.7$ nm for the first-order grating and $\Lambda_2 = 579.4$ nm for the second-order grating. (b) For the coupling from a forward-propagating TE_0 field to a backward-propagating TE_1 field, $\Delta\beta = -\beta_1 - \beta_0 = -20.8468\ \mu\text{m}^{-1}$. We find that $\Lambda_1 = 301.4$ nm and $\Lambda_2 = 602.8$ nm for the first- and second-order gratings, respectively. (c) For the coupling from a forward-propagating TE_1 field to a backward-propagating TE_1 field, $\Delta\beta = -\beta_1 - \beta_1 = -20.0072\ \mu\text{m}^{-1}$. We find that $\Lambda_1 = 314$ nm and $\Lambda_2 = 628$ nm for the first- and second-order gratings, respectively.

Coupling coefficient

As can be seen from the discussions in the preceding section and from (5.5), (5.6), and (5.7), the only important parameters for the coupling between two modes are the coupling coefficient κ, or $\kappa_{ab}(q)$ for the grating waveguide coupler discussed here, and the phase mismatch 2δ. The phase mismatch can be calculated using (5.7) once the propagation constants of the modes being coupled are known and the grating period is given. The calculation of $\kappa_{ab}(q)$ is less straightforward, however. It depends on exactly how the grating is created and where it is located in the waveguide. It also depends on the field distributions of the modes being coupled. In the following, we consider a few simple but important examples, including a grating produced by periodic index modulation, a sinusoidal corrugation grating, and a square corrugation grating, all in three-layer planar waveguides.

We assume that the unperturbed waveguides have the structure of the three-layer planar slab waveguide discussed in Chapter 2 and shown in Fig. 2.4. Combining (5.1) and (5.4), we can write

$$\kappa = \kappa_{ab}(q) = \frac{\omega}{\Lambda}\int_0^{\Lambda} \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}x\, \hat{\boldsymbol{\mathcal{E}}}_a^*(x) \cdot \Delta\epsilon(x, z) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(x) \mathrm{e}^{-\mathrm{i}qKz} \tag{5.10}$$

for coupling in a planar waveguide. The guiding layer of index n_1 has a thickness d located in the range of $-d/2 < x < d/2$. For the corrugation gratings, we consider the corrugation to be located at the interface between the guiding core and the cover layer. It is centered at the interface and has a depth of d_g, extending a maximum distance of $d_\text{g}/2$ into either side of the interface, as shown in Figs. 5.1(*c*) and (*d*).

1. **Index-modulation grating.** The geometric structure of the waveguide is not perturbed, but only the index of refraction is modulated. Because the index modulation is usually not localized but is distributed throughout the entire thickness of the guiding layer or a large portion of it, the integral in (5.10) has to be calculated with the complete field distributions $\hat{\boldsymbol{\mathcal{E}}}_a^*(x)$ and $\hat{\boldsymbol{\mathcal{E}}}_b(x)$ throughout the waveguide thickness (see Problem 5.1.1).
2. **Sinusoidal corrugation grating.** For a sinusoidal corrugation grating in an isotropic planar waveguide as shown in Fig. 5.1(*c*), the perturbation susceptibility is

$$\Delta\epsilon(x,z) = \begin{cases} \epsilon_0(n_1^2 - n_3^2), & \text{for } d/2 < x < d/2 + (d_g/2)\cos Kz, \quad \cos Kz > 0, \\ -\epsilon_0(n_1^2 - n_3^2), & \text{for } d/2 + (d_g/2)\cos Kz < x < d/2, \quad \cos Kz < 0. \end{cases} \tag{5.11}$$

Substitution of (5.11) into (5.10) yields

$$\begin{aligned} \kappa_{ab}(q) = \frac{\omega}{\Lambda} \Bigg[& \int_0^{\Lambda/4} \mathrm{d}z \int_{d/2}^{d/2+(d_g/2)\cos Kz} \mathrm{d}x \epsilon_0 (n_1^2 - n_3^2) \hat{\boldsymbol{\mathcal{E}}}_a^*(x) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(x) \mathrm{e}^{-\mathrm{i}qKz} \\ & - \int_{\Lambda/4}^{3\Lambda/4} \mathrm{d}z \int_{d/2+(d_g/2)\cos Kz}^{d/2} \mathrm{d}x \epsilon_0 (n_1^2 - n_3^2) \hat{\boldsymbol{\mathcal{E}}}_a^*(x) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(x) \mathrm{e}^{-\mathrm{i}qKz} \\ & + \int_{3\Lambda/4}^{\Lambda} \mathrm{d}z \int_{d/2}^{d/2+(d_g/2)\cos Kz} \mathrm{d}x \epsilon_0 (n_1^2 - n_3^2) \hat{\boldsymbol{\mathcal{E}}}_a^*(x) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(x) \mathrm{e}^{-\mathrm{i}qKz} \Bigg]. \end{aligned} \tag{5.12}$$

In most practical devices, $d_g \ll \lambda/n_1$. Then, we can approximate $\hat{\boldsymbol{\mathcal{E}}}_a^*(x)$ and $\hat{\boldsymbol{\mathcal{E}}}_b(x)$ in the range of $d/2 - d_g/2 < x < d/2 + d_g/2$ of the corrugation by $\hat{\boldsymbol{\mathcal{E}}}_a^*(d/2)$ and $\hat{\boldsymbol{\mathcal{E}}}_b(d/2)$, respectively, to obtain

$$\kappa_{ab}(q) \approx \omega\epsilon_0(n_1^2 - n_3^2)\hat{\boldsymbol{\mathcal{E}}}_a^*(d/2) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(d/2)\frac{d_g}{4}\left(\delta_{q,1} + \delta_{q,-1}\right), \tag{5.13}$$

where $\delta_{q,1}$ and $\delta_{q,-1}$ are Kronecker delta functions. Clearly, $\kappa_{ab}(q) \neq 0$ only for $q = 1$ or -1. Using the characteristics of the mode fields in planar slab waveguides discussed in Section 2.5, it can be shown that (5.13) yields (see Problem 5.1.3)

$$\kappa_{ab}(q) = \frac{h_a h_b}{\sqrt{\beta_a \beta_b d_a^{\mathrm{E}} d_b^{\mathrm{E}}}} \frac{d_g}{4}\left(\delta_{q,1} + \delta_{q,-1}\right) \tag{5.14}$$

for coupling between two TE modes, where h_a and h_b are the parameter h_1 defined in (2.50) and d_a^{E} and d_b^{E} are the effective waveguide thickness defined in (2.59) for the TE_a and TE_b modes, respectively. For coupling between two TM modes, we

have (see Problem 5.1.3)

$$\kappa_{ab}(q) = \frac{n_1^2 - n_3^2}{2n_1^2} \frac{\beta_a\beta_b/n_1^4 + \beta_a\beta_b/n_3^4 + 2\gamma_{3a}\gamma_{3b}/n_3^4}{\sqrt{h_a^2/n_1^4 + \gamma_{3a}^2/n_3^4}\sqrt{h_b^2/n_1^4 + \gamma_{3b}^2/n_3^4}} \frac{h_a h_b}{\sqrt{\beta_a\beta_b d_a^{\mathrm{M}} d_b^{\mathrm{M}}}} \times \frac{d_{\mathrm{g}}}{4}\left(\delta_{q,1} + \delta_{q,-1}\right), \tag{5.15}$$

where the parameters are relevant to the TM_a and TM_b modes. Because TE and TM modes do not couple in an isotropic waveguide, it is necessary to introduce birefringence in order to couple them.

3. **Square corrugation grating.** The perturbation susceptibility of a square corrugation grating in an isotropic planar waveguide as shown in Fig. 5.1(d) is

$$\Delta\epsilon(x,z) = \begin{cases} \epsilon_0(n_1^2 - n_3^2), & \text{for } d/2 < x < d/2 + d_{\mathrm{g}}/2, \ 0 < z < \xi\Lambda, \\ -\epsilon_0(n_1^2 - n_3^2), & \text{for } d/2 - d_{\mathrm{g}}/2 < x < d/2, \ \xi\Lambda < z < \Lambda, \end{cases} \tag{5.16}$$

where $0 < \xi < 1$ is the duty factor of the square corrugation. Substitution of (5.16) into (5.10) yields

$$\begin{aligned} \kappa_{ab}(q) &= \frac{\omega}{\Lambda}\left[\int_0^{\xi\Lambda} \mathrm{d}z \int_{d/2}^{d/2+d_{\mathrm{g}}/2} \mathrm{d}x\epsilon_0(n_1^2 - n_3^2)\hat{\boldsymbol{\mathcal{E}}}_a^*(x)\cdot\hat{\boldsymbol{\mathcal{E}}}_b(x)\mathrm{e}^{-\mathrm{i}qKz} \right. \\ &\qquad \left. - \int_{\xi\Lambda}^{\Lambda} \mathrm{d}z \int_{d/2-d_{\mathrm{g}}/2}^{d/2} \mathrm{d}x\epsilon_0(n_1^2 - n_3^2)\hat{\boldsymbol{\mathcal{E}}}_a^*(x)\cdot\hat{\boldsymbol{\mathcal{E}}}_b(x)\mathrm{e}^{-\mathrm{i}qKz}\right] \\ &\approx \omega\epsilon_0(n_1^2 - n_3^2)\hat{\boldsymbol{\mathcal{E}}}_a^*(d/2)\cdot\hat{\boldsymbol{\mathcal{E}}}_b(d/2)\frac{d_{\mathrm{g}}}{2}\frac{1}{\Lambda}\left[\int_0^{\xi\Lambda}\mathrm{d}z\mathrm{e}^{-\mathrm{i}qKz} - \int_{\xi\Lambda}^{\Lambda}\mathrm{d}z\mathrm{e}^{-\mathrm{i}qKz}\right] \\ &= 2\omega\epsilon_0(n_1^2 - n_3^2)\hat{\boldsymbol{\mathcal{E}}}_a^*(d/2)\cdot\hat{\boldsymbol{\mathcal{E}}}_b(d/2)\frac{d_{\mathrm{g}}}{2}\frac{\sin\xi q\pi}{q\pi}\mathrm{e}^{-\mathrm{i}\xi q\pi}. \end{aligned} \tag{5.17}$$

For coupling between two TE modes, (5.17) yields (see Problem 5.1.3)

$$\kappa_{ab}(q) = \frac{h_a h_b}{\sqrt{\beta_a\beta_b d_a^{\mathrm{E}} d_b^{\mathrm{E}}}} d_{\mathrm{g}} \frac{\sin\xi q\pi}{q\pi}\mathrm{e}^{-\mathrm{i}\xi q\pi}. \tag{5.18}$$

For coupling between two TM modes, we have (see Problem 5.1.3)

$$\kappa_{ab}(q) = \frac{n_1^2 - n_3^2}{2n_1^2} \frac{\beta_a\beta_b/n_1^4 + \beta_a\beta_b/n_3^4 + 2\gamma_{3a}\gamma_{3b}/n_3^4}{\sqrt{h_a^2/n_1^4 + \gamma_{3a}^2/n_3^4}\sqrt{h_b^2/n_1^4 + \gamma_{3b}^2/n_3^4}} \times \frac{h_a h_b}{\sqrt{\beta_a\beta_b d_a^{\mathrm{M}} d_b^{\mathrm{M}}}} d_{\mathrm{g}} \frac{\sin\xi q\pi}{q\pi}\mathrm{e}^{-\mathrm{i}\xi q\pi}. \tag{5.19}$$

We have seen that in the case where the grating is a simple sinusoidal grating proportional to $\cos Kz$ or, equivalently, $\sin Kz$, the expansion in (5.1) has only two terms, with $q = 1$ and $q = -1$. We also see from (5.18) and (5.19) that $\kappa_{ab}(q) = 0$ for a square grating if ξq is an integer. In practical situations, the grating can be rectangular, triangular, or any other shape. Then q takes on any integer value for which $\kappa_{ab}(q) \neq 0$.

It can also be seen from (5.14) and (5.18) that coupling coefficient between two TE modes does not depend on any parameters of the cover layer but only on those of the guiding layer although the grating is located at the interface between the cover and the guiding layers. This means that we would get exactly the same coupling coefficient for the two given TE modes if we instead placed the grating with the same parameters, including its period Λ, depth d_g, and shape, at the interface between the substrate and the guiding layer. This is also approximately, although not exactly, true for coupling between TM modes if the asymmetry of the waveguide is small so that $\gamma_2 \approx \gamma_3$ for each mode. This is an important conclusion for practical applications. It indicates that the same grating can be placed at either interface next to the guiding layer for the same desired coupling coefficient. This conclusion does not apply, however, to coupling between TM modes in a highly asymmetric waveguide where γ_2 and γ_3 are significantly different for a given mode.

EXAMPLE 5.2 A first-order square corrugation grating that has a depth of $d_g = 100$ nm and a duty factor of $\xi = 0.5$ is fabricated at either the upper or lower core–cladding boundary of the waveguide described in Example 2.1 to make a grating waveguide coupler for $\lambda = 1$ μm wavelength. Find the following coupling coefficients: κ_{00}^{TE} between the forward-propagating TE_0 mode field and the backward-propagating TE_0 mode field, κ_{01}^{TE} between the forward-propagating TE_0 mode field and the backward-propagating TE_1 mode field, and κ_{00}^{TM} between the forward-propagating TM_0 mode field and the backward-propagating TM_0 mode field.

Solution For the waveguide described in Example 2.1, $n_1 = 1.77$, $n_2 = 1.45$, and $n_3 = 1$ for $\lambda = 1$ μm. We first check the condition that $d_g \ll \lambda/n_1$ for the approximation made in obtaining (5.18) and (5.19) to be valid. Because $\lambda/n_1 = 565$ nm and $d_g = 100$ nm, this condition is satisfied. Therefore, we can use (5.18) to calculate the coupling coefficients for the TE modes and (5.19) for the TM modes with $q = 1$, $\xi = 0.5$, $d_g = 100$ nm, and the mode parameters given in the table in Example 2.1.

The coupling coefficient between two TE modes does not depend on whether the grating is placed in the upper or lower core–cladding boundary. From (5.18), we find that the coupling coefficient for TE_0–TE_0 coupling is

$$\kappa_{00}^{TE} = -i\frac{h_0^2}{\beta_0 d_0^E}\frac{d_g}{\pi} = -i0.014\ \mu m^{-1},$$

and the coupling coefficient for TE_0–TE_1 coupling is

$$\kappa_{01}^{\mathrm{TE}} = -\mathrm{i}\frac{h_0 h_1}{\sqrt{\beta_0 \beta_1 d_0^{\mathrm{E}} d_1^{\mathrm{E}}}}\frac{d_{\mathrm{g}}}{\pi} = -\mathrm{i}0.0277\ \mu\mathrm{m}^{-1},$$

where h_0 and h_1 are the h parameters for the TE_0 and TE_1 modes, respectively.

The coupling coefficient between two TM modes depends on the location of the grating. If the grating is placed at the upper boundary between the polymer core of index $n_1 = 1.77$ and the air cover of index $n_3 = 1$, the coupling coefficient for TM_0–TM_0 is calculated using (5.19) as

$$\kappa_{00}^{\mathrm{TM}} = -\mathrm{i}\frac{n_1^2 - n_3^2}{2n_1^2}\frac{\beta_0^2/n_1^4 + \beta_0^2/n_3^4 + 2\gamma_{30}^2/n_3^4}{h_0^2/n_1^4 + \gamma_{30}^2/n_3^4}\frac{h_0^2}{\beta_0 d_0^{\mathrm{M}}}\frac{d_{\mathrm{g}}}{\pi} = -\mathrm{i}0.0233\ \mu\mathrm{m}^{-1},$$

where all of the parameters belong to the TM_0 mode. If the grating is placed at the lower boundary between the polymer core of index $n_1 = 1.77$ and the silica substrate of index $n_2 = 1.45$, the coupling coefficient for TM_0–TM_0 is calculated by replacing γ_3 in (5.19) with γ_2 as

$$\kappa_{00}^{\mathrm{TM}} = -\mathrm{i}\frac{n_1^2 - n_2^2}{2n_1^2}\frac{\beta_0^2/n_1^4 + \beta_0^2/n_2^4 + 2\gamma_{20}^2/n_2^4}{h_0^2/n_1^4 + \gamma_{20}^2/n_2^4}\frac{h_0^2}{\beta_0 d_0^{\mathrm{M}}}\frac{d_{\mathrm{g}}}{\pi} = -\mathrm{i}0.0199\ \mu\mathrm{m}^{-1}.$$

We see that there is about 17% difference between these two coefficients for TM mode coupling because of the difference between γ_2 and γ_3.

Distributed Bragg reflector

We consider here a device of special interest that has very important applications. The function of this device is based on coupling between the forward- and backward-propagating fields of the same mode in a grating waveguide coupler. This is a special case of contradirectional coupling where $\beta_b = -\beta_a$ and $\Delta\beta = \beta_b - \beta_a$. In this case, we can define

$$\beta \equiv \beta_a = -\beta_b. \tag{5.20}$$

Then, $\Delta\beta = -2\beta$, and (5.7) becomes

$$2\delta = -2\beta + qK. \tag{5.21}$$

Thus, the phase-matching condition can be stated as the following Bragg condition:

$$\beta_{\mathrm{B}} = q\frac{K}{2}, \tag{5.22}$$

where q is the integer that allows phase matching and is the *order of coupling* between the two contrapropagating waves. The grating period required to satisfy this

phase-matching condition is

$$\Lambda = q\frac{\pi}{\beta_{\mathrm{B}}} = q\frac{\lambda_{\mathrm{B}}}{2n_{\beta}}, \tag{5.23}$$

where $\lambda_{\mathrm{B}} = 2\pi n_{\beta}/\beta_{\mathrm{B}}$ is the free-space *Bragg wavelength* of the field and n_{β} is the effective refractive index of the mode field in the waveguide. A grating with a period given by (5.23) for an integer q is called a *qth-order grating* for the mode coupling under consideration. For example, it is a *first-order grating* if $\Lambda = \lambda_{\mathrm{B}}/2n_{\beta}$ and is a *second-order grating* if $\Lambda = \lambda_{\mathrm{B}}/n_{\beta}$. A simple sinusoidal grating can only be a first-order grating because, as mentioned above, q can only be 1 or -1 in this case and thus can only have the value 1 in (5.23).

To get an idea of the size of the grating period in a practical device structure, we consider as an example the grating in an InGaAsP waveguide for an optical wavelength of $\lambda = \lambda_{\mathrm{B}} = 1.3$ μm in free space. The index of refraction for InGaAsP with a bandgap energy corresponding to 1.3 μm optical wavelength is about 3.48. Taking $n_{\beta} \approx 3.48$, we find that $\lambda_{\mathrm{B}}/n_{\beta} \approx 1.3\ \mu\mathrm{m}/3.48 \approx 374$ nm. We then have $\Lambda = 187$ nm for a first-order grating and $\Lambda = 374$ nm for a second-order grating. These are certainly very fine structures.

As discussed in the preceding section, the effect of this contradirectional coupling is an efficient transfer of power from the forward-propagating field to the backward-propagating field when $\delta^2 < |\kappa|^2$. From the input end of the grating waveguide coupler, it is seen that power is reflected back due to this coupling. This type of reflector, which relies on the coupling of waves by a distributed periodic structure, is called the *distributed Bragg reflector* (DBR). Its reflection coefficient r is that given in (4.78) with an amplitude $|r| = \eta^{1/2}$ and a phase φ_{DBR} given in (4.79). Thus, its reflectivity is simply the coupling efficiency η given by (4.91). The peak reflectivity at the Bragg wavelength is $R_{\mathrm{DBR}} = \eta_{\mathrm{PM}}$ given by (4.87) under the phase-matched condition. The reflectance R_{DBR} and the transmittance $T_{\mathrm{DBR}} = 1 - R_{\mathrm{DBR}}$ of such a reflector at its Bragg wavelength where the device has perfect phase matching are plotted in Fig. 5.2 as a function of the effective coupler length $|\kappa|l$. As can be seen from Fig. 4.6, the reflection of power, or the transfer of power from the forward-propagating mode to the backward-propagating mode, is not localized but is distributed throughout the length of the grating coupler. The phase shift of Bragg reflection at the phase-matched *Bragg frequency* $\omega_{\mathrm{B}} = 2\pi\nu_{\mathrm{B}}$, which corresponds to the Bragg wavelength λ_{B} through the relation $\nu_{\mathrm{B}} = c/\lambda_{\mathrm{B}}$, is $\varphi_{\mathrm{DBR}} = \varphi_{\mathrm{B}}$ for $\delta = 0$ in (4.79). When the optical frequency deviates from the Bragg frequency, the phase shift of Bragg reflection given in (4.79) can be expressed in terms of the variation of the propagation constant $\beta(\omega)$ away from the phase-matched value of $\beta(\omega_{\mathrm{B}}) = \beta_{\mathrm{B}}$ in the following form (see Problem 5.1.9):

$$\varphi_{\mathrm{DBR}} = \varphi_{\mathrm{B}} + 2\left[\beta(\omega) - \beta(\omega_{\mathrm{B}})\right] l_{\mathrm{DBR}}^{\mathrm{eff}}, \tag{5.24}$$

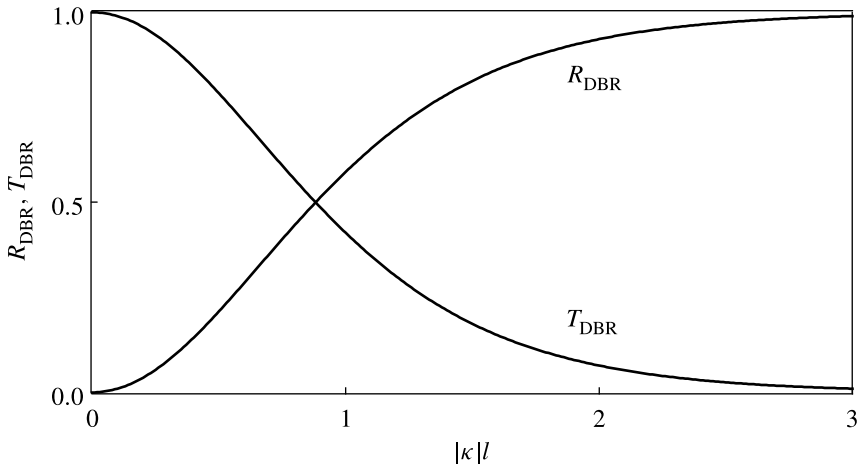


Figure 5.2 Reflectance, R_{DBR}, and transmittance, T_{DBR}, of a distributed Bragg reflector at its Bragg wavelength as a function of effective coupler length $|\kappa|l$.

where

$$l_{\mathrm{DBR}}^{\mathrm{eff}} = \frac{\tanh|\kappa|l}{2|\kappa|} = \frac{\eta_{\mathrm{PM}}^{1/2}}{2|\kappa|} = \frac{R_{\mathrm{DBR}}^{1/2}}{2|\kappa|}. \tag{5.25}$$

The parameter $l_{\mathrm{DBR}}^{\mathrm{eff}}$ is an effective length of the DBR for its reflection phase shift. These grating couplers can be used in a *distributed Bragg reflector laser* (DBR laser) or in a *distributed feedback laser* (DFB laser) to provide optical feedback without ordinary Fabry–Perot mirrors. Such lasers are discussed in Section 13.9.

A DBR can be designed to function as a narrow-band *frequency filter*. Consequently, an important characteristic of a DBR or DFB laser is its frequency selectivity and stability, which results in stable single-frequency operation of the laser if the structure is properly designed. This frequency selectivity of a DBR can be understood by considering the dispersion characteristics of a waveguide mode and the effect of phase matching, as shown in Fig. 5.3. The dispersion relations $\beta_a(\omega)$ and $\beta_b(\omega)$ for the waveguide modes are determined by the waveguide parameters and the optical frequency ω. In the case under consideration, we have $\beta_a(\omega) = \beta(\omega)$ and $\beta_b(\omega) = -\beta(\omega)$. For phase matching, we need $\beta_b + qK = \beta_a$, which can be found by shifting the dispersion curve of ω versus β horizontally by an amount qK to find the intersection between the two curves representing $\beta_b + qK$ and β_a. This procedure is illustrated in Fig. 5.3. The intersecting point of these two curves corresponds to a frequency ω_{B} at which $\delta(\omega_{\mathrm{B}}) = 0$ where phase matching is perfect. Away from this frequency, $\delta \neq 0$, and the coupling efficiency drops. The range of δ within which the modes remain well coupled is $-|\kappa| < \delta < |\kappa|$, which can be found by considering α_{c} in (4.72). For $|\delta| < |\kappa|$, α_{c} remains real, and the coupling efficiency given in (4.87) or (4.91) depends on hyperbolic functions that do not oscillate. Then, η increases monotonically as $|\kappa|l$ increases, as can be seen in Fig. 5.2.

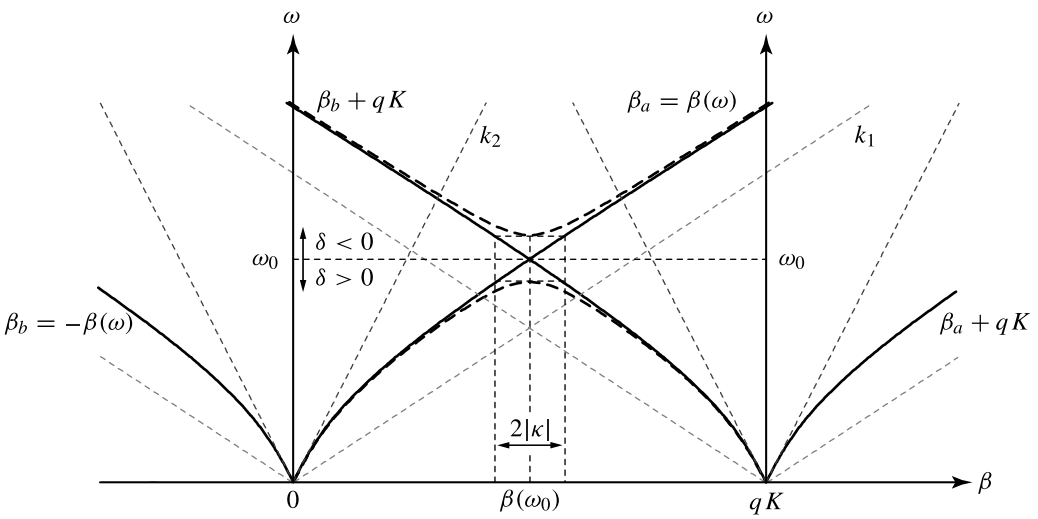


Figure 5.3 Dispersion relation showing the coupling of contradirectional modes in a grating waveguide coupler by phase matching through a qth-order grating of a grating wavenumber K. Perfect phase matching occurs at ω_0. Strong coupling exists in a range where $|\delta| < |\kappa|$.

For $|\delta| > |\kappa|$, α_c becomes purely imaginary. Then, η depends on sinusoidal functions and drops quickly as $|\delta/\kappa|$ increases, as can be seen in Fig. 4.8. As a consequence, the forward- and backward-propagating modes gradually become decoupled.

The frequency bandwidth of a DBR can be found by considering the frequency dependence of δ. Using (5.21), we have

$$\delta(\omega) = -\beta(\omega) + \frac{qK}{2} = -\beta(\omega_B) - \frac{d\beta}{d\omega}(\omega - \omega_B) + \cdots + \frac{qK}{2} \approx -\frac{d\beta}{d\omega}(\omega - \omega_B). \tag{5.26}$$

For a given value of $|\kappa|l$, the maximum efficiency is η_{PM} given by (4.87). It can be shown using (4.91) that at $\eta = \eta_{PM}/2$, we have $|\delta| > |\kappa|$ and (see Problem 5.1.11)

$$(2\coth^2|\kappa|l - 1)\sin^2\left(|\kappa|l\sqrt{|\delta/\kappa|^2 - 1}\right) = |\delta/\kappa|^2 - 1. \tag{5.27}$$

The FWHM reflectivity bandwidth $\Delta\omega$ for a DBR is given by

$$\Delta\omega = 2\left|\delta_{1/2}\frac{d\omega}{d\beta}\right|, \tag{5.28}$$

where $|\delta_{1/2}|$ is the root of (5.27) for a given l and $|\kappa|$. Its value can also be found by reading the value of $|\delta/\kappa|$ for $\eta = \eta_{PM}/2$ on the curve in Fig. 4.8 for a given value of $|\kappa|l$. For a given structure that has a fixed value of $|\kappa|$, the bandwidth $\Delta\omega$ decreases as the length l increases. However, for a fixed length l, the bandwidth increases as the coupling becomes stronger, and the value of $|\kappa|$ increases.

By taking $\mathrm{d}\beta/\mathrm{d}\omega = N_\beta/c$, where N_β is the effective group refractive index of the waveguide mode at the Bragg wavelength, the bandwidth given in (5.28) is approximately bounded within the following range (see Problem 5.1.11):

$$2\sqrt{2}\frac{|\kappa|c}{N_\beta}\coth|\kappa|l \geq \Delta\omega > 2\frac{|\kappa|c}{N_\beta}. \tag{5.29}$$

EXAMPLE 5.3 A DBR is made with the grating waveguide coupler described in Example 5.2 that has a first-order grating for phase-matched coupling of the forward-propagating TE_0 field to the backward-propagating TE_0 field at $\lambda = 1$ μm. (a) If a reflectivity of 50% is desired, what are the required length and the corresponding number of periods of the grating? (b) How much is the leakage coupling to the backward-propagating TE_1 field? (c) What is the bandwidth of this DBR?

Solution (a) From Example 4.2, we know that $|\kappa|l = 0.88$ for a phase-matched 3-dB contradirectional coupler of $\eta = 50\%$. From Example 5.2, we find that $\kappa = \kappa_{00}^{\mathrm{TE}} = -\mathrm{i}0.014\ \mu\mathrm{m}^{-1}$. Therefore, the required length of the DBR is

$$l = \frac{0.88}{|\kappa|} = 63\ \mu\mathrm{m}.$$

From Example 5.1, we know that the period of this first-order grating is $\Lambda = 289.7$ nm. The number of periods of the DBR is thus

$$N_{\mathrm{DBR}} = \frac{l}{\Lambda} = 217.$$

(b) For leakage coupling to the TE_1 mode, $\kappa_{01}^{\mathrm{TE}} = -\mathrm{i}0.0277\ \mu\mathrm{m}^{-1}$ from Example 5.2. This coupling is not phase matched. The phase mismatch can be found as

$$2\delta = \Delta\beta + qK = -\beta_1 - \beta_0 + K = 0.8396\ \mu\mathrm{m}^{-1}$$

for $\beta_1 = 10.0036\ \mu\mathrm{m}^{-1}$, $\beta_0 = 10.8432\ \mu\mathrm{m}^{-1}$, $K = 21.6864\ \mu\mathrm{m}^{-1}$, and $q = 1$. We then find that $|\delta| = 15.16|\kappa_{01}^{\mathrm{TE}}|$ and $|\kappa_{01}^{\mathrm{TE}}|l = 1.745$ for $l = 63$ μm. Plugging these numbers in (4.91) for phase-mismatched contradirectional coupling, we find that the efficiency for leakage coupling is, for $|\delta/\kappa_{01}^{\mathrm{TE}}| > 1$ in this case,

$$\eta_{01} = \frac{\sin^2\left(|\kappa_{01}^{\mathrm{TE}}|l\sqrt{|\delta/\kappa_{01}^{\mathrm{TE}}|^2 - 1}\right)}{|\delta/\kappa_{01}^{\mathrm{TE}}|^2 - \cos^2\left(|\kappa_{01}^{\mathrm{TE}}|l\sqrt{|\delta/\kappa_{01}^{\mathrm{TE}}|^2 - 1}\right)} = 0.004.$$

Therefore, there is only about 0.4% of the TE_0 mode power that is leaked to the TE_1 mode because of the large phase mismatch in the TE_0–TE_1 coupling.

(c) To find the bandwidth without solving for the entire dispersion curve of the waveguide mode, we take $N_\beta \approx n_\beta$ as an approximation. Then, for the TE_0 mode

under consideration, $N_\beta \approx \beta_0/2\pi\lambda = 10.8432/2\pi = 1.7257$ because $\lambda = 1$ μm. Using $|\kappa| = 0.014\ \mu\text{m}^{-1}$ and $\Delta\omega = 2\pi\,\Delta\nu$, we find by applying (5.29) that the bandwidth is bounded in the range

$$\sqrt{2}\frac{|\kappa|c}{\pi N_\beta}\coth|\kappa|l = 1.5494\ \text{THz} \geq \Delta\nu > \frac{|\kappa|c}{\pi N_\beta} = 774.7\ \text{GHz}.$$

Solving (5.27) exactly for $|\kappa|l = 0.88$ yields $|\delta_{1/2}| = 1.998|\kappa|$. This results in an actual bandwidth of $\Delta\nu = 1.5478$ THz. Therefore, the actual bandwidth of this DBR is very close to its upper bound.

5.2 Directional couplers

Directional couplers are multiple-waveguide couplers used for codirectional coupling. They can be used in many different applications, including power splitters, optical switches, wavelength filters, and polarization selectors. We consider in this section *two-channel directional couplers*, which consist of two parallel waveguides, as shown schematically in Fig. 5.4. For simplicity, we consider only the case where each waveguide supports only its own fundamental mode. Coupling between the two single-mode waveguides in such a two-channel directional coupler is simply described by (4.57)

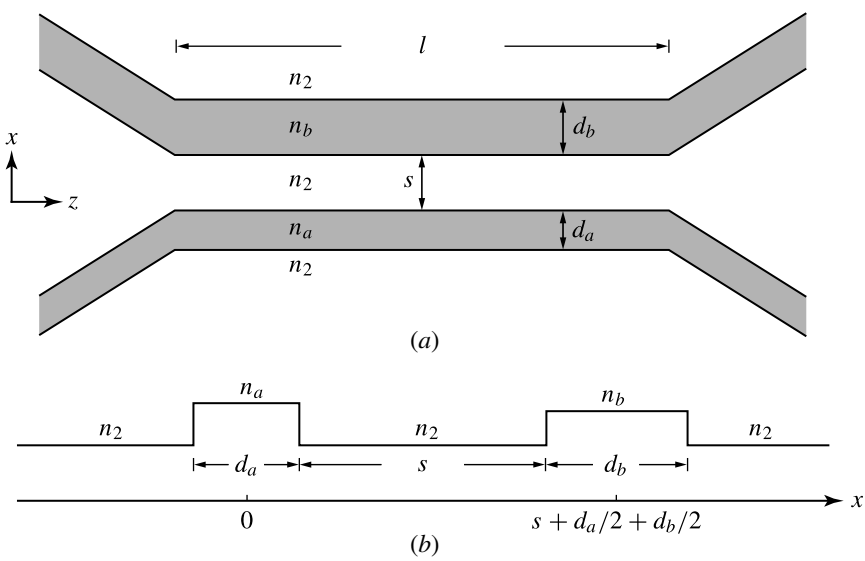


Figure 5.4 Schematic diagram of (*a*) a two-channel directional coupler and (*b*) its index profile assuming two step-index waveguides on the same substrate. The coupler is symmetric if $n_a = n_b = n_1$ and $d_a = d_b = d$.

and (4.58) with

$$2\delta = (\beta_b + \kappa_{bb}) - (\beta_a + \kappa_{aa}). \tag{5.30}$$

In addition, $c_{aa} = c_{bb} = 1$ for evaluation of the coupling coefficients using (4.47) because the waves in both waveguides are forward propagating. In general, the two waveguides are not necessarily identical. Then the directional coupler is not symmetric, and $\kappa_{ab} \neq \kappa_{ba}^*$, as discussed in Section 4.3. If the two waveguides are identical, the directional coupler is symmetric. Then, $\kappa_{ab} = \kappa_{ba}^*$, $\kappa_{aa} = \kappa_{bb}$, and $\beta_a = \beta_b$. In either case, the general solutions for codirectional coupling obtained in Section 4.3 can be used directly if the coupling coefficients and the phase mismatch are known.

Coupling coefficient

The coefficients involved in the coupling between two waveguides are given by (4.47). They are more complicated than those in the coupling between two modes in the same waveguide because of the existence of the overlap coefficient and the fact that $\kappa_{ab} \neq \kappa_{ba}^*$ in general. As a result, many parameters have to be calculated in order to obtain the coupling coefficients κ_{ab} and κ_{ba} using (4.47). Here we consider a simple example, namely, the two-channel directional coupler with step-index waveguides on the same substrate shown in Fig. 5.4. We assume that the waveguides are planar slab waveguides for simplicity although practical direction couplers are often made of channel waveguides. We also consider only isotropic waveguides where TE and TM modes do not couple.

As shown in Fig. 5.4, the two waveguides have widths d_a and d_b and guiding-layer refractive indices n_a and n_b, respectively. They are separated by a distance s between the two near edges of the guiding layers. The index of refraction of the substrate is n_2. When $n_a = n_b = n_1$ and $d_a = d_b = d$, the coupler is symmetric. Otherwise, it is asymmetric.

To calculate the relevant coefficients, we first identify the perturbation $\Delta\epsilon$ for each waveguide. For waveguide a, the susceptibility step of waveguide b above the substrate is the perturbation, and vice versa. Therefore, we have

$$\Delta\epsilon_a = \begin{cases} \epsilon_0(n_b^2 - n_2^2), & s + d_a/2 < x < s + d_a/2 + d_b, \\ 0, & \text{otherwise}, \end{cases} \tag{5.31}$$

and

$$\Delta\epsilon_b = \begin{cases} \epsilon_0(n_a^2 - n_2^2), & -d_a/2 < x < d_a/2, \\ 0, & \text{otherwise}, \end{cases} \tag{5.32}$$

where we have chosen the origin of the x coordinate to be at the center of waveguide a. Using the field distributions and the characteristic parameters of planar waveguide

modes obtained in Section 2.5, we can calculate the relevant coefficients using (4.41) and (4.42).

For coupling between TE modes, the only nonzero component for the electric field is $\hat{\mathcal{E}}_y$ given by (2.55). Therefore, we have

$$\begin{aligned}\tilde{\kappa}_{aa} &= \omega \int_{-\infty}^{\infty} \Delta\epsilon_a |\hat{\mathcal{E}}_{a,y}|^2 \mathrm{d}x \\ &= \frac{1}{\beta_a d_a^{\mathrm{E}}} \cdot \frac{h_b^2+\gamma_b^2}{h_a^2+\gamma_a^2} \cdot \frac{h_a^2}{2\gamma_a}(1-\mathrm{e}^{-2\gamma_a d_b})\mathrm{e}^{-2\gamma_a s},\end{aligned} \tag{5.33}$$

and

$$\begin{aligned}\tilde{\kappa}_{ab} &= \omega \int_{-\infty}^{\infty} \Delta\epsilon_b \hat{\mathcal{E}}_{a,y}^* \hat{\mathcal{E}}_{b,y} \mathrm{d}x \\ &= \frac{1}{\sqrt{\beta_a\beta_b d_a^{\mathrm{E}} d_b^{\mathrm{E}}}} \sqrt{\frac{h_a^2+\gamma_a^2}{h_b^2+\gamma_b^2}} \cdot \frac{h_a h_b}{h_a^2+\gamma_b^2} \left[(\gamma_a+\gamma_b)+(\gamma_a-\gamma_b)\mathrm{e}^{-\gamma_b d_a}\right]\mathrm{e}^{-\gamma_b s},\end{aligned} \tag{5.34}$$

where h_a and h_b are the parameter h_1, γ_a and γ_b are the parameter γ_2, and d_a^{E} and d_b^{E} are the effective waveguide thickness for the TE modes in waveguide a and waveguide b, respectively. The coefficient $\tilde{\kappa}_{bb}$ can be obtained by simply interchanging the indices a and b in (5.33), while $\tilde{\kappa}_{ba}$ can be obtained by interchanging the indices a and b in (5.34). It can be seen that $\tilde{\kappa}_{aa} \neq \tilde{\kappa}_{bb}$, and $\tilde{\kappa}_{ab} \neq \tilde{\kappa}_{ba}$, in general, as expected. By following a procedure that reduces (2.41) to (2.44) for TE modes, it can be shown that, for TE modes, the overlap coefficient defined by (4.41) can be reduced to

$$\begin{aligned}c_{ab} = c_{ba}^* &= \frac{\beta_a+\beta_b}{\omega\mu_0} \int_{-\infty}^{\infty} \hat{\mathcal{E}}_{a,y}^* \hat{\mathcal{E}}_{b,y} \mathrm{d}x \\ &= \frac{\beta_a+\beta_b}{\sqrt{\beta_a\beta_b d_a^{\mathrm{E}} d_b^{\mathrm{E}}}} \left[\sqrt{\frac{h_a^2+\gamma_a^2}{h_b^2+\gamma_b^2}} \frac{h_a h_b}{h_a^2+\gamma_b^2} \left(\frac{1}{\gamma_a-\gamma_b}+\frac{\mathrm{e}^{-\gamma_b d_a}}{\gamma_a+\gamma_b}\right)\mathrm{e}^{-\gamma_b s}\right. \\ &\qquad \left. + \sqrt{\frac{h_b^2+\gamma_b^2}{h_a^2+\gamma_a^2}} \frac{h_b h_a}{h_b^2+\gamma_a^2} \left(\frac{1}{\gamma_b-\gamma_a}+\frac{\mathrm{e}^{-\gamma_a d_b}}{\gamma_b+\gamma_a}\right)\mathrm{e}^{-\gamma_a s}\right].\end{aligned} \tag{5.35}$$

Similar, but somewhat more complicated, formulas can be obtained for coupling between TM modes.

We now consider the simple case of a symmetric directional coupler. In this case, $n_a = n_b = n_1$ and $d_a = d_b = d$. Therefore, we have $\beta_a = \beta_b \equiv \beta$, $h_a = h_b = h_1 \equiv h$, and $\gamma_a = \gamma_b = \gamma_2 \equiv \gamma$, and the coefficients are much simplified. For coupling between

TE modes of the same order, we have

$$\tilde{\kappa}_{aa} = \tilde{\kappa}_{bb} = \frac{1}{\beta d_{\mathrm{E}}} \cdot \frac{h^2}{2\gamma}(1 - \mathrm{e}^{-2\gamma d})\mathrm{e}^{-2\gamma s}, \tag{5.36}$$

$$\tilde{\kappa} \equiv \tilde{\kappa}_{ab} = \tilde{\kappa}^*_{ba} = \frac{2}{\beta d_{\mathrm{E}}} \cdot \frac{h^2\gamma}{h^2 + \gamma^2}\mathrm{e}^{-\gamma s}, \tag{5.37}$$

and

$$c \equiv c_{ab} = c^*_{ba} = \frac{2}{d_{\mathrm{E}}}\frac{h^2}{h^2 + \gamma^2}\left(s + \frac{\mathrm{e}^{-\gamma d}}{\gamma}\right)\mathrm{e}^{-\gamma s}. \tag{5.38}$$

Similar simplified formulas can be obtained for coupling between TM modes. The coupling coefficient used in the coupled-mode equation is

$$\kappa = \frac{\tilde{\kappa} - c^*\tilde{\kappa}_{aa}}{1 - |c|^2}. \tag{5.39}$$

Because now

$$\beta_a + \kappa_{aa} = \beta_b + \kappa_{bb}, \tag{5.40}$$

we have $\delta = 0$, and the coupling is always phase matched, as expected.

EXAMPLE 5.4 A symmetric directional coupler for $\lambda = 1$ μm wavelength, Fig. 5.5, is made by placing two strip-loaded waveguides, similar to the one described in Example 2.6, next to one another with a separation of $s = 1$ μm. Find the coupling length for the TM_{00} mode of the individual waveguide.

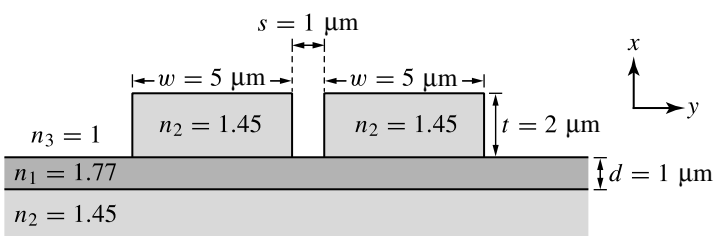


Figure 5.5 Symmetric directional coupler.

Solution Because the two waveguides are coupled to each other in the y direction, we have to examine their characteristics in that direction. In the y direction, the characteristics of TM_{00} for the strip waveguide are those of the TE_0 mode of the effective vertical waveguide in the effective index method, as discussed in Example 2.6. Therefore, the results in (5.36)–(5.38) for symmetric directional coupling between TE modes apply to this problem by using the mode parameters for TM_{00} listed in the table of Example 2.6: $\beta = 10.8120\ \mu\mathrm{m}^{-1}$, $h = 0.435\ \mu\mathrm{m}^{-1}$, and $\gamma = 0.832\ \mu\mathrm{m}^{-1}$. In applying these formulas, however, we have to replace d with $w = 5$ μm and d_{E} with $w_n = 7.4$ μm

because those are the parameters of the two identical waveguides in the y direction. Using these parameters and $s = 1$ μm, we find that $\tilde{\kappa}_{aa} = \tilde{\kappa}_{bb} = 2.7 \times 10^{-4}$ μm^{-1}, $\tilde{\kappa} = \tilde{\kappa}_{ab} = \tilde{\kappa}_{ba}^* = 1.943 \times 10^{-3}$ μm^{-1}, and $c = c_{ab} = c_{ba}^* = 2.572 \times 10^{-2}$.

From (5.39), we find the coupling coefficient to be $\kappa = 1.937 \times 10^{-3}$ μm^{-1}. Therefore, the coupling length of this directional coupler is

$$l_c = \frac{\pi}{2|\kappa|} = 811 \text{ μm}.$$

Supermodes

The variation of the field amplitudes in the two coupled waveguides of a directional coupler as a function of propagation distance is given by (4.62) and (4.63), which are the solutions for codirectionally coupled modes with initial conditions $\tilde{A}(0) \neq 0$ and $\tilde{B}(0) = 0$. The complete field profile across the directional coupler can be obtained as a combination of the two mode fields. Substituting (4.62) and (4.63) into (4.52) for the mode expansion coefficients $A(z)$ and $B(z)$, we have the total field in the directional coupler:

$$\begin{aligned}
\mathbf{E}(\mathbf{r}) &= \tilde{A}(0)\left[\hat{\boldsymbol{\mathcal{E}}}_a\left(\cos\beta_c z - \frac{\mathrm{i}\delta}{\beta_c}\sin\beta_c z\right)\mathrm{e}^{\mathrm{i}(\beta_a+\kappa_{aa}+\delta)z} + \hat{\boldsymbol{\mathcal{E}}}_b\frac{\mathrm{i}\kappa_{ba}}{\beta_c}\sin\beta_c z\,\mathrm{e}^{\mathrm{i}(\beta_b+\kappa_{bb}-\delta)z}\right] \\
&= \tilde{A}(0)\left[\frac{(\beta_c-\delta)\hat{\boldsymbol{\mathcal{E}}}_a + \kappa_{ba}\hat{\boldsymbol{\mathcal{E}}}_b}{2\beta_c}\mathrm{e}^{\mathrm{i}(\overline{\beta}+\beta_c)z} + \frac{(\beta_c+\delta)\hat{\boldsymbol{\mathcal{E}}}_a - \kappa_{ba}\hat{\boldsymbol{\mathcal{E}}}_b}{2\beta_c}\mathrm{e}^{\mathrm{i}(\overline{\beta}-\beta_c)z}\right] \\
&= \tilde{A}(0)\left[\boldsymbol{\mathcal{E}}_1(x,y)\mathrm{e}^{\mathrm{i}\beta_1 z} + \boldsymbol{\mathcal{E}}_2(x,y)\mathrm{e}^{\mathrm{i}\beta_2 z}\right],
\end{aligned} \tag{5.41}$$

where

$$\overline{\beta} = \frac{(\beta_a+\kappa_{aa}) + (\beta_b+\kappa_{bb})}{2}, \tag{5.42}$$

$$\boldsymbol{\mathcal{E}}_1 = \frac{(\beta_c-\delta)\hat{\boldsymbol{\mathcal{E}}}_a + \kappa_{ba}\hat{\boldsymbol{\mathcal{E}}}_b}{2\beta_c}, \qquad \boldsymbol{\mathcal{E}}_2 = \frac{(\beta_c+\delta)\hat{\boldsymbol{\mathcal{E}}}_a - \kappa_{ba}\hat{\boldsymbol{\mathcal{E}}}_b}{2\beta_c}, \tag{5.43}$$

and

$$\beta_1 = \overline{\beta} + \beta_c, \qquad \beta_2 = \overline{\beta} - \beta_c. \tag{5.44}$$

We see in (5.41) that the total field in the coupler is a linear combination of two independent field patterns $\boldsymbol{\mathcal{E}}_1(x, y)$ and $\boldsymbol{\mathcal{E}}_2(x, y)$ propagating with different propagation constants β_1 and β_2, respectively. They are the normal modes of the composite two-waveguide structure of the directional coupler. Such modes are known as the *supermodes* of the structure. Note that $\boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ given in (5.43) are not normalized.

The characteristics of the supermodes clearly depend on the parameters δ, κ_{ab}, and κ_{ba}. We first consider an asymmetric directional coupler with nonidentical waveguides

for which $\delta \neq 0$. When $\delta^2 \gg \kappa_{ab}\kappa_{ba}$ and $\delta > 0$, we have $\beta_c \to \delta$. As a result, $\beta_1 \to \beta_b + \kappa_{bb} \to \beta_b$ and $\beta_2 \to \beta_a + \kappa_{aa} \to \beta_a$. Therefore, the supermodes of the composite structure of the directional coupler are just those of the individual waveguides when phase mismatch is large. This situation is expected because in the limit that $\delta^2 \gg \kappa_{ab}\kappa_{ba}$, the two waveguides are effectively decoupled, and a wave propagating in either one of them is not to be affected by the existence of the other. This can be seen from the fact that in this limit, (5.43) reduces to

$$\boldsymbol{\mathcal{E}}_1 \to \frac{\kappa_{ba}}{2\delta}\hat{\boldsymbol{\mathcal{E}}}_b \approx 0, \qquad \boldsymbol{\mathcal{E}}_2 \to \hat{\boldsymbol{\mathcal{E}}}_a - \frac{\kappa_{ba}}{2\delta}\hat{\boldsymbol{\mathcal{E}}}_b \approx \hat{\boldsymbol{\mathcal{E}}}_a. \tag{5.45}$$

Therefore, the total field $\mathbf{E}(\mathbf{r})$ in (5.41) consists of approximately only the normal mode field of waveguide a propagating with β_a:

$$\mathbf{E}(\mathbf{r}) \approx \tilde{A}(0)\boldsymbol{\mathcal{E}}_2 e^{i\beta_2 z} \approx \tilde{A}(0)\hat{\boldsymbol{\mathcal{E}}}_a e^{i\beta_a z}. \tag{5.46}$$

This is a result of the fact that we have assumed the initial excitation of only waveguide a. No power is coupled to waveguide b throughout the length of the structure because of the large phase mismatch. If we assumed the initial excitation of waveguide b only, we would have $\boldsymbol{\mathcal{E}}_1 \to \hat{\boldsymbol{\mathcal{E}}}_b$ and $\boldsymbol{\mathcal{E}}_2 \to 0$. Then, the wave would propagate as the normal mode of waveguide b as if waveguide a did not exist. In the above discussion, we have assumed that $\beta_b > \beta_a$ so that $\delta > 0$. In the case of $\beta_a > \beta_b$ so that $\delta < 0$, the conclusion is the same, but with the asymptotic connection of $\boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ to $\hat{\boldsymbol{\mathcal{E}}}_b$ and $\hat{\boldsymbol{\mathcal{E}}}_a$ and that of β_1 and β_2 to β_b and β_a simply interchanged.

In a strong coupling situation, where $\kappa_{ab}\kappa_{ba} > \delta^2$, we find from (5.44) and (5.42) that

$$\beta_1 > \beta_b + \kappa_{bb} > \beta_a + \kappa_{aa} > \beta_2 \tag{5.47}$$

if $\delta > 0$. As a result, both $\boldsymbol{\mathcal{E}}_1$ and $\boldsymbol{\mathcal{E}}_2$ have significant contributions from both $\hat{\boldsymbol{\mathcal{E}}}_a$ and $\hat{\boldsymbol{\mathcal{E}}}_b$. Then, the supermodes are linear combinations of individual waveguide modes. At the input location $z = 0$, the two supermodes are in phase, and

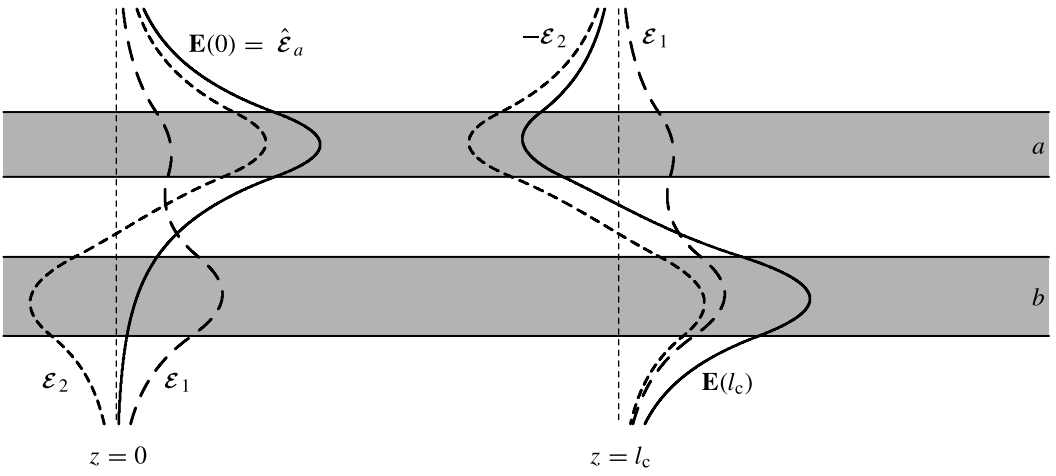
$$\mathbf{E}(x, y, 0) = \tilde{A}(0)(\boldsymbol{\mathcal{E}}_1 + \boldsymbol{\mathcal{E}}_2) = \tilde{A}(0)\hat{\boldsymbol{\mathcal{E}}}_a, \tag{5.48}$$

as expected from the fact that only waveguide a is initially excited. The maximum power transfer occurs when the two supermodes have a π phase shift. This takes place at a distance of

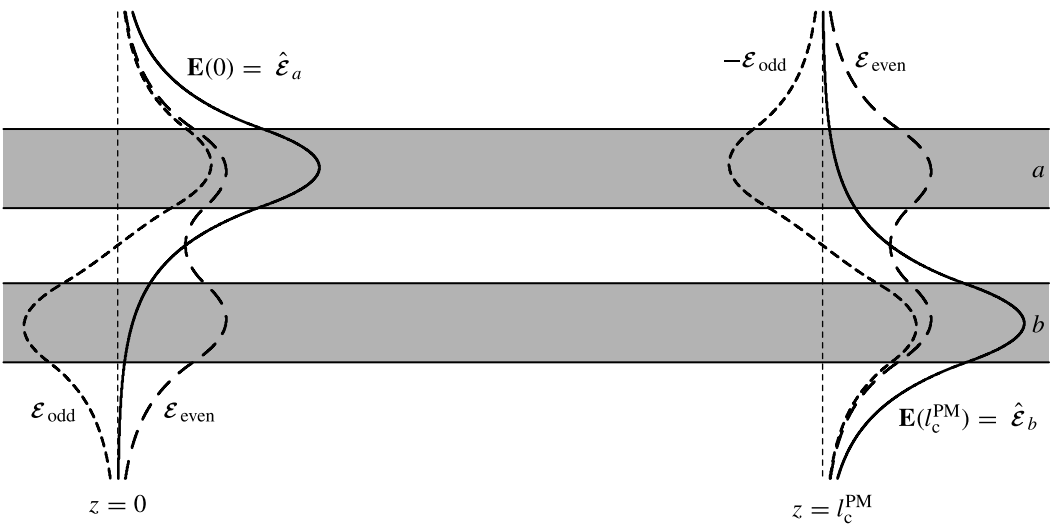
$$l_c = \frac{\pi}{\beta_1 - \beta_2} = \frac{\pi}{2\beta_c}, \tag{5.49}$$

which is exactly the coupling length given in (4.67) obtained from solution of the coupled-mode equations. At this distance, we find that the total field is

$$\mathbf{E}(x, y, l_c) = \tilde{A}(0)(\boldsymbol{\mathcal{E}}_1 - \boldsymbol{\mathcal{E}}_2)e^{i\beta_1 l_c} = \tilde{A}(0)\frac{-\delta\hat{\boldsymbol{\mathcal{E}}}_a + \kappa_{ba}\hat{\boldsymbol{\mathcal{E}}}_b}{\beta_c}e^{i\beta_1 l_c} \tag{5.50}$$



(a)



(b)

Figure 5.6 Evolution of supermode fields (dashed/dotted curves) and total fields (solid curves) in (a) an asymmetric and (b) a symmetric dual-channel directional coupler. Structural parameters used to calculate these field profiles are $n_1 = 1.5$, $n_2 = n_3 = 1.46$, $d_a = 0.8$ μm for (a) and $d_a = 1$ μm for (b), and $d_b = s = 1$ μm. Note that the coupling length for symmetric coupling is l_c^{PM}, which is larger than that for asymmetric coupling; thus $l_c^{\text{PM}} > l_c$, as seen in this figure.

because $\beta_2 l_c = \beta_1 l_c - \pi$. As expected, the power transfer to waveguide b is not complete because $\delta \neq 0$. This scenario is illustrated in Fig. 5.6(a).

We now consider the case of a symmetric coupler where the two waveguides are identical. Then, $\beta_a = \beta_b$, $\kappa_{aa} = \kappa_{bb}$, and $\kappa_{ab} = \kappa_{ba}^* \equiv \kappa$, where κ_{ab} is real and positive

as can be seen from (5.37). In addition, $\delta = 0$, $\beta_c = \kappa$, and $\overline{\beta} = \beta_a + \kappa_{aa} = \beta_b + \kappa_{bb}$. Therefore, the supermodes become the *even* and *odd* modes

$$\mathcal{E}_1 = \frac{\hat{\mathcal{E}}_a + \hat{\mathcal{E}}_b}{2} \equiv \mathcal{E}_{\text{even}}, \qquad \mathcal{E}_2 = \frac{\hat{\mathcal{E}}_a - \hat{\mathcal{E}}_b}{2} \equiv \mathcal{E}_{\text{odd}}, \tag{5.51}$$

with the following propagation constants:

$$\beta_1 = \overline{\beta} + \kappa \equiv \beta_{\text{even}}, \qquad \beta_2 = \overline{\beta} - \kappa \equiv \beta_{\text{odd}}. \tag{5.52}$$

Note again that $\mathcal{E}_{\text{even}}$ and $\mathcal{E}_{\text{odd}}$ as given in (5.51) are not normalized. The total field in the coupler following initial input to waveguide a only is then

$$\mathbf{E}(\mathbf{r}) = \tilde{A}(0)[\mathcal{E}_{\text{even}}(x, y)e^{i\beta_{\text{even}}z} + \mathcal{E}_{\text{odd}}(x, y)e^{i\beta_{\text{odd}}z}]. \tag{5.53}$$

It can be seen that the coupling length is now given by

$$l_c^{\text{PM}} = \frac{\pi}{\beta_{\text{even}} - \beta_{\text{odd}}} = \frac{\pi}{2\kappa}, \tag{5.54}$$

which is consistent with that given in (4.86) for phase-matched codirectional coupling. Complete power transfer between the two waveguides is now accomplished at the coupling length because of perfect phase matching. Figure 5.6(*b*) illustrates the evolution of supermode fields in this situation. It can be seen from (5.51) and Fig. 5.6(*b*) that the even supermode has a symmetric field pattern while the odd supermode has an antisymmetric field pattern. Because κ is positive real, we also have $\beta_{\text{even}} > \beta_{\text{odd}}$.

For two waveguides of given structural parameters and index profiles, the coupling coefficient depends solely on the proximity between them, as can be seen from (5.33)–(5.38). As the spacing between the two waveguides is reduced, the coupling becomes stronger and, according to (5.52), the disparity between the propagation constants of the two supermodes is increased, resulting in a shorter coupling length.

EXAMPLE 5.5 Find the propagation constants for the even and odd supermodes of the directional coupler described in Example 5.4. What is the coupling length found from the beat length of these two supermodes?

Solution Because $c_{aa} = c_{bb} = 1$ for a directional coupler, the self-coupling coefficients can be found using (4.47) and the parameters found in Example 5.4 as

$$\kappa_{aa} = \kappa_{bb} = \frac{\tilde{\kappa}_{aa} - c\tilde{\kappa}}{1 - |c|^2} = 2.2 \times 10^{-4}\ \mu\text{m}^{-1}.$$

Because $\beta_a = \beta_b = 10.8120\ \mu\text{m}^{-1}$ and $\kappa_{aa} = \kappa_{bb}$, we find that $\overline{\beta} = \beta_a + \kappa_{aa} = \beta_b + \kappa_{bb} = 10.812\ 22\ \mu\text{m}^{-1}$. Therefore, the propagation constants for the even and odd supermodes are, respectively,

$$\beta_{\text{even}} = \overline{\beta} + \kappa = 10.814\ 157\ \mu\text{m}^{-1} \quad \text{and} \quad \beta_{\text{odd}} = \overline{\beta} - \kappa = 10.810\ 063\ \mu\text{m}^{-1},$$

where $\kappa = 1.937 \times 10^{-3}\ \mu\text{m}^{-1}$ found in Example 5.4. The beat length is

$$l_c^{\text{PM}} = \frac{\pi}{\beta_{\text{even}} - \beta_{\text{odd}}} = \frac{\pi}{2\kappa} = 811\ \mu\text{m},$$

which is exactly the coupling length found in Example 5.4, as expected.

Asymmetric directional couplers

In a dual-channel directional coupler, the phase mismatch δ is determined by the symmetry between the two individual waveguides. In an asymmetric directional coupler, the two waveguides are not identical. In this case, $\delta \neq 0$ in general. Nevertheless, it is possible to make $\delta = 0$ for two nonidentical waveguides at a particular optical frequency, but not at all frequencies, by compensating for the difference in their thicknesses with a proper difference in their index profiles. Because of the different dispersion characteristics of the two nonidentical waveguides, $\delta \neq 0$ at frequencies away from the phase-matching frequency. Therefore, such a coupler can be used as a frequency filter similar to the function of a DBR, but for copropagating waves rather than for contrapropagating waves as in the case of a DBR.

Similarly to the discussions for the DBR, the frequency selectivity of a phase-matched asymmetric directional coupler can be illustrated by considering the dispersion characteristics shown in Fig. 5.7. The waveguides in the coupler are designed so that $\beta_b + \kappa_{bb} = \beta_a + \kappa_{aa}$ at a desired optical frequency ω_0 located at the crossing of the two dispersion curves in Fig. 5.7. Therefore, $\delta(\omega_0) = 0$ for perfect phase matching is

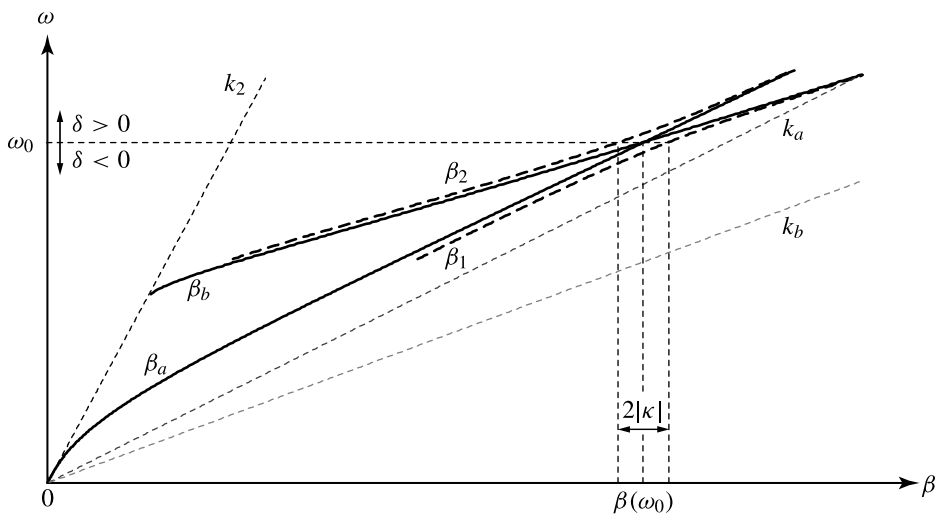


Figure 5.7 Dispersion relation showing coupling of the fields in an asymmetric dual-channel directional coupler. In the range $|\delta| < |\kappa|$, the two waveguides are well coupled. Mixing of the waveguide modes corresponding to β_a and β_b then results in supermodes corresponding to β_1 and β_2.

accomplished at ω_0. The two modes $\hat{\boldsymbol{\mathcal{E}}}_a$ and $\hat{\boldsymbol{\mathcal{E}}}_b$ are relatively well coupled within the range $|\delta| \leq |\kappa|$, where $|\kappa| = \sqrt{\kappa_{ab}\kappa_{ba}}$ for coupling in the asymmetric directional coupler. It can be seen from Fig. 5.7 that within this phase-matched range, mixing between $\hat{\boldsymbol{\mathcal{E}}}_a$ and $\hat{\boldsymbol{\mathcal{E}}}_b$ results in supermodes whose propagation constants β_1 and β_2 are different from either β_a or β_b. Outside this range, the modes in the two waveguides are effectively decoupled, and the supermodes effectively reduce to the individual waveguide modes. It can be seen from Fig. 5.7 that in the frequency range where $\beta_b + \kappa_{bb} > \beta_a + \kappa_{aa}$ for $\delta > 0$, $\beta_1 \to \beta_b$ and $\beta_2 \to \beta_a$ at large phase mismatch, whereas in the range where $\beta_b + \kappa_{bb} < \beta_a + \kappa_{aa}$ for $\delta < 0$, $\beta_1 \to \beta_a$ and $\beta_2 \to \beta_b$ at large phase mismatch. This observation is consistent with that discussed above regarding the asymptotic behavior of the supermodes at large phase mismatches.

The frequency bandwidth of the coupler can be found by considering the frequency dependence of δ. We have

$$\begin{aligned}\delta(\omega) &= \frac{[\beta_b(\omega) + \kappa_{bb}] - [\beta_a(\omega) + \kappa_{aa}]}{2} \\ &= \frac{[\beta_b(\omega_0) + \kappa_{bb}] - [\beta_a(\omega_0) + \kappa_{aa}]}{2} + \frac{1}{2}\left(\frac{\mathrm{d}\beta_b}{\mathrm{d}\omega} - \frac{\mathrm{d}\beta_a}{\mathrm{d}\omega}\right)(\omega - \omega_0) + \cdots \\ &\approx \frac{1}{2}\left(\frac{\mathrm{d}\beta_b}{\mathrm{d}\omega} - \frac{\mathrm{d}\beta_a}{\mathrm{d}\omega}\right)(\omega - \omega_0),\end{aligned} \tag{5.55}$$

where the frequency dependence of κ_{aa} and κ_{bb} in the Taylor series expansion is ignored. It can be included if necessary. At the phase-matching point, the length of the coupler needed for complete transfer of power from waveguide a to waveguide b is one of the odd multiples of $l_{\mathrm{c}}^{\mathrm{PM}}$:

$$l = (2n+1)l_{\mathrm{c}}^{\mathrm{PM}}, \qquad n = 0, 1, 2, \ldots \tag{5.56}$$

If the length is chosen to be one given in (5.56), we have $\eta = 1$ at $\delta = 0$. Then, $\delta_{1/2}$ for $\eta = 1/2$ can be found from the root of the following equation (see Problem 5.2.6):

$$2\sin^2\left(|\kappa| l\sqrt{|\delta/\kappa|^2 + 1}\right) = |\delta/\kappa|^2 + 1, \tag{5.57}$$

where l is one of those given in (5.56). The FWHM frequency bandwidth is then given by

$$\Delta\omega = 4\left|\frac{\delta_{1/2}}{\mathrm{d}\beta_b/\mathrm{d}\omega - \mathrm{d}\beta_a/\mathrm{d}\omega}\right|. \tag{5.58}$$

Rather than solving (5.57), the value of $|\delta_{1/2}|$ can also be found by reading the value of $|\delta/\kappa|$ for $\eta = 1/2$ on the curve in Fig. 4.7(b) corresponding to a given length of the coupler. It is seen that for a fixed value of $|\kappa|$, the frequency bandwidth is narrower for a length corresponding to a higher multiple of $l_{\mathrm{c}}^{\mathrm{PM}}$. For example, a coupler with $l = 3l_{\mathrm{c}}^{\mathrm{PM}}$ has a narrower bandwidth than one with $l = l_{\mathrm{c}}^{\mathrm{PM}}$.

By taking $|\mathrm{d}\beta_b/\mathrm{d}\omega - \mathrm{d}\beta_a/\mathrm{d}\omega| = \Delta N_\beta/c$, where ΔN_β is the effective group index difference between the two waveguide modes at the coupling wavelength, the bandwidth given in (5.58) is approximately bounded within the range (see Problem 5.2.6)

$$3.2\frac{|\kappa|c}{\Delta N_\beta} \geq \Delta\omega > 0, \tag{5.59}$$

where the equals sign for the upper limit of the bandwidth applies only when $l = l_c^{PM}$ for $|\kappa|l = \pi/2$.

Symmetric directional couplers

In an ideal symmetric directional coupler, the two waveguides are identical, and the modes are always phase matched. The coupling efficiency is then simply that given by (4.85). For a desired coupling efficiency η_{PM}, the length of the coupler has to be

$$l = \frac{1}{\kappa}\left(n\pi \pm \sin^{-1}\sqrt{\eta_{PM}}\right) = 2\left(n \pm \frac{1}{\pi}\sin^{-1}\sqrt{\eta_{PM}}\right) l_c^{PM}. \tag{5.60}$$

For complete transfer of power, the length of the coupler has to be exactly one of the odd multiples of l_c^{PM} given in (5.56), as discussed above. It is interesting to see that for a 50% coupling efficiency, we need $l = (n + 1/2)l_c^{PM}$, where $n = 0, 1, 2, \ldots$, and the shortest coupler length needed is exactly $l_c^{PM}/2$, as can be seen in Fig. 4.4(*b*). Launching optical power into one waveguide of such a coupler at its input end results in equal division of power between the two waveguides at the output end. Thus, the device functions as a 3-dB coupler or as a 50 : 50 *power divider*. Any desired coupling efficiency can be obtained by properly choosing the length of the coupler for a given value of κ or by choosing a proper value of κ through the design of the coupler for a given length.

In many applications, it is often necessary to vary the coupling efficiency in the operation of a device for certain purposes. In some situations, this objective can be accomplished by altering the effective length of the coupler or the spacing between the waveguides. This approach is possible if the coupler is not integrated. An example is a directional coupler made of two closely placed optical fibers whose interaction length and spacing can be adjusted. If the coupler has an integrated structure, it is certainly not easy to vary the length of the coupler at will, nor is it convenient to vary the spacing between the waveguides. In this situation, changes in the coupling efficiency of a coupler can be accomplished either by altering the value of κ or by varying the propagation constants β_a and β_b in the two waveguides by different amounts to introduce a finite phase mismatch δ in the coupler. In fact, a change in κ will necessarily result in changes in the propagation constants, and vice versa. Nevertheless, changing the value of κ, and thus the values of β_a and β_b, does not necessarily result in a phase mismatch although creating a phase mismatch in an originally symmetric coupler will

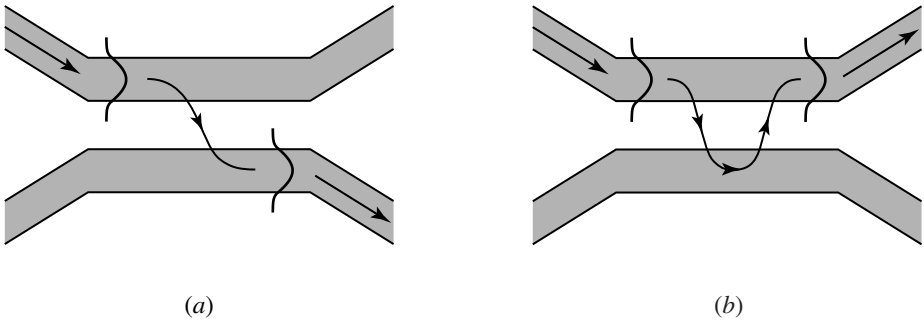


Figure 5.8 (*a*) Cross state and (*b*) parallel state of a directional-coupler optical switch.

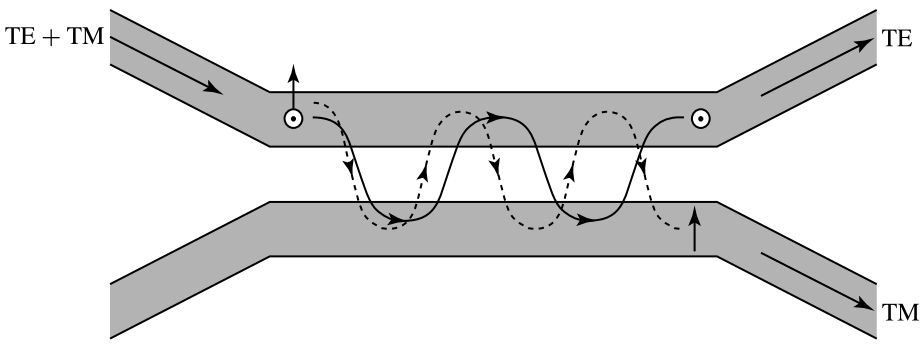


Figure 5.9 Schematic illustration of the use of a directional coupler as a TE–TM polarization splitter. In this example, TE polarization is in the parallel state, while TM polarization is in the cross state.

certainly cause changes in the coupling coefficients, as indicated by (4.44). In practical devices, these changes can be created through the electro-optic effect, thus being controllable with an externally applied voltage, through nonlinear optical effects, thus being controllable with the optical power in the waveguides, or through any other effects that cause changes in the refractive index of the medium of the coupler.

The ability to vary the coupling efficiency of a directional coupler through a control signal results in many useful applications. An important example is an *optical switch* that functions between the *cross state*, in which power is completely transferred from the input channel to the other channel at the output, and the *parallel state*, in which power is completely passed through the input channel at the output without any transfer to the other channel. These states are illustrated in Fig. 5.8. The cross state is denoted by $\otimes$, while the parallel state is denoted by $\ominus$. The parallel state is also called the *bar state*. The function of electro-optically controlled directional coupler switches is discussed in Section 6.4.

Another interesting example is the TE–TM *polarization splitter* illustrated in Fig. 5.9. It is possible to create a difference between the coupling coefficients, κ_{EE} and κ_{MM}, of the TE and TM modes, respectively, even though the coupler has a symmetric structure.

This can be accomplished by fabricating the coupler in a birefringent medium such as $LiNbO_3$ or a nonbirefringent electro-optic material such as GaAs and by applying a voltage to adjust properly the different refractive indices seen by the TE and TM fields. For a coupler of length l, polarization splitting as shown in Fig. 5.9, where TE polarization is in the parallel state while TM polarization is in the cross state, is achieved when

$$l = \frac{n\pi}{\kappa_{EE}} = \frac{(2n \pm 1)\pi}{2\kappa_{MM}} \tag{5.61}$$

for an integer n. This is possible if there is a difference between the coupling coefficients of the two different polarizations of the amount

$$\Delta\kappa \equiv |\kappa_{MM} - \kappa_{EE}| = \frac{\kappa_{EE}}{2n}. \tag{5.62}$$

For polarization splitting resulting in TE polarization in the cross state and TM polarization in the parallel state, we need $\Delta\kappa = \kappa_{MM}/2n$ and $l = n\pi/\kappa_{MM}$ instead.

5.3 Surface input and output couplers

In a system, it is always necessary to couple light from sources, such as lasers or light-emitting diodes, to transmission components, which are usually dielectric waveguides or fibers, to various functional devices, such as optical switches, power dividers, amplifiers, and modulators, possibly through transmission components again, and ultimately to photodetectors. The approaches to coupling light in and out of optical waveguides, including fibers, are basically classified into two categories: (1) *surface coupling*, also called *longitudinal coupling*, and (2) *end coupling*, also called *end-fire coupling* or *transverse coupling*. The first approach relies on the coupling of an optical wave in or out of a waveguide through the longitudinal surface of the waveguide. In the second approach, the optical wave is coupled directly through an exposed cross section at one end of the waveguide. In this section, we examine the surface couplers.

From the viewpoint of coupled-mode theory, coupling of an optical wave through the longitudinal surface of a waveguide into a guided mode is an effect of coupling between the radiation modes of the waveguide and the guided mode. Any unguided propagating field such as that of the incident optical wave to be coupled into the waveguide can be expanded in terms of the radiation modes of the waveguide in the form of (4.23) and (4.24) with the summation replaced by integration over all radiation modes. For efficient coupling, the phase-matching condition has to be satisfied, as we have seen time and again in the discussions of the preceding sections. Phase matching is not possible, however, without some special arrangements to perturb the system because the longitudinal propagation constant of any radiation mode is always smaller than that of a guided mode, as discussed in Section 2.1 and illustrated in Fig. 2.3. This difficulty

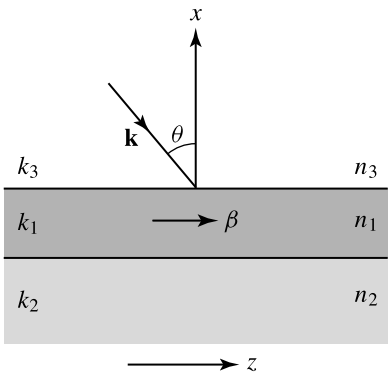


Figure 5.10 Illustration of the phase mismatch between a guided mode and a free-propagating field incident at an angle θ on the surface of the waveguide.

can be seen from the illustration in Fig. 5.10. For a beam incident on the surface of the guiding layer at an angle θ, the largest propagation constant is that of a plane wave, $k = k_3$. Its longitudinal propagation constant is the z component, which satisfies

$$k_z = k_3 \sin\theta < k_3 < \beta, \tag{5.63}$$

where β is the propagation constant of a guided mode. By the same argument and by the reciprocity theorem for electromagnetic waves, it is equally impossible to couple the field in a guided mode out of the waveguide through the surface of the waveguide without some special arrangements.

The task of a surface coupler is to change this situation and to accomplish phase matching so that a radiation field can be coupled to the guided mode. Naturally, the same coupler can be used as an output coupler to couple a guided field out to a radiation field.

Prism couplers

One approach to accomplishing phase matching is to use a prism of high index of refraction, n_p, as a surface coupler, as shown in Fig. 5.11. For this arrangement, the cover of the waveguide is usually air or some low-index fluid filling the gap s between the prism and the waveguide core. For $k_\mathrm{p} = n_\mathrm{p}\omega/c$, the phase-matching condition is then

$$k_\mathrm{p} \sin\theta = \beta, \tag{5.64}$$

which can be accomplished if

$$n_\mathrm{p} > n_1 > n_3 \quad \text{and} \quad \theta > \theta_\mathrm{c} = \sin^{-1}\frac{n_3}{n_\mathrm{p}}. \tag{5.65}$$

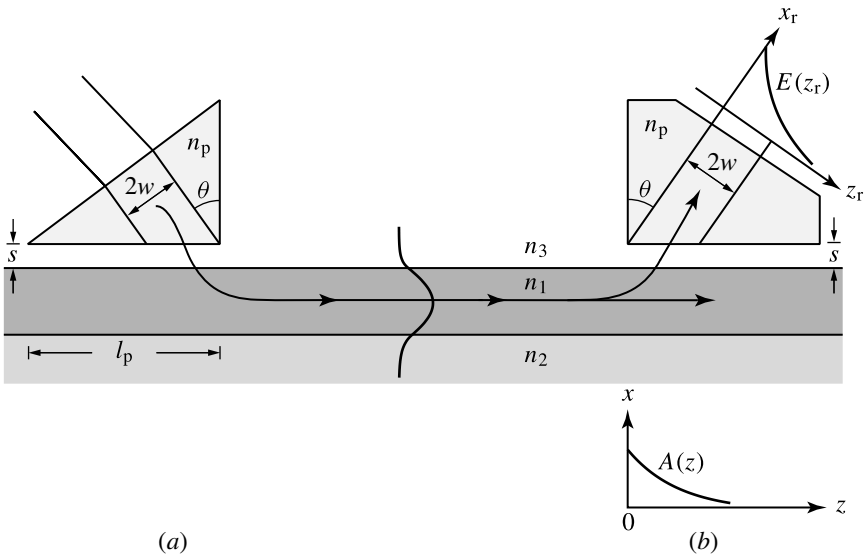


Figure 5.11 (*a*) Input and (*b*) output coupling using prism couplers.

Note that the angle θ is now measured inside the prism. It is necessary that total internal reflection occurs for the incident field inside the prism. As a result, the field does not propagate freely in the gap between the prism and the waveguide core. Then, the difficulty of phase matching due to (5.63) can be avoided. The total internal reflection results in an exponentially decaying evanescent field in the gap between the prism and the waveguide. Coupling to the waveguide mode occurs through *optical tunneling* when this evanescent field overlaps with the field of a guided mode if the phase-matching condition in (5.64) is satisfied.

EXAMPLE 5.6 A rutile prism is used for surface coupling of optical waves at $\lambda = 1$ μm into the guided modes of the waveguide described in Example 2.1. Rutile is a positive uniaxial crystal that has $n_o = 2.4585$ and $n_e = 2.7495$ at $\lambda = 1$ μm. The optical wave to be coupled into the waveguide is incident on the prism surface at an incident angle of θ_i, as shown in Fig. 5.12. The prism is cut with an angle of $\zeta = 45^\circ$ and with its optical axis perpendicular to the plane of incidence as shown. What are the polarization of the incident optical wave and the angle θ_i for coupling into the TE_0 mode? What are they for coupling into the TM_0 mode?

Solution The propagation constants for the TE_0 and TM_0 modes are $\beta_{TE} = 10.8432$ μm^{-1} and $\beta_{TM} = 10.7800$ μm^{-1}, both found in Example 2.1. At $\lambda = 1$ μm in rutile, $k_p^o = 2\pi n_o/\lambda = 15.6187$ μm^{-1} for the ordinary wave and $k_p^e = n_e/\lambda = 17.2756$ μm^{-1} for the extraordinary wave. The optical wave is incident to the prism from the air with $n_3 = 1$.

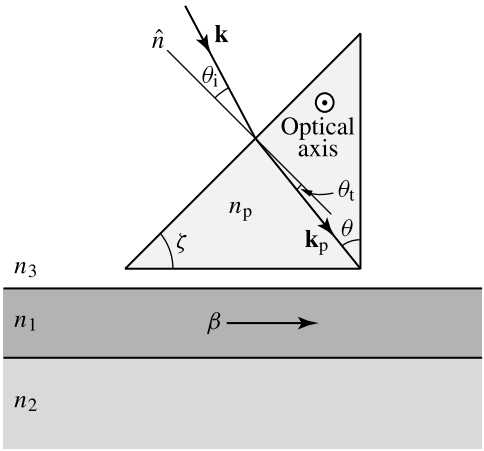


Figure 5.12 Prism for surface coupling.

For coupling into the TE_0 mode, the incident wave has to be TE polarized, in a direction perpendicular to the plane of incidence, in order to match the polarization of the guided TE_0 mode. This polarization is the extraordinary polarization in the given rutile prism. Therefore, k_p^e is used for $k_p^e \sin\theta = \beta_{TE}$ to find that $\theta = 38.87^\circ$ with the given parameters. We check that the critical angle $\theta_c^e = \sin^{-1}(n_3/n_e) = 21.33^\circ < \theta$, as expected. From Fig. 5.12, we find that the angle of refraction θ_t for the wave transmitted into the prism is $\theta_t = \zeta - \theta = 6.13^\circ$. Therefore, the angle of incidence is

$$\theta_i = \sin^{-1}\left(\frac{n_e}{n_3}\sin\theta_t\right) = 17.07^\circ.$$

For coupling into the TM_0 mode, the incident wave has to be TM polarized, in the plane of incidence, in order to match the polarization of the guided TM_0 mode. This polarization is the ordinary polarization in the given rutile prism. Therefore, k_p^o is used for $k_p^o \sin\theta = \beta_{TM}$ to find that $\theta = 43.65^\circ$ with the given parameters. We check that the critical angle $\theta_c^o = \sin^{-1}(n_3/n_o) = 23.72^\circ < \theta$, as expected. We then find that $\theta_t = \zeta - \theta = 1.35^\circ$. Therefore, the angle of incidence is

$$\theta_i = \sin^{-1}\left(\frac{n_o}{n_3}\sin\theta_t\right) = 3.36^\circ.$$

The coupling efficiency and other characteristics of the prism coupler can be analyzed with coupled-mode theory or other approaches of field analysis such as a leaky-wave analysis. The mathematics is quite involved; therefore, we only discuss some key characteristics without detailed derivations.

It is more straightforward to understand the output coupling from a guided mode to the radiation field through a prism coupler shown in Fig. 5.11(*b*) rather than the

input coupling shown in Fig. 5.11(*a*). We thus consider the output coupler first. For this purpose, we take $z = 0$ to be at the edge of the output-coupling prism. For a guided mode that has an amplitude of $A(0)$ at $z = 0$, the analysis using coupled-mode or leaky-wave theory indicates that the amplitude distribution along the z direction for $z > 0$ is

$$A(z) = A(0)\mathrm{e}^{-\alpha_\mathrm{r} z}, \tag{5.66}$$

where α_r is the *radiation decay constant*. The field coupled out to the prism has a pattern directly proportional to $A(z)$. For convenience, we assume that the prism is cut in such a manner that the output radiation field is normal to the exit surface of the prism as shown in Fig. 5.11(*b*). Then the output field has the following transverse pattern:

$$E_\mathrm{out}(z_\mathrm{r}) = \begin{cases} 0, & \text{for} \quad z_\mathrm{r} < 0, \\ E_0 \exp(-\alpha_\mathrm{r} z_\mathrm{r}/\cos\theta), & \text{for} \quad z_\mathrm{r} > 0, \end{cases} \tag{5.67}$$

where $z_\mathrm{r} = z\cos\theta$ is the coordinate transverse to the direction of propagation of the output field. The radiation decay constant has a complicated form but has the following key characteristics:

$$\alpha_\mathrm{r} \propto \frac{1}{d_\mathrm{eff}}\mathrm{e}^{-2\gamma_3 s}, \tag{5.68}$$

where d_eff is the effective waveguide thickness, $d_\mathrm{eff} = d_\mathrm{E}$ for a TE mode and $d_\mathrm{eff} = d_\mathrm{M}$ for a TM mode, and γ_3 is the transverse mode field parameter in the cover region defined by (2.52).

When the length of the prism facing the waveguide is l_p, the maximum output-coupling efficiency is

$$\eta_\mathrm{out} = 1 - \mathrm{e}^{-2\alpha_\mathrm{r} l_\mathrm{p}}, \tag{5.69}$$

which can be obtained from (5.66). If l_p is sufficiently long that $\alpha_\mathrm{r} l_\mathrm{p} \gg 1$, an output-coupling efficiency that is nearly 100% can be achieved.

For input coupling, the situation is slightly more complicated than that of output coupling. Owing to the reciprocity theorem, input coupling is just the reverse of output coupling. Therefore, an input efficiency of nearly 100% can be accomplished if an input beam propagating reversely with respect to the output beam shown in Fig. 5.11(*b*) has exactly the same exponentially decaying profile as that given in (5.67). In practice, however, most of the input fields do not have such a profile. As a result, the coupling efficiency is reduced by the *overlap factor* Γ of the two field patterns:

$$\eta_\mathrm{in} = \Gamma \eta_\mathrm{out}, \tag{5.70}$$

where

$$\Gamma = \frac{\left[\int_{-\infty}^{\infty} E_{\text{out}}^{*}(z_{\text{r}})E_{\text{in}}(z_{\text{r}})\text{d}z_{\text{r}}\right]^2}{\int_{-\infty}^{\infty} |E_{\text{out}}(z_{\text{r}})|^2\text{d}z_{\text{r}} \int_{-\infty}^{\infty} |E_{\text{in}}(z_{\text{r}})|^2\text{d}z_{\text{r}}}, \tag{5.71}$$

and $E_{\text{in}}(z_{\text{r}})$ is the transverse pattern of the input field. The overlap factor Γ depends on the profiles of E_{in} and E_{out}. For E_{out} having a pattern as given in (5.67), $\Gamma = 80.1\%$ if E_{in} has a Gaussian profile and $\Gamma = 81.4\%$ if E_{in} has a rectangular profile. Therefore, the maximum input-coupling efficiency is 80.1% for a Gaussian input beam and is 81.4% for a rectangular input beam profile. However, in order to obtain the maximum input-coupling efficiency, alignment is very critical. In addition to sending the beam exactly at an angle required by phase matching, the parallel position of the beam has to be such that the front boundary of the beam exactly intercepts the prism corner, as shown in Fig. 5.11(*a*). Also, the length l_{p} of the prism has to be such that $l_{\text{p}} > 2w/\cos\theta$ to intercept the entire beam and that $\alpha_{\text{r}} l_{\text{p}} \gg 1$.

For a given input or output coupler, the coupling efficiency depends on α_{r}. It can be seen from (5.68) that α_{r} depends sensitively on the spacing s between the prism and the waveguide. For efficient coupling, s typically has to be kept less than half the optical wavelength in the medium filling the gap. This small gap is usually accomplished by applying a large pressure to clamp the prism onto the waveguide. The sensitivity of the coupling efficiency to the variation of this gap requires critical adjustment to control coupling. This is one of the disadvantages of prism couplers. Furthermore, for good phase matching and good alignment, the input beam has to be well collimated. Another disadvantage is that the prism material must have a refractive index higher than that of the waveguide material. This requirement is particularly difficult to meet for a waveguide of high refractive index, such as one based on a semiconductor. The advantages of prism couplers are that they are noninvasive and that no permanent structures have to be fabricated on the waveguides. Another advantage is that it is easy to excite different waveguide modes selectively by choosing appropriate incident angles for phase matching. Prism couplers can be used for channel waveguides as well as planar waveguides. Therefore, prism couplers are most often used in the characterization of optical waveguides.

Grating couplers

Another approach to obtaining phase matching for the coupling between the radiation field and a guided mode is the use of a grating. Figure 5.13 shows input and output coupling of a planar waveguide using a grating surface coupler. Basically, the function of the grating is to provide an extra phase factor qK in a manner similar to that of

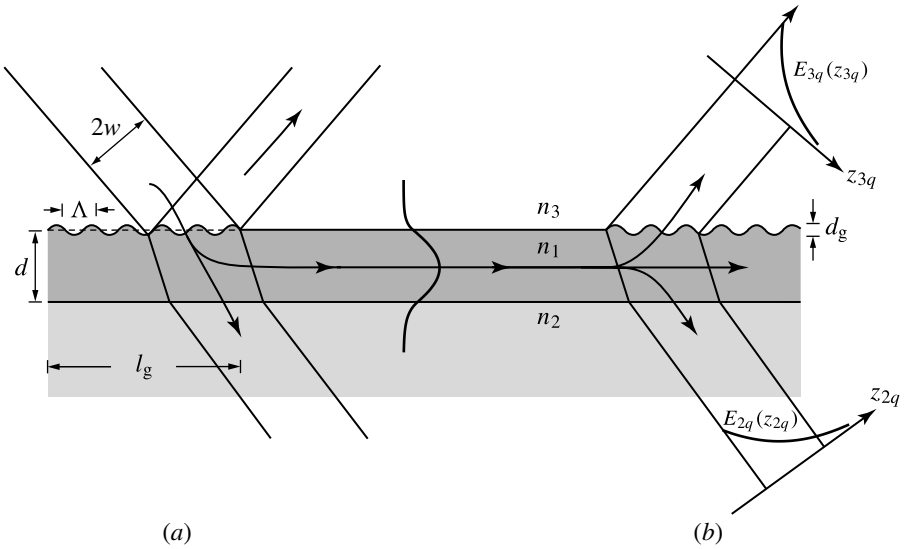


Figure 5.13 (*a*) Input and (*b*) output coupling using grating couplers.

the grating waveguide couplers discussed in Section 5.1 so that modes of different propagation constants can be phase matched and efficiently coupled. However, because the radiation fields are not restricted to a single propagation direction, the situation here is more complicated than that of the coupling between guided modes discussed in Section 5.1. The radiation fields can exist simultaneously in both the cover and the substrate regions, and the grating can scatter the light into different diffraction orders. Because the grating is periodic only along the z direction, the extra phase factor qK it provides is also only in the z direction. Therefore, the phase-matching condition for a guided mode with a propagation constant β is

$$k_{iq,z} + qK = \beta, \qquad i = 2, 3, \tag{5.72}$$

or

$$k_i \sin\theta_{iq} + qK = \beta, \qquad i = 2, 3, \tag{5.73}$$

where $i = 2$ for the radiation fields in the substrate region of refractive index n_2, $i = 3$ for those in the cover region of refractive index n_3, and θ_{iq} is the incident angle of $\mathbf{k}_{iq}$. Note that because $k_i \sin\theta_{iq} = k_{iq,z}$, the sign of θ_{iq} is positive if $k_{iq,z} = \mathbf{k}_{iq} \cdot \hat{z} > 0$ and is negative if $k_{iq,z} = \mathbf{k}_{iq} \cdot \hat{z} < 0$. Because $k_1 > \beta > k_2 > k_3$, phase matching is possible only for $q \geq 1$ in (5.73).

Similarly to the treatment of the prism coupler, we consider output coupling first, which is also more straightforward than input coupling. The field of a guided mode incident upon the grating region can be coupled to radiation fields in different diffraction orders that satisfy the phase-matching condition given by (5.73). The result can be output with multiple beams, two beams, or one beam, shown in Figs. 5.14(*a*), (*c*),

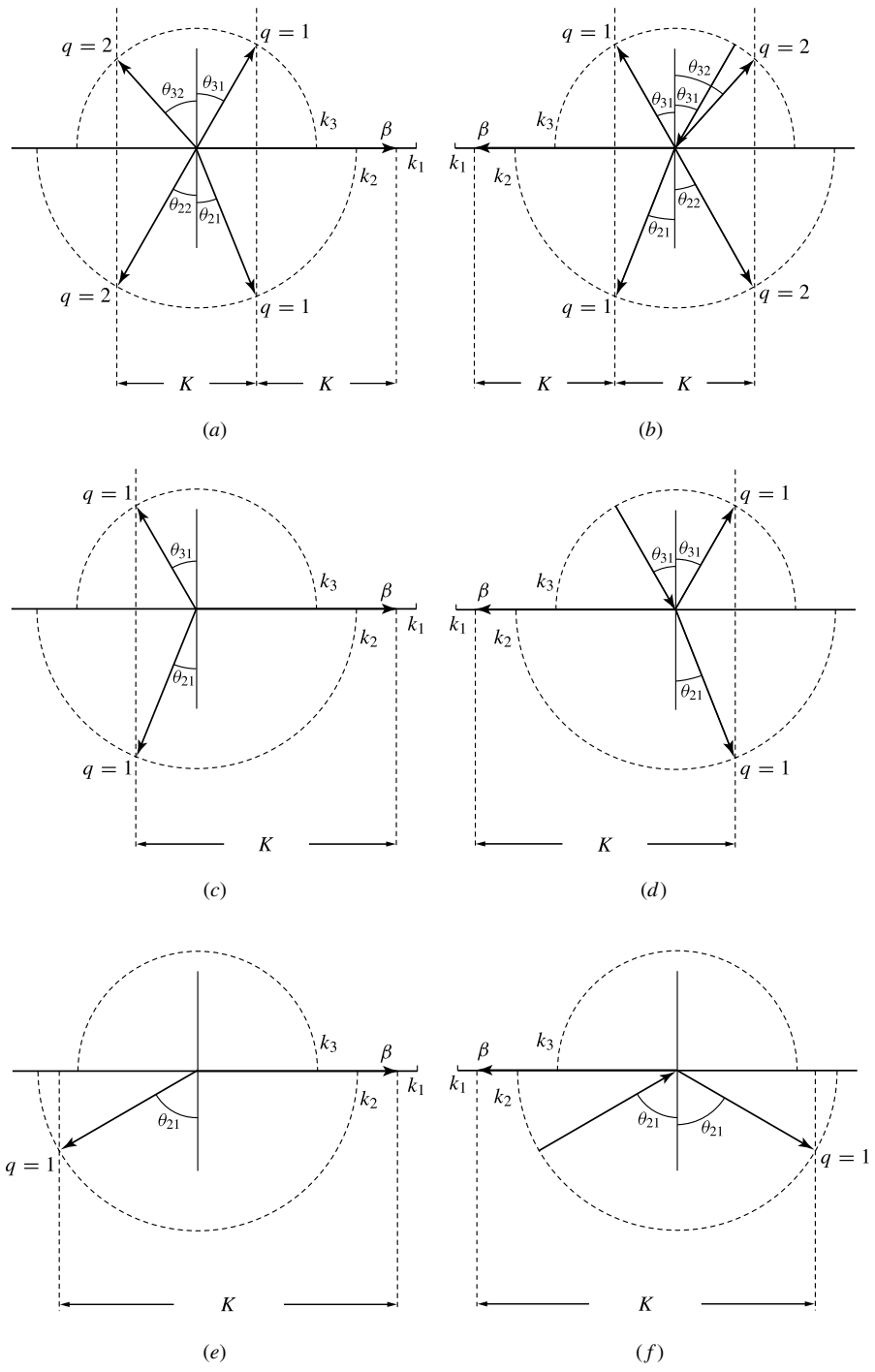


Figure 5.14 Phase-matched coupling between a guided mode and radiation fields: (*a*), (*c*), and (*e*) illustrate the conditions for output coupling into multiple beams, two beams, and one beam, respectively, which are determined by the value of K; (*b*), (*d*), and (*f*) show the conditions for input coupling corresponding to the reverse of (*a*), (*c*), and (*e*), respectively.

and (e), respectively. For a given waveguide and a given guided mode, the values of k_1, k_2, k_3, and β are all fixed. Therefore, the number of phase-matched output beams is simply determined by the grating wavenumber K or, equivalently, the grating period $\Lambda = 2\pi/K$. Because $n_2 > n_3$, a single-beam output as shown in Fig. 5.14(e) can be obtained on the substrate side by making K so large, or Λ so small, that phase matching is not allowed for any diffraction order in the cover region. A single-beam output on the cover side can be obtained in the same manner by making $n_3 > n_2$.

EXAMPLE 5.7 A first-order grating is fabricated on the surface of a polymer core layer of the waveguide described in Example 2.1 for surface coupling of an optical wave of 1 μm wavelength at an incident angle of 20° into the forward-propagating TE_0 mode of the waveguide. What is the period of the grating? How many diffracted beams are found on the cover and substrate sides? What are the directions of the diffracted beams?

Solution We have $\beta = 10.8432\ \mu m^{-1}$, $k_2 = 9.1106\ \mu m^{-1}$, and $k_3 = 6.2832\ \mu m^{-1}$ found in Example 2.1. The condition for first-order input coupling at an incident angle θ_i is $k_3 \sin\theta_i + K = \beta$. For $\theta_i = 20°$, we find that $K = \beta - k_3 \sin\theta_i = 8.6942\ \mu m^{-1}$. Therefore, the grating period is

$$\Lambda = \frac{2\pi}{K} = 722.7 \text{ nm}.$$

The diffracted beams are generated by output coupling of the guided mode field due to scattering by the grating. Therefore, they can be found by finding the phase-matched output coupling from the guided mode. On the cover side, the condition is $k_3 \sin\theta_{3q} + qK = \beta$, which leads to

$$6.2832 \sin\theta_{3q} = 10.8432 - 8.6942q.$$

This condition has only one solution: $\theta_{31} = 20°$ for $q = 1$. Therefore, there is only one diffracted beam on the cover side. At $\theta_{31} = 20°$, this beam is in the specular reflection direction of the incident beam as shown in Fig. 5.15. On the substrate side, the condition is $k_2 \sin\theta_{2q} + qK = \beta$, which leads to

$$9.1106 \sin\theta_{2q} = 10.8432 - 8.6942q.$$

This condition has two solutions: $\theta_{21} = 13.64°$ for $q = 1$, and $\theta_{22} = -45.9°$ for $q = 2$. Therefore, there are two diffracted beams on the substrate side, one from first-order diffraction and the other from second-order diffraction of the guided mode field by the grating. The same results can be obtained by graphic solution, as is shown in Fig. 5.15.

Similarly to the output coupling with a prism coupler, the amplitude of the guided mode also decays exponentially in the grating region:

$$A(z) = A(0)e^{-\alpha_r z}, \tag{5.74}$$

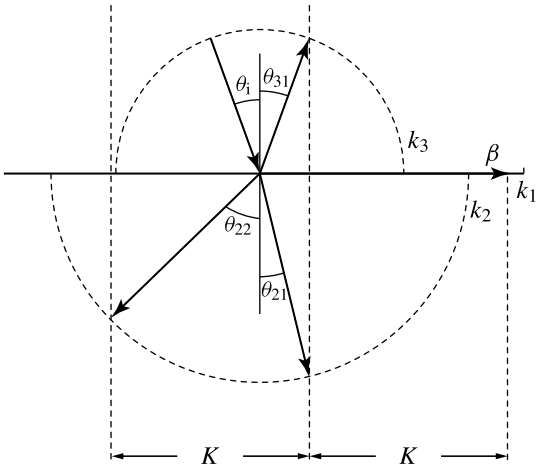


Figure 5.15 Phase-matching diagram of surface input coupling through a first-order grating.

where $z = 0$ marks the beginning of the grating, as shown in Fig. 5.13(b). In case there is more than one output beam, each output beam also has a pattern directly proportional to $A(z)$ along the surface of the grating in the z direction. Because the direction of each output beam is determined by a different angle θ_{iq}, the transverse field pattern across the cross section of an output beam is given by

$$E_{iq}^{\text{out}}(z_{iq}) = \begin{cases} 0, & \text{for} \quad z_{iq} < 0, \\ E_{iq}(0)\exp(-\alpha_r z_{iq}/\cos\theta_{iq}), & \text{for} \quad z_{iq} > 0, \end{cases} \tag{5.75}$$

where z_{iq} is the transverse coordinate along the cross section of the iq output beam. Meanwhile, the total radiation decay constant is the sum of the radiation decay constants for the coupling of all existing output beams:

$$\alpha_r = \sum_{i,q} \alpha_{iq}. \tag{5.76}$$

The radiation decay constants can be obtained from a coupled-mode or leaky-wave analysis. They also have a complicated form. However, their characteristics are very different from those of the prism coupler, as can be expected. The key characteristics are that for a grating of corrugation depth d_g,

$$\alpha_{iq} \propto \frac{d_g^2}{d_{\text{eff}}}, \quad \text{if } d_g \leq \frac{1}{\gamma_3}, \tag{5.77}$$

where d_{eff} is the effective waveguide thickness used in (5.68), and that the values of α_{iq} saturate for $d_g > 1/\gamma_3$. The dependence of α_{iq} on the indices i and q is determined by the shape and symmetry of the grating.

For a grating of length l_g, the output coupling efficiency to the beam of the indices i and q is

$$\eta_{iq}^{\text{out}} = \frac{\alpha_{iq}}{\alpha_r}\left(1 - e^{-2\alpha_r l_g}\right). \tag{5.78}$$

When $\alpha_r l_g \gg 1$, all of the power in a guided mode is coupled out but is divided among the existing output beams propagating in different directions.

Input coupling to a guided mode through a grating coupler can also be considered as the reverse of output coupling owing to the reciprocity theorem. However, again the situation is more complicated than that of the prism coupler because of the possibility of multiple diffraction orders. As a result, (1) there can be more than one incident direction for input coupling, and (2) the multiple-beam distribution for input coupling may look different from that of output coupling. For example, in the case when phase matching is possible for multiple beams, as shown in Fig. 5.14(a), input coupling can be accomplished by sending in the beam in a reverse direction to any one of the output beams or by sending in more than one beam along any or all of these directions. However, the pattern of the distribution of the diffracted beams, such as that shown in Fig. 5.14(b), looks different from that in the case of output coupling shown in Fig. 5.14(a). Figures 5.14(d) and (f) show examples of input coupling corresponding to the output coupling shown in Figs. 5.14(c) and (e), respectively. In any event, if an input field is incident along the reverse path of an output beam of field E_{iq} given by (5.75), the input coupling efficiency is simply

$$\eta_{iq}^{\text{in}} = \Gamma_{iq}\eta_{iq}^{\text{out}}, \tag{5.79}$$

where

$$\Gamma_{iq} = \frac{\left[\int_{-\infty}^{\infty} E_{iq}^*(z_{iq})E_{\text{in}}(z_{iq})\mathrm{d}z_{iq}\right]^2}{\int_{-\infty}^{\infty} |E_{iq}(z_{iq})|^2\mathrm{d}z_{iq}\int_{-\infty}^{\infty} |E_{\text{in}}(z_{iq})|^2\mathrm{d}z_{iq}}. \tag{5.80}$$

For an exponential beam profile of E_{iq} given in (5.75), $\Gamma_{iq} = 80.1\%$ if E_{in} is Gaussian and $\Gamma_{iq} = 81.4\%$ if it is rectangular.

In general, $\alpha_{2q} \approx \alpha_{3q}$ if the shape of the grating teeth is not highly asymmetric. Therefore, an output beam on the substrate side has about the same power as that of the same order on the cover side if both can be phase matched to the guided mode. If this grating is used as an input coupler, then about one-half of the incident power will be lost before reduction by the overlap factor Γ_{iq}, resulting in a low coupling efficiency. This shortcoming can be circumvented by using the one-beam coupler. One approach is that shown in Figs. 5.14(e) and (f). However, there are a few practical difficulties associated with this backward coupling through the substrate because the

substrate tends to have a high refractive index close to that of the waveguide core. A more practical approach is to use a *blazed grating* with highly asymmetric teeth, such as those of saw-tooth shape, to cause the major portion of power to be coupled to a single order on the cover side. Alignment of the input beam is not extremely critical, but it is necessary to align the front boundary of the beam to intersect with the edge of the grating for maximum efficiency, as shown in Fig. 5.13(a).

In comparison to the prism coupler, grating couplers are compact and stable because they are integrated with the waveguide structure. They are compatible with the integration technology of photonic and optoelectronic devices. With proper design, they can be used as efficient input and output couplers in innovative applications, such as vertical output coupling from a semiconductor laser without conventional mirrors. Grating-coupled surface-emitting lasers based on this concept are discussed in Section 13.9.

PROBLEMS

5.1.1 An index-modulation grating characterized by $\Delta n(x, z)$ that is periodic in z is incorporated into a planar waveguide defined by an index profile $n(x)$ to make a grating waveguide coupler for optical waves at an optical wavelength λ. The grating has a period Λ and a wavenumber $K = 2\pi/\Lambda$ as defined in (5.3).

a. Show that the qth Fourier component of the coupling coefficient as defined in (5.4) can be expressed as

$$\kappa_{ab}(q) = \frac{2\omega}{\Lambda}\epsilon_0 \int_0^{\Lambda} \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}x\; n(x)\Delta n(x, z)\hat{\boldsymbol{\mathcal{E}}}_a^*(x) \cdot \hat{\boldsymbol{\mathcal{E}}}_b(x)\mathrm{e}^{-\mathrm{i}qKz}. \tag{5.81}$$

b. If the waveguiding effect is very weak so that $n(x) \approx n$, show that the coupling coefficient between two TE modes for a coupler that has a sinusoidal index modulation of $\Delta n(x, z) = \Delta n \cos kz$ is

$$\kappa_{ab}(q) = \frac{\Delta n\pi}{\lambda}(\delta_{q,1} + \delta_{q,-1})\delta_{ab}. \tag{5.82}$$

c. With $n(x) \approx n$ and a square index modulation of duty factor ξ characterized by

$$\Delta n(x, z) = \begin{cases} \Delta n, & \text{for } \ 0 < z < \xi\Lambda, \\ -\Delta n, & \text{for } \ \xi\Lambda < z < \Lambda, \end{cases} \tag{5.83}$$

show that the coupling coefficient between two TE modes is

$$\kappa_{ab}(q) = \frac{4\Delta n}{\lambda}\frac{\sin \xi q\pi}{q}\mathrm{e}^{-\mathrm{i}\xi q\pi}\delta_{ab}. \tag{5.84}$$

5.1.2 A DBR mirror for a vertical-cavity surface-emitting laser (VCSEL) has the index profiles described in Problem 5.1.1(c). The index grating is a first-order grating for the wavelength λ that has a duty factor of $\xi = 1/2$.
   a. If a reflectivity of 99% is desired, find the length l and the number of grating periods of the DBR mirror as a function of λ, n, and Δn. Find the length l and the number of periods for $\lambda = 870$ nm, $n = 3.5$, and $\Delta n = 0.3$.
   b. Answer the same questions in (a) for a reflectivity of 99.9%.
   c. Answer the same questions in (a) and (b) for a first-order sinusoidal grating.

5.1.3 Show that (5.14) and (5.15) are, respectively, the coupling coefficients between two TE modes and between two TM modes for a grating waveguide coupler with a sinusoidal corrugation grating. Show also that (5.18) and (5.19) are the coupling coefficients between two TE modes and between two TM modes, respectively, for a grating waveguide coupler with a square corrugation grating. What is the coupling coefficient between a TE mode and a TM mode in each of the two waveguide couplers?

5.1.4 Why is it not possible to use a high-order sinusoidal grating for a grating waveguide coupler? Is it possible to use a second-order square corrugation grating that has a duty factor of $\xi = 1/2$ for a grating waveguide coupler? When using a qth-order square corrugation grating, what is the best choice of the duty factor for the largest coupling coefficient? What is the worst choice?

5.1.5 If the 3-dB grating coupler for the TE_0 mode described in Example 5.3 is used as a DBR for the TM_0 mode, what is the efficiency that can be obtained?

5.1.6 Answer the same questions as those raised in Example 5.3 if a second-order grating that has a duty factor of $\xi = 1/4$ is used instead of the first-order grating for the grating coupler.

5.1.7 Estimate the grating period of a second-order DBR for a guided wave at 1.55 μm in a symmetric slab semiconductor waveguide where $n_1 = 3.5$ and $n_2 = 3.45$.

5.1.8 The propagation constant at $\lambda = 1.3$ μm is $\beta_{TE} = 1.65 \times 10^7$ m^{-1} for the TE mode of an InGaAsP/InP slab waveguide that supports only fundamental TE and TM modes. The waveguide has a symmetric structure with $n_1 = 3.53$ for the waveguide core and $n_2 = 3.4$ for the cladding layers.
   a. Find the grating period of the second-order DBR for the TE mode of the waveguide.
   b. Estimate the grating period of the second-order DBR for the TM mode within the smallest range possible with the given information on the waveguide.

5.1.9 Show that for small deviations of optical frequency from the Bragg frequency, the Bragg diffraction phase shift can be expressed in terms of the variation of the propagation constant $\beta(\omega)$ away from the phase-matched value of $\beta(\omega_B)$ in

the form of (5.24) with an effective length of the DBR for its reflection phase shift defined in (5.25).

5.1.10 In this problem, we design an InGaAsP/InP DFB laser for 1.55 μm wavelength. The laser waveguide is a symmetric slab waveguide with $n_1 = 3.54$ in the waveguide core and $n_2 = 3.4$ in the cladding regions. The thickness d of the waveguide core is to be chosen so that the structure is single moded with a confinement factor of $\Gamma = 0.67$ for the TE_0 mode. A square grating of period Λ, depth d_g, and duty factor $\xi = 1/2$, as shown in Fig. 5.16, is to be fabricated at the lower core–cladding boundary. For simplicity, we consider only the operation of the TE mode.

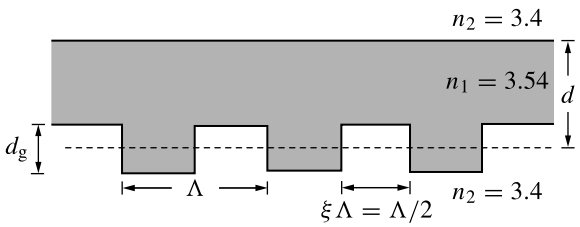


Figure 5.16 InGaAsP/InP DFB waveguide.

a. Find the thickness, d, of the waveguide core.
b. What is the propagation constant β of the TE_0 mode?
c. If a first-order grating is fabricated, what is the grating period Λ?
d. Find the length of the grating required for a reflectivity of 50% if the corrugation depth of the grating is $d_g = 100$ nm.
e. What happens to the reflectivity if the grating period has a $\pm 10\%$ uncertainty caused by fabrication error?
f. Can a second-order grating of the same shape and duty factor be used to accomplish the same function? What is the period of this second-order grating?
g. If a sinusoidal grating is used, what should its period be? Again, if the sinusoidal grating has the same length as that found in (c) for the square grating, what should its corrugation depth be in order to maintain a reflectivity of 50%?

5.1.11 In this problem, we examine the frequency bandwidth of a DBR structure.
a. Show that $|\delta| > |\kappa|$ at $\eta = \eta_{PM}/2$ for any given value of $|\kappa|l$, where η_{PM} is the coupling efficiency with $\delta = 0$ for the given value of $|\kappa|l$.
b. Show that the equation in (5.27) determines the value of $|\delta_{1/2}|$ for $\eta = \eta_{PM}/2$ for a given value of $|\kappa|l$.
c. Show that for a given value of $|\kappa|l$, the value of $|\delta_{1/2}|$ is bounded as

$$\sqrt{2}|\kappa| \coth |\kappa|l \geq |\delta_{1/2}| > |\kappa|. \tag{5.85}$$

d. What are the upper and lower bounds of the frequency bandwidth? How does the bandwidth vary with length l for a fixed coupling coefficient $|\kappa|$? How does it vary with $|\kappa|$ for a fixed l?
e. What is the strategy for designing a DBR with high reflectivity and narrow bandwidth? Is it possible to get an arbitrarily small bandwidth for a given reflectivity?

5.1.12 A DBR structure has a fixed coupling coefficient κ and dispersion characteristics $\beta(\omega)$ such that the product $|\kappa \mathrm{d}\omega/\mathrm{d}\beta| = 2\pi \times 100$ GHz at a given optical wavelength $\lambda = 1$ μm. We investigate the dependence of its bandwidth on its length.
a. Suppose the length of the structure is chosen to be $l = l_c^{\mathrm{PM}} = \pi/2|\kappa|$. What is the FWHM frequency bandwidth $\Delta\nu \equiv \Delta\omega/2\pi$? What is the reflectivity?
b. With the given structure, what is the narrowest frequency bandwidth that can be obtained? What is the shortest length needed to obtain such a bandwidth? What is the reflectivity under this condition?
c. Show that the bandwidth cannot be reduced further below what is obtained in (b) no matter how long the structure is. If you really need a narrower bandwidth, what can you do to get it?

5.2.1 Show that the coupling coefficients $\tilde{\kappa}_{ab}$ and $\tilde{\kappa}_{ba}$ obtained from (5.34) and the overlap coefficients c_{ab} and c_{ba} obtained from (5.35) for the coupling between two TE modes in the asymmetric two-channel directional coupler shown in Fig. 5.4 satisfy the relation given in (4.44) when $\tilde{\kappa}_{ab} \neq \tilde{\kappa}_{ba}^*$ in the case of asymmetric coupling, as required.

5.2.2 Find $\tilde{\kappa}_{aa}$, $\tilde{\kappa}_{bb}$, $\tilde{\kappa}_{ab}$, and $\tilde{\kappa}_{ba}$, as well as c_{ab} and c_{ba}, for the coupling between TM modes in the asymmetric two-channel directional coupler shown in Fig. 5.4. What are the simplified formulas for these coefficients in the case of a symmetric directional coupler? Do the coefficients satisfy the relation in (4.44) in the case of asymmetric coupling?

5.2.3 If the symmetric directional coupler described in Example 5.4 is used to couple the TM_{01} modes of the waveguides at $\lambda = 1$ μm, what is the coupling length? What are the propagation constants for the even and odd supermodes?

5.2.4 In this problem, we examine the effect of separation s on the coupling of two waveguides in a symmetric directional coupler by considering the directional coupler described in Example 5.4 for TM_{00} modes at $\lambda = 1$ μm.
a. If all of the waveguide parameters remain unchanged from those used in Example 5.4 except that the separation between the two waveguides is reduced to $s = 0.5$ μm, what is the coupling length? What are the propagation constants for the even and odd supermodes, respectively?
b. If the separation is increased to $s = 2$ μm, what is the coupling length? What are the propagation constants for the even and odd supermodes?

5.2.5 Consider the coupling between two identical symmetric slab waveguides of thickness 0.5 μm, separated by 0.5 μm, as shown in Fig. 5.17. At $\lambda = 1.3$ μm, the indices of refraction are $n_1 = 3.46$ and $n_2 = 3.44$.

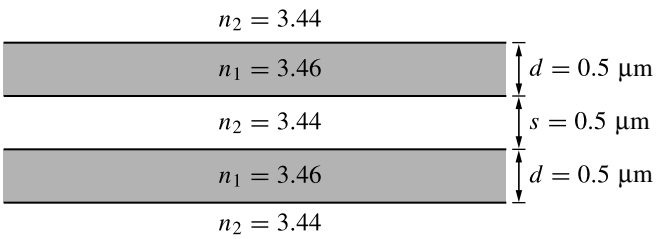


Figure 5.17 Codirectional coupler consisting of two identical symmetric slab waveguides.

a. Find the coupling coefficient for the TE_0 mode.
b. What is the 3-dB coupling length for the TE_0 mode?
c. Repeat (a) and (b) for the TM_0 mode.

5.2.6 In this problem, we examine the frequency bandwidth of an asymmetric codirectional coupler that has a length l that is an odd multiple of l_c^{PM}.
a. Show that the equation in (5.57) determines the value of $|\delta_{1/2}|$ for $\eta = \eta_{PM}/2$ for a given value of $|\kappa|l$ that is an odd multiple of $\pi/2$.
b. Show that for a given value of $|\kappa|l$ that is an odd multiple of $\pi/2$, the value of $|\delta_{1/2}|$ is bounded as

$$0.8|\kappa| \geq |\delta_{1/2}| > 0. \tag{5.86}$$

c. What are the upper and lower bounds of the frequency bandwidth? How does the bandwidth vary with the length l for a fixed coupling coefficient $|\kappa|$ while keeping the value of $|\kappa|l$ to be an odd multiple of $\pi/2$? How does it vary with $|\kappa|$ for a fixed l?

5.2.7 Consider the use of a dual-channel asymmetric directional coupler as a frequency filter. The structure is fixed so that $|\kappa| = \sqrt{\kappa_{ab}\kappa_{ba}}$ and $\beta_a(\omega)$ and $\beta_b(\omega)$ are given. The only parameter that can be varied is the length of the coupler. The given parameters combined give a value $|\kappa/(\mathrm{d}\beta_b/\mathrm{d}\omega - \mathrm{d}\beta_a/\mathrm{d}\omega)| = 2\pi \times 20$ THz at an optical wavelength $\lambda = 600$ nm.
a. If the length of the coupler is chosen to be $l = l_c^{PM}$, what is the FWHM bandwidth in terms of $\Delta\lambda$?
b. If we desire a bandwidth of $\Delta\lambda < 10$ nm, what is the minimum length of the coupler?
c. Show that the bandwidth can continue to narrow if the length of the coupler continues to increase. However, it is to be noted that the length has to be one of the odd multiples of l_c^{PM} for maximum efficiency.

5.2.8 If a grating is placed in the spacing between the two waveguides of a dual-channel asymmetric directional coupler as shown in Fig. 5.18, what should the grating period Λ be for complete power transfer between the two channels to be possible? The propagation constants in the two waveguides are β_a and β_b, respectively, and the coupling coefficients without the grating are κ_{aa}, κ_{bb}, κ_{ab}, and κ_{ba}.

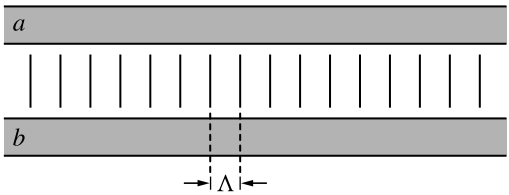


Figure 5.18 Dual-channel asymmetric directional coupler with a grating of period Λ.

5.2.9 A fiber-optic frequency filter is made of two single-mode fibers of different mode propagation constants. They are placed in close contact for a length l as shown in Fig. 5.19. At 1.3 μm optical wavelength, the effective indices for the two fiber modes are $n_\beta^a = 1.466$ and $n_\beta^b = 1.484$, respectively, and the coupling coefficient between the two fiber modes is $\kappa_{ab} = 10\ \text{cm}^{-1}$. A sinusoidal fiber grating of period Λ is built into the fibers in the coupling section. The input port of the device is port 1. The device is to function as an optical filter for separating the 1.3 μm wavelength from other wavelengths.

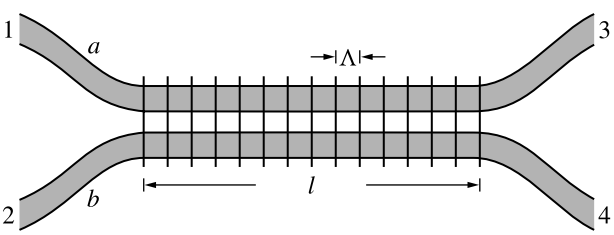


Figure 5.19 Fiber-optic frequency filter consisting of two different single-mode fibers modulated by a fiber grating of period Λ.

a. If the device is to function in such a way as to direct all of the optical power at 1.3 μm to port 4 and to dump all other wavelengths to port 3, what values of Λ and l should be selected?
b. If the device is to direct the power at 1.3 μm to port 2, what should the grating period Λ be? In this case, if the length l of the coupler remains the same as that found in (a), what is the efficiency of directing the 1.3 μm light from port 1 to port 2?

c. For the grating periods of (a) and (b), respectively, is it possible to have light at any wavelength reflected back to port 1? If it is possible, what is this wavelength?

d. Compare the two types of devices in (a) and (b) to decide which one is a better device. Why?

5.2.10 Two optical waves of exactly the same wavelength and the same power are simultaneously injected into the two input ports of a 3-dB directional coupler as shown in Fig. 5.20. What are the possible power ratios between the two output ports? What factor determines this ratio?

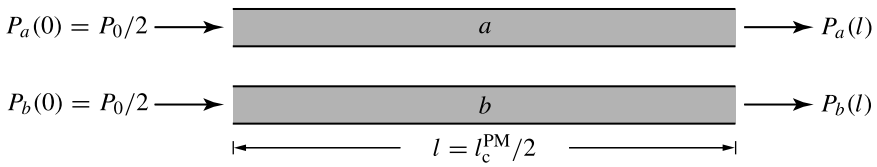


Figure 5.20 3-dB directional coupler.

5.2.11 A directional coupler can be used for multiplexing or demultiplexing optical beams at different wavelengths. In this problem, we consider directional couplers for such applications. For each of the following questions, clearly state the basic operation principle of the device and the quantitative conditions for the functioning of the device.

a. A symmetric directional coupler is used as a wavelength demultiplexer for 1.3 and 1.55 μm wavelengths. The two wavelengths enter the device from the same channel at the input end, but they are split at the output end by leaving the device from different channels with 1.3 μm in the parallel state and 1.55 μm in the cross state.

b. An asymmetric directional coupler is used to accomplish the same function as that described in (a).

c. A wavelength demultiplexer in the structure of a directional coupler is designed to select a particular wavelength, say 1.3 μm, from many wavelengths that enter the device from the same channel at the input end. At the output end, the selected wavelength alone leaves the device from one channel while the other wavelengths all leave from the other channel.

5.3.1 In using a surface prism coupler to couple an optical wave from free space to a waveguide mode, what are the three factors that need to be considered for efficient coupling?

5.3.2 The angle ζ of the prism coupler shown in Fig. 5.12 cannot be arbitrarily chosen but has to satisfy a certain condition in order for the prism to be useful in coupling an optical beam from outside the prism into the waveguide, as illustrated in

Example 5.6. Show that the condition is

$$\theta + \theta_c > \zeta > \theta - \theta_c, \tag{5.87}$$

where θ is the coupling angle defined by (5.64) and θ_c is the critical angle given in (5.65). Show that $\zeta = 45^\circ$, as chosen for the prism in Example 5.6, satisfies this condition for the coupling of both TE_0 and TM_0 modes described in Example 5.6.

5.3.3 A prism is used to couple light of 1 μm wavelength into a glass slab waveguide, as shown in Fig. 5.21. The thickness of the waveguide film is 1 μm. Find the angle θ shown in the figure at which the fundamental TE mode can be excited.

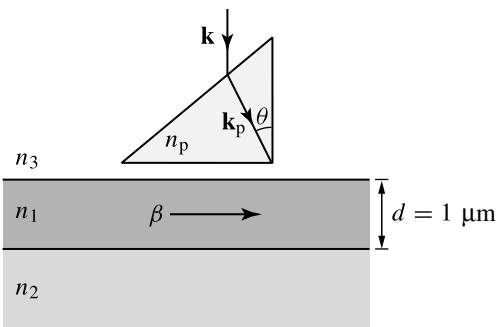


Figure 5.21 Prism surface coupler.

5.3.4 In coupling light into a planar waveguide using a prism coupler, you have done everything correctly in terms of choosing a high-index prism and a correct angle of incidence to ensure phase matching, but there is very little light coupled into the waveguide. What is the most likely problem you have overlooked? What is the second most likely problem?

5.3.5 For a surface grating coupler, what determines the grating period? Which factors determine the coupling efficiency?

5.3.6 Answer the questions raised in Example 5.7 if an incident angle of $\theta_i = 45^\circ$ instead of 20° is desired for the input beam at $\lambda = 1$ μm.

5.3.7 Answer the questions raised in Example 5.7 for a first-order grating that allows normal incidence of the input beam at $\theta_i = 0^\circ$.

5.3.8 A grating surface coupler is fabricated on the surface of a thin-film polymer waveguide of $n_1 = 1.55$ on a glass substrate of $n_2 = 1.5$, as shown in Fig. 5.22. At the optical wavelengths considered in this problem, the fundamental mode of the waveguide is far from its cutoff point.

a. If a laser beam at 532 nm wavelength is to be coupled into the waveguide from normal incidence to the surface of the thin film, how should the grating period be chosen?

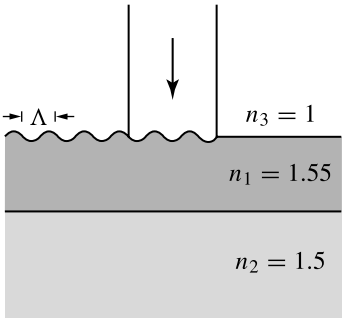


Figure 5.22 Grating surface coupler for input coupling at normal incidence.

b. Assume that dispersion of the waveguide is negligible. What grating period should be chosen if laser beams at both 532 nm and 1.064 μm are to be coupled into the waveguide from normal incidence?
c. If the grating chosen in (b) is used for output coupling, how many beams will appear on the air side for the cases of 532 nm and 1.064 μm, respectively?

5.3.9 A grating surface coupler is designed for input and output coupling of the GaAs waveguide shown in Fig. 5.23. The indices of refraction for the three layers are $n_3 = 1$, $n_1 = 3.6$, and $n_2 = 3.4$ at the 850 nm optical wavelength of interest. Answer the following questions without actually solving for the propagation constants of the guided modes.

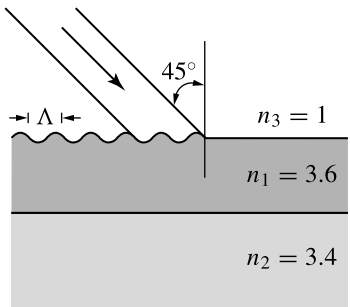


Figure 5.23 Grating surface coupler on a GaAs waveguide.

a. What is the range of the grating period that guarantees one and only one output beam coupled from any guided mode? Which side does this beam appear?
b. Answer the questions in (a) if two and only two output beams are desired.
c. If the grating is designed so that input coupling can be accomplished through 45° incidence from the air side as shown in the figure, what is the direction of propagation of the guided mode that is excited by this input? What is the approximate value of the grating period needed for this coupling?

Give the upper and lower bounds for this value with the highest accuracy possible.

d. Because we did not know the propagation constants exactly, the actual input-coupling angle for the excitation of a guided mode is unlikely to be exactly 45° if a grating period is arbitrarily picked within the bounds obtained in (c). According to this uncertainty, what is the range within which one can be sure to find the actual phase-matched incident angle or angles if a grating period within these bounds is arbitrarily picked?

SELECT BIBLIOGRAPHY

Buckman, A. B., *Guided-Wave Photonics*. Fort Worth, TX: Sauders College Publishing, 1992.

Ebeling, K. J., *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*. Berlin: Springer-Verlag, 1993.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Marcuse, D., *Theory of Dielectric Optical Waveguides*, 2nd edn. Boston, MA: Academic Press, 1991.

Mentzer, M. A., *Principles of Optical Circuit Engineering*. New York: Marcel Dekker, 1990.

Nishihara, H., Haruna, M. and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Tamir, T., ed., *Integrated Optics*. New York: Springer-Verlag, 1982.

ADVANCED READING LIST

Gaylord, T. K. and Moharam, M. G., "Analysis and applications of optical diffraction by gratings," *Proceedings of the IEEE* **73**(5): 894–937, May 1985.

Ladouceur, F., "Roughness, inhomogeneity, and integrated optics," *Journal of Lightwave Technology* **15**(6): 1020–1025, June 1997.

Parriaux, O., Sychugov, V. A. and Tishchenko, A. V., "Coupling gratings as waveguide functional elements," *Pure and Applied Optics* **5**(4): 453–469, July 1996.

Payne, F. P., "Fused single-mode optical fibre couplers," *Journal of the Institution of Electronics and Telecommunication Engineers* **32**(4): 319–326, July–Aug. 1986.

Raburn, M., Liu, B., Rauscher, K., Okuno, Y., Dagli, N. and Bowers, J. E., "3-D photonic circuit technology," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(4): 935–942, July–Aug. 2002.

Tien, P. K., "Integrated optics and new wave phenomena in optical waveguides," *Reviews of Modern Physics*, **49**(2): 361–420, Apr. 1977.

Yariv, A. and Nakamura, M., "Periodic structures for integrated optics," *IEEE Journal of Quantum Electronics* **QE-13**(4/2): 233–252, Apr. 1977.

Part III

Nonlinear photonics

6 Electro-optic devices

In many applications, it is necessary to control optical waves with externally applied signals to perform such functions as modulation, switching, deflection, isolation, frequency shifting, and polarization rotation of optical signals. Depending on the nature of the external control signal, these functions can be accomplished through the interactions of an optical wave with an electric field, a magnetic field, an acoustic wave, or another optical wave. Generally speaking, these interactions are all nonlinear optical phenomena. Nevertheless, electro-optic, magneto-optic, and acousto-optic effects each have very specific characteristics. Many useful devices have been developed based specifically on these effects for many important functions, such as optical modulation and optical switching. The electro-optic, magneto-optic, and acousto-optic devices are discussed separately in this and the following two chapters. The discussions in Chapter 9 then focus on nonlinear optical devices based solely on the interactions between optical waves.

6.1 Electro-optic effects

The optical property of a dielectric material can be changed through an electro-optic effect in the presence of a static or low-frequency electric field $\boldsymbol{E}_0$. The result is a field-dependent susceptibility and thus a field-dependent electric permittivity:

$$\mathbf{P}(\omega, \boldsymbol{E}_0) = \epsilon_0 \boldsymbol{\chi}(\omega, \boldsymbol{E}_0) \cdot \mathbf{E}(\omega) = \epsilon_0 \boldsymbol{\chi}(\omega) \cdot \mathbf{E}(\omega) + \epsilon_0 \Delta\boldsymbol{\chi}(\omega, \boldsymbol{E}_0) \cdot \mathbf{E}(\omega) \tag{6.1}$$

and

$$\mathbf{D}(\omega, \boldsymbol{E}_0) = \boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) \cdot \mathbf{E}(\omega) = \boldsymbol{\epsilon}(\omega) \cdot \mathbf{E}(\omega) + \Delta\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) \cdot \mathbf{E}(\omega), \tag{6.2}$$

where field-independent $\boldsymbol{\chi}(\omega) = \boldsymbol{\chi}(\omega, \boldsymbol{E}_0 = 0)$ and $\boldsymbol{\epsilon}(\omega) = \boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0 = 0)$ represent the intrinsic linear response of the material at the optical frequency ω, while $\Delta\boldsymbol{\chi}$ and $\Delta\boldsymbol{\epsilon}$ represent changes induced by the low-frequency field $\boldsymbol{E}_0$. We can write $\mathbf{D}(\omega, \boldsymbol{E}_0) = \mathbf{D}(\omega) + \Delta\mathbf{P}(\omega, \boldsymbol{E}_0)$, where $\Delta\mathbf{P}(\omega, \boldsymbol{E}_0) = \Delta\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) \cdot \mathbf{E}(\omega)$. The total permittivity of the material in the presence of the applied field is then

$$\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) = \boldsymbol{\epsilon}(\omega) + \Delta\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) = \boldsymbol{\epsilon}(\omega) + \epsilon_0 \Delta\boldsymbol{\chi}(\omega, \boldsymbol{E}_0). \tag{6.3}$$

The dielectric permittivity tensor $\epsilon(\omega)$ in the absence of an applied electric field is diagonal in the coordinate system defined by the intrinsic principal dielectric axes, $\hat{x}$, $\hat{y}$, and $\hat{z}$, of the dielectric material. The electro-optically induced changes usually generate off-diagonal elements in addition to changing the diagonal elements,

$$\boldsymbol{\epsilon}(\omega) = \begin{bmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_y & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix}, \quad \text{while} \quad \boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0) = \begin{bmatrix} \epsilon_x + \Delta\epsilon_{xx} & \Delta\epsilon_{xy} & \Delta\epsilon_{xz} \\ \Delta\epsilon_{yx} & \epsilon_y + \Delta\epsilon_{yy} & \Delta\epsilon_{yz} \\ \Delta\epsilon_{zx} & \Delta\epsilon_{zy} & \epsilon_z + \Delta\epsilon_{zz} \end{bmatrix}, \tag{6.4}$$

in the coordinate system of $\hat{x}$, $\hat{y}$, and $\hat{z}$ axes. As discussed in Section 1.6, $\boldsymbol{\epsilon}$ for a dielectric material is a symmetric tensor. This remains true for an electro-optic material subject to an applied electric field. Therefore,

$$\epsilon_{ij} = \epsilon_{ji} \quad \text{and} \quad \Delta\epsilon_{ij} = \Delta\epsilon_{ji} \tag{6.5}$$

for field-dependent permittivity tensors.

The electro-optically induced nondiagonal permittivity tensor, $\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0)$ given in (6.4), can be diagonalized. Its orthonormalized eigenvectors, $\hat{X}$, $\hat{Y}$, and $\hat{Z}$, are the new principal dielectric axes of the material in the presence of an applied electric field $\boldsymbol{E}_0$. In general, they depend on the direction of $\boldsymbol{E}_0$. If the unit vectors $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ are expressed in terms of $\hat{x}$, $\hat{y}$, and $\hat{z}$ as

$$\hat{X} = a_1\hat{x} + b_1\hat{y} + c_1\hat{z}, \qquad \hat{Y} = a_2\hat{x} + b_2\hat{y} + c_2\hat{z}, \qquad \hat{Z} = a_3\hat{x} + b_3\hat{y} + c_3\hat{z}, \tag{6.6}$$

then transformation between the old coordinate system defined by $\hat{x}$, $\hat{y}$, and $\hat{z}$ and the new coordinate system defined by $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ can be carried out using the following transformation matrix:

$$\mathbf{T} = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}. \tag{6.7}$$

Because both sets of vectors, $\{\hat{x}, \hat{y}, \hat{z}\}$ and $\{\hat{X}, \hat{Y}, \hat{Z}\}$, that define the transformation matrix $\mathbf{T}$ are *orthonormal unit vectors*, the transformation characterized by the matrix $\mathbf{T}$ is an *orthogonal transformation* with the convenient characteristic that $\mathbf{T}^{-1} = \tilde{\mathbf{T}}$, where $\tilde{\mathbf{T}}$ is the transpose of $\mathbf{T}$.

The relation in (6.6) between old and new principal axes can be written

$$\begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \end{bmatrix} = \mathbf{T} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \tilde{\mathbf{T}} \begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \end{bmatrix}. \tag{6.8}$$

The transformation of the coordinates of any vector $\mathbf{r} = x\hat{x} + y\hat{y} + z\hat{z} = X\hat{X} + Y\hat{Y} + Z\hat{Z}$ in space is given by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{T} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{6.9}$$

or

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{T}^{-1} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \tilde{\mathbf{T}} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} a_1X + a_2Y + a_3Z \\ b_1X + b_2Y + b_3Z \\ c_1X + c_2Y + c_3Z \end{bmatrix}. \tag{6.10}$$

Accordingly, the field components in the two coordinate systems are related through

$$\begin{bmatrix} E_X \\ E_Y \\ E_Z \end{bmatrix} = \mathbf{T} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}, \qquad \begin{bmatrix} D_X \\ D_Y \\ D_Z \end{bmatrix} = \mathbf{T} \begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix}, \tag{6.11}$$

and so on. Diagonalization of $\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0)$ to obtain its eigenvalues can be carried out using $\mathbf{T}$ as

$$\mathbf{T}\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0)\mathbf{T}^{-1} = \mathbf{T}\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0)\tilde{\mathbf{T}} = \begin{bmatrix} \epsilon_X & 0 & 0 \\ 0 & \epsilon_Y & 0 \\ 0 & 0 & \epsilon_Z \end{bmatrix}. \tag{6.12}$$

The propagation characteristics of an optical wave in the presence of an electro-optic effect are then determined by ϵ_X, ϵ_Y, and ϵ_Z with the following new principal indices of refraction:

$$n_X = \sqrt{\frac{\epsilon_X}{\epsilon_0}}, \qquad n_Y = \sqrt{\frac{\epsilon_Y}{\epsilon_0}}, \qquad n_Z = \sqrt{\frac{\epsilon_Z}{\epsilon_0}}. \tag{6.13}$$

The discussions above describe a formal and systematic approach to treating an electro-optic effect in terms of changes in the permittivity tensor. However, electro-optic effects are traditionally defined in terms of the changes in the elements of the relative impermeability tensor as $\boldsymbol{\eta}(\boldsymbol{E}_0) = \boldsymbol{\eta} + \Delta\boldsymbol{\eta}(\boldsymbol{E}_0)$, which is expanded in the following form:

$$\eta_{ij}(\boldsymbol{E}_0) = \eta_{ij} + \Delta\eta_{ij}(\boldsymbol{E}_0) = \eta_{ij} + \sum_k r_{ijk}E_{0k} + \sum_{k,l} s_{ijkl}E_{0k}E_{0l} + \cdots, \tag{6.14}$$

where the first term $\eta_{ij} = \eta_{ij}(0)$ is the field-independent component, the elements of the r_{ijk} tensor are the *linear electro-optic coefficients* known as the *Pockels coefficients*, and those of the s_{ijkl} tensor are the *quadratic electro-optic coefficients* known as the *Kerr coefficients*. The *first-order electro-optic effect* characterized by the linear dependence of $\eta_{ij}(\boldsymbol{E}_0)$ on $\boldsymbol{E}_0$ through the coefficients r_{ijk} is called the *linear electro-optic effect*, also known as the *Pockels effect*. The *second-order electro-optic effect* characterized by the quadratic field dependence through the coefficients s_{ijkl} is called the *quadratic*

electro-optic effect, also known as the *Kerr effect*. Both linear and quadratic electro-optic effects are nonlinear optical effects, as discussed above.

The Pockels effect does not exist in a *centrosymmetric* material, which is a material that possesses *inversion symmetry*. The structure and properties of such a material remain unchanged under the transformation of space inversion, which changes the signs of all rectangular spatial coordinates from (x, y, z) to $(-x, -y, -z)$, and those of all polar vectors. As discussed in Section 1.1, an electric field vector is a polar vector that changes sign under the transformation of space inversion. By simply considering the effect of space inversion, it is clear that the electro-optically induced changes in the optical property of a centrosymmetric material are not affected by the sign change in the applied field from $\boldsymbol{E}_0$ to $-\boldsymbol{E}_0$, meaning that $\eta_{ij}(\boldsymbol{E}_0) = \eta_{ij}(-\boldsymbol{E}_0)$. As can be seen from (6.14), this condition requires that the Pockels coefficients r_{ijk} vanish. It can also be seen that the condition does not require vanishing of the Kerr coefficients s_{ijkl}. Consequently, the Pockels effect exists only in *noncentrosymmetric* materials, while the Kerr effect exists in all materials, including centrosymmetric ones.

In (6.14), indices i and j are associated with optical fields, while indices k and l are associated with the low-frequency applied field. Because $\eta_{ij} = \eta_{ji}$ and $\Delta\eta_{ij} = \Delta\eta_{ji}$, indices i and j can be contracted using the index contraction rule of (1.115), thus reducing (6.14) to

$$\eta_\alpha(\boldsymbol{E}_0) = \eta_\alpha + \Delta\eta_\alpha(\boldsymbol{E}_0) = \eta_\alpha + \sum_k r_{\alpha k} E_{0k} + \sum_{k,l} s_{\alpha kl} E_{0k} E_{0l} + \cdots, \tag{6.15}$$

where $\alpha = 1, 2, \ldots, 6$ with the meaning defined in (1.115).

From the relation that $\boldsymbol{\eta} = (\boldsymbol{\epsilon}/\epsilon_0)^{-1}$ defined in (1.111), it can be seen that $\boldsymbol{\eta}$ in the absence of $\boldsymbol{E}_0$ is a diagonal tensor in the coordinate system defined by $\hat{x}$, $\hat{y}$, and $\hat{z}$ with the following eigenvalues:

$$\eta_x = \frac{\epsilon_0}{\epsilon_x} = \frac{1}{n_x^2}, \qquad \eta_y = \frac{\epsilon_0}{\epsilon_y} = \frac{1}{n_y^2}, \qquad \eta_z = \frac{\epsilon_0}{\epsilon_z} = \frac{1}{n_z^2}, \tag{6.16}$$

where n_x, n_y, and n_z are the principal indices of refraction of the material in the absence of an applied electric field. In the presence of an applied field, $\boldsymbol{\eta}(\boldsymbol{E}_0)$ is generally not diagonal in this coordinate system. Using the relation $\boldsymbol{\eta} \cdot \boldsymbol{\epsilon}/\epsilon_0 = 1$, the relation between $\Delta\boldsymbol{\epsilon}$ and $\Delta\boldsymbol{\eta}$ can be found:

$$\Delta\boldsymbol{\epsilon} = -\frac{1}{\epsilon_0}\boldsymbol{\epsilon} \cdot \Delta\boldsymbol{\eta} \cdot \boldsymbol{\epsilon} \quad \text{and} \quad \Delta\boldsymbol{\eta} = -\frac{1}{\epsilon_0}\boldsymbol{\eta} \cdot \Delta\boldsymbol{\epsilon} \cdot \boldsymbol{\eta}. \tag{6.17}$$

When $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ in the absence of $\boldsymbol{E}_0$ are diagonalized, the relations in (6.17) can be written explicitly as

$$\Delta\epsilon_{ij} = -\epsilon_0 \frac{\Delta\eta_{ij}}{\eta_i \eta_j} = -\epsilon_0 n_i^2 n_j^2 \Delta\eta_{ij} \quad \text{and} \quad \Delta\eta_{ij} = -\epsilon_0 \frac{\Delta\epsilon_{ij}}{\epsilon_i \epsilon_j} = -\frac{\Delta\epsilon_{ij}}{\epsilon_0 n_i^2 n_j^2}. \tag{6.18}$$

In the absence of an electric field, the index ellipsoid of a material is that given by (1.117) with its principal axes aligned with $\hat{x}$, $\hat{y}$, and $\hat{z}$. Changes in the optical property

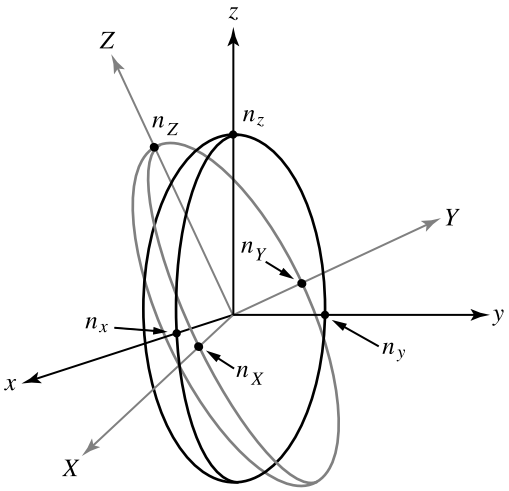


Figure 6.1 Transformation of index ellipsoid by an electro-optic effect. An electro-optic effect transforms an index ellipsoid originally aligned with the x, y, and z coordinates that are defined by the original principal axes $\hat{x}$, $\hat{y}$, and $\hat{z}$ into a new one aligned with the X, Y, and Z coordinates that are defined by the new principal axes $\hat{X}$, $\hat{Y}$, and $\hat{Z}$. Meanwhile, the principal indices of refraction have been changed from n_x, n_y, and n_z to n_X, n_Y, and n_Z.

of the material induced by an electro-optic effect deform the index ellipsoid into a new one described by

$$(\eta_1 + \Delta\eta_1)x^2 + (\eta_2 + \Delta\eta_2)y^2 + (\eta_3 + \Delta\eta_3)z^2 + 2\Delta\eta_4 yz + 2\Delta\eta_5 zx + 2\Delta\eta_6 xy = 1, \tag{6.19}$$

whose principal axes no longer line up with $\hat{x}$, $\hat{y}$, and $\hat{z}$ unless $\Delta\eta_4 = \Delta\eta_5 = \Delta\eta_6 = 0$. To find the principal axes of this new ellipsoid and their corresponding principal indices of refraction, we can perform a coordinate rotation in space to eliminate the cross-product terms containing yz, zx, and xy. From the discussions above, it can be seen that this procedure is the same as the coordinate rotation used to diagonalize $\boldsymbol{\epsilon}$. Thus, we can use (6.9) to transform (6.19) into

$$\frac{X^2}{n_X^2} + \frac{Y^2}{n_Y^2} + \frac{Z^2}{n_Z^2} = 1, \tag{6.20}$$

where n_X, n_Y, and n_Z are the same as those given in (6.13). The principal axes of this ellipsoid are simply the same $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ as those found from the eigenvectors of $\boldsymbol{\epsilon}$ and given in (6.6). Figure 6.1 illustrates the concept described here.

6.2 Pockels effect

The majority of electro-optic devices are based on the Pockels effect. *Structurally isotropic materials, including all gases, liquids, and amorphous solids such as glass,*

show no Pockels effect because they are centrosymmetric. Among the 32 point groups in the 7 crystal systems, 11 are centrosymmetric, and the remaining 21 are noncentrosymmetric. It is important to note that the linear optical property of a crystal is determined only by its crystal system, as mentioned in Section 1.6 and summarized in Table 1.2, but its nonlinear optical properties, including its Pockels coefficients, further depend on its point group. An instructive example is that all cubic crystals have isotropic linear optical properties but not isotropic crystal structures. Two cubic crystals belonging to different point groups can have very different nonlinear optical properties. Among the cubic crystals, C, Si, and Ge are centrosymmetric materials of diamond structure that show no Pockels effect, whereas GaAs, InP, and other III–V semiconductors are noncentrosymmetric materials that have nonvanishing Pockels coefficients.

For the Pockels effect,

$$\Delta\eta_\alpha = \sum_k r_{\alpha k} E_{0k}, \tag{6.21}$$

which can be written explicitly in matrix form:

$$\begin{bmatrix} \Delta\eta_1 \\ \Delta\eta_2 \\ \Delta\eta_3 \\ \Delta\eta_4 \\ \Delta\eta_5 \\ \Delta\eta_6 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{bmatrix} \begin{bmatrix} E_{0x} \\ E_{0y} \\ E_{0z} \end{bmatrix}. \tag{6.22}$$

Even for a noncentrosymmetric material, the number of nonvanishing independent elements in its $r_{\alpha k}$ matrix is generally reduced by its symmetry. Table 6.1 shows the matrix form of the Pockels coefficients for the 21 noncentrosymmetric point groups. Some crystal point groups are of particular interest.

1. **Cubic $\bar{4}3m$.** Most III–V semiconductors, such as GaAs, InP, AlAs, and GaP, and many II–VI compounds, such as ZnTe, ZnSe, CdTe, and HgSe, are cubic crystals of $\bar{4}3m$ symmetry. They have isotropic linear optical properties with $n_x = n_y = n_z = n_\mathrm{o}$. The $r_{\alpha k}$ matrix has only three nonvanishing elements with the same value: $r_{41} = r_{52} = r_{63}$.
2. **Tetragonal $\bar{4}2m$.** Crystals possessing tetragonal $\bar{4}2m$ symmetry include many commonly used nonlinear optical crystals, such as KH_2PO_4 (KDP), KD_2PO_4 (KD^*P), $NH_4H_2PO_4$ (ADP), $ND_4D_2PO_4$ (AD^*P), CsH_2AsO_4 (CDA), $AgGaS_2$, and $AgGaSe_2$. These are uniaxial crystals with $n_x = n_y = n_\mathrm{o}$ and $n_z = n_\mathrm{e}$. The $r_{\alpha k}$ matrix has only three nonvanishing elements with two independent values: $r_{41} = r_{52} \neq r_{63}$.
3. **Trigonal $3m$.** The very useful nonlinear optical crystals $LiNbO_3$, $LiTaO_3$, and β-BaB_2O_4 (BBO) of trigonal $3m$ symmetry are uniaxial with $n_x = n_y = n_\mathrm{o}$ and $n_z = n_\mathrm{e}$. The $r_{\alpha k}$ matrix has eight nonvanishing elements with four independent values: $r_{13} = r_{23}$, $r_{12} = r_{61} = -r_{22}$, r_{33}, and $r_{42} = r_{51}$.

Table 6.1 *Matrix form of Pockels coefficients for noncentrosymmetric point groups*[a]

Triclinic	1	$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{bmatrix}$
Monoclinic	2 $(2 \parallel \hat{y})$	$\begin{bmatrix} 0 & r_{21} & 0 \\ 0 & r_{22} & 0 \\ 0 & r_{23} & 0 \\ r_{41} & 0 & r_{43} \\ 0 & r_{52} & 0 \\ r_{61} & 0 & r_{63} \end{bmatrix}$
	m $(m \perp \hat{y})$	$\begin{bmatrix} r_{11} & 0 & r_{13} \\ r_{21} & 0 & r_{23} \\ r_{31} & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{51} & 0 & r_{53} \\ 0 & r_{62} & 0 \end{bmatrix}$
Orthorhombic	222	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{52} & 0 \\ 0 & 0 & r_{63} \end{bmatrix}$
	$mm2$	$\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{23} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{51} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
Tetragonal	4	$\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ r_{41} & r_{42} & 0 \\ r_{42} & -r_{41} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
	$\bar{4}$	$\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & -r_{13} \\ 0 & 0 & 0 \\ r_{41} & r_{42} & 0 \\ -r_{42} & r_{41} & 0 \\ 0 & 0 & r_{63} \end{bmatrix}$
	422	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & -r_{41} & 0 \\ 0 & 0 & 0 \end{bmatrix}$
	$4mm$	$\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{42} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
	$\bar{4}2m$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{41} & 0 \\ 0 & 0 & r_{63} \end{bmatrix}$
Trigonal	3	$\begin{bmatrix} r_{11} & -r_{22} & r_{13} \\ -r_{11} & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ r_{41} & r_{42} & 0 \\ r_{42} & -r_{41} & 0 \\ -r_{22} & -r_{11} & 0 \end{bmatrix}$

(*continued*)

Table 6.1 (*Cont.*)

	32 $\begin{bmatrix} r_{11} & 0 & 0 \\ -r_{11} & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & -r_{41} & 0 \\ 0 & -r_{11} & 0 \end{bmatrix}$	$3m$ $\begin{bmatrix} 0 & -r_{22} & r_{13} \\ 0 & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{42} & 0 & 0 \\ -r_{22} & 0 & 0 \end{bmatrix}$
Hexagonal	6 $\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ r_{41} & r_{42} & 0 \\ r_{42} & -r_{41} & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\bar{6}$ $\begin{bmatrix} r_{11} & -r_{22} & 0 \\ -r_{11} & r_{22} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -r_{22} & -r_{11} & 0 \end{bmatrix}$
622 $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & -r_{41} & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$6mm$ $\begin{bmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{42} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\bar{6}m2$ $(m \perp \hat{x})$ $\begin{bmatrix} 0 & -r_{22} & 0 \\ 0 & r_{22} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -r_{22} & 0 & 0 \end{bmatrix}$
Cubic	432 $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	23 and $\bar{4}3m$ $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{41} & 0 \\ 0 & 0 & r_{41} \end{bmatrix}$

[a] From Kaminow, I. P., *An Introduction to Electrooptic Devices*. Orlando, FL: Academic Press, 1974, pp. 110–111.

4. **Orthorhombic** $mm2$**.** The nonlinear crystals $KTiOPO_4$ (KTP), $KTiOAsO_4$ (KTA), LiB_3O_5 (LBO), $KNbO_3$, and $Ba_2NaNb_5O_{15}$ have orthorhombic $mm2$ symmetry. They are biaxial crystals with $n_x \neq n_y \neq n_z$, and they have five independent nonvanishing Pockels coefficients in the $r_{\alpha k}$ matrix: r_{13}, r_{23}, r_{33}, r_{42}, and r_{51}.

A secondary effect due to the existence of *piezoelectricity* causes complexity in the determination of the Pockels coefficients of a crystal. A *stress* applied to a noncentrosymmetric polar crystal can induce an *electric polarization* in the crystal. This effect is called the *direct piezoelectric effect*. In the *converse piezoelectric effect*, an *electric field* applied to the same crystal can induce a *strain* in the crystal. The piezoelectric effect and the Pockels effect have similar symmetry properties: both vanish in centrosymmetric materials, and both are restricted by crystal symmetry in similar manners. Consequently, the piezoelectric effect exists in a crystal that shows the

Table 6.2 *Properties of representative electro-optic crystals*[a]

Point group	Material	Pockels coefficients (pm V^{-1})	Refractive index (at 1 μm)
$\bar{4}3m$	GaAs	(S)$r_{41} = 1.2$	$n_o = 3.50$
	ZnTe	(S)$r_{41} = 4.3$	$n_o = 2.76$
$\bar{4}2m$	KDP[b]	(T)$r_{41} = 8.8$, (T)$r_{63} = 10.5$, (S)$r_{63} = 9.7$	$n_o = 1.51$, $n_e = 1.47$
	ADP	(T)$r_{41} = 24.5$, (T)$r_{63} = 8.5$, (S)$r_{63} = 5.5$	$n_o = 1.52$, $n_e = 1.48$
$3m$	$LiNbO_3$	(S)$r_{13} = 8.6$, (S)$r_{22} = 3.4$, (S)$r_{33} = 30.8$, (S)$r_{42} = 28$	$n_o = 2.238$, $n_e = 2.159$
	$LiTaO_3$	(S)$r_{13} = 8.5$, (S)$r_{22} = 1$, (S)$r_{33} = 30.5$, (S)$r_{42} = 20$	$n_o = 2.131$, $n_e = 2.134$
$mm2$	KTP	(S)$r_{13} = 8.8$, (S)$r_{23} = 13.8$, (S)$r_{33} = 35$, (S)$r_{42} = 8.8$, (S)$r_{51} = 6.9$	$n_x = 1.742$, $n_y = 1.750$, $n_z = 1.832$
	KTA[c]	(S)$r_{13} = 15$, (S)$r_{23} = 21$, (S)$r_{33} = 40$	$n_x = 1.783$, $n_y = 1.789$, $n_z = 1.870$

[a] Data are collected from various sources in the literature.

[b] KTP properties from Bierlein, J. D. and Vanherzeele, H., *Journal of the Optical Society of America B* **6**: 622–633, 1989.

[c] KTA properties from Bierlein, J. D. and Vanherzeele, H., *Applied Physics Letters* **54**: 783–785, 1989.

Pockels effect. The strain generated in a crystal by an applied electric field can induce index changes through the *photoelastic effect* discussed in Chapter 7. In a free crystal, which is allowed to strain in response to the applied electric field, this secondary effect is comparable in magnitude to the primary effect that accounts for the index changes directly caused by the applied electric field. Pockels coefficients measured at constant strain (indicated by S) with a crystal clamped reflect only the primary effect, whereas those measured at constant stress (indicated by T) with a crystal free and unclamped reflect the sum of the primary and secondary effects.

Table 6.2 lists the properties of some representative electro-optic materials. In practical device applications, an electro-optic crystal is not clamped, but its electro-optic coefficient is a function of the modulation frequency. At low modulation frequencies, the electro-optic response of the crystal is that of a free crystal at constant stress because the photoelastic response can follow the low-frequency modulation signal. At high modulation frequencies, however, the photoelastic effect vanishes because the strain in the crystal cannot respond quickly enough to follow the modulation signal. Consequently, the Pockels coefficients measured at constant stress have to be used for low-frequency modulation, but those measured at constant strain have to be used for high-frequency modulation. Besides their dependence on the modulation frequency,

the Pockels coefficients are also a function of temperature and optical wavelength. Because of these complications, only typical values of the Pockels coefficients measured at constant strain are listed in Table 6.2 except for those of KDP and ADP crystals, for which the r_{41} coefficient at constant strain is not available.

The technologically most important electro-optic materials are the III–V semiconductors, particularly GaAs and InP and related compounds, and the 3*m* crystals, such as $LiNbO_3$ and $LiTaO_3$. Electro-optic devices based on $LiNbO_3$ are the most extensively studied and most well developed. Those based on the III–V semiconductors are also intensively studied because they can be monolithically integrated with other optoelectronic devices, including semiconductor lasers, amplifiers, and detectors. It can be seen from Table 6.2 that the Pockels coefficients of the III–V semiconductors are relatively small compared with those of other important electro-optic materials. However, this disadvantage is generally compensated by using advanced semiconductor processing technologies. For example, small waveguide structures with optimized overlap of the applied electric field and the optical field can be made in a III–V semiconductor to maximize the electro-optic modulation efficiency. The intrinsic electro-optic effect in a III–V material can also be substantially enhanced by incorporating artificially tailored structures, such as *quantum-well* structures, in the material.

Index changes and rotation of principal axes

Depending on the symmetry of a specific material being used and the direction of the electric field being applied to the material, the index changes induced by the Pockels effect may or may not be accompanied by a rotation of principal axes. This fact is best illustrated through real examples.

We first consider $LiNbO_3$, which is a negative uniaxial crystal of 3*m* symmetry. The following analysis applies equally to other 3*m* crystals although some of them, such as $LiTaO_3$, are positive uniaxial crystals.

1. The electric field is applied along the optical axis: $E_{0x} = E_{0y} = 0$, $E_{0z} \neq 0$. In this case, the changes induced by the Pockels effect are $\Delta\eta_1 = r_{13}E_{0z}$, $\Delta\eta_2 = r_{13}E_{0z}$, and $\Delta\eta_3 = r_{33}E_{0z}$. The index ellipsoid becomes

$$\left(\frac{1}{n_0^2} + r_{13}E_{0z}\right)x^2 + \left(\frac{1}{n_0^2} + r_{13}E_{0z}\right)y^2 + \left(\frac{1}{n_e^2} + r_{33}E_{0z}\right)z^2 = 1. \tag{6.23}$$

Equivalently, by using (6.4) and (6.18), the field-dependent dielectric permittivity tensor can be found:

$$\boldsymbol{\epsilon}(\boldsymbol{E}_0) = \epsilon_0 \begin{bmatrix} n_o^2 - n_o^4 r_{13}E_{0z} & 0 & 0 \\ 0 & n_o^2 - n_o^4 r_{13}E_{0z} & 0 \\ 0 & 0 & n_e^2 - n_e^4 r_{33}E_{0z} \end{bmatrix}. \tag{6.24}$$

The principal axes are not rotated: $\hat{X} = \hat{x}$, $\hat{Y} = \hat{y}$, and $\hat{Z} = \hat{z}$. The crystal remains uniaxial with the same optical axis, but the indices of refraction are changed. Since the induced changes are generally so small that $|r_{13}E_{0z}| \ll n_{\mathrm{o}}^{-2}$ and $|r_{33}E_{0z}| \ll n_{\mathrm{e}}^{-2}$, the new principal indices of refraction are

$$n_X = n_Y \approx n_{\mathrm{o}} - \frac{n_{\mathrm{o}}^3}{2} r_{13} E_{0z}, \qquad n_Z \approx n_{\mathrm{e}} - \frac{n_{\mathrm{e}}^3}{2} r_{33} E_{0z}. \tag{6.25}$$

2. The electric field is applied along the y axis: $E_{0x} = E_{0z} = 0$, $E_{0y} \neq 0$. Then the induced changes are $\Delta\eta_1 = -r_{22}E_{0y}$, $\Delta\eta_2 = r_{22}E_{0y}$, and $\Delta\eta_4 = r_{42}E_{0y}$. The index ellipsoid becomes

$$\left(\frac{1}{n_{\mathrm{o}}^2} - r_{22}E_{0y}\right) x^2 + \left(\frac{1}{n_{\mathrm{o}}^2} + r_{22}E_{0y}\right) y^2 + \frac{1}{n_{\mathrm{e}}^2} z^2 + 2r_{42}E_{0y}yz = 1. \tag{6.26}$$

The corresponding dielectric permittivity tensor is

$$\boldsymbol{\epsilon}(\boldsymbol{E}_0) = \epsilon_0 \begin{bmatrix} n_{\mathrm{o}}^2 + n_{\mathrm{o}}^4 r_{22} E_{0y} & 0 & 0 \\ 0 & n_{\mathrm{o}}^2 - n_{\mathrm{o}}^4 r_{22} E_{0y} & -n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42} E_{0y} \\ 0 & -n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42} E_{0y} & n_{\mathrm{e}}^2 \end{bmatrix}. \tag{6.27}$$

Because of the existence of the yz term in (6.26), which corresponds to the existence of the off-diagonal terms of $\boldsymbol{\epsilon}(\boldsymbol{E}_0)$ in (6.27), the new principal axes $\hat{Y}$ and $\hat{Z}$ are rotated away from $\hat{y}$ and $\hat{z}$ while $\hat{X}$ remains the same as $\hat{x}$:

$$\hat{X} = \hat{x}, \qquad \hat{Y} = \hat{y}\cos\theta + \hat{z}\sin\theta, \qquad \hat{Z} = -\hat{y}\sin\theta + \hat{z}\cos\theta. \tag{6.28}$$

The angle of rotation θ and the new principal indices of refraction can be found by eliminating the yz term in (6.26) or, equivalently, by diagonalizing $\boldsymbol{\epsilon}(\boldsymbol{E}_0)$ in (6.27) through a transformation matrix $\mathbf{T}$ defined by (6.6) and (6.7). For $LiNbO_3$, since $n_{\mathrm{o}} > n_{\mathrm{e}}$ and $n_{\mathrm{o}}^2 - n_{\mathrm{e}}^2 \gg |n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42} E_{0y}| > |n_{\mathrm{o}}^4 r_{22} E_{0y}|$ for any E_{0y} below the material breakdown field of the order of 100 MV m^{-1}, it can be shown that

$$\theta \approx -\tan^{-1} \frac{n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42} E_{0y}}{n_{\mathrm{o}}^2 - n_{\mathrm{e}}^2} \tag{6.29}$$

and

$$\begin{aligned} n_X &\approx n_{\mathrm{o}} + \frac{n_{\mathrm{o}}^3}{2} r_{22} E_{0y}, \\ n_Y &\approx n_{\mathrm{o}} - \frac{n_{\mathrm{o}}^3}{2} r_{22} E_{0y} + \frac{1}{2} \frac{n_{\mathrm{o}}^3 n_{\mathrm{e}}^4}{n_{\mathrm{o}}^2 - n_{\mathrm{e}}^2} \left(r_{42} E_{0y}\right)^2, \\ n_Z &\approx n_{\mathrm{e}} - \frac{1}{2} \frac{n_{\mathrm{e}}^3 n_{\mathrm{o}}^4}{n_{\mathrm{o}}^2 - n_{\mathrm{e}}^2} \left(r_{42} E_{0y}\right)^2. \end{aligned} \tag{6.30}$$

The crystal becomes biaxial in the presence of an electric field applied in the y direction. Note that not only do the index changes depend on the applied field, but the angle of rotation of the principal axes is a function of E_{0y} as well.

3. The electric field is applied along the x axis: $E_{0x} \neq 0$, $E_{0y} = E_{0z} = 0$. The induced changes are $\Delta\eta_5 = r_{51}E_{0x} = r_{42}E_{0x}$ and $\Delta\eta_6 = -r_{22}E_{0x}$. Then, we have

$$\frac{1}{n_{\mathrm{o}}^2}x^2 + \frac{1}{n_{\mathrm{o}}^2}y^2 + \frac{1}{n_{\mathrm{e}}^2}z^2 + 2r_{42}E_{0x}zx - 2r_{22}E_{0x}xy = 1 \tag{6.31}$$

and

$$\boldsymbol{\epsilon}(\boldsymbol{E}_0) = \epsilon_0 \begin{bmatrix} n_{\mathrm{o}}^2 & n_{\mathrm{o}}^4 r_{22}E_{0x} & -n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42}E_{0x} \\ n_{\mathrm{o}}^4 r_{22}E_{0x} & n_{\mathrm{o}}^2 & 0 \\ -n_{\mathrm{o}}^2 n_{\mathrm{e}}^2 r_{42}E_{0x} & 0 & n_{\mathrm{e}}^2 \end{bmatrix}. \tag{6.32}$$

In this case, all three new principal axes $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ are rotated away from the original principal axes. The crystal also becomes biaxial. Because $n_{\mathrm{o}} \neq n_{\mathrm{e}}$, the angles of rotation depend on the magnitude of E_{0x}. Again, the new principal axes and their corresponding principal indices of refraction in the presence of E_{0x} can be found by eliminating the zx and xy terms in (6.31) or by diagonalizing $\boldsymbol{\epsilon}(\boldsymbol{E}_0)$ in (6.32). This problem is left as homework for the reader.

Because r_{33} is the largest electro-optic coefficient of $LiNbO_3$, the largest index change is obtained in n_Z when the electric field is applied along the z axis.

Another important example is the Pockels effect in a III–V semiconductor of $\bar{4}3m$ symmetry, such as GaAs or InP. A similar effect is seen when an electric field is applied along any of the original principal axes because $n_x = n_y = n_z = n_{\mathrm{o}}$ and the only nonvanishing Pockels coefficients are $r_{41} = r_{52} = r_{63}$ for such a crystal. We therefore consider only the case when the field is applied along the z axis: $E_{0x} = E_{0y} = 0$, $E_{0z} \neq 0$. Then, we only have $\Delta\eta_6 = r_{41}E_{0z}$. The index ellipsoid becomes

$$\frac{1}{n_{\mathrm{o}}^2}x^2 + \frac{1}{n_{\mathrm{o}}^2}y^2 + \frac{1}{n_{\mathrm{o}}^2}z^2 + 2r_{41}E_{0z}xy = 1, \tag{6.33}$$

and the dielectric permittivity tensor becomes

$$\boldsymbol{\epsilon}(\boldsymbol{E}_0) = \epsilon_0 \begin{bmatrix} n_{\mathrm{o}}^2 & -n_{\mathrm{o}}^4 r_{41}E_{0z} & 0 \\ -n_{\mathrm{o}}^4 r_{41}E_{0z} & n_{\mathrm{o}}^2 & 0 \\ 0 & 0 & n_{\mathrm{o}}^2 \end{bmatrix}. \tag{6.34}$$

The crystal has the following new principal axes:

$$\hat{X} = \frac{1}{\sqrt{2}}(\hat{x} + \hat{y}), \qquad \hat{Y} = \frac{1}{\sqrt{2}}(-\hat{x} + \hat{y}), \qquad \hat{Z} = \hat{z}, \tag{6.35}$$

with the following new principal indices:

$$n_X \approx n_o - \frac{n_o^3}{2} r_{41} E_{0z}, \qquad n_Y \approx n_o + \frac{n_o^3}{2} r_{41} E_{0z}, \qquad n_Z = n_o. \tag{6.36}$$

In the above, we have considered the simple cases where the electric field is applied only along one of the principal axes of the crystal. In certain practical situations, however, the applied electric field may not line up with any of the principal axes. The index changes and the rotation of principal axes can be found by following the same general procedure as illustrated above although the mathematics may be somewhat more complicated.

EXAMPLE 6.1 Find the index changes and the birefringence at $\lambda = 1$ μm caused by an electric field of $E_0 = 1$ MV m^{-1} applied to $LiNbO_3$ and GaAs, respectively, in a direction along the z principal axis of the crystal.

Solution The values of the Pockels coefficients and the refractive indices for both $LiNbO_3$ and GaAs are listed in Table 6.2. For $LiNbO_3$, an electric field applied along its z axis does not rotate its principal axes but only causes changes in its refractive indices. The crystal remains uniaxial. From (6.25), we find that the change in the ordinary index is

$$\Delta n_o = -\frac{n_o^3}{2} r_{13} E_{0z} = -\frac{2.238^3}{2} \times 8.6 \times 10^{-12} \times 1 \times 10^6 = -4.82 \times 10^{-5},$$

while the change in the extraordinary index is

$$\Delta n_e = -\frac{n_e^3}{2} r_{33} E_{0z} = -\frac{2.159^3}{2} \times 30.8 \times 10^{-12} \times 1 \times 10^6 = -1.55 \times 10^{-4}.$$

The electro-optically induced birefringence is $\Delta n_o - \Delta n_e \approx 1 \times 10^{-4}$, which is almost three orders of magnitude smaller than the intrinsic birefringence of $n_o - n_e = 0.08$ for $LiNbO_3$. In normal device applications, the applied electric field typically falls in the range between 0.1 and 10 MV m^{-1}. Because the index changes are linearly proportional to the applied electric field, the electro-optically induced birefringence is typically two to three orders of magnitude smaller than the intrinsic birefringence of $LiNbO_3$.

For GaAs, which is originally nonbirefringent, an electric field applied along its z axis causes a rotation of its x and y principal axes and a birefringence between them. From (6.36), we find that the electro-optically induced index changes are

$$\Delta n_Y = -\Delta n_X = \frac{n_o^3}{2} r_{41} E_{0z} = \frac{3.5^3}{2} \times 1.2 \times 10^{-12} \times 1 \times 10^6 = 2.57 \times 10^{-5}.$$

The electro-optically induced birefringence is $n_Y - n_X = \Delta n_Y - \Delta n_X = 5.15 \times 10^{-5}$. Although this birefringence is smaller than that in the case of $LiNbO_3$, it is significant because GaAs is originally nonbirefringent.

6.3 Electro-optic modulators

The index changes induced by the Pockels effect can be utilized to construct a variety of electro-optic modulators, in either bulk or waveguide structures. An electro-optically induced rotation of principal axes is not required for the functioning of an electro-optic modulator though it often accompanies the index changes. However, the directions of the principal axes in the presence of an applied electric field, whether rotated or not, have to be taken into consideration in the design and operation of an electro-optic modulator. In this section, we consider the operation principles of basic electro-optic modulators. Although some of the concepts, such as transverse phase modulation, that are considered in this section for bulk devices can be directly applied to guided-wave devices, specific guided-wave electro-optic devices are discussed in the next section.

Phase modulators

The phase of an optical wave can be electro-optically modulated. For this type of application, the optical wave is linearly polarized in a direction that is parallel to one of the principal axes, $\hat{X}$, $\hat{Y}$, or $\hat{Z}$, of the crystal in the presence of a modulation field. The preferred choice is a principal axis that has a large electro-optically induced index change but remains in a fixed direction as the magnitude of the modulation electric field varies. In $LiNbO_3$, this can be accomplished by applying the electric field along the z axis, as shown in Figure 6.2. In this case, $\hat{X} = \hat{x}$, $\hat{Y} = \hat{y}$, and $\hat{Z} = \hat{z}$, as discussed earlier. There are two possible arrangements: *transverse modulation*, where the optical wave propagates in a direction perpendicular to the modulation field, as shown in Fig. 6.2(*a*), and *longitudinal modulation*, where the modulation field is parallel to the direction of optical wave propagation, as shown in Fig. 6.2(*b*).

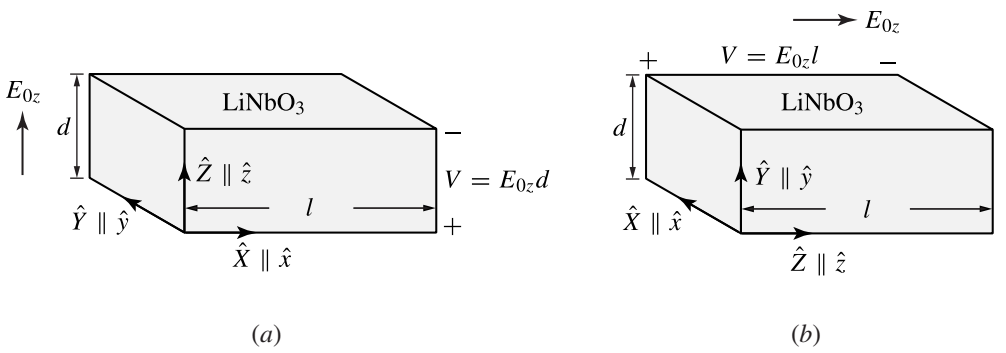


Figure 6.2 (*a*) $LiNbO_3$ transverse electro-optic phase modulator. (*b*) $LiNbO_3$ longitudinal electro-optic phase modulator. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the original principal axes of the crystal, and $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ represent its new principal axes.

Transverse phase modulators

We first consider the situation of the *transverse phase modulator* shown in Fig. 6.2(*a*), where the optical wave propagates in the X direction. In this case, the optical wave can be polarized in either the Z or Y direction. If it is linearly polarized in the Z direction, its space and time dependence can be written as

$$\mathbf{E}(X,t) = \hat{Z}\mathcal{E}\exp(\mathrm{i}k^Z X - \mathrm{i}\omega t) = \hat{Z}\mathcal{E}\exp(\mathrm{i}\varphi_Z - \mathrm{i}\omega t). \tag{6.37}$$

For propagation over a crystal of length l, the total phase shift is

$$\varphi_Z = k^Z l = \frac{\omega}{c} n_Z l = \frac{\omega}{c}\left(n_\mathrm{e} l - \frac{n_\mathrm{e}^3}{2} r_{33} E_{0z} l\right) = \frac{\omega}{c}\left(n_\mathrm{e} l - \frac{n_\mathrm{e}^3}{2} r_{33} V \frac{l}{d}\right), \tag{6.38}$$

where $V = E_{0z} d$ is the voltage applied to the modulator shown in Fig. 6.2(*a*).

For a sinusoidal modulation of a frequency $f = \Omega/2\pi$, the modulation voltage can be written as

$$V(t) = V_\mathrm{pk} \sin \Omega t, \tag{6.39}$$

which has a peak value of V_pk. The optical field at the output plane, $X = l$, of the crystal is

$$\mathbf{E}(l,t) = \hat{Z}\mathcal{E}\mathrm{e}^{\mathrm{i}\omega n_\mathrm{e} l/c} \exp\left[-\mathrm{i}(\omega t + \varphi_\mathrm{pk} \sin \Omega t)\right], \tag{6.40}$$

where

$$\varphi_\mathrm{pk} = \frac{\omega}{c}\frac{n_\mathrm{e}^3}{2} r_{33} V_\mathrm{pk} \frac{l}{d} = \frac{\pi n_\mathrm{e}^3}{\lambda} r_{33} V_\mathrm{pk} \frac{l}{d} \tag{6.41}$$

is the peak phase shift known as the *phase modulation depth* for the Z-polarized optical field. Using the Bessel-function identities

$$\exp(-\mathrm{i}\varphi_\mathrm{pk} \sin \Omega t) = \sum_{q=-\infty}^{\infty} J_q(\varphi_\mathrm{pk}) \mathrm{e}^{-\mathrm{i}q\Omega t} \tag{6.42}$$

and $J_{-q} = (-1)^q J_q$, we find that

$$\mathbf{E}(l,t) = \hat{Z}\mathcal{E}\mathrm{e}^{\mathrm{i}\omega n_\mathrm{e} l/c} \left\{ J_0(\varphi_\mathrm{pk}) \mathrm{e}^{-\mathrm{i}\omega t} + \sum_{q=1}^{\infty} J_q(\varphi_\mathrm{pk}) \left[\mathrm{e}^{-\mathrm{i}(\omega + q\Omega)t} + (-1)^q \mathrm{e}^{-\mathrm{i}(\omega - q\Omega)t}\right] \right\}. \tag{6.43}$$

Thus, a series of side bands at the harmonics of the modulation frequency are generated on both high- and low-frequency sides of the optical carrier frequency by the sinusoidal phase modulation.

If the optical field is instead linearly polarized in the Y direction, the phase shift after propagation through the crystal is

$$\varphi_Y = k^Y l = \frac{\omega}{c} n_Y l = \frac{\omega}{c}\left(n_\mathrm{o} l - \frac{n_\mathrm{o}^3}{2} r_{13} E_{0z} l\right) = \frac{\omega}{c}\left(n_\mathrm{o} l - \frac{n_\mathrm{o}^3}{2} r_{13} V \frac{l}{d}\right). \tag{6.44}$$

The phase modulation depth is then

$$\varphi_{\mathrm{pk}} = \frac{\omega}{c}\frac{n_{\mathrm{o}}^3}{2} r_{13} V_{\mathrm{pk}} \frac{l}{d} = \frac{\pi n_{\mathrm{o}}^3}{\lambda} r_{13} V_{\mathrm{pk}} \frac{l}{d} \tag{6.45}$$

for the modulation voltage given in (6.39). Since $n_{\mathrm{o}} \approx n_{\mathrm{e}}$ but $r_{33} \approx 3.6 r_{13}$, it can be seen by comparison of (6.45) with (6.41) that for a desired modulation depth, the modulation voltage required for a Y-polarized optical wave is about 3.6 times that for a Z-polarized wave.

Longitudinal phase modulators

For the *longitudinal phase modulator* shown in Fig. 6.2(*b*), an optical wave of any polarization in the XY plane will experience the same amount of phase shift because $n_X = n_Y$. For a crystal of length l as shown in Fig. 6.2(*b*), we have

$$\varphi_X = \varphi_Y = \frac{\omega}{c}\left(n_{\mathrm{o}} l - \frac{n_{\mathrm{o}}^3}{2} r_{13} E_{0z} l\right) = \frac{\omega}{c}\left(n_{\mathrm{o}} l - \frac{n_{\mathrm{o}}^3}{2} r_{13} V\right), \tag{6.46}$$

where $V = E_{0z} l$ for the longitudinal modulator. Therefore, with a sinusoidal modulation voltage as given in (6.39), the modulation depth of the longitudinal phase modulator is

$$\varphi_{\mathrm{pk}} = \frac{\omega}{c}\frac{n_{\mathrm{o}}^3}{2} r_{13} V_{\mathrm{pk}} = \frac{\pi n_{\mathrm{o}}^3}{\lambda} r_{13} V_{\mathrm{pk}}, \tag{6.47}$$

which is independent of crystal length l.

It is seen that the voltage required for a given modulation depth is independent of the physical dimensions of the modulator in the case of longitudinal modulation, while it is proportional to d/l in the case of transverse modulation. One advantage of transverse modulation is that the required modulation voltage can be substantially lowered by reducing the d/l dimensional ratio of a transverse modulator. Another advantage is that the electrodes of a transverse modulator can be made with standard techniques and can be patterned if desired, while those of a longitudinal modulator have to be made of transparent conductors that can be very difficult, if not impossible, to fabricate in the dimensions of a typical optical waveguide. However, if a large input and output aperture is desired such that $d/l > 1$, it becomes advantageous to use longitudinal modulation rather than transverse modulation.

The relative advantages and disadvantages of transverse versus longitudinal modulation discussed above also hold true for the polarization and intensity modulators discussed in the following.

EXAMPLE 6.2 As a practical example, consider the $LiNbO_3$ transverse and longitudinal phase modulators shown in Figs. 6.2(*a*) and (*b*), respectively, where the modulation

voltage is applied along the z axis of the crystal. Find the required voltage $V_{\rm pk}$ for a phase modulation depth of $\varphi_{\rm pk} = \pi$ at $\lambda = 1$ μm for optical waves of different polarizations.

Solution We first consider the transverse modulator shown in Fig. 6.2(*a*). Using (6.41) and the parameters of $r_{13} = 8.6$ pm V^{-1}, $r_{33} = 30.8$ pm V^{-1}, $n_{\rm e} = 2.159$, and $n_{\rm o} = 2.238$ given in Table 6.2 for $LiNbO_3$, the peak voltage required to have $\varphi_{\rm pk} = \pi$ for a Z-polarized wave is found to be

$$V_{\rm pk} = \frac{\lambda}{n_{\rm e}^3 r_{33}} \frac{d}{l} = \frac{1 \times 10^{-6}}{2.159^3 \times 30.8 \times 10^{-12}} \frac{d}{l} \text{ V} = 3.23 \frac{d}{l} \text{ kV}.$$

Using (6.45), we find that the required peak voltage for a Y-polarized wave is

$$V_{\rm pk} = \frac{\lambda}{n_{\rm o}^3 r_{13}} \frac{d}{l} = \frac{1 \times 10^{-6}}{2.238^3 \times 8.6 \times 10^{-12}} \frac{d}{l} \text{ V} = 10.4 \frac{d}{l} \text{ kV}.$$

For a bulk modulator where d and l are generally of the same order of magnitude, the required modulation voltage is on the order of kilovolts. However, for a waveguide modulator of typical waveguide dimensions, d/l is of the order of 10^{-3}. For example, in a transverse waveguide modulator that has dimensions of $d = 5$ μm and $l = 5$ mm, the peak voltage required is reduced to 3.23 and 10.4 V for Z- and Y-polarized waves, respectively.

For the longitudinal modulator shown in Fig. 6.2(*b*), the optical wave is polarized in the XY plane. From (6.47), we find that the peak voltage required for $\varphi_{\rm pk} = \pi$ is always

$$V_{\rm pk} = \frac{\lambda}{n_{\rm o}^3 r_{13}} = \frac{1 \times 10^{-6}}{2.238^3 \times 8.6 \times 10^{-12}} \text{ V} = 10.4 \text{ kV},$$

irrespective of the dimensions of the longitudinal modulator or the polarization of the optical wave.

Polarization modulators

In the operation of an electro-optic polarization modulator, the optical wave is not linearly polarized in a direction that is parallel to any of the principal axes in the presence of the modulation field. The optical field can be decomposed into two linearly polarized normal modes. If the two normal modes see different field-induced indices of refraction, there is a field-dependent *phase retardation* between the two modes. The polarization of the optical wave at the output of the crystal can then be controlled by the modulation field.

The $LiNbO_3$ transverse modulator discussed above becomes a polarization modulator if the input optical field polarized in the YZ plane is parallel to neither $\hat{Y}$ nor $\hat{Z}$:

$$\mathbf{E}(0, t) = (\hat{Y}\mathcal{E}_Y + \hat{Z}\mathcal{E}_Z)\mathrm{e}^{-\mathrm{i}\omega t}, \tag{6.48}$$

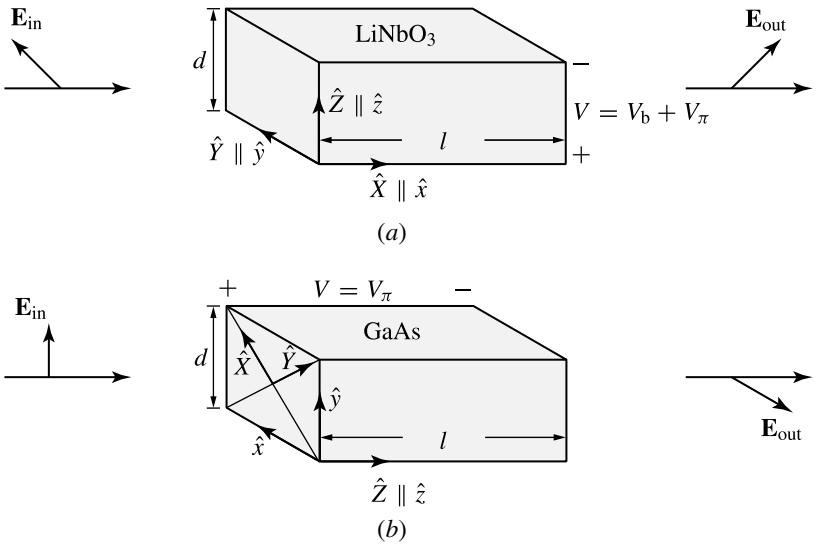


Figure 6.3 (*a*) $LiNbO_3$ transverse electro-optic polarization modulator. (*b*) GaAs longitudinal electro-optic polarization modulator. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the original principal axes of the crystal, and $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ represent its new principal axes.

where $\mathcal{E}_Y \neq 0$ and $\mathcal{E}_Z \neq 0$, as is shown in Fig. 6.3(*a*). At the output, we have

$$\mathbf{E}(l,t) = \left(\hat{Y}\mathcal{E}_Y e^{ik^Y l} + \hat{Z}\mathcal{E}_Z e^{ik^Z l}\right) e^{-i\omega t} = \left(\hat{Y}\mathcal{E}_Y e^{i\Delta\varphi} + \hat{Z}\mathcal{E}_Z\right) e^{ik^Z l - i\omega t}, \tag{6.49}$$

where

$$\Delta\varphi = (k^Y - k^Z)l \tag{6.50}$$

is the phase retardation between the Y and Z components. Using (6.25), we have

$$\Delta\varphi = \frac{\pi}{\lambda}\left[2(n_o - n_e)l + (n_e^3 r_{33} - n_o^3 r_{13})V\frac{l}{d}\right]. \tag{6.51}$$

The intrinsic birefringence of a uniaxial $LiNbO_3$ crystal causes a voltage-independent background phase retardation of $\Delta\varphi_0 = 2\pi(n_o - n_e)l/\lambda$ at $V = 0$ in the absence of an applied field. For a given crystal length, the background phase retardation is fixed, and the function of the device cannot be varied if no modulation field is applied. By properly choosing the value of l, the device can function as a quarter-wave or half-wave plate, as discussed in Section 1.6. An applied field causes an additional voltage-dependent phase retardation. The output polarization state of a given input optical wave with nonvanishing Y- and Z-field components can be varied by varying the modulation voltage. Thus, the device functions as a voltage-controlled polarization modulator.

For proper operation of the device as a voltage-controlled polarization modulator, the background phase retardation can be compensated by a fixed bias voltage. To find the compensation bias voltage, we note that any phase retardation that is an integral

multiple of 2π has no net effect on the polarization of the optical wave at the output of the device. Therefore, the net effect of the background phase retardation can be evaluated by expressing $\Delta\varphi_0$ as

$$\Delta\varphi_0 = \frac{2\pi}{\lambda}(n_\mathrm{o} - n_\mathrm{e})l = 2m\pi + \frac{2\pi}{\lambda}(n_\mathrm{o} - n_\mathrm{e})\Delta l, \tag{6.52}$$

where m is a properly chosen integer for

$$-\frac{\lambda}{2(n_\mathrm{o} - n_\mathrm{e})} < \Delta l = l - \frac{m\lambda}{n_\mathrm{o} - n_\mathrm{e}} \leq \frac{\lambda}{2(n_\mathrm{o} - n_\mathrm{e})} \tag{6.53}$$

so that

$$-\pi < \Delta\varphi_0 - 2m\pi = \frac{2\pi}{\lambda}(n_\mathrm{o} - n_\mathrm{e})\Delta l \leq \pi. \tag{6.54}$$

The polarization of the output wave is actually only determined by the following differential phase retardation:

$$\Delta\varphi - 2m\pi = \frac{2\pi}{\lambda}(n_\mathrm{o} - n_\mathrm{e})\Delta l + \frac{\pi}{\lambda}(n_\mathrm{e}^3 r_{33} - n_\mathrm{o}^3 r_{13})V\frac{l}{d}. \tag{6.55}$$

A fixed bias voltage for compensation of the background phase retardation can then be chosen as

$$V_\mathrm{b} = \frac{2(n_\mathrm{e} - n_\mathrm{o})}{n_\mathrm{e}^3 r_{33} - n_\mathrm{o}^3 r_{13}}\frac{d}{l}\Delta l = \frac{2(n_\mathrm{e} - n_\mathrm{o})}{\lambda}\Delta l V_\pi, \tag{6.56}$$

where

$$V_\pi = \frac{\lambda}{n_\mathrm{e}^3 r_{33} - n_\mathrm{o}^3 r_{13}}\frac{d}{l} \tag{6.57}$$

is the *half-wave voltage*, which can also be denoted as $V_{\lambda/2}$. The voltage-controlled phase retardation can then be recast in the following form:

$$\Delta\varphi - 2m\pi = \frac{V - V_\mathrm{b}}{V_\pi}\pi. \tag{6.58}$$

From (6.53) and (6.56), we find that the bias voltage can always be chosen within a range of $-V_\pi \leq V_\mathrm{b} < V_\pi$. At $V - V_\mathrm{b} = \pm V_\pi$, the device functions as a half-wave plate that has a phase retardation of $\Delta\varphi = 2m\pi \pm \pi$. At $V - V_\mathrm{b} = \pm V_\pi/2$, the device functions as a quarter-wave plate with a phase retardation of $\Delta\varphi = 2m\pi \pm \pi/2$. Therefore, the *quarter-wave voltage* $V_{\pi/2}$, or $V_{\lambda/4}$, is half that of the half-wave voltage, both measured with respect to the bias point.

The background phase retardation contributed by the intrinsic birefringence is the major drawback of the $LiNbO_3$ transverse polarization modulator discussed here. Although it can be compensated by a bias voltage that falls within the range of $\pm V_\pi$, the requirement of such a DC bias voltage complicates the operation of the device, particularly when it is modulated at a high frequency. Because the bias voltage depends on the length and the refractive indices of the device, it is susceptible to changes in the

operating condition, such as temperature variations caused by operation of the device. Furthermore, the bias voltage is also a function of optical wavelength because Δl varies as the optical wavelength varies. In practice, the bias voltage has to be carefully adjusted for each individual device in a given operating condition due to small variations in device length and refractive indices.

EXAMPLE 6.3 A $LiNbO_3$ transverse polarization modulator for $\lambda = 1$ μm as shown in Fig. 6.3(a) has dimensions of $d = 5$ μm and $l = 5$ mm. Find the half-wave voltage V_π and the bias voltage V_b required for compensating the background phase retardation from the intrinsic birefringence of $LiNbO_3$. If the length varies by ± 5 μm due to fabrication errors or changes in the operating condition, what are the changes in V_π and V_b, respectively?

Solution Using the parameters of $LiNbO_3$ given in Table 6.2, we find from (6.57) that $V_\pi \approx 4.68$ V. With $n_o - n_e = 0.079$ at $\lambda = 1$ μm found from the data in Table 6.2, we find that $\Delta l = 0$ with $m = 395$ for $l = 5$ mm. Therefore, $V_b = 0$ from (6.56).

A length variation of ± 5 μm amounts to a change of $\pm 0.1\%$ in the total length. From (6.57), we find that it causes a change of only about $\mp 0.1\%$ in V_π. However, it results in $\Delta l = \pm 5$ μm, also with $m = 395$. From (6.56), we find that the required compensation bias voltage is $V_b \approx \mp 3.70$ V. We therefore see that a small variation in the length of the device causes a similarly small change in V_π, but it can lead to a large change in the compensation bias voltage.

The $LiNbO_3$ longitudinal modulator shown in Fig. 6.2(b) cannot function as a polarization modulator because $n_X = n_Y$. Instead, we consider the GaAs longitudinal modulator shown in Fig. 6.3(b). In this case, the principal axes and their corresponding indices of refraction in the presence of a modulation field are those given in (6.35) and (6.36), respectively. The optical wave to be modulated propagates in the Z direction and has both X and Y field components. At the input end, it can be written as

$$\mathbf{E}(0, t) = \left(\hat{X}\mathcal{E}_X + \hat{Y}\mathcal{E}_Y\right) e^{-i\omega t}. \tag{6.59}$$

After propagating through the crystal, the optical field is

$$\mathbf{E}(l, t) = \left(\hat{X}\mathcal{E}_X e^{ik^X l} + \hat{Y}\mathcal{E}_Y e^{ik^Y l}\right) e^{-i\omega t} = \left(\hat{X}\mathcal{E}_X + \hat{Y}\mathcal{E}_Y e^{i\Delta\varphi}\right) e^{ik^X l - i\omega t}, \tag{6.60}$$

where

$$\Delta\varphi = (k^Y - k^X)l \tag{6.61}$$

is the phase retardation between the Y and X components of the optical field. Using (6.36) and the fact that $V = E_{0z}l$ for the longitudinal modulator, we have

$$\Delta\varphi = \frac{2\pi}{\lambda} n_o^3 r_{41} V = \frac{V}{V_\pi}\pi, \tag{6.62}$$

where the half-wave voltage is

$$V_\pi = \frac{\lambda}{2n_o^3 r_{41}}. \tag{6.63}$$

It can be seen from (6.62) that no bias voltage is needed for this GaAs modulator because both x and y axes are ordinary axes in the absence of an applied field. However, because of the longitudinal modulation scheme, V_π is a constant independent of both dimensions l and d. Therefore, in comparison with the $LiNbO_3$ transverse modulator, the advantage of this modulator in requiring no bias voltage is completely offset by the disadvantage due to its longitudinal modulation scheme. In a GaAs transverse polarization modulator, both problems can be eliminated (see Problem 6.3.8).

EXAMPLE 6.4 For $\lambda = 1$ μm, the parameters for GaAs given in Table 6.2 yield $V_\pi \approx 9.72$ kV from (6.63), which is independent of the dimensions of the GaAs longitudinal modulator. Though no bias voltage is needed because GaAs has no intrinsic birefringence, this half-wave voltage cannot be reduced by varying the dimensions of the modulator. For $\Delta\varphi$ to vary in the range between 0 and π, the modulation voltage has to be varied between 0 and 9.72 kV.

Amplitude modulators

An electro-optic amplitude modulator can be constructed by simply placing a polarization modulator between a polarizer at the input end and another, often referred to as an *analyzer*, at the output end. Usually, the axis of the polarizer and that of the analyzer are arranged to be orthogonally crossed, although other arrangements are possible.

Figure 6.4 shows a typical setup for a GaAs longitudinal amplitude modulator. In this arrangement, the polarizer ensures that the input optical wave is linearly polarized in the y direction while the analyzer passes only the x component of the optical wave at the output end. The input field is $\mathbf{E}(0, t) = \hat{y}\mathcal{E}e^{-i\omega t}$, which can be written in the form of (6.59) with $\mathcal{E}_X = \mathcal{E}_Y = \mathcal{E}/\sqrt{2}$. Then, from (6.60), the field at the output end of the crystal is

$$\mathbf{E}(l, t) = \frac{\mathcal{E}}{\sqrt{2}}(\hat{X} + \hat{Y}e^{i\Delta\varphi})e^{ik^X l - i\omega t} = \frac{\mathcal{E}}{2}\left[\hat{x}(1 - e^{i\Delta\varphi}) + \hat{y}(1 + e^{i\Delta\varphi})\right]e^{ik^X l - i\omega t}, \tag{6.64}$$

where $\Delta\varphi$ is the same as that given in (6.61). Because the analyzer passes only the x component of the optical field, the transmittance of the amplitude modulator is

$$T = \frac{I_{out}}{I_{in}} = \sin^2\frac{\Delta\varphi}{2} = \frac{1}{2}(1 - \cos\Delta\varphi). \tag{6.65}$$

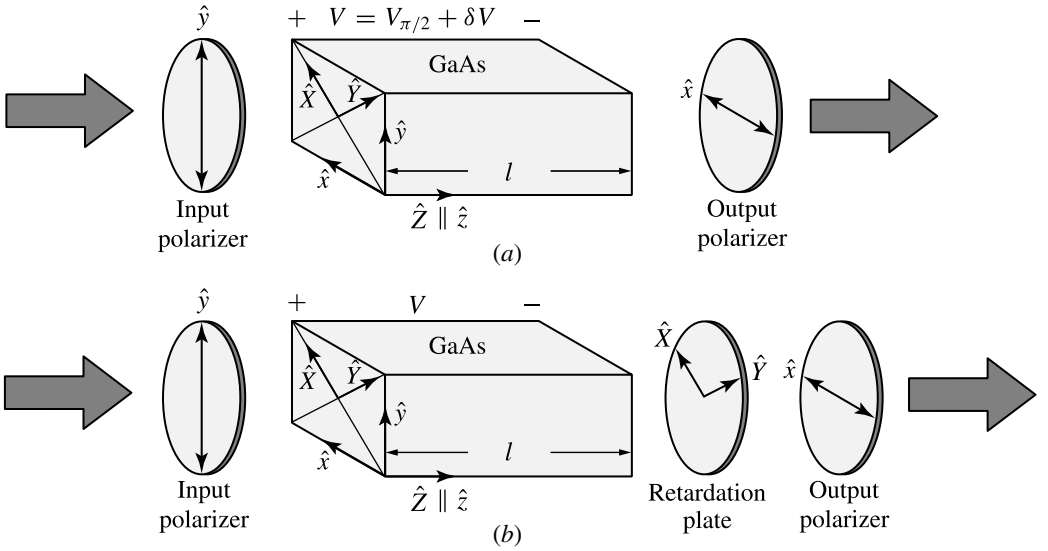


Figure 6.4 GaAs longitudinal electro-optic amplitude modulator. A bias phase retardation of $\pi/2$ can be introduced with (*a*) a bias voltage $V_b = V_{\pi/2}$ or (*b*) a properly oriented quarter-wave plate. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the original principal axes of the crystal, and $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ represent its new principal axes.

A similar result is obtained for an amplitude modulator constructed by placing any other polarization modulator, such as the $LiNbO_3$ transverse polarization modulator shown in Fig. 6.3(*a*), between a pair of properly oriented polarizer and analyzer.

Because $\Delta\varphi$ varies linearly with applied voltage V, the amplitude modulator would have a linear response if its transmittance T varied linearly with $\Delta\varphi$. It can be seen from (6.65) that this is not generally true. However, for small variations of $\Delta\varphi$, it is approximately true near $\Delta\varphi = \pi/2$, as can be seen by substituting

$$\Delta\varphi = \frac{\pi}{2} + \delta\varphi \tag{6.66}$$

in (6.65) to get

$$T = \frac{1}{2}(1 + \sin\delta\varphi) \approx \frac{1}{2}(1 + \delta\varphi) \tag{6.67}$$

for $|\delta\varphi| \ll \pi$. By setting the operating point at a bias phase retardation of $\Delta\varphi_b = \pi/2$, the device has a linear small-signal response, as shown in Fig. 6.5. Then, with a voltage such as that given in (6.39), the output intensity will be sinusoidally modulated. This bias phase retardation can be obtained either by operating the device with a fixed bias voltage of $V_b = V_{\pi/2}$, as shown in Fig. 6.4(*a*), or by inserting a properly oriented quarter-wave plate between the modulator crystal and the analyzer to introduce an extra fixed phase retardation of $\pi/2$, as shown in Fig. 6.4(*b*).

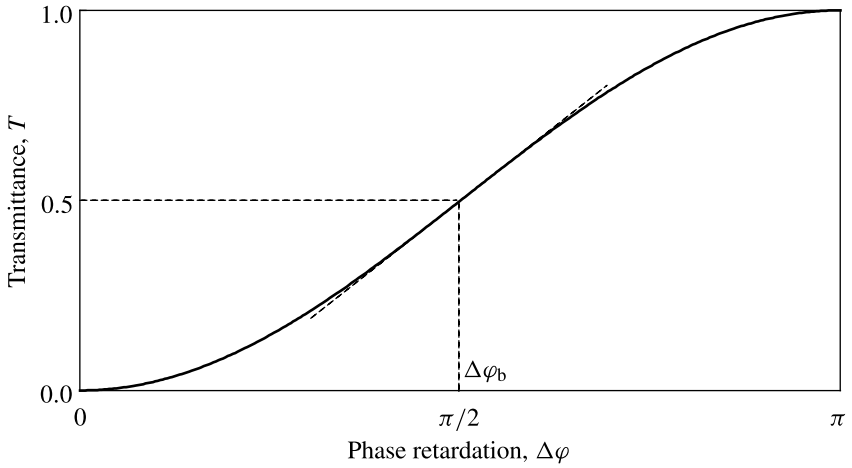


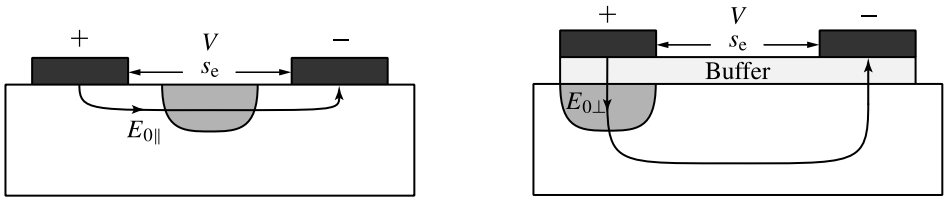
Figure 6.5 Transmission characteristics of the electro-optic amplitude modulator shown in Fig. 6.4. The response is nearly linear for small variations near the operating point at $\Delta\varphi_b = \pi/2$.

For large-signal applications, the response of the amplitude modulator is nonlinear. The device is then often used as an electro-optically controlled ON–OFF modulator. In these types of applications, a bias is neither useful nor necessary.

6.4 Guided-wave electro-optic modulators

Optical waveguides possess many unique characteristics that do not exist in bulk optics. An important one is their ability to guide optical waves within a small cross-sectional area over a long distance. This allows for the possibility of using the transverse modulation scheme to realize very efficient modulators at very low modulation voltages. In bulk optics, the ratio of the length to the cross-sectional dimensions is limited by the diffraction effect, limiting the advantage that can be realized using transverse modulation. This limitation does not exist in waveguide optics. Another unique characteristic is the existence of waveguide modes. This results in many phenomena that have no counterpart in bulk optics, such as mode coupling, mode conversion, and modal dispersion. These unique features are the basis of many devices that take advantage of the waveguide configuration. In addition, guided-wave electro-optic devices are important building-block components of integrated optical and integrated optoelectronic systems.

The modulation electric field in a waveguide is usually the fringe field around surface electrodes or, in some cases of semiconductor waveguides, the field resulting from a junction voltage drop. Figure 6.6 shows the two commonly used approaches for buried waveguides, particularly the Ti-diffused $LiNbO_3$ waveguides, using surface-loading electrodes. In the configuration shown in Fig. 6.6(*a*), the electrodes are placed on two



(*a*) (*b*)

Figure 6.6 Configurations for applying a modulation field to a buried waveguide through surface-loading electrodes. In (*a*), $E_{0\parallel}$ is applied to the waveguide. In (*b*), $E_{0\perp}$ is applied. The buffer layer in (*b*) is required to reduce loss to TM-like modes.

sides of the waveguide, and the horizontal electric field $E_{0\parallel}$ is applied. In the configuration shown in Fig. 6.6(*b*), one of the electrodes is placed directly over the waveguide, and the applied electric field is the vertical $E_{0\perp}$. The buried waveguide shown in Fig. 6.6 is a channel waveguide, but the index step at the air–crystal interface along the vertical direction is much higher than those at other waveguide boundaries. Therefore, modes with electric fields polarized mainly parallel to the air–crystal interface are called *TE-like modes*, whereas those with electric fields polarized mainly perpendicular to this interface are called *TM-like modes*. When an electrode is placed directly over a waveguide, an insulating buffer layer, usually SiO_2 or Al_2O_3, between the electrode and the substrate crystal is needed to ensure low loss for TM-like modes, as also shown in Fig. 6.6(*b*).

In a waveguide, the modulation electric field applied to a particular waveguide mode depends on a number of parameters, including the geometric dimensions of the waveguide structure and the optical field distribution of the mode. In general, the modulation field is not uniformly distributed across the mode field distribution. The effect of electro-optic modulation in a waveguide can be calculated using the coupled-mode theory discussed in Section 4.2. For modulation on a single mode, the effect is to introduce a change in the propagation constant of the mode. This change is equal to the self-coupling coefficient of the mode given by

$$\Delta\beta_\nu = \kappa_{\nu\nu} = \omega \int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\nu^* \cdot \Delta\boldsymbol{\epsilon} \cdot \hat{\boldsymbol{\mathcal{E}}}_\nu \mathrm{d}\boldsymbol{\rho}, \tag{6.68}$$

where $\Delta\boldsymbol{\epsilon}$ is the electro-optically induced change in the dielectric permittivity tensor and $\boldsymbol{\rho}$ is the two-dimensional vector in the cross-sectional plane of the waveguide. As an example, we consider phase modulation of a TE-like mode in a waveguide modulator that is fabricated in a $LiNbO_3$ crystal with the crystal surface perpendicular to the x principal axis and the longitudinal direction of the waveguide parallel to the y principal axis. This arrangement is shown in Fig. 6.7(*a*) and is referred to as *y propagating* in an *x-cut* crystal. The modulation field appearing in the waveguide area is $E_{0\parallel}$, which is

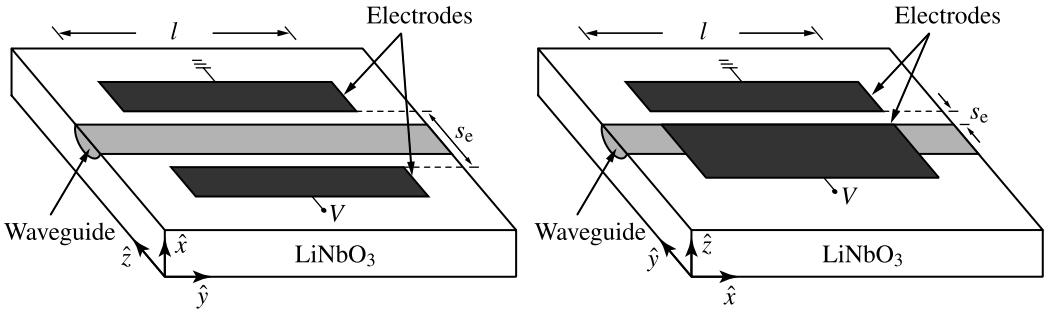


Figure 6.7 Waveguide phase modulators in (*a*) an x-cut, y-propagating $LiNbO_3$ crystal and (*b*) a z-cut, x-propagating $LiNbO_3$ crystal. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the principal axes of the crystal.

E_{0z} in this configuration. Because a TE-like mode of this waveguide is predominantly polarized in the z direction and $\Delta\epsilon_{zz} = -n_e^4 r_{33} E_{0z}$ from (6.24), we have

$$\begin{aligned}\Delta\beta_{TE} &= -n_e^4 r_{33}\omega\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{0z}(x,z)|\hat{\boldsymbol{\mathcal{E}}}_{TE}(x,z)|^2 \mathrm{d}x\mathrm{d}z \\ &= -\frac{n_e^4 r_{33}}{2}\frac{V}{s_e}\frac{\omega^2\mu_0\epsilon_0}{\beta_{TE}}\Gamma_{TE} \\ &\approx -\frac{\pi}{\lambda}n_e^3 r_{33}\frac{V}{s_e}\Gamma_{TE},\end{aligned} \tag{6.69}$$

where V is the applied voltage, s_e is the separation between the electrodes, β_{TE} is approximated by $n_e\omega/c$, (2.44) is used to normalize the mode field, and

$$\begin{aligned}\Gamma_{TE} &= \frac{s_e}{V}\frac{2\beta_{TE}}{\omega\mu_0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{0z}(x,z)|\hat{\boldsymbol{\mathcal{E}}}_{TE}(x,z)|^2 \mathrm{d}x\mathrm{d}z \\ &= \frac{s_e}{V}\frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_{0z}(x,z)|\hat{\boldsymbol{\mathcal{E}}}_{TE}(x,z)|^2 \mathrm{d}x\mathrm{d}z}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{\boldsymbol{\mathcal{E}}}_{TE}(x,z)|^2 \mathrm{d}x\mathrm{d}z}\end{aligned} \tag{6.70}$$

is the *overlap factor*, which accounts for the overlap between the modulation electric field and the optical mode. The overlap factor has a value between 0 and 1. The total electro-optically induced phase shift of this mode over length l of the modulator is simply $\Delta\varphi = \Delta\beta_{TE} l$.

Comparing the result obtained above with that in (6.41) for the bulk phase modulator, we find that the net effect for a waveguide mode ν can be approximated by using a single uniform effective modulation electric field given by

$$E_{\text{eff}} = \Gamma_\nu \frac{V}{s_e}, \tag{6.71}$$

where Γ_ν is evaluated using the appropriate modulation field component for the device configuration under consideration. For example, $E_{0\parallel}$ is used for the electrode configuration in Fig. 6.6(*a*), while $E_{0\perp}$ is used for the configuration in Fig. 6.6(*b*). The value of Γ_ν depends on the electrode configuration and is different for different waveguide modes in the same structure. For a given configuration and a given waveguide mode, it increases monotonically as the ratio of the electrode separation to the horizontal waveguide width increases.

EXAMPLE 6.5 An x-cut, y-propagating $LiNbO_3$ single-mode waveguide phase modulator for $\lambda = 1.3$ μm, as shown in Fig. 6.7(*a*), has a gap separation of $s_e = 20$ μm between its electrodes and an overlap factor of $\Gamma_{TE} = 0.57$ for its TE-like mode. It is modulated with an applied voltage of $V = 12$ V. What is the effective modulation electric field strength? Find the electro-optically induced change in the propagation constant of the TE-like mode. If an electro-optically controlled phase shift of π is desired, what is the required length of the device?

Solution According to (6.71), the effective modulation electric field is

$$E_{\text{eff}} = \Gamma_{TE} \frac{V}{s_e} = 0.57 \times \frac{12}{20 \times 10^{-6}}\ \text{V m}^{-1} = 342\ \text{kV m}^{-1}.$$

From (1.191) in Problem 1.9.2, we find that $n_e = 2.145$ for $LiNbO_3$ at $\lambda = 1.3$ μm. We then find, using (6.69), that

$$\begin{aligned}\Delta\beta_{TE} &= -\frac{\pi}{1.3 \times 10^{-6}} \times 2.145^3 \times 30.8 \times 10^{-12} \times \frac{12}{20 \times 10^{-6}} \times 0.57\ \text{m}^{-1} \\ &= -251.23\ \text{m}^{-1}.\end{aligned}$$

For $\Delta\varphi = \pi$, the required length of the device is

$$l = \frac{\pi}{|\Delta\beta_{TE}|} = 12.5\ \text{mm}.$$

In the following, we discuss a few important electro-optic waveguide devices. The principle of phase modulation discussed in the preceding section can be applied directly to guided-wave phase modulators. Amplitude modulation and switching functions using guided-wave devices are typically realized using either waveguide interferometers or directional couplers. Polarization modulation is accomplished through

electro-optically controlled coupling and conversion between guided modes of different polarizations. Other functions, such as frequency filtering, are also possible using guided-wave devices. Preferably, single-mode waveguide devices are used to attain the best performance at the lowest modulation voltage.

Mach–Zehnder waveguide interferometers

Guided-wave electro-optic phase modulators can be used to construct waveguide interferometers for effective amplitude modulation of guided optical waves. A Mach–Zehnder waveguide interferometer consists of two parallel waveguides connected at the input and output ends, respectively, by beam-splitting and beam-combining optical couplers. These couplers can be *Y-junction* waveguides, as in the devices shown in Fig. 6.8, or directional couplers, as shown in Fig. 6.9. Modulation electric fields are

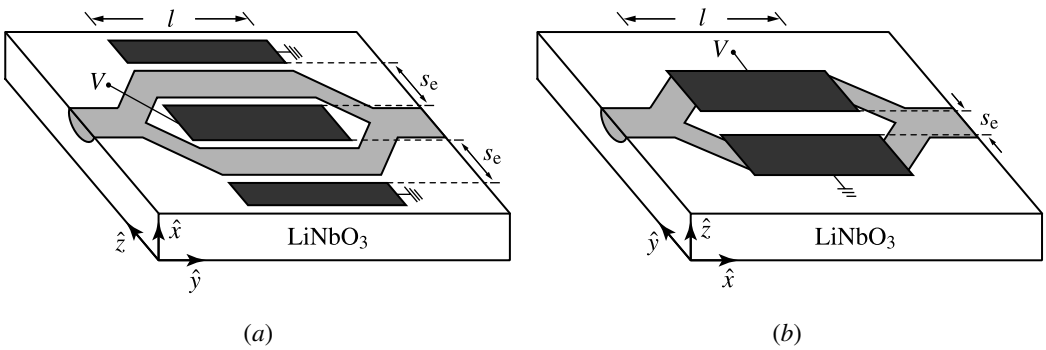


Figure 6.8 Mach–Zehnder waveguide interferometric modulator using Y junctions fabricated on (*a*) an *x*-cut, *y*-propagating $LiNbO_3$ substrate and (*b*) a *z*-cut, *x*-propagating $LiNbO_3$ substrate. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the principal axes of the crystal.

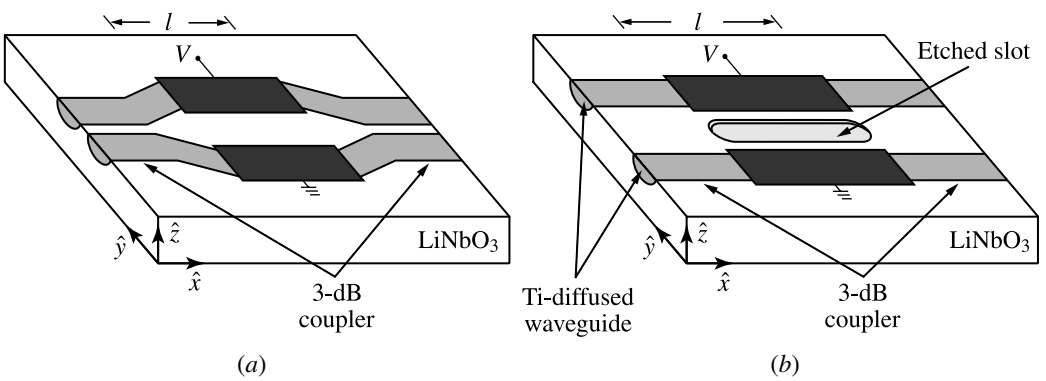


Figure 6.9 Balanced-bridge interferometers fabricated on *z*-cut, *x*-propagating $LiNbO_3$ substrates using (*a*) bent waveguides and (*b*) straight waveguides. Isolation between the two arms of the interferometer is accomplished with a large separation in (*a*) and an etched slot in (*b*). The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the principal axes of the crystal.

applied to two parallel waveguides, which form the two arms of the interferometer and are sufficiently separated to avoid direct coupling between them. Each waveguide by itself functions as an electro-optic phase modulator. Constructive or destructive interference occurs at the output coupler if the phase difference between the two arms is, respectively, an even or odd multiple of π. By electro-optically controlling this phase difference through the applied voltage, the amplitude of the guided optical field at the output can be modulated.

The Mach–Zehnder waveguide interferometer shown in Fig. 6.8(*a*) uses Y-junction couplers and is fabricated in an x-cut, y-propagating $LiNbO_3$ crystal. To use the largest electro-optic coefficient r_{33} in $LiNbO_3$, both the modulation electric field and the optical field have to be polarized in the z direction. This requirement can be fulfilled by using the electrodes shown in Fig. 6.8(*a*) for TE-like modes. In this electrode configuration, the modulation voltage is applied to the central electrode while the outer electrodes are grounded. The modulation electric fields appearing in the two arms point in opposite directions, resulting in a *push–pull operation* with equal but opposite phase shifts in the optical waves propagating through the two arms. For an interferometer with identical arms, any other background phase shifts are exactly canceled. Thus the total phase difference is twice the electro-optically induced phase shift in each arm. If the two arms are identical single-mode waveguides, the phase difference induced by a modulation voltage V for a TE-like mode is

$$\Delta\varphi = \frac{2\pi}{\lambda} n_e^3 r_{33} \Gamma_{TE} \frac{l}{s_e} V = \pi \frac{V}{V_\pi}, \tag{6.72}$$

where

$$V_\pi = \frac{\lambda}{2n_e^3 r_{33} \Gamma_{TE}} \frac{s_e}{l} \tag{6.73}$$

is the half-wave voltage corresponding to a phase difference of π between the two arms. For a TM-like mode, we have

$$V_\pi = \frac{\lambda}{2n_o^3 r_{13} \Gamma_{TM}} \frac{s_e}{l}. \tag{6.74}$$

The half-wave voltage for a TM-like mode is more than three times that for a TE-like mode of a similar overlap factor. Therefore, this particular interferometer favors operation with a TE-like mode.

EXAMPLE 6.6 A Mach–Zehnder waveguide interferometric modulator for $\lambda = 1.3$ μm using Y junctions as shown in Fig. 6.8(*a*) consists of two parallel x-cut, y-propagating $LiNbO_3$ single-mode waveguide phase modulators with $s_e = 20$ μm in a push–pull configuration. Both waveguides have $\Gamma_{TE} = \Gamma_{TM} = 0.57$ and $l = 12.5$ mm, like the

one discussed in Example 6.5. Find the half-wave voltages for the TE-like and TM-like modes, respectively.

Solution We find from (6.73) that V_π for the TE-like mode is

$$V_\pi = \frac{1.3 \times 10^{-6}}{2 \times 2.145^3 \times 30.8 \times 10^{-12} \times 0.57} \times \frac{20 \times 10^{-6}}{12.5 \times 10^{-3}} \text{ V} = 6 \text{ V},$$

which is half of the value of V_π for the phase modulator in Example 6.5 because of the push–pull operation of the interferometer. From (1.190) in Problem 1.9.2, we find that $n_\text{o} = 2.222$ for $LiNbO_3$ at $\lambda = 1.3$ μm. Therefore, from (6.74), we find that V_π for the TM-like mode is

$$V_\pi = \frac{1.3 \times 10^{-6}}{2 \times 2.222^3 \times 8.6 \times 10^{-12} \times 0.57} \times \frac{20 \times 10^{-6}}{12.5 \times 10^{-3}} \text{ V} = 19.3 \text{ V},$$

which is larger than V_π for the TE-like mode because $r_{13} < r_{33}$ for $LiNbO_3$.

In comparison, Fig. 6.8(*b*) shows a Mach–Zehnder interferometer fabricated on a z-cut, x-propagating $LiNbO_3$ substrate. In this configuration, the electrodes have to be placed directly over the waveguides in order to use r_{33}. For a push–pull operation with equal but opposite phase shifts in the two arms of this interferometer, only two electrodes are needed with one receiving the modulation voltage and the other grounded, as illustrated in Fig. 6.8(*b*). This interferometer favors a TM-like mode, which has a lower half-wave voltage of

$$V_\pi = \frac{\lambda}{2n_\text{e}^3 r_{33} \Gamma_\text{TM}} \frac{s_\text{e}}{l} \tag{6.75}$$

than the half-wave voltage of

$$V_\pi = \frac{\lambda}{2n_\text{o}^3 r_{13} \Gamma_\text{TE}} \frac{s_\text{e}}{l} \tag{6.76}$$

for a TE-like mode of a similar overlap factor.

For a Mach–Zehnder waveguide interferometer using Y junctions, if both input and output Y junctions are ideal 3-dB couplers, the power transmittance for a specific guided mode is

$$T = \frac{P_\text{out}}{P_\text{in}} = \cos^2 \frac{\Delta\varphi}{2} = \frac{1}{2}(1 + \cos \Delta\varphi). \tag{6.77}$$

For applications as a small-signal amplitude modulator, the device can be operated with a fixed bias voltage of $V_\text{b} = V_\pi/2$ or $-V_\pi/2$ for linear response. For applications as an ON–OFF modulator, no bias is needed. The maximum transmittance in the ON state versus the minimum transmittance in the OFF state is defined as the *extinction ratio*. It

is usually measured in decibels:

$$\mathrm{ER} = -10 \log \frac{T_{\min}}{T_{\max}} \quad \text{(dB)}. \tag{6.78}$$

One important advantage of the waveguide interferometer is that a very high extinction ratio can be accomplished with single-mode structures at a low modulation voltage. The major source of incomplete extinction in the OFF state comes from the imbalance between the two arms due to small fabrication errors, but a single-mode waveguide interferometer is very tolerant of this small imbalance. In a multimode structure, however, different modes have different V_π because Γ is different for different modes. Because this variation results in different $\Delta\varphi$ for different modes at a specific modulation voltage, a high extinction ratio is difficult to accomplish when more than one mode is excited.

Instead of Y junctions, 3-dB directional couplers can be used at both input and output ends of a Mach–Zehnder waveguide interferometer. This type of Mach–Zehnder interferometer is called the *balanced-bridge interferometer*. Figure 6.9 shows two examples. The phase-shifter section consists of two decoupled identical phase modulators. It has exactly the same function as that in a Mach–Zehnder interferometer using Y junctions. However, both input and output ports now have two channels. If straight waveguides are used, as in the example shown in Fig. 6.9(*b*), the waveguides have to be closely spaced to allow coupling in the coupler sections. Crosstalk due to coupling in the phase-shifter section has to be eliminated. This objective can be accomplished by etching a slot in the gap, as shown in Fig. 6.9(*b*), or by mismatching the two waveguides.

With input to only one channel, the straight transmission through the same channel at the output is

$$T = \sin^2 \frac{\Delta\varphi}{2} = \frac{1}{2}(1 - \cos \Delta\varphi), \tag{6.79}$$

where $\Delta\varphi$ is the electro-optically induced phase difference between the two arms of the interferometer discussed above. The crossover efficiency to the other output channel is

$$\eta = 1 - T = \cos^2 \frac{\Delta\varphi}{2} = \frac{1}{2}(1 + \cos \Delta\varphi). \tag{6.80}$$

When used as a modulator, a balanced-bridge interferometer has two complementary output channels. Otherwise, it has similar characteristics to those of an interferometer using Y junctions. In addition, a balanced-bridge interferometer can also be used as an optical switch. When $\Delta\varphi$ is 0 or any integral multiple of 2π, the interferometer is in the cross state because $\eta = 1$. This feature is expected because in this situation the phase-shifter section has null net effect, and the function of the interferometer is simply that of two serially connected 3-dB directional couplers. When $\Delta\varphi$ is equal to any odd integral multiple of π, the interferometer is in the parallel state with $T = 1$. By switching the control voltage for $\Delta\varphi$ to have even or odd multiples of π, the device can be electrically switched between the two switch states.

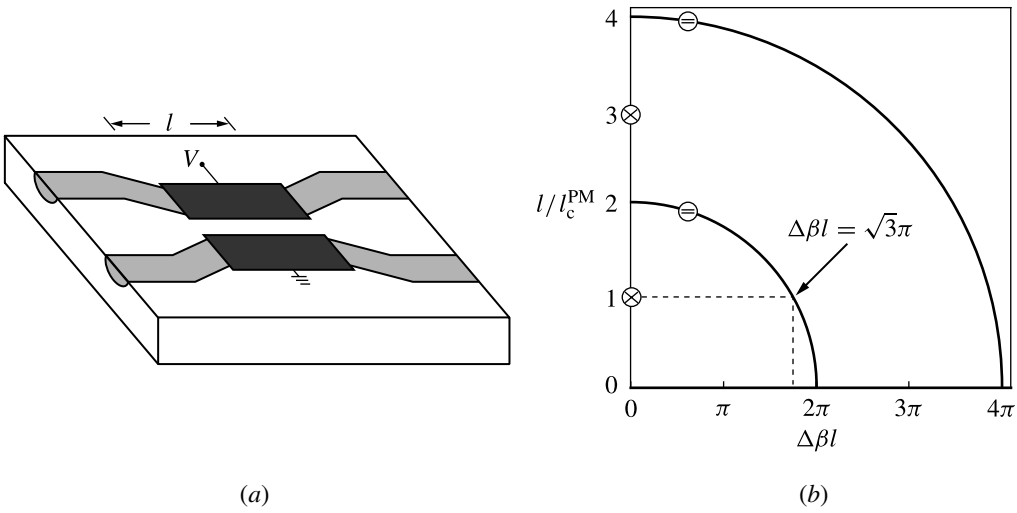


Figure 6.10 (*a*) Schematic structure and (*b*) switching diagram of an electro-optic uniform-$\Delta\beta$ directional coupler switch. Note that a trivial solution of $l = 0$ exists for the parallel state.

Directional coupler switches

A very important practical application of directional couplers is to use them as optical switches, which can be switched between cross and parallel states. This function requires switching the coupling efficiency η between the values of 1 and 0. For a fabricated device with fixed geometric parameters, this can be done by varying the phase mismatch or the coupling coefficient between the two waveguides through electro-optically induced changes in the refractive index of the waveguide material. For simplicity, we consider in the following only symmetric directional couplers where the two waveguide channels have identical geometric and material parameters. Any differences between the two channels are thus induced solely by the applied electric field through the electro-optic effect. Similar concepts can be applied to asymmetric directional couplers as well.

Figure 6.10(*a*) shows the basic structure of an electro-optic directional coupler switch. This structure has a two-electrode configuration similar to that of the two-electrode Mach–Zehnder interferometers. The only difference is that the two waveguides in a directional coupler are coupled while those in an interferometer are isolated. If the device is fabricated on a z-cut, x-propagating $LiNbO_3$ substrate, the effect of the applied voltage is also to induce a phase difference of $\Delta\varphi = \pi V/V_\pi$, with V_π given by (6.75) and (6.76) for TM-like and TE-like modes, respectively.

Without an applied voltage, the symmetric coupler is perfectly phase matched with $\beta_a = \beta_b$, $\kappa_{aa} = \kappa_{bb}$, and $\kappa_{ab} = \kappa_{ba}^* \equiv \kappa$, where κ is real and positive. The effective propagation constant for each waveguide is $\beta = \beta_a + \kappa_{aa} = \beta_b + \kappa_{bb}$, and the phase-matched coupling length is $l_c^{PM} = \pi/2\kappa$ given in (5.54). The phase difference $\Delta\varphi$

induced by the applied voltage results in a phase mismatch of

$$2\delta = \Delta\beta = \frac{\Delta\varphi}{l} \tag{6.81}$$

between the two waveguides of the coupler. To first order, the changes in the coupling coefficient κ induced by the electric field can be neglected in this two-electrode configuration. More significant changes in κ are possible using a three-electrode configuration.

With a phase mismatch given by (6.81), the coupling efficiency of this switch is that given by (4.88):

$$\eta = \frac{1}{1 + \delta^2/\kappa^2} \sin^2\left(\kappa l \sqrt{1 + \delta^2/\kappa^2}\right). \tag{6.82}$$

Because $\Delta\varphi$ and hence the value of δ are linearly proportional to the applied voltage, the switching function of the device can be accomplished by electrically switching δ between the values corresponding to $\eta = 1$ and 0. However, it can be seen from (6.82) that if $\delta \neq 0$, it is not possible to have $\eta = 1$ though $\eta = 0$ is possible. Therefore, to allow access to both cross and parallel states, it is necessary to design the device to be in the cross state when there is no applied voltage, $V = 0$ and thus $\delta = 0$. This requirement means that the coupler has to be perfectly symmetric and its length has to be exactly one of the odd integral multiples of l_c^{PM}:

$$l = (2n + 1) l_c^{PM}, \qquad n = 0, 1, 2, \ldots \tag{6.83}$$

The parallel state can then be reached with an applied voltage to induce a δ that satisfies $\kappa l \sqrt{1 + \delta^2/\kappa^2} = m\pi$ so that $\eta = 0$. Using (6.81) and (6.83), this condition can be cast in the following form:

$$\left(\frac{l}{l_c^{PM}}\right)^2 + \left(\frac{\Delta\beta l}{\pi}\right)^2 = 4m^2, \qquad m = 0, 1, 2, \ldots \tag{6.84}$$

These conditions for the cross and the parallel state are plotted in Fig. 6.10(b). As an example, we see that if $l = l_c^{PM}$, the parallel state can be first reached with $\Delta\varphi = \Delta\beta l = \sqrt{3}\pi$. This phase shift corresponds to a switching voltage $V_s = \sqrt{3} V_\pi$, which is $\sqrt{3}$ times that needed for the balanced-bridge interferometer switch discussed above to reach the first parallel state. As can be seen from Fig. 6.10(b), for $l = l_c^{PM}$, the parallel state can be further reached with $\Delta\beta l = \sqrt{4m^2 - 1}\pi$ at higher applied voltages, but the cross state exists only when $\Delta\beta l = 0$.

EXAMPLE 6.7 A uniform-$\Delta\beta$ directional coupler switch as shown in Fig. 6.10 for $\lambda = 1.3$ μm is fabricated on a z-cut, x-propagating $LiNbO_3$ substrate. The gap separation between the electrodes is $s_e = 5$ μm, and the overlap factor for the TM-like mode is found to be $\Gamma_{TM} = 0.247$ for optical waveguides of 6 μm width. A desired coupling coefficient between the two parallel waveguides can be obtained by properly choosing the spacing between the parallel waveguides in the coupling section covered by the

electrodes. If a switching voltage of 5 V for the TM-like mode is desired, what are the required coupling coefficient and the length of the coupling section?

Solution The length of the coupling section can be chosen to be $l = l_c^{PM} = \pi/2\kappa$, which is the shortest length required to have access to both cross and parallel states. Then the switching voltage is $V_s = \sqrt{3}V_\pi$. For a switching voltage of 5 V, we find that $V_\pi = 2.89$ V. For this z-cut device in a push–pull configuration, V_π is that given by (6.75). Therefore, we find that the required length for $V_\pi = 2.89$ V is

$$l = \frac{\lambda}{2n_e^3 r_{33}\Gamma_{TM}}\frac{s_e}{V_\pi} = \frac{1.3 \times 10^{-6}}{2 \times 2.145^3 \times 30.8 \times 10^{-12} \times 0.247} \times \frac{5 \times 10^{-6}}{2.89}\ \text{m} = 15\ \text{mm}.$$

The required coupling coefficient is

$$\kappa = \frac{\pi}{2l} = 104.72\ \text{m}^{-1}.$$

The function of the directional coupler switch shown in Fig. 6.10 depends critically on the accuracy of fabrication because its cross state cannot be reached by tuning the applied voltage but requires precise symmetry and the exact length of the coupler. Any slight deviation in the symmetry or in the length results in crosstalk to the other channel in the cross state. This limitation can be overcome by using the *reversed-*$\Delta\beta$ configuration shown in Fig. 6.11(*a*). In this configuration, voltages of equal magnitude but opposite polarities are applied to the split electrodes. If the electro-optically induced phase mismatch in the first section of a length of $l/2$ is $\Delta\beta = 2\delta$, that in the second

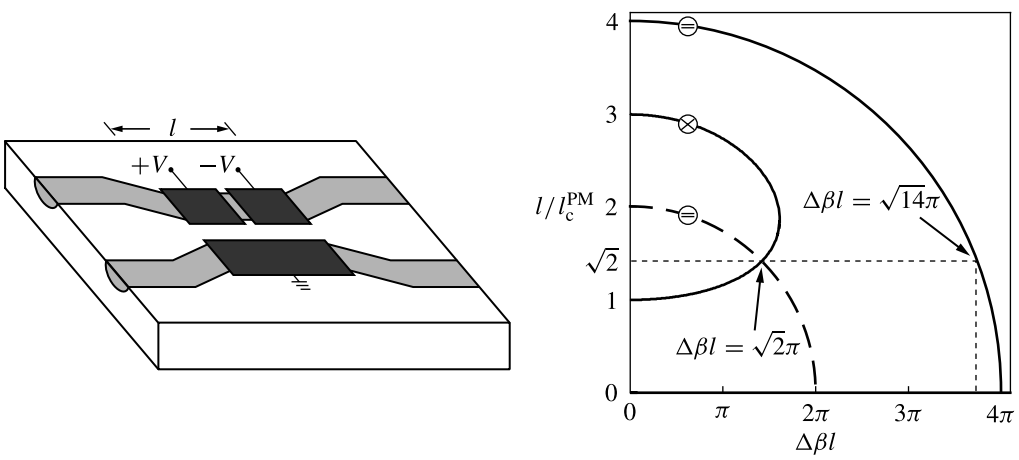


Figure 6.11 (*a*) Schematic structure and (*b*) switching diagram of a reversed-$\Delta\beta$ directional coupler switch. The solid curves are the solutions for the reversed-$\Delta\beta$ directional coupler switch, while the dashed curve is that of the uniform-$\Delta\beta$ directional coupler switch shown in Fig. 6.10(*b*). Note that a trivial solution of $l = 0$ exists for the parallel state.

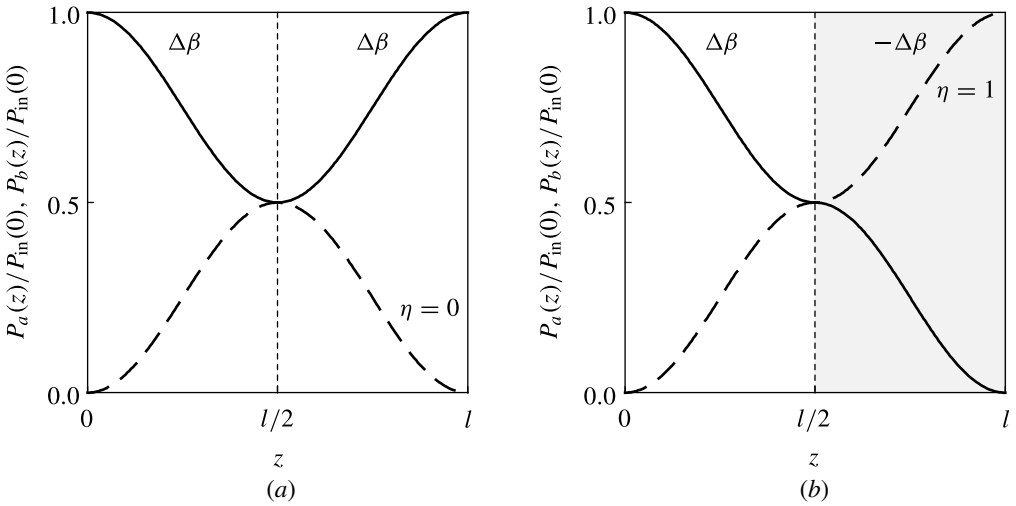


Figure 6.12 Evolution of power flow in (*a*) a two-section directional coupler with uniform $\Delta\beta$ and (*b*) a reversed-$\Delta\beta$ directional coupler for $\Delta\beta = 2\kappa$. The solid curves represent $P_a(z)/P_{in}(0)$, and the dashed curves represent $P_b(z)/P_{in}(0)$.

section, also of a length of $l/2$, is $-\Delta\beta = -2\delta$. By considering the device as two couplers in tandem and by solving the coupled-mode equations, it can be shown that the total coupling efficiency of this device is (see Problem 6.4.8)

$$\eta = \frac{4\kappa^2}{\beta_c^2} \sin^2 \frac{\beta_c l}{2} \left(1 - \frac{\kappa^2}{\beta_c^2} \sin^2 \frac{\beta_c l}{2}\right), \tag{6.85}$$

where $\beta_c = (\kappa^2 + \delta^2)^{1/2}$, as defined in (4.61). By expressing (6.85) in terms of l/l_c^{PM} and $\Delta\beta l/\pi$, the conditions for the parallel and the cross state, which have $\eta = 0$ and $\eta = 1$, respectively, can be plotted in the switching diagram shown in Fig. 6.11(*b*).

It is now possible to reach both the parallel and the cross state by controlling the applied voltage if l/l_c^{PM} is chosen to be within a proper range such as $1 \le l/l_c^{PM} \le 3$. As an example, consider $l = \sqrt{2} l_c^{PM}$. In this case, the cross state with $\eta = 1$ is reached when $\Delta\beta = 2\kappa$, thus $\Delta\beta l = \sqrt{2}\pi$, and the parallel state with $\eta = 0$ is reached when $\Delta\beta = 2\sqrt{7}\kappa$, thus $\Delta\beta l = \sqrt{14}\pi$. This example is also illustrated in Fig. 6.11(*b*). The possibility of reaching the cross state in this condition can be understood intuitively with the illustration shown in Fig. 6.12 for the conditions of $l = \sqrt{2} l_c^{PM}$ and $\Delta\beta = 2\kappa$. For a coupler with a uniform $\Delta\beta$ across the two sections, we find from (6.82) that $\eta = 0$, and the device is in the parallel state rather than in the cross state. We also find by substituting l in (6.82) with $l/2$ that at the end of the first section, the input power from one channel is equally divided between the two channels. Therefore, each section alone functions as a 3-dB coupler. Under the conditions considered here, the second section of the uniform-$\Delta\beta$ coupler couples all of the power back to the original channel, resulting in $\eta = 0$ and the parallel state, as shown in Fig. 6.12(*a*). However, if the sign of phase mismatch is reversed in the second section, as is the case in the reversed-$\Delta\beta$

coupler, the process in the second section is also reversed, resulting in $\eta = 1$ and the cross state. This process is shown in Fig. 6.12(b).

Waveguide polarization modulators

The function of a waveguide polarization modulator depends on the coupling between modes of different polarizations. As is true for any coupled-mode device, the key parameters to be considered are also the phase mismatch and the coupling coefficient between the two polarization modes. Unlike the situation in the directional couplers discussed above, however, the two polarization modes are generally phase mismatched and are originally uncoupled. Therefore, the task of the modulation electric field is to create a sufficiently strong field-dependent coupling coefficient while simultaneously reducing or eliminating the phase mismatch if necessary. For two orthogonally polarized modes to couple to one another, it is necessary to induce the corresponding off-diagonal elements in the dielectric permittivity tensor. For instance, if a $LiNbO_3$ waveguide is fabricated with its structural axes lined up with the principal axes of the crystal, a modulation electric field E_{0y} will couple y- and z-polarized mode fields, as can be seen from (6.27), whereas an E_{0x} creates coupling between x- and y-polarized modes and that between x- and z-polarized modes, as can be seen from (6.32). In contrast, an E_{0z} cannot create such coupling between any two orthogonal polarizations, as can be seen from (6.24).

As an example, we consider the coupling between fundamental TE-like and TM-like modes in an x-cut, y-propagating $LiNbO_3$ waveguide shown in Fig. 6.13. Because the birefringence of $LiNbO_3$ is relatively large, phase mismatch between the TE-like and the TM-like modes is contributed primarily by the fact that they have quite different indices, approximately n_e and n_o, respectively:

$$\Delta\beta = \beta_{TM} - \beta_{TE} \approx \frac{2\pi}{\lambda}(n_o - n_e). \tag{6.86}$$

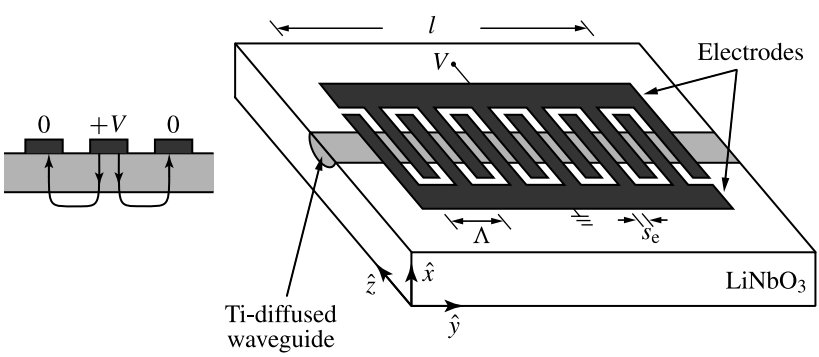


Figure 6.13 Waveguide polarization modulator fabricated on an x-cut, y-propagating $LiNbO_3$ substrate using a periodic electrode for phase matching between TE-like and TM-like modes. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the principal axes of the crystal.

This large phase mismatch has to be compensated before any significant coupling between these two modes can take place. The required phase matching can be accomplished by using a grating of a proper period Λ, as discussed in Section 5.1 and indicated by (5.7). For perfect phase matching with $\delta = 0$, we need $qK = -\Delta\beta$. Therefore, the grating period has to be one of those given by

$$\Lambda = -q\frac{2\pi}{\Delta\beta} \approx -q\frac{\lambda}{n_\mathrm{o} - n_\mathrm{e}}, \qquad q = -1, -2, \ldots, \tag{6.87}$$

where q takes on negative integral values because $n_\mathrm{o} > n_\mathrm{e}$ for $LiNbO_3$ and the value of Λ has to be positive. For a first-order grating ($q = -1$), the phase-matching condition requires that $\Lambda \approx 7$, 12.5, and 18 μm for $\lambda = 0.6$, 1, and 1.3 μm, respectively. Such a grating can be generated by using the periodic electrode shown in Fig. 6.13 to create a periodic electro-optically modulated $\Delta\epsilon$ along the propagation direction of the waveguide. The coupling coefficient is then given by

$$\kappa_\mathrm{EM}(q) = \kappa^*_\mathrm{ME}(q) = \frac{\omega}{\Lambda}\int_0^\Lambda \mathrm{d}y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z \hat{\boldsymbol{\mathcal{E}}}^*_\mathrm{TE}(x,z)\cdot\Delta\boldsymbol{\epsilon}(x,y,z)\cdot\hat{\boldsymbol{\mathcal{E}}}_\mathrm{TM}(x,z)\mathrm{e}^{-\mathrm{i}qKy}, \tag{6.88}$$

where $\Delta\epsilon$ is periodic in y for the device configuration shown in Fig. 6.13.

Since the TE-like and TM-like modes are mainly polarized in the z and x directions, respectively, they are coupled primarily through the off-diagonal zx and xz terms of $\Delta\epsilon$ in (6.32) that are contributed by the effect of the E_{0x} component of the periodic modulation electric field. Although a periodic E_{0y} component also exists, its contribution is not significant. For a first-order grating, this leads to the following coupling coefficient:

$$\begin{aligned}
\kappa &= \kappa_\mathrm{EM}(q=-1) = \kappa^*_\mathrm{ME}(q=-1)\\
&\approx \frac{\omega}{\Lambda}\int_0^\Lambda \mathrm{d}y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z \hat{\mathcal{E}}^*_{\mathrm{TE},z}(x,z)(-\epsilon_0 n_\mathrm{o}^2 n_\mathrm{e}^2 r_{42})E_{0x}(x,y,z)\hat{\mathcal{E}}_{\mathrm{TM},x}(x,z)\mathrm{e}^{\mathrm{i}Ky}\\
&\approx -\frac{\pi}{\lambda} n_\mathrm{o}^{3/2} n_\mathrm{e}^{3/2} r_{42}\Gamma_\mathrm{EM}\frac{V}{s_\mathrm{e}},
\end{aligned} \tag{6.89}$$

where

$$\begin{aligned}
\Gamma_\mathrm{EM} &= \frac{2\beta_\mathrm{TE}^{1/2}\beta_\mathrm{TM}^{1/2}}{\omega\mu_0}\frac{s_\mathrm{e}}{V}\frac{1}{\Lambda}\int_0^\Lambda \mathrm{d}y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z E_{0x}\hat{\mathcal{E}}^*_{\mathrm{TE},z}\hat{\mathcal{E}}_{\mathrm{TM},x}\mathrm{e}^{\mathrm{i}Ky}\\
&\approx \frac{s_\mathrm{e}}{V}\frac{\dfrac{1}{\Lambda}\displaystyle\int_0^\Lambda \mathrm{d}y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z E_{0x}\hat{\mathcal{E}}^*_{\mathrm{TE},z}\hat{\mathcal{E}}_{\mathrm{TM},x}\mathrm{e}^{\mathrm{i}Ky}}{\left(\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z|\hat{\mathcal{E}}_{\mathrm{TE},z}|^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}z|\hat{\mathcal{E}}_{\mathrm{TM},x}|^2\right)^{1/2}}
\end{aligned} \tag{6.90}$$

is the overlap factor for electro-optic coupling of TE-like and TM-like modes. If perfect phase matching is accomplished by accurate selection of the grating period, the coupling efficiency is simply that given by (4.85): $\eta_{\mathrm{PM}} = \sin^2 |\kappa| l$. Because $|\kappa|$ can now be controlled by the modulation voltage, the coupling length $l_{\mathrm{c}}^{\mathrm{PM}} = \pi/2|\kappa|$ can be varied by varying the voltage. For a given device of length l, the coupling efficiency can then be varied between 0 and 1 through variation of the modulation voltage. In general, the device can be used to modulate or control the polarization of the guided optical wave, for example, to convert an elliptically polarized input beam into a linearly polarized output beam, or vice versa. When the voltage is set at a value for $\eta_{\mathrm{PM}} = 1$, it is possible to convert a TE-like mode to a TM-like mode completely, or vice versa. In this instance, the device functions as a *TE–TM mode converter*.

EXAMPLE 6.8 An x-cut, y-propagating $LiNbO_3$ waveguide TE–TM mode converter as shown in Fig. 6.13 for $\lambda = 1.3$ μm consists of a periodic interdigital electrode that has a period of $\Lambda = 18$ μm for phase matching. The interelectrode gap is $s_{\mathrm{e}} = 4$ μm, and the overlap factor for TE–TM electro-optic coupling is $\Gamma_{\mathrm{EM}} = 0.1$. The total length of the electrode section is $l = 10$ mm. Find the mode-conversion voltage that is required for complete conversion between TE and TM modes.

Solution The required electro-optically controlled coupling coefficient for complete TE–TM mode conversion is $|\kappa| = \pi/2l$ for the device of length l. By using (6.89), the mode-conversion voltage can be expressed as

$$V_{\mathrm{EM}} = \frac{\lambda |\kappa| s_{\mathrm{e}}}{\pi n_{\mathrm{o}}^{3/2} n_{\mathrm{e}}^{3/2} r_{42} \Gamma_{\mathrm{EM}}} = \frac{\lambda}{2 n_{\mathrm{o}}^{3/2} n_{\mathrm{e}}^{3/2} r_{42} \Gamma_{\mathrm{EM}}} \frac{s_{\mathrm{e}}}{l}. \tag{6.91}$$

With the given parameters of the device, we find that

$$V_{\mathrm{EM}} = \frac{1.3 \times 10^{-6}}{2 \times 2.222^{3/2} \times 2.145^{3/2} \times 28 \times 10^{-12} \times 0.1} \times \frac{4 \times 10^{-6}}{10 \times 10^{-3}} \mathrm{V} = 8.92 \mathrm{V}.$$

Phase matching for the polarization modulator shown in Fig. 6.13 cannot be adjusted electrically but has to be accomplished by accurate selection and fabrication of the electrode period. Figure 6.14 shows a few other configurations of waveguide polarization modulators, which are fabricated on z-propagating $LiNbO_3$ substrates. In the configuration shown in Fig. 6.14(a), a y-cut substrate is used. A two-electrode structure provides the horizontal modulation field component E_{0x} for coupling between TE-like and TM-like modes, which are polarized mainly in the x and y directions, respectively. The two polarizations have the same ordinary index because the propagation direction is along the optical axis of the crystal. The only phase mismatch between the TE-like and the TM-like modes is caused by modal dispersion. Because modal dispersion is small in a weakly guiding waveguide, it need not be intentionally compensated if the coupling coefficient is made sufficiently large so that $|\kappa| \gg \delta$. Therefore, a fixed bias

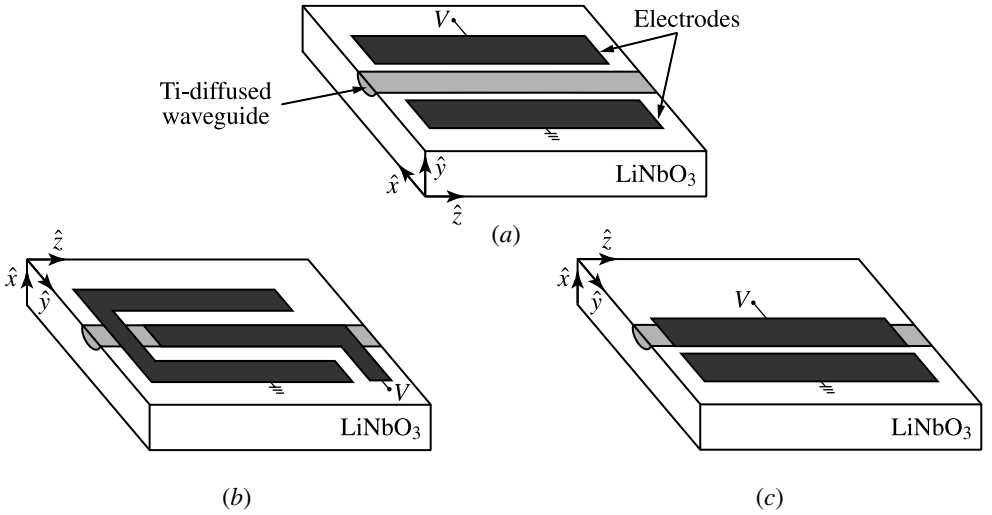


Figure 6.14 Three configurations for z-propagating waveguide polarization modulators (a) on y-cut $LiNbO_3$, (b) and (c) on x-cut $LiNbO_3$. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the principal axes of the crystal.

voltage can be used to provide a large bias $|\kappa|$ to reduce the effect of the phase mismatch effectively. A similar situation applies to polarization modulators fabricated in III–V semiconductors, as can be seen from (6.34). In fact, because a III–V semiconductor is not birefringent, the waveguide can be fabricated along any direction, not necessarily parallel to a crystal axis, for modes with different polarizations to suffer only modal dispersion.

In z-propagating $LiNbO_3$ waveguides, the slight modal dispersion can be further compensated, if desired, by applying an E_{0y} component in addition to E_{0x}. This can be done using a three-electrode configuration with asymmetrically applied voltages, as shown in Fig. 6.14(b), or using an asymmetrically placed two-electrode configuration with one electrode directly on top of the waveguide, as shown in Fig. 6.14(c). Compensation of the phase mismatch can be accomplished by adjusting E_{0y} through the applied voltage because E_{0y} induces equal but opposite changes in the refractive indices along x and y directions, as can be seen from (6.27).

6.5 Traveling-wave modulators

At a low modulation frequency, the time it takes for the optical wave to travel through an electro-optic modulator is short compared to the modulation period. This situation is characterized by the condition that $f\tau_{tr} < 1$, where f is the modulation frequency and τ_{tr} is the transit time for the optical wave to propagate through the modulator.

In this case, the modulator can be considered as a lumped device because its length is small compared to the wavelength of the modulation field. The 3-dB *modulation bandwidth*, $f_{3\text{dB}}$, of a lumped electro-optic modulator is determined by both the transit time τ_{tr} of the optical wave and the RC time constant τ_{RC} of the lumped driving circuit including the loading effects of the modulator. Because $\tau_{\text{RC}} > \tau_{\text{tr}}$ for most lumped modulators, the modulation bandwidth of a lumped modulator is usually determined by its RC time constant so that

$$f_{3\text{dB}} = \frac{1}{2\pi\tau_{\text{RC}}}. \tag{6.92}$$

The value of τ_{RC} for a modulator of a given resistance, such as 50 Ω, increases with the length l of its electrode because the capacitance increases with this length. For a lumped $LiNbO_3$ waveguide modulator, the product $f_{3\text{dB}}l$ typically falls in the range of 1–3 GHz cm. Therefore, the modulation bandwidth of a lumped $LiNbO_3$ is limited to a few gigahertz at best.

In many applications, however, the modulation frequency is in the microwave or millimeter-wave range to take advantage of the high-bandwidth capacity of the optical carrier wave. The modulation efficiency of a lumped modulator drops drastically at high modulation frequencies because of its RC-limited frequency response. This problem can be overcome by using a traveling-wave configuration for the modulator and by matching the phase velocity of the microwave modulation field to that of the optical wave in the *traveling-wave modulator*.

The electrodes of a traveling-wave modulator are made of strip transmission lines, as shown in Fig. 6.15. The electrodes are specifically designed for traveling-wave interactions. The high-frequency modulation signal is injected at one end, propagates along the same direction as the optical wave, and terminates at the end of the electrode transmission line. The traveling-wave configuration inherently requires the use of the transverse modulation scheme. This, however, is consistent with the configurations of most guided-wave devices. Therefore, traveling-wave modulation can be applied to a large variety of guided-wave devices, including single-waveguide phase modulators, Mach–Zehnder interferometers, and directional coupler switches, to meet the demand

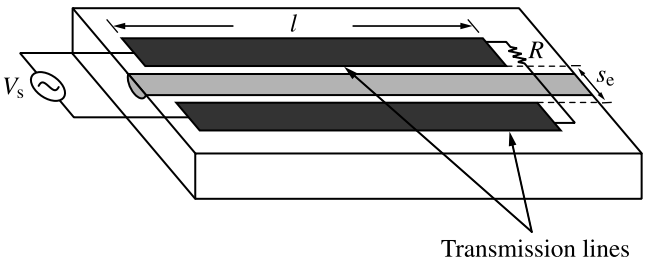


Figure 6.15 Traveling-wave phase modulator.

for high-frequency modulation and switching very often found in the applications of guided-wave devices.

Two key factors determine the modulation bandwidth of a traveling-wave modulator: (1) phase-velocity mismatch between the optical wave in the waveguide and the microwave in the transmission line and (2) frequency-dependent attenuation of the microwave modulation signal as it propagates along the transmission line. These parameters are determined by the waveguide material, the details of the waveguide structure, and the design of the transmission line. A transmission line has a characteristic impedance, Z, given by

$$Z = \sqrt{\frac{L}{C}}, \tag{6.93}$$

where L and C are, respectively, the inductance and capacitance per unit length of the transmission line. The phase velocity, $v_{\mathrm{p}}^{\mathrm{m}}$, of a microwave electrical signal propagating in a transmission line is

$$v_{\mathrm{p}}^{\mathrm{m}} = \frac{1}{\sqrt{LC}} = \frac{c}{n_{\mathrm{m}}}, \tag{6.94}$$

where n_{m} is the refractive index of the microwave in the transmission line. For a microwave transmission line on a $LiNbO_3$ substrate, $n_{\mathrm{m}} \approx 4.225$ with variations around this value caused by variations in the structure and parameters of the transmission line. The microwave traveling in the transmission line suffers a loss characterized by a frequency-dependent *power attenuation* coefficient of α_{m}, thus a *voltage attenuation* coefficient of $\alpha_{\mathrm{m}}/2$. The modulation signal sent from the input end of the modulator is then characterized by the following space- and time-dependent traveling microwave voltage throughout the electrode:

$$V(z,t) = V_{\mathrm{pk}} \mathrm{e}^{-\alpha_{\mathrm{m}} z/2} \cos\left[2\pi f\left(\frac{z}{v_{\mathrm{p}}^{\mathrm{m}}} - t\right)\right], \tag{6.95}$$

where V_{pk} is the peak modulation voltage and f is the modulation frequency of the microwave signal. The phase velocity, $v_{\mathrm{p}}^{\mathrm{o}}$, of a guided optical wave is determined by its frequency ω and propagation constant β as

$$v_{\mathrm{p}}^{\mathrm{o}} = \frac{\omega}{\beta} = \frac{c}{n_\beta}, \tag{6.96}$$

where n_β is the effective refractive index of the guided optical wave.

Because an optical wavefront entering the input end of the electrode at time t arrives at location z at a later time of $t + z/v_{\mathrm{p}}^{\mathrm{o}}$, it sees a space- and time-varying voltage of $V(z, t + z/v_{\mathrm{p}}^{\mathrm{o}})$, rather than $V(z, t)$, as it travels through the waveguide in the modulator. Therefore, the electro-optically induced change in the propagation constant of the guided optical wave as a function of modulation frequency varies with space and time

as follows:

$$\Delta\beta(f,z,t) = \Delta\beta_{\rm pk}{\rm e}^{-\alpha_{\rm m}z/2}\cos\left[2\pi f\left(\frac{\tau_{\rm VM}}{l}z - t\right)\right], \tag{6.97}$$

where $\Delta\beta_{\rm pk}$ is the change corresponding to a constant peak voltage $V_{\rm pk}$ and

$$\tau_{\rm VM} = \left|\frac{l}{\upsilon_{\rm p}^{\rm m}} - \frac{l}{\upsilon_{\rm p}^{\rm o}}\right| = \frac{l}{c}|n_{\rm m} - n_\beta| \tag{6.98}$$

is an effective temporal walk-off between the wavefronts of the optical wave and the microwave due to the *velocity mismatch* between them. The electro-optically induced phase shift for the optical wave at the output end of the modulator as a function of the modulation frequency and time can be found as

$$\Delta\varphi(f,t) = \int_0^l \Delta\beta(f,z,t){\rm d}z = \Delta\varphi_{\rm pk}(0)\frac{1}{l}\int_0^l {\rm e}^{-\alpha_{\rm m}z/2}\cos\left[2\pi f\left(\frac{\tau_{\rm VM}}{l}z - t\right)\right]{\rm d}z, \tag{6.99}$$

where $\Delta\varphi_{\rm pk}(0) = \Delta\beta_{\rm pk}l$ is the phase shift induced by a constant voltage $V_{\rm pk}$ at $f = 0$. Instead of carrying out the integration in (6.99), we examine two important limiting cases in the following.

1. **Bandwidth limited by velocity mismatch.** If $\tau_{\rm VM}$ is significant but $\alpha_{\rm m} \approx 0$, the integral in (6.99) yields

   $$\Delta\varphi(f,t) = \Delta\varphi_{\rm pk}(f)\cos\left[2\pi f\left(t - \frac{\tau_{\rm VM}}{2}\right)\right], \tag{6.100}$$

   where

   $$\Delta\varphi_{\rm pk}(f) = \Delta\varphi_{\rm pk}(0)\frac{\sin\pi f\tau_{\rm VM}}{\pi f\tau_{\rm VM}} \tag{6.101}$$

   is the modulation-frequency-dependent peak phase shift. By setting $\Delta\varphi_{\rm pk}(f_{\rm 3dB}) = \Delta\varphi_{\rm pk}(0)/2$, the 3-dB modulation bandwidth limited by velocity mismatch is found to be

   $$f_{\rm 3dB}^{\rm VM} \approx \frac{2}{\pi\tau_{\rm VM}} = \frac{2c}{\pi l|n_{\rm m} - n_\beta|}. \tag{6.102}$$

2. **Bandwidth limited by microwave attenuation.** If the phase velocities of the optical wave and the microwave signal are perfectly matched, we have $\tau_{\rm VM} = 0$. The modulation bandwidth is then only limited by the increasing loss of the strip line at high frequencies characterized by $\alpha_{\rm m}(f)$ as a function of the modulation frequency. In this limit, the integral in (6.99) yields

   $$\Delta\varphi(f,t) = \Delta\varphi_{\rm pk}(f)\cos 2\pi f t, \tag{6.103}$$

where

$$\Delta\varphi_{\rm pk}(f) = \Delta\varphi_{\rm pk}(0)\frac{1 - {\rm e}^{-\alpha_{\rm m}(f)l/2}}{\alpha_{\rm m}(f)l/2} \tag{6.104}$$

is a function of the modulation frequency through the frequency dependence of $\alpha_{\rm m}$. By setting $\Delta\varphi_{\rm pk}(f_{\rm 3dB}) = \Delta\varphi_{\rm pk}(0)/2$, the 3-dB modulation bandwidth limited by attenuation in the transmission line is found to be determined by

$$\alpha_{\rm m}(f_{\rm 3dB}) \approx \frac{3.2}{l}. \tag{6.105}$$

The frequency dependence of the attenuation coefficient of a transmission line is usually characterized by $\alpha_{\rm m} = af^{1/2}$, where a is a constant. Then the 3-dB modulation bandwidth limited by attenuation can be expressed as

$$f_{\rm 3dB}^{\rm att} \approx \left(\frac{3.2}{al}\right)^2. \tag{6.106}$$

A figure of merit for a modulator is its power–bandwidth ratio, $P/f_{\rm 3dB}$, which measures the power cost of imposing a unit bandwidth of information on the optical carrier. The required microwave power to drive the modulator depends on the impedance $Z_{\rm s}$ of the microwave source and the impedance Z of the modulator. The standard impedance for the microwave source is $Z_{\rm s} = 50\ \Omega$. It is desired that the impedance Z of the modulator matches $Z_{\rm s}$ as closely as possible for the most efficient delivery of the microwave power to the modulator, but perfect match is often not possible because of design constraints. In the general situation when $Z \neq Z_{\rm s}$, the power required from the microwave source for a peak modulation voltage $V_{\rm pk}$ on the modulator is

$$P = \frac{V_{\rm pk}^2}{2Z_{\rm s}}\frac{1}{1 - [(Z_{\rm s} - Z)/(Z_{\rm s} + Z)]^2}. \tag{6.107}$$

For a modulator such as a Mach–Zehnder interferometer that requires a total phase variation from 0 to π for the full range of its operation, only a peak voltage of $V_{\rm pk} = V_\pi/2$ is required by biasing the device at $V_{\rm b} = V_\pi/2$ so that the full swing of the microwave voltage from $-V_{\rm pk}$ to $V_{\rm pk}$ provides the voltage variations from 0 to V_π. The $P/f_{\rm 3dB}$ ratio for a properly designed traveling-wave modulator is generally much smaller than that for a comparable lumped modulator, reflecting a much improved performance.

EXAMPLE 6.9 Properly designed transmission lines are used for the electrodes of the x-cut, y-propagating $LiNbO_3$ Mach–Zehnder waveguide interferometric modulator shown in Fig. 6.8(a) and discussed in Example 6.6. The transmission line electrodes have an impedance of $Z = 30\ \Omega$; a microwave index of $n_{\rm m} = 4.225$; and a frequency-dependent

microwave power attenuation coefficient of $\alpha_m = af^{1/2}$, with $a = 2\,\text{dB cm}^{-1}\,\text{GHz}^{-1/2}$. It is driven by a microwave source of impedance $Z_s = 50\ \Omega$. Find the 3-dB modulation bandwidth and the P/f_{3dB} ratio of this traveling-wave Mach–Zehnder waveguide interferometric modulator for the TE-like mode at $\lambda = 1.3\ \mu\text{m}$.

Solution From Example 6.6, we find that the length of the electrodes is $l = 12.5$ mm, and the half-wave voltage for the TE-like mode is $V_\pi = 6$ V. We also find that $n_\beta \approx n_e = 2.145$ for the TE-like mode of this device. We then find that $\tau_{VM} = 86.7$ ps, and that the bandwidth limited by velocity mismatch is

$$f_{3dB}^{VM} = \frac{2 \times 3 \times 10^8}{\pi \times 12.5 \times 10^{-3} \times |4.225 - 2.145|}\ \text{Hz} = 7.3\ \text{GHz}$$

using (6.102). To find the bandwidth limited by attenuation, we first convert the attenuation coefficient measured in decibels per centimeter into that measured per centimeter using the relation in (3.91). Therefore, $a = 2\,\text{dB cm}^{-1}\,\text{GHz}^{-1/2} = 0.46\,\text{cm}^{-1}\,\text{GHz}^{-1/2}$. We then find that the bandwidth limited by attenuation is

$$f_{3dB}^{att} = \left(\frac{3.2}{0.46 \times 1.25}\right)^2\ \text{GHz} = 31\ \text{GHz}$$

using (6.106). Because $f_{3dB}^{att} > f_{3dB}^{VM}$, the bandwidth of this modulator is limited by velocity mismatch to be $f_{3dB} = 7.3$ GHz.

By biasing the modulator at $V_b = V_\pi/2 = 3$ V, the device can be modulated with a peak voltage of $V_{pk} = V_\pi/2 = 3$ V. With $Z = 30\ \Omega$ and $Z_s = 50\ \Omega$, it is then found that $P = 96$ mW using (6.107). Thus, the modulator has a power–bandwidth ratio of $P/f_{3dB} = 13.15\,\text{mW GHz}^{-1}$.

PROBLEMS

6.2.1 Is it possible to make a Pockels cell for electro-optic modulation using silicon or ordinary glass? Explain.

6.2.2 Verify the angle of rotation of principal axes and the principal indices of refraction given in (6.29) and (6.30), respectively, which are induced by an electric field E_{0y} through the Pockels effect in $LiNbO_3$. Use the parameters listed in Table 6.2 to evaluate the rotation angle θ of the y and z axes and the changes in the principal indices at $\lambda = 1\ \mu\text{m}$ induced by an applied field of 1 MV m^{-1}.

6.2.3 Find the principal axes and their corresponding principal indices of refraction as a result of the Pockels effect in $LiNbO_3$ when the electric field is applied along the x axis. Use the parameters listed in Table 6.2 to evaluate the changes at $\lambda = 1\ \mu\text{m}$ in the principal indices and the crystal birefringence caused by an applied field of 1 MV m^{-1}.

6.2.4 Consider a uniaxial crystal of 622 symmetry such as CdS, in which the only nonvanishing Pockels coefficients are $r_{52} = -r_{41}$. The uniaxial optical axis is taken to be the z axis. The ordinary index of refraction is n_o, and the extraordinary index is n_e.

a. With an applied DC voltage polarized in an arbitrary direction, write down the nonvanishing elements of the $\Delta\chi$ tensor in terms of the nonvanishing electro-optic coefficients.

b. Show that by applying a DC voltage along the x axis, the crystal becomes biaxial. Find the new refractive indices as a function of the applied DC electric field.

6.2.5 KTP is a biaxial crystal of $mm2$ symmetry that has five nonvanishing Pockels coefficients, r_{13}, r_{23}, r_{33}, r_{42}, and r_{51}. An electric field E_0 is applied along its z principal axis.

a. Write down the index ellipsoid equation and the dielectric permittivity tensor as a function of the applied electric field.

b. Find the principal axes and their corresponding principal indices as a function of the applied electric field as a result of the Pockels effect.

c. Use the parameters of KTP listed in Table 6.2 to find the changes at $\lambda = 1\ \mu\text{m}$ in the principal indices and in the birefringence of the crystal caused by an applied field of $E_0 = 1\ \text{MV m}^{-1}$.

6.2.6 Answer the questions in Problem 6.2.5 for the KTP crystal in the case when E_0 is applied along the y principal axis.

6.2.7 Answer the questions in Problem 6.2.5 for the KTP crystal in the case when E_0 is applied along the x principal axis.

6.2.8 A $\bar{4}3m$ crystal, such as GaAs, has cubic crystal symmetry and isotropic linear optical properties. It can become birefringent due to the Pockels effect when a DC electric field is applied.

a. A DC electric field of $\boldsymbol{E}_0 = E_0(\hat{y}\cos\theta + \hat{z}\sin\theta)$ is applied to the crystal in the yz plane, where θ is the angle between the DC field polarization and the y principal axis. Find the new principal axes and their corresponding new principal refractive indices.

b. With this DC field applied to the crystal at a given angle θ, how do you arrange the polarization of a linearly polarized optical beam propagating along the x principal axis so that it remains linearly polarized along its entire path through the crystal?

c. If the linearly polarized optical beam propagates along the z principal axis, how would you arrange the directions of polarization for the DC and optical fields so that the beam remains linearly polarized along its entire path through the crystal?

6.2.9 $BaTiO_3$ is a uniaxial crystal of $4mm$ symmetry with $n_o = 2.44$ and $n_e = 2.37$ at the optical wavelength of 546 nm. Take $\hat{z}$ to be the optical axis. Its only

nonvanishing electro-optic coefficients are $r_{33} = 23$ pm V^{-1}, $r_{13} = r_{23} = 8$ pm V^{-1}, and $r_{42} = r_{51} = 820$ pm V^{-1}.

a. Consider a beam at 546 nm wavelength traveling through a $BaTiO_3$ crystal of 1 mm thickness along the y axis of the crystal. If the wave is linearly polarized in the xz plane but not along the x or z axis, what is its state of polarization at its exit from the crystal?

b. In principle we can apply a DC electric field on the crystal to keep the beam linearly polarized along its entire path of propagation through the crystal. Along which direction should this voltage be applied for this purpose? What is this voltage? Do you see any practical problem with this proposal?

c. What happens to the optical axis of the crystal if we apply a DC electric field along the y direction? Explain.

6.3.1 What is the advantage of transverse modulation over longitudinal modulation for an electro-optic modulator?

6.3.2 A KTP transverse phase modulator for $\lambda = 1$ μm has the configuration shown in Fig. 6.2(a) with the modulation voltage applied along its z axis. Answer each of the following questions for optical waves polarized in x, y, and z directions, respectively, under different possible arrangements.

a. Find the phase modulation depth φ_{pk} as a function of the parameters of KTP, the dimensions of the modulator, and the peak modulation voltage V_{pk}.

b. Use the parameters given in Table 6.2 to find the peak modulation voltage required for a phase modulation depth of $\varphi_{pk} = \pi$ for a bulk modulator of dimensions of $d = 3$ mm and $l = 6$ mm.

c. Find the peak modulation voltage required for a phase modulation depth $\varphi_{pk} = \pi$ for a waveguide modulator that has dimensions of $d = 3$ μm and $l = 6$ mm.

6.3.3 A KTP longitudinal phase modulator for $\lambda = 1$ μm has the configuration shown in Fig. 6.2(b) with the modulation voltage applied along its z axis. Answer each of the following questions for optical waves polarized in x, y, and z directions, respectively, if possible, under different possible arrangements.

a. Find the phase modulation depth φ_{pk} as a function of the parameters of KTP, the dimensions of the modulator, and the peak modulation voltage V_{pk}.

b. Use the parameters given in Table 6.2 to find the peak modulation voltage required for a phase modulation depth of $\varphi_{pk} = \pi$ for a bulk modulator that has dimensions of $d = 3$ mm and $l = 6$ mm.

c. Compare the result obtained in (b) with that in Problem 6.3.2(b) of the transverse modulator. Explain the difference.

6.3.4 A KTP transverse polarization modulator for $\lambda = 1$ μm has the configuration shown in Fig. 6.3(a) with the modulation voltage applied along its z axis and the optical wave propagating in the x direction.

a. Express the phase retardation $\Delta\varphi$ between the y- and z-polarized components of the optical wave as a function of the parameters of KTP, the dimensions of the modulator, and the modulation voltage V.
b. Use the parameters given in Table 6.2 to find the bias voltage V_b required for compensation of the intrinsic birefringence of KTP and the voltage differential, $V_\pi - V_b$, between the half-wave voltage and the bias voltage for a waveguide modulator that has dimensions of $d = 5$ μm and $l = 5$ mm.

6.3.5 A III–V semiconductor is an electro-optic crystal that has $\bar{4}3m$ symmetry and an index of refraction n_o. A DC electric field E_0 is applied to the crystal in the [001] crystallographic direction. A light beam at a wavelength λ travels in the [110] direction. Its polarization makes an angle of 45° with respect to the [001] axis.
a. Show that the plane of polarization is rotated by 90° after the light has traveled a distance of

$$l = \frac{\lambda}{n_o^3 r_{41} E_0}.$$

b. What is changed if the DC electric field is turned by 90° but is still transverse to the direction of wave propagation, i.e., if the DC field is now parallel to the $[1\bar{1}0]$ direction?

6.3.6 KDP is a uniaxial crystal of $\bar{4}2m$ symmetry. Its optical axis is the z axis.
a. In an application of KDP for electro-optic phase modulation, describe how you would arrange the principal axes of the crystal with respect to the DC and optical field polarization directions in the longitudinal and transverse modulation schemes, respectively. Show the arrangements with sketches.
b. Find the half-wave voltage for an optical beam at $\lambda = 1$ μm wavelength in the longitudinal modulation scheme.
c. What arrangement is necessary so that the half-wave voltage in the transverse modulation scheme is half that of the longitudinal modulation scheme found in (b)?

6.3.7 The electro-optic crystal ADP is uniaxial, with the z axis being the optical axis. A DC voltage is applied in the z direction.
a. It is used as a polarization modulator in a longitudinal modulation scheme with a linearly polarized input beam at 1 μm wavelength propagating in the z direction. How do you arrange the polarization of the input beam with respect to the principal axes of the crystal so that it will be rotated by 90° after it passes through the crystal if the magnitude of the DC voltage is properly adjusted? What is this voltage? If the voltage is then reduced by one half, what happens to the polarization of the output light? If rotation of

the linear polarization by 60° rather than 90° is desired, what should you do?

b. Answer the same questions in (a) if the optical wave propagates in the xy plane while any possible walk-off is avoided.

6.3.8 Design GaAs transverse electro-optic polarization and intensity modulators that require no bias voltage. Compare them to $LiNbO_3$ transverse modulators and GaAs longitudinal modulators for the same functions.

6.4.1 Give the names of three important types of interferometers and sketch their basic structures.

6.4.2 In this problem, we consider the effect of imbalance between the two arms of a waveguide interferometer. The waveguides in the two arms support only the dominant TE mode. However, due to fabrication errors, the input and output Y-junction couplers deviate slightly from ideal 3-dB couplers. As a result, the input coupler splits the input power with a ratio of $(1 + \delta_1)/(1 - \delta_1)$ between the two arms, while the output coupler has an imbalance of $(1 + \delta_2)/(1 - \delta_2)$ between the two arms.

a. Show that the extinction ratio for this interferometer is given by

$$\mathrm{ER} = 20 \log \frac{1 + [(1 - \delta_1^2)(1 - \delta_2^2)]^{1/2} + \delta_1 \delta_2}{\delta_1 + \delta_2}. \tag{6.108}$$

b. Assume that $\delta_1 = \delta_2 = \delta$. For an extinction ratio of 30 dB, how much imperfection can the device tolerate?

6.4.3 A z-cut, x-propagating $LiNbO_3$ Mach–Zehnder waveguide interferometer as shown in Fig. 6.8(b) is used as an electro-optic amplitude modulator for an optical wave at 1.55 μm wavelength. The two arms of the interferometer are single-mode waveguides at this wavelength. The waveguides are weakly guiding such that the propagation constants of the guided modes are very close to those determined by the material properties alone. The electrodes are separated by a gap of $s_e = 18$ μm and are designed to cover most of the waveguide regions such that the overlap factors for the TE-like and TM-like modes are $\Gamma_{TE} = \Gamma_{TM} = 0.45$. At 1.55 μm, we have $n_o = 2.213$, $n_e = 2.137$, $r_{33} = 30.8$ pm V^{-1}, and $r_{13} = 8.6$ pm V^{-1} for $LiNbO_3$.

a. If the electrodes have an equal length of $l = 1$ cm, what are the half-wave voltages for the TE-like and TM-like modes, respectively?

b. If the electrodes are designed such that the interferometer functions as a lumped modulator, what are the highest modulation frequencies for the TE-like and TM-like modes, respectively?

c. If the electrodes are designed as transmission lines with an impedance $Z = 50$ Ω, what are the power–bandwidth ratios for the TE-like and TM-like modes of this device, respectively?

6.4.4 A symmetric Mach–Zehnder electro-optic waveguide modulator with 3-dB couplers at the input and output ends, as shown in Fig. 6.16, is fabricated on GaAs. At the optical wavelength of interest, $\lambda = 1.3$ μm, the single-mode waveguides have a confinement factor of $\Gamma = 0.75$ for both TE-like and TM-like modes. The linear refractive index is $n_o = 3.48$, and the nonvanishing Pockels coefficients are $r_{41} = r_{52} = r_{63} = 1.2$ pm V^{-1} for GaAs at this wavelength. The gap between the electrodes is $s_e = 3$ μm, and the length of them is $l = 4.25$ mm. For the following questions, assume that an input optical wave at 1.3 μm is launched into port 1.

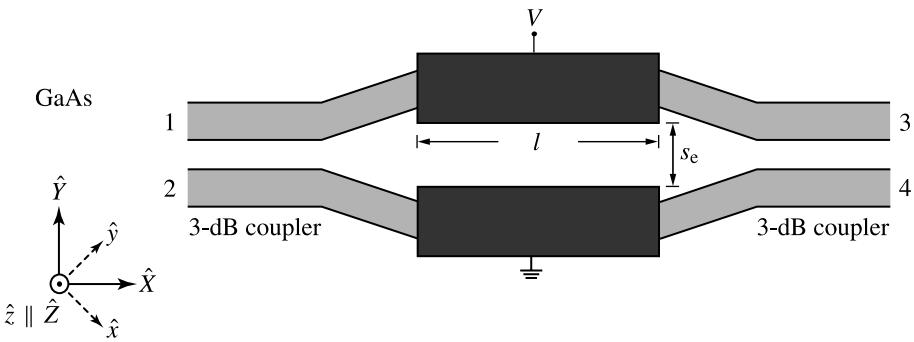


Figure 6.16 Symmetric Mach–Zehnder electro-optic waveguide modulator with 3-dB couplers. The $\hat{x}$, $\hat{y}$, and $\hat{z}$ unit vectors represent the original principal axes of the crystal, and $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ represent its new principal axes.

a. If the input light is launched into the TE-like mode, what is the minimum voltage that is needed to have equal intensity at output ports 3 and 4?
b. If the device is used as a switch for the TE-like mode, what is the switching voltage between the parallel state and the cross state?
c. What is the highest modulation frequency of the device if it is a lumped modulator with a capacitance of 1 pF and a resistance of 100 Ω?
d. If the device is designed in the form of a traveling-wave modulator with $Z = 50$ Ω for the transmission-line electrodes, what is the average power for high-frequency switching operation?
e. Can the device be used as an electro-optic switch for the TM-like mode? If you answer positively, what is the switching voltage? If you answer negatively, give an explanation.

6.4.5 A single-pole-double-throw electro-optic switch consists of an electro-optically modulated Mach–Zehnder interferometer that is connected to a 50 : 50 power-splitting Y-junction waveguide at the input end and to a 3-dB directional coupler at the output end, as shown in Fig. 6.17. The two channels of the device are identical single-mode waveguides, and the 3-dB coupler at the output end is

perfectly phase matched. A modulation voltage is applied to the lower arm of the interferometer to cause a phase difference $\Delta\varphi$ between the waves traveling through the two separate arms of the interferometer. Therefore, at the beginning of the output 3-dB coupler, the fields in the two channels are of equal magnitude but that in channel b has a phase shift of $\Delta\varphi$ with respect to that in channel a.

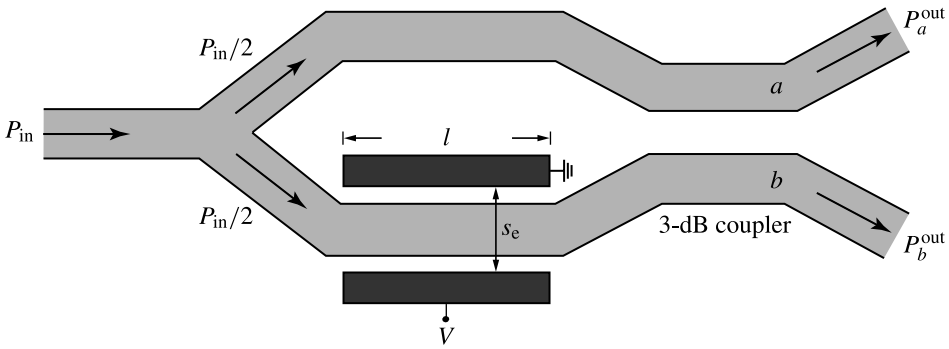


Figure 6.17 Single-pole-double-throw electro-optic switch with Y-junction input and 3-dB coupler output.

a. If the coupling coefficient of the 3-dB coupler is κ, which is a real parameter, what is the shortest possible length of the coupler?
b. Show that the output power in channel a is given by

$$P_a^{\text{out}} = \frac{P_{\text{in}}}{2}(1 - \sin\Delta\varphi), \tag{6.109}$$

and that in channel b is given by

$$P_b^{\text{out}} = \frac{P_{\text{in}}}{2}(1 + \sin\Delta\varphi), \tag{6.110}$$

where P_{in} is the power launched into the device at the input end.

6.4.6 Consider a single-pole-double-throw electro-optic switch as shown in Fig. 6.17 and considered in Problem 6.4.5 that is fabricated on an x-cut, y-propagating $LiNbO_3$ substrate. The length of the electrodes is $l = 6$ mm, and the gap between them is $s_{\text{e}} = 5$ μm. The effective overlap factor is approximately the same for the TE-like and TM-like modes: $\Gamma_{\text{TE}} = \Gamma_{\text{TM}} = 0.7$. At the desired operating wavelength of $\lambda = 1.55$ μm, we find $n_{\text{o}} = 2.213$, $n_{\text{e}} = 2.137$, $r_{33} = 30.8$ pm V^{-1}, and $r_{13} = 8.6$ pm V^{-1} for $LiNbO_3$.

a. What is the minimum voltage needed for the output power to be concentrated in only one channel if the device is operated in the TE-like mode? What is the minimum voltage change required in order to switch the output power completely from one channel to the other?
b. Answer the questions in (a) for the operation in TM-like mode.

6.4.7 What is the shortcoming of an ordinary electro-optic directional coupler switch with uniform electrodes? What can be done to overcome this limitation?

6.4.8 By solving the coupled-mode equations discussed in Section 4.2, show that the coupling efficiency of the reversed-$\Delta\beta$ directional coupler shown in Fig. 6.11(a) is that given by (6.85). Note that because the second section of the coupler has a different phase mismatch from that of the first section, the input conditions for the second section are $\tilde{A}_2(0) = \tilde{A}_1(l/2)e^{i(\beta_a+\kappa_{aa})l/2}$ and $\tilde{B}_2(0) = \tilde{B}_1(l/2)e^{i(\beta_b+\kappa_{bb})l/2}$, where β_a and β_b include the induced phase mismatch such that $\delta = (\beta_b - \beta_a)/2$.

6.4.9 A reversed-$\Delta\beta$ directional coupler switch as shown in Fig. 6.11(a) for the TM-like mode at $\lambda = 1.3$ μm is fabricated on a z-cut, x-propagating $LiNbO_3$ substrate with the same waveguide and electrode parameters as those of the uniform-$\Delta\beta$ directional coupler switch described in Example 6.7. Find the voltages to reach the cross and parallel states, respectively, if the length of the device is chosen to be (a) $l = l_c^{PM}$, (b) $l = \sqrt{2}l_c^{PM}$, and (c) $l = 2l_c^{PM}$, where l_c^{PM} is the coupling length found for the uniform-$\Delta\beta$ directional coupler switch in Example 6.7.

6.4.10 The concept of the reversed-$\Delta\beta$ directional coupler discussed in Section 6.4 can be extended to have any even number of $2m$ sections of periodically reversed $\Delta\beta$. If all sections have an equal length of $l/2m$ and η_s is the coupling efficiency of a single section, which can be obtained by replacing l with $l/2m$ in (6.82), the output coupling efficiency of the entire coupler of length l can be expressed as

$$\eta = \sin^2 \kappa_{eff} l, \tag{6.111}$$

where

$$\kappa_{eff} = \frac{2m}{l} \sin^{-1}\sqrt{\eta_s}. \tag{6.112}$$

Show that the coupling efficiency given by (6.85) for a two-section device can be expressed in this form with $m = 1$.

6.4.11 What is the basic requirement for the functioning of an electro-optic waveguide polarization modulator? What else is often also necessary in order to make such a device efficient?

6.4.12 Find the coupling coefficient for the polarization modulator shown in Fig. 6.14(a) using the coupled-mode theory. A $LiNbO_3$ device has electrodes of an equal length of $l = 2$ cm, a gap of $s_e = 8$ μm between the two electrodes, a waveguide geometry of $\Gamma_{EM} = 2/\pi$, and a phase mismatch of $\delta = 1.25$ cm^{-1} between its TE-like and TM-like modes. Plot the relative TE and TM output powers as a function of the modulation voltage for this device when it is operated with a TE-polarized input beam at $\lambda = 0.632$ μm. What are the voltage and the percentage of power conversion to the TM polarization at the first peak? What

is the required voltage for a power conversion efficiency of more than 90%? What improvement can be realized if the phase mismatch is reduced by 80% to $\delta = 0.5\ \text{cm}^{-1}$?

6.4.13 In this problem, we compare the principles and operating conditions of the three types of guided-wave electro-optic devices: the Mach–Zehnder waveguide interferometer, the directional coupler switch, and the TE–TM mode converter based on x-cut, y-propagating $LiNbO_3$.

a. What is the basic operation principle of each device? What are the differences in the concepts of these devices?

b. What is the function of the applied voltage in the operation of each device?

c. How is phase matching accomplished for each device?

6.5.1 What is the factor that fundamentally limits the modulation bandwidth of a lumped electro-optic modulator? What are the primary factors that determine the modulation bandwidth of a traveling-wave electro-optic modulator?

6.5.2 Find the 3-dB modulation bandwidth and the $P/f_{3\text{dB}}$ ratio for the TM-like mode of the traveling-wave Mach–Zehnder waveguide interferometric modulator considered in Example 6.9.

6.5.3 If the device considered in Problem 6.4.6 functions as a traveling-wave modulator with its electrodes made of microstrip lines of $n_{\text{m}} = 4.225$, what are the maximum modulation frequencies for the TE-like and TM-like modes, respectively?

SELECT BIBLIOGRAPHY

Boyd, R. W., *Nonlinear Optics*. Boston, MA: Academic Press, 1992.

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge University Press, 1996.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Iizuka, K., *Elements of Photonics in Free Space and Special Media*, Vol. I. New York: Wiley, 2002.

Iizuka, K., *Engineering Optics*. New York: Springer-Verlag, 1987.

Kaminow, I. P., *An Introduction to Electrooptic Devices.* Orlando, FL: Academic Press, 1974.

Nishihara, H., Haruna, M., and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Nye, J. F., *Physical Properties of Crystals*. London: Oxford University Press, 1957.

Parker, S. P., *Optical Source Book*. New York: McGraw-Hill, 1987.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Sirotin, Yu. I. and Shaskolskaya, M. P., *Fundamentals of Crystal Physics*. Moscow: Mir Publishers, 1982.

Syms, R. and Cozens, J., *Optical Guided Waves and Devices*. London: McGraw-Hill, 1992.

Tamir, T., ed., *Integrated Optics*. New York: Springer-Verlag, 1982.

Willardson, R. K. and Beer, A. C., eds., *Semiconductors and Semimetals*, Vol. 22, W. T. Tsang, ed. *Lightwave Communications Technology, Part E, Integrated Optoelectronics*. New York: Academic Press, 1985.

Yariv, A., *Optical Electronics in Modern Communications*, 5th edn. Oxford: Oxford University Press, 1997.

Yariv, A., and Yeh, P., *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: Wiley, 1984.

ADVANCED READING LIST

Alferness, R. C., "Waveguide electrooptic modulators," *IEEE Transactions on Microwave Theory and Techniques* **MTT-30**(8): 1121–1137, Aug. 1982.

Kawano, K., "High-speed Ti : $LiNbO_3$ and semiconductor optical modulators," *IEICE Transactions on Electronics* **E76-C**(2): 183–190, Feb. 1993.

7 Magneto-optic devices

Magneto-optic materials have unique physical properties that offer the opportunity of constructing devices with many special functions not possible from other photonic devices. The most significant of these properties are that the linear magneto-optic effect can produce circular birefringence and that, unlike other optical effects in dielectric media, it is nonreciprocal. All practical magneto-optic devices exploit one or both of these two properties. Important applications of these devices include polarization control, optical isolation, optical modulation, and magneto-optic recording. The basic principles of magneto-optic effects, as well as the functions of various magneto-optic devices based on these effects, are considered in this chapter.

7.1 Magneto-optic effects

Because magneto-optic effects are intimately connected to the magnetic properties of materials, we first briefly summarize the fundamental magnetic properties of materials. To a certain degree there is a parallelism between the electric and the magnetic properties of materials, but this parallelism is not complete. We shall pay attention to the similarities and differences between these properties in order to gain an appreciation of the uniqueness of magneto-optic devices.

Following the similarity between (1.1) and (1.2), a *magnetic susceptibility tensor*, χ_m, analogous to the electric susceptibility tensor χ can be defined to describe the magnetization induced by a magnetic field. In parallel with (1.54) and (1.55) for electric fields, we have, for magnetic fields,

$$\mathbf{M}(\mathbf{k}, \omega) = \chi_m(\mathbf{k}, \omega) \cdot \mathbf{H}(\mathbf{k}, \omega) \tag{7.1}$$

and

$$\mathbf{B}(\mathbf{k}, \omega) = \mu_0 \mathbf{H}(\mathbf{k}, \omega) + \mu_0 \mathbf{M}(\mathbf{k}, \omega) = \boldsymbol{\mu}(\mathbf{k}, \omega) \cdot \mathbf{H}(\mathbf{k}, \omega). \tag{7.2}$$

In general, magnetic permeability

$$\boldsymbol{\mu} = \mu_0(1 + \chi_m) \tag{7.3}$$

is a tensor for an anisotropic material. As mentioned in Section 1.1, in any material there is no physical basis to specify a magnetization induced by the magnetic component of an electromagnetic field at an optical frequency. Because the electric and magnetic components in an electromagnetic field are mutually coupled, it is not physically meaningful to define an optically induced magnetization that is separate from the optically induced electric polarization. Consequently, at an optical frequency ω, $\boldsymbol{\chi}_{\mathrm{m}}(\omega) = 0$ and $\boldsymbol{\mu}(\omega) = \mu_0$ in all materials, including anisotropic ones.

In magnetostatics, however, a nonvanishing magnetization in response to an externally applied magnetic field can appear in a material. In this situation, a nonvanishing magnetic susceptibility tensor $\boldsymbol{\chi}_{\mathrm{m}}$ is meaningful and $\boldsymbol{\mu}$ clearly differs from μ_0. Both $\boldsymbol{\chi}_{\mathrm{m}}$ *and* $\boldsymbol{\mu}$ *are symmetric tensors*. They can be diagonalized with real, orthogonal principal axes in a manner similar to the diagonalization of the symmetric $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ tensors of a nonmagnetic dielectric material that shows no optical activity. A fundamental difference between the electric and the magnetic properties of materials is that *while the principal dielectric susceptibilities are always positive*, as mentioned in Section 1.6, *the principal magnetic susceptibilities of lossless magnetic materials can be either positive or negative*. A material is referred to as *paramagnetic* if its principal magnetic susceptibilities are positive and as *diamagnetic* if they are negative. While the dielectric susceptibilities of ordinary materials are typically on the order of 1–10, the magnetic susceptibilities of paramagnetic and diamagnetic materials are extremely small, typically on the order of $\pm 10^{-5}$.

In a diamagnetic material there are no intrinsic magnetic dipole moments. The negative magnetization in such a material results from the magnetic dipole moments induced by an external magnetic field, which are always aligned in opposition to the inducing field. In contrast, a paramagnetic material consists of atoms or ions that have intrinsic magnetic dipole moments. Above a certain temperature that is characteristic of a particular paramagnetic material, these magnetic dipoles are randomly oriented in thermal equilibrium. In the presence of an externally applied magnetic field, these dipoles tend to align in the direction of the magnetic field and overshadow all diamagnetic effects in the material, resulting in a net positive magnetization. A few crystals are paramagnetic along one principal axis but diamagnetic along another, however.

In some paramagnetic solids, the intrinsic magnetic dipole moments can become orderly oriented in the absence of an external magnetic field due to their mutual interactions if the temperature is reduced below a certain critical value. Such solids are called *magnetically ordered*. A magnetically ordered solid whose intrinsic magnetic dipoles tend to line up in the same direction has a *spontaneous magnetization* and is called a *ferromagnetic* material, or a *ferromagnet*. Examples of ferromagnets are Fe, Co, Ni, Gd, Dy, and the alloy MnBi. The intrinsic magnetic dipoles in a magnetically ordered solid can also assume alternate antiparallel directions. This ordering can still result in a net spontaneous magnetization if the alternating dipoles are different and their moments do not cancel. Materials with this kind of property are called *ferrimagnetic* materials,

or *ferrimagnets*. The most important ferrimagnetic materials are the rare-earth iron garnets, particularly $Y_3Fe_5O_{12}$ (YIG), and the rare-earth transition-metal (RE–TM) alloys, such as GdFe, GdCo, TbFe, GdTbFe, TbFeCo, etc. If antiparallel alignment of the alternating dipoles in a solid results in complete cancellation of their magnetic moments, there is no net spontaneous magnetization though the solid is magnetically ordered. Such a solid is called an *antiferromagnetic* material, or an *antiferromagnet*. Examples of antiferromagnets are Cr, FeO, CoO, NiO, FeF_2, CoF_2, $CoCO_3$, and many garnets, such as $Ca_3Cr_2Ge_3O_{12}$, $Ca_3Mn_2Ge_3O_{12}$, and $Mn_3Al_2Si_3O_{12}$. The *critical temperature*, T_c, is called the *Curie temperature* for a ferromagnetic or ferrimagnetic material and the *Néel temperature* for an antiferromagnetic material. Above T_c, these materials are paramagnetic with small magnetic susceptibilities. In a ferromagnetic or ferrimagnetic material the value of the magnetic susceptibility diverges at T_c as the spontaneous magnetization appears. In an antiferromagnetic material, however, no spontaneous magnetization occurs. The magnetic susceptibility of an antiferromagnet merely reaches a finite maximum value at a temperature slightly above T_c.

Macroscopic *magnetic domains* having different magnetization orientations normally exist in a ferromagnetic or ferrimagnetic material below its Curie temperature. Consequently, the material appears to be only weakly magnetized or completely unmagnetized. By applying an external magnetic field, the domains that are oriented along the field grow at the expense of the adversely oriented ones, thus increasing the total net magnetization of the material. This process is reversible in weak magnetic fields but shows the characteristics of a hysteresis in strong fields. The largest magnetization reached is the *saturation magnetization*, M_s, beyond which the magnetization does not increase further with increasing magnetic field. Properties analogous to ferromagnetism also exist in some dielectric materials called *ferroelectrics*. In a ferroelectric material, a spontaneous polarization appears below its Curie temperature. Examples of ferroelectric crystals are KDP, $BaTiO_3$, $LiNbO_3$, and $KNbO_3$.

Because $\chi_m = 0$ and $\boldsymbol{\mu} = \mu_0$ at optical frequencies, the response of a material, irrespective of whether it is magnetic or nonmagnetic, to an optical field at a frequency ω is fully described by its electric susceptibility $\boldsymbol{\chi}(\omega)$ and, equivalently, by its electric permittivity $\boldsymbol{\epsilon}(\omega)$. A material does respond to a static magnetic field, $\boldsymbol{H}_0$, however. Its optical properties can be changed by its response to $\boldsymbol{H}_0$, resulting in various magneto-optic effects. The electric susceptibility and electric permittivity at an optical frequency ω thus become a function of $\boldsymbol{H}_0$:

$$\mathbf{P}(\omega, \boldsymbol{H}_0) = \epsilon_0 \boldsymbol{\chi}(\omega, \boldsymbol{H}_0) \cdot \mathbf{E}(\omega) = \epsilon_0 \boldsymbol{\chi}(\omega) \cdot \mathbf{E}(\omega) + \epsilon_0 \Delta\boldsymbol{\chi}(\omega, \boldsymbol{H}_0) \cdot \mathbf{E}(\omega) \tag{7.4}$$

and

$$\mathbf{D}(\omega, \boldsymbol{H}_0) = \boldsymbol{\epsilon}(\omega, \boldsymbol{H}_0) \cdot \mathbf{E}(\omega) = \boldsymbol{\epsilon}(\omega) \cdot \mathbf{E}(\omega) + \Delta\boldsymbol{\epsilon}(\omega, \boldsymbol{H}_0) \cdot \mathbf{E}(\omega), \tag{7.5}$$

where $\boldsymbol{\chi}(\omega) = \boldsymbol{\chi}(\omega, \boldsymbol{H}_0 = 0)$ and $\boldsymbol{\epsilon}(\omega) = \boldsymbol{\epsilon}(\omega, \boldsymbol{H}_0 = 0)$ represent the intrinsic properties of the medium in the absence of the magnetic field. In the case of a ferromagnetic

or ferrimagnetic material, in which a static magnetization $\boldsymbol{M}_0$ exists, the properties of the medium at an optical frequency are dependent on $\boldsymbol{M}_0$. Then, instead of (7.4) and (7.5), we have

$$\mathbf{P}(\omega, \boldsymbol{M}_0) = \epsilon_0 \boldsymbol{\chi}(\omega, \boldsymbol{M}_0) \cdot \mathbf{E}(\omega) = \epsilon_0 \boldsymbol{\chi}(\omega) \cdot \mathbf{E}(\omega) + \epsilon_0 \Delta \boldsymbol{\chi}(\omega, \boldsymbol{M}_0) \cdot \mathbf{E}(\omega) \tag{7.6}$$

and

$$\mathbf{D}(\omega, \boldsymbol{M}_0) = \boldsymbol{\epsilon}(\omega, \boldsymbol{M}_0) \cdot \mathbf{E}(\omega) = \boldsymbol{\epsilon}(\omega) \cdot \mathbf{E}(\omega) + \Delta \boldsymbol{\epsilon}(\omega, \boldsymbol{M}_0) \cdot \mathbf{E}(\omega). \tag{7.7}$$

While $\boldsymbol{\chi}$ and $\boldsymbol{\epsilon}$ are changed in the presence of $\boldsymbol{H}_0$ or $\boldsymbol{M}_0$, the magnetic permeability of the material at an optical frequency remains the constant μ_0, and the relation between $\mathbf{B}(\omega)$ and $\mathbf{H}(\omega)$ remains independent of $\boldsymbol{H}_0$ or $\boldsymbol{M}_0$:

$$\mathbf{B}(\omega) = \mu_0 \mathbf{H}(\omega). \tag{7.8}$$

Therefore, magneto-optic effects are completely characterized by $\boldsymbol{\epsilon}(\omega, \boldsymbol{H}_0)$, if no internal magnetization is present, or by $\boldsymbol{\epsilon}(\omega, \boldsymbol{M}_0)$, if an internal magnetization is present. In general, these effects are weak perturbations to the optical properties of the material. The *first-order, or linear, magneto-optic effect* is characterized by a linear dependence of $\boldsymbol{\epsilon}$ on $\boldsymbol{H}_0$ or $\boldsymbol{M}_0$, and the *second-order, or quadratic, magneto-optic effect* results from a quadratic dependence of $\boldsymbol{\epsilon}$ on $\boldsymbol{H}_0$ or $\boldsymbol{M}_0$. Note that like electro-optic effects, both first- and second-order magneto-optic effects are nonlinear optical effects.

The general description of magneto-optic effects in terms of $\boldsymbol{\epsilon}(\omega, \boldsymbol{H}_0)$ or $\boldsymbol{\epsilon}(\omega, \boldsymbol{M}_0)$ is analogous to the general description of electro-optic effects in terms of $\boldsymbol{\epsilon}(\omega, \boldsymbol{E}_0)$. The classification of first- and second-order magneto-optic effects is also analogous to that of first- and second-order electro-optic effects. However, there are many important fundamental differences between magneto-optic and electro-optic effects. These differences originate from basic distinctions in the electric and the magnetic characteristics of materials and are mostly tied to the fact that electric and magnetic fields follow different rules of transformation under space inversion and time reversal, as described in Section 1.1. The major differences and their implications are summarized below.

1. **Space-inversion symmetry.** Materials with the space-inversion symmetry are centrosymmetric. In such materials, no spontaneous electric polarization can exist, and the first-order electro-optic effect also vanishes. However, *neither a spontaneous magnetization nor the first-order magneto-optic effect is not forbidden in a centrosymmetric material*. This difference is due to the fact that under space inversion, the polar vectors **P** and **E** change sign, but the axial vectors **M** and **H** do not. Therefore, amorphous solids can be ferromagnetic or ferrimagnetic but cannot be ferroelectric. The first-order magneto-optic effect appears in gases and liquids, as well as in amorphous solids and nonpolar cubic crystals, where the Pockels effect does not exist.

2. **Time-reversal symmetry.** Materials with the time-reversal symmetry are lossless and reciprocal. A lossless dielectric material not subject to an external magnetic field possesses time-reversal symmetry because it is reciprocal. Time-reversal symmetry is lost in a dielectric material when it has a loss or gain, or when an external magnetic field is applied to it. A magnetically ordered material, regardless of whether it is lossless or not, is nonreciprocal and thus does not possess time-reversal symmetry.
3. **Reciprocity.** In an optical system that has time-reversal symmetry, an optical signal can be run backward in time without changing the reality of the physics. This is not possible if the medium involved is nonreciprocal or if it has a loss or gain. However, there is a difference between the two possibilities that causes the time-reversal symmetry to break down. In a nonreciprocal medium, there is no symmetry in the interchange of the source and the detector of an optical signal. In contrast, the symmetry in such an interchange exists in a reciprocal medium that has a loss or gain. Therefore, only the magneto-optic system, being nonreciprocal, can provide the function of *optical isolation* discussed in later sections. A lossy dielectric system, despite its lack of time-reversal symmetry, is not capable of such a function.
4. **Magnetic symmetry.** The symmetry properties of dielectric materials discussed in Chapters 1 and 6 are based on the considerations of spatial transformations only. They are in fact the *electric symmetry properties* of materials. The *magnetic symmetry properties* of materials have to be determined by considering spatial transformations in combination with time-reversal transformation because magnetic structures do not have time-reversal symmetry. The result is *magnetic symmetry groups*, called *Shubnikov's groups*, which are much more complicated than the ordinary symmetry groups based solely on the electric structures of crystals. Therefore, general symmetry considerations for the magneto-optic effects in magnetically ordered crystals, particularly in anisotropic ones, are quite complicated.

We first consider the magneto-optic effects in a material that is not magnetically ordered, i.e., a paramagnet or a diamagnet. In such a material, operation of the time-reversal transformation yields

$$\epsilon_{ij}(\omega, \boldsymbol{H}_0) = \epsilon_{ji}(\omega, -\boldsymbol{H}_0) \tag{7.9}$$

when the material is subject to an external magnetic field $\boldsymbol{H}_0$. This relation characterizes the magneto-optic effects in a magnetically nonordered material. It is generally true regardless of the symmetry property of the material. If the material is lossless, then its dielectric permittivity tensor is Hermitian:

$$\epsilon_{ij}(\omega, \boldsymbol{H}_0) = \epsilon^*_{ji}(\omega, \boldsymbol{H}_0). \tag{7.10}$$

If we express the real and imaginary parts of $\boldsymbol{\epsilon}$ explicitly by writing $\epsilon_{ij} = \epsilon'_{ij} + \mathrm{i}\epsilon''_{ij}$, we

find from combination of these two relations that

$$\epsilon'_{ij}(\omega, \boldsymbol{H}_0) = \epsilon'_{ij}(\omega, -\boldsymbol{H}_0) = \epsilon'_{ji}(\omega, \boldsymbol{H}_0) = \epsilon'_{ji}(\omega, -\boldsymbol{H}_0), \tag{7.11}$$

$$\epsilon''_{ij}(\omega, \boldsymbol{H}_0) = -\epsilon''_{ij}(\omega, -\boldsymbol{H}_0) = -\epsilon''_{ji}(\omega, \boldsymbol{H}_0) = \epsilon''_{ji}(\omega, -\boldsymbol{H}_0). \tag{7.12}$$

As a result, the magneto-optic effects in a magnetically nonordered, lossless material can be generally described as

$$\epsilon_{ij}(\boldsymbol{H}_0) = \epsilon_{ij} + \Delta\epsilon_{ij}(\boldsymbol{H}_0) = \epsilon_{ij} + \mathrm{i}\epsilon_0 \sum_k f_{ijk} H_{0k} + \epsilon_0 \sum_{k,l} c_{ijkl} H_{0k} H_{0l} + \cdots, \tag{7.13}$$

where f_{ijk} and c_{ijkl} are real quantities that satisfy the following relations:

$$f_{ijk} = -f_{jik}, \qquad c_{ijkl} = c_{jikl} = c_{ijlk} = c_{jilk}. \tag{7.14}$$

The linear dependence of $\epsilon_{ij}(\boldsymbol{H}_0)$ on the magnetic field appears only as antisymmetric, imaginary components in the off-diagonal elements of the permittivity tensor. This first-order magneto-optic effect results in *circular birefringence*; it manifests itself as, notably, the *Faraday effect* and the *magneto-optic Kerr effect* discussed in the following sections. *The first-order magneto-optic effect and the phenomena resulting from it are nonreciprocal*. In any material, even a centrosymmetric one, that has a spontaneous magnetization or is subject to an external magnetic field, the first-order magneto-optic effect always exists, resulting in the nonreciprocity of such a material. In contrast, the quadratic dependence on the magnetic field appears as symmetric, real components in the permittivity tensor elements. *This second-order magneto-optic effect is reciprocal* and is called the *Cotton–Mouton effect*. It causes a *linear birefringence* in the material and is analogous to, but much weaker than, the electro-optic Kerr effect.

Note that in expressing the magneto-optic effects in terms of (7.13) we have expanded the elements of the permittivity tensor $\boldsymbol{\epsilon}$ instead of expanding those of the relative impermeability tensor $\boldsymbol{\eta}$, as done in (6.14) for electro-optic effects. The reason is that it is convenient to use the index ellipsoid in dealing with electro-optic effects but not in treating magneto-optic effects. Instead, as we shall see later in this chapter, it is convenient to use the permittivity tensor directly to treat magneto-optic effects. As demonstrated in Section 6.1, the choice of using $\boldsymbol{\epsilon}$ or $\boldsymbol{\eta}$ does not make a difference in the final result and is only a matter of convenience.

The magneto-optic effects are relatively weak in comparison to, and tend to be obscured by, any natural or structural birefringence that might exist in a material. Fortunately, both first- and second-order magneto-optic effects exist in isotropic materials, including noncrystals and cubic crystals. For these reasons, materials of particular interest and practical importance for magneto-optic effects and their applications are those in which any birefringence originated from other effects, such as material anisotropy or inhomogeneity, does not exist or, if it exists, does not dominate the particular magneto-optic effect of interest. Such materials include isotropic materials and, in some cases,

uniaxial crystals subject to a magnetic field that is parallel to the optical axis. For magneto-optic effects in these materials, we can take the direction of $\boldsymbol{H}_0$ to be the z direction without loss of generality. Then, $\boldsymbol{H}_0 = H_{0z}\hat{z}$, and (7.13) yields

$$\boldsymbol{\epsilon}(\boldsymbol{H}_0) = \epsilon_0 \begin{bmatrix} n_o^2 + c_\perp H_{0z}^2 & \mathrm{i}fH_{0z} & 0 \\ -\mathrm{i}fH_{0z} & n_o^2 + c_\perp H_{0z}^2 & 0 \\ 0 & 0 & n_e^2 + c_\parallel H_{0z}^2 \end{bmatrix}, \tag{7.15}$$

where $f = f_{123}, c_\perp = c_{1133} = c_{2233}$, and $c_\parallel = c_{3333}$. Clearly, the Cotton–Mouton effect results in a linear birefringence, which is insignificant unless $(c_\parallel - c_\perp)H_{0z}^2 \geq n_e^2 - n_o^2$. This condition usually means that the material has to be isotropic or nearly isotropic in order for the Cotton–Mouton effect to be observable. Note also that in order to observe the Cotton–Mouton effect, the propagation direction of an optical wave has to be perpendicular to the magnetic field direction so that an optical field component parallel to $\boldsymbol{H}_0$ exists.

The magneto-optic effects in magnetically ordered crystals can be rather complicated due to the magnetic symmetry properties of such crystals. However, for the same reasons as described above, magnetically ordered materials of practical importance for device applications are also isotropic materials and, in some cases, uniaxial crystals with the magnetic field or the magnetization parallel to the optical axis. For a magnetically ordered, lossless material that falls into one of these categories, (7.15) applies if it is antiferromagnetic. In a ferromagnetic or ferrimagnetic material, the magneto-optic effects are determined by its magnetization $\boldsymbol{M}_0$, rather than by any externally applied magnetic field $\boldsymbol{H}_0$, despite the fact that the value and direction of $\boldsymbol{M}_0$ can be varied by $\boldsymbol{H}_0$. Then, for $\boldsymbol{M}_0 = M_{0z}\hat{z}$, $\boldsymbol{\epsilon}(\boldsymbol{M}_0)$ can be expressed as

$$\boldsymbol{\epsilon}(\boldsymbol{M}_0) = \epsilon_0 \begin{bmatrix} n_\perp^2 & \mathrm{i}\xi & 0 \\ -\mathrm{i}\xi & n_\perp^2 & 0 \\ 0 & 0 & n_\parallel^2 \end{bmatrix}, \tag{7.16}$$

where ξ, representing the first-order effect, is linearly dependent on M_{0z} with the symmetry of $\xi(M_{0z}) = -\xi(-M_{0z})$, and $n_\perp^2$ and $n_\parallel^2$, accounting for the second-order effect, are functions of M_{0z}^2. In the case of an isotropic material, $n_\parallel^2 - n_\perp^2$ is proportional to M_{0z}^2, creating a magnetically induced linear birefringence. Because of this *magnetic linear birefringence*, a ferromagnet or ferrimagnet that has an isotropic structure does not really have isotropic optical properties. For the same reason, the so-called cubic ferromagnetic or cubic ferrimagnetic crystals, such as YIG and other magnetic garnets, are never really cubic. However, the magnetic linear birefringence is generally very small, with $n_\parallel - n_\perp$ on the order of a few $\times 10^{-5}$ at room temperature for most magnetic garnets. In the case of a uniaxial crystal, the magnetic linear birefringence is dominated by, and is difficult to separate from, the nonmagnetic natural birefringence of the crystal.

In the rest of this chapter, we shall restrict our discussions to magneto-optic effects in isotropic materials or uniaxial crystals with $\boldsymbol{H}_0$, or $\boldsymbol{M}_0$, parallel to the optical axis.

7.2 Faraday effect

The Faraday effect is a phenomenon based on the propagation and transmission of an optical wave through a material in the presence of a magnetic field. For the convenience of a general discussion, we consider the $\boldsymbol{\epsilon}$ tensor in the presence of a magnetic field or a magnetization of the form given by (7.16). When there is an applied magnetic field but no spontaneous magnetization, we identify the tensor elements, ξ, $n_\perp$, and $n_\parallel$, with the corresponding elements in (7.15) as $\xi = f H_{0z}$, $n_\perp^2 = n_\text{o}^2 + c_\perp H_{0z}^2$, and $n_\parallel^2 = n_\text{e}^2 + c_\parallel H_{0z}$. When there is a spontaneous magnetization, we follow the definitions of $\xi(M_{0z})$, $n_\perp^2(M_{0z}^2)$, and $n_\parallel^2(M_{0z}^2)$ in (7.16) because the effect is then completely determined by the magnetization regardless of whether there is an applied magnetic field or not.

The eigenvalues and the corresponding eigenvectors of $\boldsymbol{\epsilon}$ can be found by diagonalizing $\boldsymbol{\epsilon}$ through the normal procedure (see Problem 7.2.1). By so doing, we find that the eigenvalues are

$$\epsilon_+ = \epsilon_0(n_\perp^2 - \xi), \qquad \epsilon_- = \epsilon_0(n_\perp^2 + \xi), \qquad \epsilon_z = \epsilon_0 n_\parallel^2, \tag{7.17}$$

and the eigenvectors are, correspondingly,

$$\hat{e}_+ = \frac{1}{\sqrt{2}}(\hat{x} + \mathrm{i}\hat{y}), \qquad \hat{e}_- = \frac{1}{\sqrt{2}}(\hat{x} - \mathrm{i}\hat{y}), \qquad \hat{z}. \tag{7.18}$$

The complex eigenvectors, $\hat{e}_+$ and $\hat{e}_-$, respectively, are the left- and right-circularly polarized unit vectors defined in Section 1.4. These two complex unit vectors appear as eigenvectors because the $\boldsymbol{\epsilon}$ tensor in the presence of a magnetic field or a magnetization is not symmetric. The eigenvalues are all real because $\boldsymbol{\epsilon}$ is Hermitian. It is clearly not possible to attach the meaning of the principal axes in real space to these complex eigenvectors. Nonetheless, these eigenvectors still define the principal normal modes of polarization for proper decomposition of the electric field components of an optical wave that propagates in the medium:

$$D_+ = \epsilon_+ E_+, \qquad D_- = \epsilon_- E_-, \qquad D_z = \epsilon_z E_z. \tag{7.19}$$

Therefore, ϵ_+/ϵ_0, ϵ_-/ϵ_0, and ϵ_z/ϵ_0 are the principal dielectric constants for the three normal modes. They define the following three principal indices of refraction:

$$n_+ = \sqrt{n_\perp^2 - \xi} \approx n_\perp - \frac{\xi}{2n_\perp}, \qquad n_- = \sqrt{n_\perp^2 + \xi} \approx n_\perp + \frac{\xi}{2n_\perp}, \qquad n_z = n_\parallel. \tag{7.20}$$

The propagation constants for the normal modes are given by

$$k^+ = \frac{n_+\omega}{c}, \qquad k^- = \frac{n_-\omega}{c}, \qquad k^z = \frac{n_z\omega}{c}. \tag{7.21}$$

When an optical wave propagates along the z axis, in either the positive z or the negative z direction, the normal modes are the circularly polarized modes $\hat{e}_+$ and $\hat{e}_-$, with propagation constants k^+ and k^-, respectively. As mentioned in Section 1.4, if the wave propagates in the positive z direction, $\hat{e}_+$ is the left-circular polarization and $\hat{e}_-$ is the right-circular polarization. If the wave propagates in the negative z direction, $\hat{e}_+$ becomes the right-circular polarization while $\hat{e}_-$ becomes the left-circular polarization. In either situation, however, n_+ and k^+ defined above remain with $\hat{e}_+$, and n_- and k^- still belong to $\hat{e}_-$. With a fixed z direction, the wavevectors for the two circularly polarized normal modes are $\mathbf{k}^+ = k^+\hat{z}$ and $\mathbf{k}^- = k^-\hat{z}$, respectively, for forward propagation in the positive z direction and are $\mathbf{k}^+ = -k^+\hat{z}$ and $\mathbf{k}^- = -k^-\hat{z}$, respectively, for backward propagation in the negative z direction. Furthermore, we see from (7.20) that the values of n_+ and n_-, thus also those of k^+ and k^-, do not depend on the wave propagation direction. Instead, they depend only on the direction of $\boldsymbol{H}_0$, or that of $\boldsymbol{M}_0$ if an internal magnetization exists. If an optical wave is initially circularly polarized, either left or right, it is in one of the normal modes. It propagates with a single propagation constant belonging to the circular polarization and maintains the same polarization state throughout its path in the medium. If it is reflected to propagate in the opposite direction, its handedness changes, but not its unit vector or its propagation constant. If an optical wave is initially linearly or elliptically polarized, its field is a superposition of the two circularly polarized normal modes. This field then decomposes into two circularly polarized orthogonal components that propagate with different propagation constants, k^+ and k^-. This phenomenon is called *circular birefringence*. It is known as *magnetic circular birefringence* because it is caused by the magneto-optic effect.

A case of special interest is the propagation of a linearly polarized optical wave in such a medium. Assume, without loss of generality, that the wave is initially linearly polarized in the x direction at an arbitrary initial position $z = 0$:

$$\mathbf{E}(0, t) = \hat{x}\mathcal{E}\mathrm{e}^{-\mathrm{i}\omega t} = \frac{\mathcal{E}}{\sqrt{2}}(\hat{e}_+ + \hat{e}_-)\mathrm{e}^{-\mathrm{i}\omega t}, \tag{7.22}$$

with $\mathcal{E}_+ = \mathcal{E}_- = \mathcal{E}/\sqrt{2}$. Both circularly polarized components propagate as normal modes with their respective propagation constants. When the wave propagates a distance l in the positive z direction, we have

$$\begin{aligned}
\mathbf{E}(l, t) &= \hat{e}_+\mathcal{E}_+ \exp[\mathrm{i}\mathbf{k}^+ \cdot \hat{z}(l - 0) - \mathrm{i}\omega t] + \hat{e}_-\mathcal{E}_- \exp[\mathrm{i}\mathbf{k}^- \cdot \hat{z}(l - 0) - \mathrm{i}\omega t] \\
&= \hat{e}_+\mathcal{E}_+ \exp(\mathrm{i}k^+ l - \mathrm{i}\omega t) + \hat{e}_-\mathcal{E}_- \exp(\mathrm{i}k^- l - \mathrm{i}\omega t) \\
&= \frac{\mathcal{E}}{2}\left[\hat{x}\left(\mathrm{e}^{\mathrm{i}k^+ l} + \mathrm{e}^{\mathrm{i}k^- l}\right) + \mathrm{i}\hat{y}\left(\mathrm{e}^{\mathrm{i}k^+ l} - \mathrm{e}^{\mathrm{i}k^- l}\right)\right]\mathrm{e}^{-\mathrm{i}\omega t} \\
&= \mathcal{E}\left(\hat{x}\cos\frac{k^- - k^+}{2}l + \hat{y}\sin\frac{k^- - k^+}{2}l\right)\exp\left(\mathrm{i}\frac{k^+ + k^-}{2}l - \mathrm{i}\omega t\right).
\end{aligned} \tag{7.23}$$

The optical field clearly remains linearly polarized because its x and y components are

in phase, but its plane of polarization is rotated by an angle of

$$\theta_{\mathrm{F}} = \tan^{-1}\frac{\mathcal{E}_y}{\mathcal{E}_x} = \frac{k^- - k^+}{2}l = \frac{\pi}{\lambda}(n_- - n_+)l \approx \frac{\pi\xi}{\lambda n_\perp}l. \tag{7.24}$$

This magnetically induced rotation of the plane of polarization of a linearly polarized optical wave is called *Faraday rotation*, and this phenomenon is known as the *Faraday effect*. It can be shown that the plane of polarization rotates by the same amount in the same sense if the wave propagates in the negative z direction for the same distance l. Therefore, *the sense of Faraday rotation is independent of the direction of wave propagation* (see Problem 7.2.4). A positive value for θ corresponds to a rotation from the positive x axis to the positive y axis, which is counterclockwise rotation when viewed facing against the direction of propagation. A device that provides the function of the Faraday rotation is called a *Faraday rotator*.

In a paramagnetic or diamagnetic material, which has no internal magnetization, the Faraday rotation for a linearly polarized wave propagating over a distance l is linearly proportional to the externally applied magnetic field. The Faraday rotation angle in this case is generally expressed as

$$\theta_{\mathrm{F}} = V H_{0z} l, \tag{7.25}$$

where

$$V = \frac{\omega f}{2cn_\perp} = \frac{\pi f}{\lambda n_\perp} \tag{7.26}$$

is the *Verdet constant* (measured in radians per ampere). In the literature, the Verdet constant is often quoted in Gaussian units (minutes per oersted per centimeter). The conversion between Gaussian and SI units is 1 min $\mathrm{Oe}^{-1}\ \mathrm{cm}^{-1} = 2.094 \times 10^{-2}$ deg $\mathrm{A}^{-1} = 3.655 \times 10^{-4}$ rad A^{-1}. The Verdet constant defined in terms of (7.25) and given in radians per ampere or degrees per ampere is convenient when the magnetic field is generated by a current. In many practical situations, however, the magnetic field is provided by a permanent magnet. Then, the Faraday rotation angle is often written in terms of the magnetic induction as $\theta_{\mathrm{F}} = V B_{0z} l$, and the unit of the Verdet constant is correspondingly quoted as degrees per gauss per centimeter in the Gaussian system or radians per tesla per meter in the SI system. The conversion between Gaussian and SI units is $1^\circ\ \mathrm{G}^{-1}\ \mathrm{cm}^{-1} = 10^{6\circ}\ \mathrm{T}^{-1}\ \mathrm{m}^{-1} = 1.745 \times 10^4$ rad $\mathrm{T}^{-1}\ \mathrm{m}^{-1}$. The conversion of units for the Verdet constant from one defined in terms of B_{0z} to one defined in terms of H_{0z} is $1^\circ\ \mathrm{G}^{-1}\ \mathrm{cm}^{-1} \rightarrow 1^\circ\ \mathrm{Oe}^{-1}\ \mathrm{cm}^{-1}$ for Gaussian units and 1 rad $\mathrm{T}^{-1}\ \mathrm{m}^{-1} \rightarrow 4\pi \times 10^{-7}$ rad A^{-1} for SI units. The Verdet constant has positive values for diamagnetic materials and negative values for paramagnetic materials. The Verdet constants of some materials of interest are listed in Table 7.1. The Verdet constant of a given material is a function of both optical wavelength and temperature. In the optical spectral region, its absolute value usually increases when the optical wavelength or the temperature decreases.

Table 7.1 *Verdet constants of representative paramagnetic and diamagnetic materials at* 300 K

Material	Wavelength λ (nm)	Verdet constant V (rad A^{-1})	Verdet constant V (rad T^{-1} m^{-1})
Water[a]	589.3	4.79×10^{-6}	3.81
Diamond	589.3	5.88×10^{-6}	4.68
Quartz	589.3	6.07×10^{-6}	4.84
Light flint glass	589.3	1.16×10^{-5}	9.23
CS_2	589.3	1.55×10^{-5}	12.3
Pr^{3+}–B glass	670	-8.88×10^{-5}	−70.7
Pr^{3+}–Al–Si glass	700	-7.24×10^{-5}	−57.6
Tb^{3+}–Al–Si glass	700	-7.89×10^{-5}	−62.8
Dy^{3+}–Al–Si glass	700	-9.94×10^{-5}	−79.1
Pr^{3+}–P glass	700	-4.50×10^{-5}	−35.8
Tb^{3+}–P glass	700	-5.48×10^{-5}	−43.6
Ce^{3+}–P glass	500	-1.19×10^{-4}	−94.7
	700	-4.82×10^{-5}	−38.4
Pure silica glass[b]	532	6.00×10^{-6}	4.77
	632.8	3.93×10^{-6}	3.13
	785	3.24×10^{-6}	2.58
TGG[c]	500	-2.74×10^{-4}	−218
	532	-2.39×10^{-4}	−190
	632.8	-1.68×10^{-4}	−134
	750	-1.01×10^{-4}	−80
	800	-8.17×10^{-5}	−65
	1064	-5.03×10^{-5}	−40
	1300	-2.51×10^{-5}	−20

[a] *Handbook of Optics*, New York: McGraw-Hill, 1978, pp. 17-20, 17-21. Same source for diamond, quartz, light flint glass, and CS_2.
[b] Tan, C. Z. and Arndt, J., *Journal of Non-Crystalline Solids* **222**: 391–395, 1997; *Journal of Physics & Chemistry of Solids* **60**: 1689–1692, 1999.
[c] $Tb_3Ga_5O_{12}$: Barnes, N. P. and Petway, L. B., *Journal of the Optical Society of America B* **9**: 1912–1915, 1992; Chen, X., Lavorel, B., Boquillon, J. P., Saint-Loup, R., and Jannin, M., *Solid-State Electronics* **42**: 1765–1766, 1998; and assorted other sources.

In a ferromagnetic or ferrimagnetic material, which has an internal magnetization, ξ is determined by the magnetization rather than by the applied magnetic field. The total Faraday rotation angle for an optical wave traveling over a distance l through such a material is simply

$$\theta_{\mathrm{F}} = \rho_{\mathrm{F}} \frac{M_{0z}}{M_{\mathrm{s}}} l, \tag{7.27}$$

where $M_{0z} \leq M_s$ is the existing magnetization in the material and M_s is the saturation magnetization of the material. The Faraday rotation can be small if the material is not sufficiently magnetized; it is maximized only when the material is fully magnetized to reach its saturation magnetization. The Faraday rotation is then characterized by the following *specific Faraday rotation*, or *rotatory power*:

$$\rho_F = \frac{\omega\xi(M_s)}{2cn_\perp} = \frac{\pi\xi(M_s)}{\lambda n_\perp}, \tag{7.28}$$

which is the amount of rotation per unit length traversed by the optical wave in the material at saturation magnetization. It has the unit of radians per meter, but is often quoted in the unit of degrees per centimeter. The conversion between them is $1^\circ\ \text{cm}^{-1} = 1.745\ \text{rad m}^{-1}$. The specific Faraday rotation can have either positive or negative values. Many metallic ferromagnetic materials, such as Fe, Co, and Ni, have very large values of specific Faraday rotation, but they also have very large absorption coefficients and, consequently, are not very useful in many device applications that require optical transmission. Therefore, a figure of merit for these materials is ρ_F/α, often quoted in degrees per decibel, which measures the amount of Faraday rotation in a medium for a certain amount of attenuation. The specific Faraday rotations of some ferromagnetic and ferrimagnetic materials, together with their absorption coefficients and figures of merit, are listed in Table 7.2. Both the specific Faraday rotation and the absorption coefficient of a material are highly dependent on the optical wavelength and the temperature. Similar to the Verdet constant of a paramagnetic or diamagnetic material, the specific Faraday rotation of a ferromagnetic or ferrimagnetic material normally increases when the temperature or the wavelength decreases. Significant variations can occur in either direction, however, near the optical frequencies corresponding to ferromagnetic resonances in such materials, sometimes even changing the sign of the specific Faraday rotation in a given material at certain resonance frequencies.

The Faraday effect is nonreciprocal. It has the characteristic that *the sense of the Faraday rotation in a particular material is independent of the direction of wave propagation but is determined only by the direction of the external magnetic field, or that of the magnetization if the material is ferromagnetic or ferrimagnetic*. The expression for θ_F in (7.25) holds true for propagation in either the parallel or the antiparallel direction with respect to $\boldsymbol{H}_0$, and that for ρ_F in (7.28) is also valid for propagation in either direction with respect to $\boldsymbol{M}_0$. The amount of the Faraday rotation is doubled, rather than canceled, when an optical wave passing through a magneto-optic material is reflected to retrace its original path in the opposite direction back to the starting point. This phenomenon is a direct consequence of the fact discussed above that the propagation constant associated with each circularly polarized eigenvector is independent of the wave propagation direction and, therefore, is not changed by reflection.

Table 7.2 *Specific Faraday rotation of representative ferromagnetic and ferrimagnetic materials at* 300 K

Material	Wavelength λ (nm)	Specific rotation ρ_F ($^\circ$ cm^{-1})	Absorption coefficient α (dB cm^{-1})	Figure of merit ρ_F/α ($^\circ$ dB^{-1})
Fe[a]	546	3.5×10^5	3.3×10^6	0.11
Co	546	3.6×10^5	3.7×10^6	0.10
Ni	400	7.2×10^5	9.1×10^5	0.79
MnBi	632.8	5.3×10^5	3.3×10^6	0.16
YIG[b]	1064	280	65	4.3
	1150	250	54	4.6
	1200	240	50	4.8
	1310	224	35	6.4
	1550	216	23.8	9.1
YbBi : YIG[c]	1310	760	38	20
	1550	404	15.7	25.8
Bi : YIG[d]	1550	−1250	2.7	463
Ce : YIG[e]	1310	−2510	9.8	256
	1550	−1310	2.7	486

[a] Freiser, M. J., *IEEE Transactions on Magnetics* **MAG-4**: 152–161, 1968. Same source for Co and Ni.

[b] $Y_3Fe_5O_{12}$: Sekijima, T., Fuji, T., Wakino, K., and Okada, M., *IEEE Transactions on Microwave Theory and Techniques* **47**: 2294–2298, 1999; Zhao, W., *Sensors and Actuators A* **89**: 250–254, 2001; and assorted other sources. The absorption coefficients of YIG cited here are much higher than those reported in old literature.

[c] $Yb_yBi_xY_{3-x-y}Fe_5O_{12}$ with $x = 1.03$, $y = 1.12$: Zhao, W., *Sensors and Actuators A* **89**: 250–254, 2001. The properties of YbBi : YIG vary with Yb and Bi concentrations.

[d] $Bi_xY_{3-x}Fe_5O_{12}$: Sekijima, T., Fuji, T., Wakino, K., and Okada, M., *IEEE Transactions on Microwave Theory and Techniques* **47**: 2294–2298, 1999. The properties of Bi : YIG vary with Bi concentration.

[e] $Ce_xY_{3-x}Fe_5O_{12}$ with $x = 0.5$: Sekijima, T., Fuji, T., Wakino, K., and Okada, M., *IEEE Transactions on Microwave Theory and Techniques* **47**: 2294–2298, 1999. The properties of Ce : YIG vary with Ce concentration.

The Faraday rotation is positive when the value of θ_F, or that of ρ_F, is positive, meaning that the rotation is counterclockwise when viewed in the direction against that of $\boldsymbol{H}_0$, or that of $\boldsymbol{M}_0$ when an internal magnetization exists. Therefore, *the sense of positive Faraday rotation is the same as the electric current that generates* $\boldsymbol{H}_0$ *or, in the case of ferromagnets and ferrimagnets, the current that can be conceptually associated with* $\boldsymbol{M}_0$. Using the right-hand rule, the axial vector corresponding to a positive Faraday rotation points in the same direction as that of the $\boldsymbol{H}_0$ or $\boldsymbol{M}_0$ causing the Faraday effect. For negative Faraday rotation, the sense of rotation is opposite to that for positive Faraday rotation. Figure 7.1 summarizes these concepts.

The Faraday rotation in a diamagnetic material is positive because its Verdet constant is positive, whereas that in a paramagnetic material is negative because its Verdet

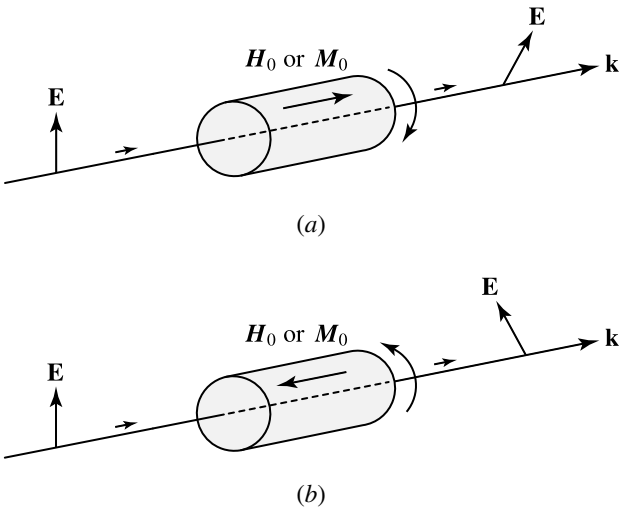


Figure 7.1 Positive Faraday rotation for an optical wave propagating in (*a*) a parallel direction and (*b*) an antiparallel direction with respect to $\boldsymbol{H}_0$, or $\boldsymbol{M}_0$. The sense of positive rotation is the same as the electric current that can be associated with $\boldsymbol{H}_0$, or $\boldsymbol{M}_0$. For negative Faraday rotation, the sense of rotation is just the opposite.

constant is negative. The Faraday rotation in ferromagnets and ferrimagnets can be either positive or negative.

EXAMPLE 7.1 A Faraday rotator consists of a TGG crystal in a magnetic field that has a flux density of $B_0 = 0.5$ T along the longitudinal axis of the crystal. If a Faraday rotation angle of 45° is desired for a linearly polarized optical beam at 800 nm wavelength traveling through the crystal, what should the crystal length be? In which sense is the polarization of the wave rotated? If the beam is reflected back at the output end of the crystal, what is the polarization direction of the reflected wave at the input end?

Solution From Table 7.1, $V = -65$ rad T^{-1} m^{-1} at 800 nm wavelength for TGG. Because the value of V is negative but that of B_0 is positive in this problem, the Faraday rotation angle is negative. The sense of rotation for a negative Faraday angle is clockwise when viewed facing against the direction of wave propagation. The desired Faraday rotation angle is $\theta_F = -45° = -\pi/4$ rad. Therefore, the required length of the crystal is

$$l = \frac{\theta_F}{V B_0} = \frac{-\pi/4}{-65 \times 0.5} \text{ m} = 0.024 \text{ m} = 24 \text{ mm}.$$

The total Faraday rotation angle of the reflected wave is double that of the single-pass rotation angle. Thus, the reflected wave returning to the input end is linearly polarized at 90° with respect to the polarization direction of the incident wave.

Besides attenuating the transmission of an optical wave, the optical absorption of a material has some very interesting consequences on the Faraday effect. In the presence

of absorption losses, both n_+ and n_- become complex. When the imaginary parts of n_+ and n_- do not have equal values, the two circularly polarized normal modes experience different degrees of attenuation. This phenomenon is called *circular dichroism*, as distinct from the linear dichroism between two linearly polarized modes. In the presence of circular dichroism, a linearly polarized wave undergoing a Faraday rotation does not remain linearly polarized but becomes elliptically polarized with a Faraday rotation angle, θ_F, and a *Faraday ellipticity*, ε_F, given by (see Problem 7.2.8)

$$\theta_F = \mathrm{Re}\left[\frac{\pi}{\lambda}(n_- - n_+)l\right] \approx \mathrm{Re}\left[\frac{\pi\xi}{\lambda n_\perp}l\right], \tag{7.29}$$

and

$$\varepsilon_F = \tan^{-1}\tanh \mathrm{Im}\left[\frac{\pi}{\lambda}(n_- - n_+)l\right] \approx \tan^{-1}\tanh \mathrm{Im}\left[\frac{\pi\xi}{\lambda n_\perp}l\right]. \tag{7.30}$$

The absorption that is directly related to the magneto-optic effect makes ξ a complex quantity with an imaginary part, ξ''. If the background material is relatively lossless so that $n'_\perp \gg |n''_\perp|$, the circular dichroism is solely contributed by ξ'' and is known as *magnetic circular dichroism*.

A material in which the normal modes of optical wave propagation are circularly polarized is referred to as being *optically active* or *optically gyroscopic*. Such a material exhibits circular birefringence. Certain nonmagnetic materials, such as quartz and sugar solutions, possess *natural optical activity* in the absence of a magnetic field or a magnetization. In analogy, the existence of a magnetically induced circular birefringence in an otherwise optically nonactive material is sometimes called *artificial*, or *induced*, *optical activity*. The similarities between these two phenomena are that both have circularly polarized normal modes and both can cause circular birefringence and circular dichroism. The plane of polarization of a linearly polarized wave can also be rotated while propagating through a naturally optically active medium in a way similar to Faraday rotation. The fundamental difference between them is that natural optical activity is reciprocal, as mentioned in Section 1.4, whereas magnetically induced optical activity is nonreciprocal. In the simplest case, natural optical activity can be described by an ϵ tensor in the form of (7.16) but with $\xi = \gamma\hat{k}\cdot\hat{z}$, where γ is a characteristic constant of the medium. Because of this dependence on wavevector, the values of k^+ and k^- associated with $\hat{e}_+$ and $\hat{e}_-$ are exchanged when the propagation direction is reversed. This characteristic, in contrast to the discussions immediately following (7.21), manifests the fundamental difference between natural circular birefringence and magnetic circular birefringence. As a result, when a linearly polarized optical wave traverses a naturally optically active medium twice along the same path but in opposite directions, the angle of rotation of its polarization gained in the forward pass is exactly canceled by that obtained in the backward pass, thus returning the wave back to its exact original polarization direction (see Problem 7.2.11). *Whereas magnetically induced optical activity exists in all materials, natural optical activity cannot*

exist in centrosymmetric materials. In an otherwise centrosymmetric medium, such as a liquid, the addition of molecules, such as sugar molecules, that cause optical activity breaks the centrosymmetry of the system.

7.3 Magneto-optic Kerr effect

Reflection of a polarized optical wave from the surface of a material with an internal magnetization or from that of one subject to an external magnetic field results in a change of the polarization state and/or the reflectivity that is dependent on the magnetization or the magnetic field. This phenomenon is known as the *magneto-optic Kerr effect*. It is totally unrelated to, and should not be confused with, the electro-optic Kerr effect discussed in Chapter 6. The only connection between the two is that both were discovered by J. C. Kerr. The magneto-optic Kerr effect stems from the same physical origin as the Faraday effect. Both are first-order magneto-optic effects. The distinction between the two is that the Faraday effect is associated with light in transmission whereas the Kerr effect is associated with light in reflection. The Kerr effect can be observed from the surface of a material that has no internal magnetization but is subject to an external magnetic field or from the surface of a ferromagnetic or ferrimagnetic material. The latter is more important than the former in practical applications, and we shall take it to be the case in the following discussions. The general concepts and the results obtained can be applied to the former case in a similar fashion.

There are three configurations, shown in Fig. 7.2, of the magneto-optic Kerr effect. In the *polar Kerr effect*, the magnetization $\boldsymbol{M}_0$ is normal to the surface of the magnetic medium from which the optical wave is reflected. In the *longitudinal Kerr effect*, also known as the *meridional Kerr effect*, $\boldsymbol{M}_0$ is parallel to the surface of the medium and lies in the plane of incidence. In the *transverse Kerr effect*, also known as the *equatorial Kerr effect*, $\boldsymbol{M}_0$ is normal to the plane of incidence.

Consider the simple and most common situation in which the incident optical wave traveling in an isotropic, nonmagnetic medium is reflected from the surface of an isotropic magnetic medium. The TE and TM polarizations are both normal modes for

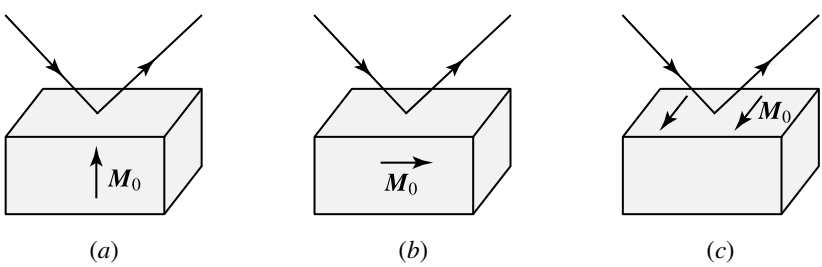


Figure 7.2 Three configurations of magneto-optic Kerr effect: (*a*) polar Kerr effect, (*b*) longitudinal Kerr effect, and (*c*) transverse Kerr effect.

the incident and the reflected optical waves traveling in the isotropic, nonmagnetic medium. However, they are not normal modes for the wave traveling in the magnetic medium because of circular birefringence of the medium. Consequently, upon reflection from the magnetic surface, the TE-polarized component of an incident wave can be coupled to the TM-polarized component, and vice versa. All of the three Kerr effects can then be generally expressed in terms of a tensor of reflection coefficients:

$$\begin{bmatrix} E_s^r \\ E_p^r \end{bmatrix} = \begin{bmatrix} r_{ss} & r_{sp} \\ r_{ps} & r_{pp} \end{bmatrix} \begin{bmatrix} E_s^i \\ E_p^i \end{bmatrix}, \tag{7.31}$$

where the subscripts s and p represent TE and TM polarizations, for s and p waves, respectively, and the superscripts i and r indicate incident and reflected fields, respectively.

From the discussions in the preceding section, we know that only the electric field components that are perpendicular to the magnetization interact with the magnetization and see the circular birefringence. For polar and longitudinal Kerr effects, both TE and TM polarizations interact with $\boldsymbol{M}_0$ because both have nonvanishing components perpendicular to $\boldsymbol{M}_0$. This interaction results in nonvanishing r_{sp} and r_{ps} due to coupling between the TE- and TM-polarized fields through circular birefringence of the magnetic medium. The values of r_{sp} and r_{ps} are proportional to ξ. Their phases are different from those of r_{ss} and r_{pp}. Consequently, a linearly polarized incident wave is turned into an elliptically polarized wave with the direction of its major axis rotated away from the initial plane of polarization. Because r_{sp} and r_{ps} are linearly dependent on $\boldsymbol{M}_0$, they change sign when the direction of $\boldsymbol{M}_0$ is reversed. Consequently, *both the angle of rotation and the ellipticity of the elliptically polarized reflected wave change sign upon reversing the direction of* $\boldsymbol{M}_0$. For the transverse Kerr effect, only TM polarization interacts with $\boldsymbol{M}_0$ because the electric field of the TE polarization is parallel to $\boldsymbol{M}_0$. Then, $r_{sp} = r_{ps} = 0$, and r_{ss} is independent of $\boldsymbol{M}_0$. The net effect is only a magnetically induced change in r_{pp}. This change is linearly proportional to ξ. It also changes sign when the direction of $\boldsymbol{M}_0$ is reversed.

The Kerr effect in each configuration varies with the angle of incidence. In a majority of cases, particularly in those with small incident angles, the polar configuration has a larger effect than both longitudinal and transverse configurations. The Kerr effect in the polar configuration increases as the angle of incidence decreases. At normal incidence, the Kerr effect reaches its maximum in the polar configuration, but it vanishes in both longitudinal and transverse configurations. Therefore, among the three different configurations in combination with all possible angles of incidence, *the polar configuration at normal incidence has the most pronounced Kerr effect* and is most useful for practical applications such as magneto-optic recording. It is also the simplest to analyze. In the following, we consider this specific case.

We consider the polar Kerr effect at normal incidence from free space on the surface of a ferromagnet or ferrimagnet. The magnetic material is assumed to be isotropic or

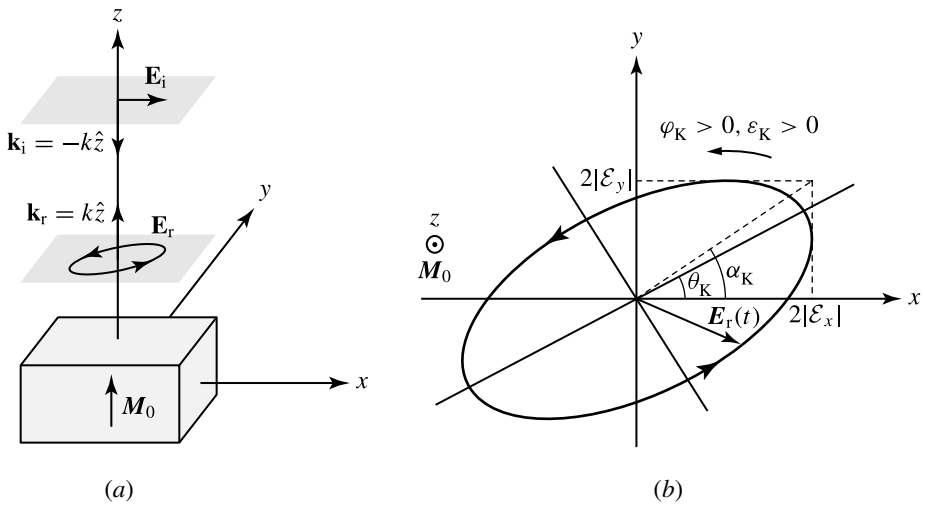


Figure 7.3 Polar Kerr effect at normal incidence: (*a*) configuration showing the designation of the coordinates and the direction of the magnetization and (*b*) sense of Kerr rotation and that of Kerr ellipticity with respect to the direction of magnetization for positive values of θ_K and ε_K.

uniaxial with its magnetization parallel to the optical axis along the z direction. The incident wave propagates in the negative z direction while the reflected wave propagates in the positive z direction, as is shown in Fig. 7.3(*a*). Without loss of generality in the case under consideration, a linearly polarized incident beam can be assumed to be polarized in the x direction:

$$\mathbf{E}_i(z,t) = \hat{x}\mathcal{E}e^{-ikz-i\omega t} = \frac{\mathcal{E}}{\sqrt{2}}(\hat{e}_+ + \hat{e}_-)e^{-ikz-i\omega t}. \tag{7.32}$$

In free space, $\hat{e}_+$ and $\hat{e}_-$ are also normal modes of propagation like $\hat{x}$, all of which have the same propagation constant k. In addition, $\mathcal{E}_+ = \mathcal{E}_- = \mathcal{E}/\sqrt{2}$. In the magnetic material, $\hat{x}$ is not a normal mode any more, and the normal modes $\hat{e}_+$ and $\hat{e}_-$ have different propagation constants. As discussed in the preceding section, when the direction of $\boldsymbol{M}_0$ is fixed, the values of the propagation constants k^+ and k^-, as well as those of the principal indices of refraction n_+ and n_-, for the normal modes $\hat{e}_+$ and $\hat{e}_-$, respectively, are independent of whether the wave propagates in a parallel or an antiparallel direction with respect to the direction of $\boldsymbol{M}_0$. Only the handedness of $\hat{e}_+$ and $\hat{e}_-$ is interchanged when the propagation direction is reversed. As independent eigenmodes with characteristic indices of refraction, the circularly polarized modes do not couple to each other on reflection. They have the following clearly defined reflection coefficients (see Problem 7.3.2):

$$r_+ = \frac{1-n_+}{1+n_+}, \qquad r_- = \frac{1-n_-}{1+n_-}, \tag{7.33}$$

for the $\hat{e}_+$ and $\hat{e}_-$ modes, respectively. According to the configuration defined in Fig. 7.3,

the reflected wave propagates in the positive z direction. The wave reflected from the magnetic surface back into free space, where $\hat{e}_+$ and $\hat{e}_-$ modes have the same propagation constant k, is then described by

$$\begin{aligned}\mathbf{E}_{\mathrm{r}}(z,t) &= \hat{e}_+ r_+ \mathcal{E}_+ \mathrm{e}^{\mathrm{i}kz-\mathrm{i}\omega t} + \hat{e}_- r_- \mathcal{E}_- \mathrm{e}^{\mathrm{i}kz-\mathrm{i}\omega t} \\ &= \frac{\mathcal{E}}{\sqrt{2}}(\hat{e}_+ r_+ + \hat{e}_- r_-)\mathrm{e}^{\mathrm{i}kz-\mathrm{i}\omega t} \\ &= \frac{\mathcal{E}}{2}\left[\hat{x}(r_+ + r_-) + \mathrm{i}\hat{y}(r_+ - r_-)\right]\mathrm{e}^{\mathrm{i}kz-\mathrm{i}\omega t}. \end{aligned} \tag{7.34}$$

The x and y components of this reflected field are $\mathcal{E}_x^{\mathrm{r}} = \mathcal{E}(r_+ + r_-)/2$ and $\mathcal{E}_y^{\mathrm{r}} = \mathrm{i}\mathcal{E}(r_+ - r_-)/2$, respectively. Clearly, the reflected wave is elliptically polarized.

As described in Section 1.4, the characteristics of an elliptically polarized field can be completely specified by either of the two sets of parameters: (α, φ) and (θ, ε). Using (1.62), the parameters α_{K} and φ_{K} that characterize the elliptically polarized field in (7.34) are given by

$$\tan\alpha_{\mathrm{K}}\, \mathrm{e}^{\mathrm{i}\varphi_{\mathrm{K}}} = \frac{\mathcal{E}_y^{\mathrm{r}}}{\mathcal{E}_x^{\mathrm{r}}} = \mathrm{i}\frac{r_+ - r_-}{r_+ + r_-}, \tag{7.35}$$

where the subscript K for the parameters indicates their association with the Kerr effect. Because it is generally true in practical magnetic materials that $|r_+ - r_-| \ll |r_+ + r_-|$, the value of α_{K} is very small. So are the values of the corresponding θ_{K} and ε_{K} parameters. Using the relations in (1.66) and (1.67) in the limit of small values for these parameters, we have

$$\tan\alpha_{\mathrm{K}}\, \mathrm{e}^{\mathrm{i}\varphi_{\mathrm{K}}} \approx \alpha_{\mathrm{K}} \mathrm{e}^{\mathrm{i}\varphi_{\mathrm{K}}} \approx \theta_{\mathrm{K}} + \mathrm{i}\varepsilon_{\mathrm{K}}. \tag{7.36}$$

Combining (7.35) and (7.36), we find

$$\theta_{\mathrm{K}} + \mathrm{i}\varepsilon_{\mathrm{K}} = \mathrm{i}\frac{r_+ - r_-}{r_+ + r_-} = \mathrm{i}\frac{n_+ - n_-}{n_+ n_- - 1} \approx -\mathrm{i}\frac{\xi}{n_\perp(n_\perp^2 - 1)}. \tag{7.37}$$

Therefore, the *Kerr rotation angle*, θ_{K}, and the *Kerr ellipticity*, ε_{K}, are given by

$$\theta_{\mathrm{K}} = \mathrm{Im}\left[\frac{\xi}{n_\perp(n_\perp^2 - 1)}\right], \qquad \varepsilon_{\mathrm{K}} = -\mathrm{Re}\left[\frac{\xi}{n_\perp(n_\perp^2 - 1)}\right], \tag{7.38}$$

respectively. In case optical loss of nonmagnetic origin is small in the magnetic medium so that $|n_\perp''| \ll n_\perp'$, θ_{K} is caused solely by magnetic circular dichroism and is completely determined by ξ'' while ε_{K} is caused by magnetic circular birefringence and is determined by ξ'. Note the distinction between the Faraday effect and the Kerr effect. *Magnetic circular birefringence is the cause of Faraday rotation and Kerr ellipticity, whereas magnetic circular dichroism is the mechanism behind Kerr rotation and Faraday ellipticity* (see Problem 7.2.8).

As can be seen from (7.38), both θ_{K} and ε_{K} are linearly dependent on magnetization. Both depend strongly on the optical wavelength. They are intrinsic properties of a

magnetic material that are generally determined by direct measurement. Their values are small, normally on the order of $\pm 1^\circ$ or less for most materials. Their signs are defined in a manner similar to how the sign of Faraday rotation is defined: θ_K is positive if the major axis of the ellipse describing the polarization of the reflected wave is rotated away from the direction of linear polarization of the incident wave in a sense that can be described by an axial vector pointing in the direction of $\boldsymbol{M}_0$, and ε_K is positive if the elliptically polarized field at a fixed point in space rotates in the same sense as the electric current that can be associated with $\boldsymbol{M}_0$. This convention is illustrated in Fig. 7.3(*b*).

EXAMPLE 7.2 At 546 nm wavelength, iron at room temperature has a complex index of refraction of $n = 2.73 + \mathrm{i}3.3$ and, when fully magnetized, a complex linear magneto-optic constant of $\xi = -0.18 + \mathrm{i}0.74$. A linearly polarized optical wave at 546 nm is normally incident on a fully magnetized iron surface in polar configuration, as shown in Fig. 7.3. Find the Kerr rotation angle and the Kerr ellipticity of the reflected wave.

Solution Neglecting second-order magneto-optic effects, we can take $n_\perp = n = 2.73 + \mathrm{i}3.3$. Then,

$$\frac{\xi}{n_\perp(n_\perp^2 - 1)} = \frac{-0.18 + \mathrm{i}0.74}{(2.73 + \mathrm{i}3.3)[(2.73 + \mathrm{i}3.3)^2 - 1]} = 0.0061 - \mathrm{i}0.0074.$$

Using (7.38), we find that $\theta_K = -7.4$ mrad $\approx -0.42^\circ$ and $\varepsilon_K = -6.1$ mrad $\approx -0.35^\circ$. Because $|\theta_K| \ll 1$ and $|\varepsilon_K| \ll 1$, the approximations made in (7.36) and (7.37) to obtain (7.38) are valid.

7.4 Optical isolators and circulators

In an optical system, reflections and backscattering of light often cause serious problems ranging from noise in the photodetectors to instabilities in the light sources. A feedback, even at an extremely low level, to a laser usually has a significant effect on the laser characteristics. It can change the laser frequency, increase the laser noise, create fluctuations in the laser intensity, lock the laser to a different mode of operation, or drive the laser into instability, even chaos. A feedback to a photodetector or other parts of an optical system also has many undesirable effects. Sometimes the problem is so severe that it renders the entire system useless.

Optical isolators are needed to avoid such problems. An *optical isolator* is a nonreciprocal device that transmits an optical wave in one direction but blocks it in the reverse direction much as the function of a diode in an electric circuit. The key specifications of an optical isolator are the *insertion loss*, the *return loss*, and the *reverse isolation* of the device. The insertion loss is the attenuation of an optical signal propagating in

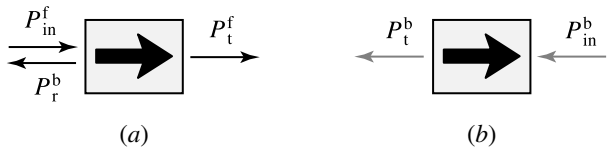


Figure 7.4 Diagrammatic illustration of an optical isolator with (*a*) a forward-propagating input signal, P_{in}^{f}, and (*b*) a backward-propagating input signal, P_{in}^{b}, together with the transmitted and reflected components for defining the key specifications.

the forward direction through the optical isolator. The return loss specifies how well the reflection of the forward-propagating signal is eliminated. Referring to Fig. 7.4(*a*), they are defined as

$$\text{Insertion loss} = -10 \log \frac{P_{t}^{f}}{P_{in}^{f}}, \tag{7.39}$$

$$\text{Return loss} = -10 \log \frac{P_{r}^{b}}{P_{in}^{f}}. \tag{7.40}$$

The reverse isolation is a measure of the isolation function of the device and is defined as the attenuation of a backward-propagating optical signal through the isolator. Referring to Fig. 7.4(*b*), it is given by

$$\text{Reverse isolation} = -10 \log \frac{P_{t}^{b}}{P_{in}^{b}}. \tag{7.41}$$

For a good isolator, it is desired that the value of the insertion loss be as low as possible while those of the return loss and the reverse isolation be as high as possible. Furthermore, the return loss has to be higher than the reverse isolation for a device to be functionally useful.

In some systems, such as fiber-optic transmission systems, bidirectional transmission with isolation from backscattering and reflections is necessary. This function can be accomplished by an *optical circulator*, which loops an optical signal through successive ports while blocking backscattered and reflected light. The diagrams in Fig. 7.5 illustrate the function of a four-port optical circulator. A *true optical circulator* connects all ports in an endless loop, as shown in Fig. 7.5(*b*). A *quasi-optical circulator* loops an optical signal through successive ports but is not able to transmit it from the last port to the first port. As an example of the application of optical circulators, Fig. 7.6 shows bidirectional transmission in a single-fiber transmission line using one circulator on each end. Both true circulators and quasi-circulators are acceptable for this particular application.

The key components of optical isolators and circulators are the Faraday rotators because only the Faraday effect has the nonreciprocity in transmission required by these devices. The most commonly used materials for Faraday rotators in these applications are the YIG crystal and bismuth-substituted rare-earth iron garnet films, such as

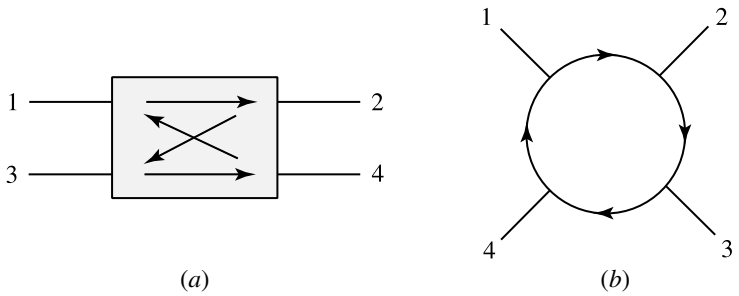


Figure 7.5 Diagrammatic illustration of (*a*) a four-port optical circulator and (*b*) its looping function.

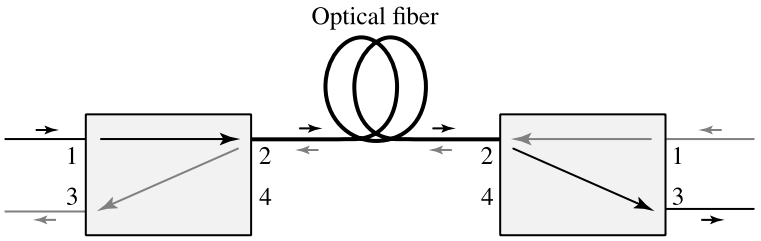


Figure 7.6 Schematic illustration of bidirectional transmission in a single fiber transmission line using two circulators.

$Bi_xY_{3-x}Fe_5O_{12}$ (Bi : YIG), $Gd_{3-x}Bi_xFe_5O_{12}$, $(YbTbBi)_3Fe_5O_{12}$, and many other different compositions. These ferrimagnetic films have very large values of specific Faraday rotation and are used for making very compact isolators and circulators. Paramagnetic garnets, such as $Tb_3Ga_5O_{12}$ (TGG) and $Tb_3Al_5O_{12}$, and paramagnetic glasses, such as terbium-doped glasses, are sometimes also used. Whether the material used is ferrimagnetic or paramagnetic, it is normally placed in the magnetic field of a permanent magnet. In the case of a ferrimagnetic garnet, the magnetic field keeps the garnet magnetized, preferably at saturation magnetization for maximum efficiency.

Polarization-dependent isolators

The basic structure of an optical isolator consists of a Faraday rotator of total Faraday rotation angle $\theta_F = 45^\circ$ and two linear polarizers, as shown in Fig. 7.7(*a*). The axis of the input polarizer can be arbitrarily oriented, but the axis of the output polarizer has to be rotated by $\theta_p = 45^\circ$ with respect to that of the input one in the same direction as the polarization rotation caused by the Faraday rotator. An optical wave entering the device in the forward direction through the input polarizer becomes linearly polarized by this polarizer. The Faraday rotator then rotates its plane of polarization by 45° into a direction parallel to the axis of the output polarizer. Therefore, the linearly polarized wave emerging from the Faraday rotator is transmitted by the output polarizer without

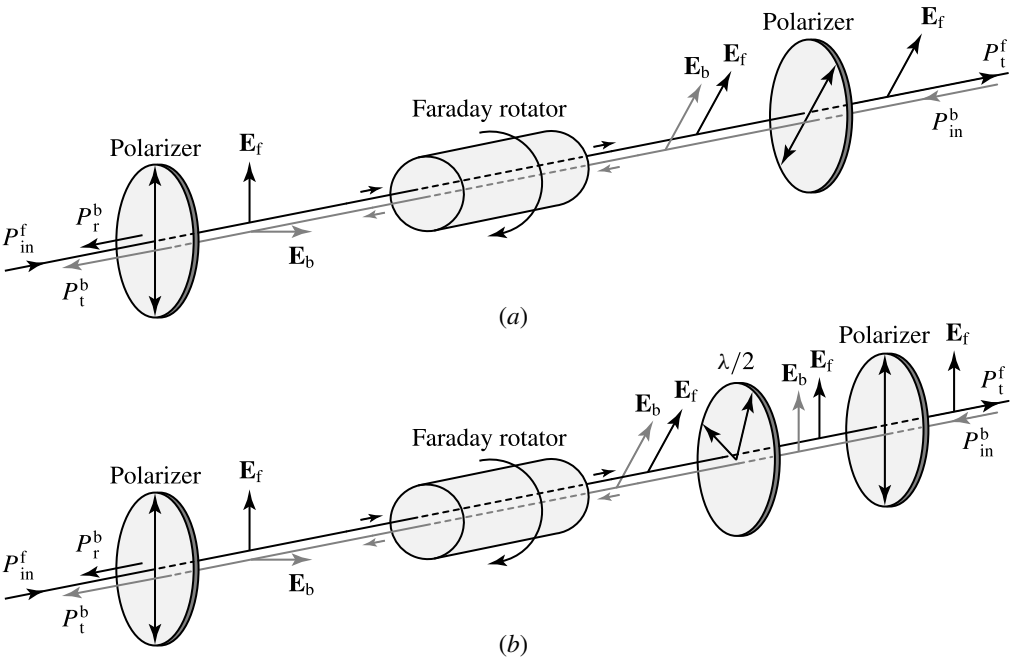


Figure 7.7 Basic structure and principle of polarization-dependent optical isolators. (*a*) The basic structure of an optical isolator changes the polarization direction at the output. (*b*) The addition of a properly oriented half-wave plate in the isolator restores the polarization back to its original direction. $\lambda/2$ labels a half-wave plate.

attenuation. For reverse isolation, an optical wave of any polarization entering from the output end is polarized by the output polarizer. Because Faraday rotation is independent of wave propagation direction, the backward-propagating wave emerging from the Faraday rotator has a linear polarization orthogonal to the axis of the input polarizer and is thus blocked. To minimize the insertion loss in the application of this isolator, the input optical wave clearly has to be linearly polarized in a direction parallel to the axis of the input polarizer. After passing through the isolator, its plane of polarization is rotated by 45° though it is still linearly polarized. Therefore, the basic isolator shown in Fig. 7.7(*a*) changes the polarization direction of the optical wave. This problem can be eliminated by adding a properly oriented half-wave plate and reorienting the output polarizer, as shown in Fig. 7.7(*b*). Because the half-wave plate is a reciprocal element, it does not interfere with the function of the isolator while restoring the polarization direction. For both devices shown in Figs. 7.7(*a*) and (*b*), the function of the reverse isolation is independent of the polarization of the backward-propagating wave.

To obtain a high return loss, the surfaces of every component along the optical path in an isolator have to be antireflection coated. To eliminate any further residual reflection, the facets of the components are sometimes cut or tilted at a small angle, typically somewhere between 1° and 10°. The antireflection coatings also serve the purpose of

reducing the insertion loss. The limiting factor for insertion loss is the absorption of the optical components, particularly that of the Faraday rotator. That part of the absorption that contributes to circular dichroism also reduces the reverse isolation by making a linearly polarized input wave elliptically polarized after its passage through the rotator. Therefore, a Faraday rotator that has both a large specific rotation and a low absorption coefficient is most desirable. In order to have a low insertion loss and a high reverse isolation, it is also very important to choose polarizers of very high extinction ratios. The extinction ratio of a polarizer is an intrinsic property of the polarizer that is defined as

$$\mathrm{ER} = -10 \log \frac{T_\perp}{T_\parallel} = -10 \log \sigma, \tag{7.42}$$

where $T_\parallel$ is the transmittance of an optical wave that is linearly polarized in a direction parallel to the axis of the polarizer, $T_\perp$ is that of its orthogonal polarization, and $\sigma = T_\perp / T_\parallel$ is the extinction ratio in the linear scale. Other major factors that affect both the insertion loss and the reverse isolation are the Faraday rotation angle θ_F and the Faraday ellipticity ε_F generated by the Faraday rotator, as well as the angle θ_p between the axis of the output polarizer and that of the input polarizer. For an isolator of the basic structure shown in Fig. 7.7(*a*), where the input and output polarizers have extinction ratios σ_in and σ_out, respectively, the insertion loss and the reverse isolation can be expressed, respectively, as (see Problem 7.4.1)

$$\text{Insertion loss} = L_0 - 10 \log \left[\cos^2(\theta_\mathrm{F} - \theta_\mathrm{p}) + \sigma_\mathrm{out} \sin^2(\theta_\mathrm{F} - \theta_\mathrm{p}) + (1 + \sigma_\mathrm{out}) \varepsilon_\mathrm{F}^2 \right], \tag{7.43}$$

$$\text{Reverse isolation} = L_0 - 10 \log \left[\cos^2(\theta_\mathrm{F} + \theta_\mathrm{p}) + \sigma_\mathrm{in} \sin^2(\theta_\mathrm{F} + \theta_\mathrm{p}) + (1 + \sigma_\mathrm{in}) \varepsilon_\mathrm{F}^2 \right], \tag{7.44}$$

where L_0 is the background optical loss including absorption losses of the Faraday rotator and the polarizers and residual reflection losses due to imperfect antireflection coating of the optical surfaces in the system. It can be seen from (7.43) and (7.44) that the insertion loss is minimized while the reverse isolation is maximized by choosing $\theta_\mathrm{F} = \theta_\mathrm{p} = 45^\circ$. The effect on the reverse isolation caused by a deviation of θ_F from 45° can be removed by adjusting θ_p for a counter deviation of the same amount, albeit at the expense of increasing the insertion loss. Similar relations, with proper modifications according to a particular structure, can be written for the isolator shown in Fig. 7.7(*b*), as well as for those of other structures.

EXAMPLE 7.3 An optical isolator of the structure shown in Fig. 7.7(*a*) has a background optical loss of 0.5 dB. Both the input and output polarizers have the same extinction ratio of 40 dB. The Faraday ellipticity generated by the Faraday rotator is negligibly small. If the Faraday rotation angle has the ideal value of $\theta_\mathrm{F} = 45^\circ$, what are the minimum

insertion loss and the maximum reverse isolation? If the Faraday rotation angle is off by only 1° from this ideal value, what is the reverse isolation when the insertion loss is minimized? What is the insertion loss when the reverse isolation is maximized? Is it better to minimize the insertion loss or to maximize the reverse isolation in this situation?

Solution From (7.42), we find that $\sigma = 10^{-4}$ for both input and output polarizers of a 40-dB extinction ratio. When $\theta_F = 45°$, the minimum insertion loss and the maximum reverse isolation are simultaneously achieved by setting $\theta_p = 45°$. Because $L_0 = 0.5$ dB and $\varepsilon_F \approx 0$, we find from (7.43) and (7.44) that

$$\text{Minimum insertion loss} = L_0 = 0.5 \text{ dB},$$

$$\text{Maximum reverse isolation} = L_0 - 10 \log \sigma = 40.5 \text{ dB}.$$

When the Faraday rotation angle is off by 1°, we have $\theta_F = 45 \pm 1°$. To minimize the insertion loss, we choose $\theta_p = \theta_F = 45 \pm 1°$ so that $\theta_F - \theta_p = 0$, but then $\theta_F + \theta_p = 90 \pm 2°$. With this choice of θ_p, the insertion loss remains at its minimum value of 0.5 dB according to (7.43), but the reverse isolation becomes

$$L_0 - 10 \log \left[\cos^2(90° \pm 2°) + \sigma \sin^2(90° \pm 2°)\right] \approx 29.3 \text{ dB}.$$

To maximize the reverse isolation, we have to choose $\theta_p = 45 \mp 1°$ so that $\theta_F + \theta_p = 90°$, but then $\theta_F - \theta_p = \pm 2°$. With this choice of θ_p, the reverse isolation remains at its maximum value of 40.5 dB according to (7.44), but the insertion loss becomes

$$L_0 - 10 \log \left[\cos^2(\pm 2°) + \sigma \sin^2(\pm 2°)\right] = 0.505 \text{ dB}.$$

We find that there is a significant reduction of 11.2 dB in the reverse isolation when the insertion loss is minimized, but there is only a negligibly small increase of 0.005 dB in the insertion loss when the reverse isolation is maximized. One clearly should choose to maximize the reverse isolation by maintaining $\theta_F + \theta_p = 90°$ when θ_F deviates from its ideal value of 45°.

For most applications, a reverse isolation of 30 dB or higher is usually required. Isolators with a reverse isolation of 30–50 dB, an insertion loss of 1–2 dB or less, and a return loss higher than 60 dB are commercially available. Some applications require a reverse isolation of 60 dB or more. In this case, two separate isolators in tandem or a *two-stage cascaded optical isolator*, such as the one shown in Fig. 7.8, can be used to increase the reverse isolation, but at the expense of increasing the insertion loss.

Though high-quality polarizers and coatings of relatively large bandwidths are common, an optical isolator generally has a narrow bandwidth because the rotation angle of a given Faraday rotator depends strongly on the optical wavelength. However, wavelength tunability is possible. It can be done by moving the rotator rod into or out of the surrounding magnet for more or less exposure to the magnetic field, thus maintaining

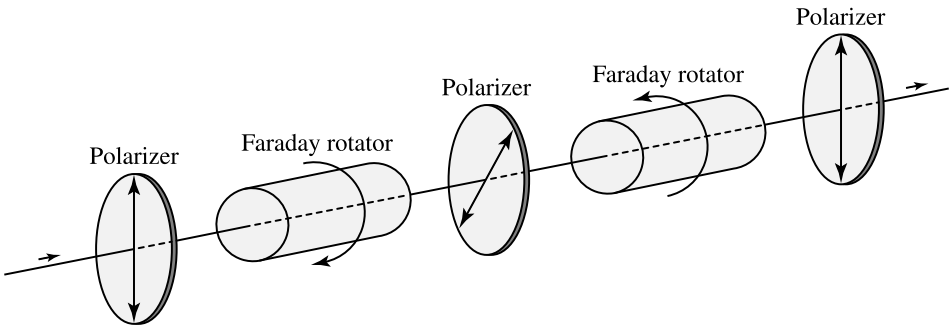


Figure 7.8 Two-stage cascaded optical isolator.

the Faraday rotation angle at 45° for varying wavelengths. In the case of the two-stage cascaded isolator shown in Fig. 7.8, the amount of the Faraday rotation can alternatively be controlled by varying the separation between the two rotators to adjust the influence of the stray magnetic field from each stage on the rotator rod of the other. The performance of an isolator also varies with temperature because of the temperature dependence of the Faraday rotator. To some extent, this temperature dependence can be compensated by applying similar techniques.

Polarization-independent isolators

Although the isolators discussed above extinguish light of any polarization in a reverse direction, their transmission in the forward direction is polarization dependent because of the input polarizer. In some applications, the sensitivity of an isolator to the polarization of the input optical signal is a major drawback. For instance, the polarization state of an optical wave transmitted through a non-polarization-preserving optical fiber not only is uncertain but also changes with environmental conditions. It is therefore highly desirable that polarization-independent isolators be used in fiber-optic transmission systems, as well as in the applications of in-line isolation for fiber amplifiers. In some other instances, light sources are capable of emitting in different polarization states, sometimes for very useful applications such as in the cases of polarization-switching and polarization-bistable lasers. Clearly, for optical isolation in systems containing such sources, polarization-independent isolators are absolutely necessary.

The function of an isolator inherently relies on the manipulation of the polarization state of an optical wave using a Faraday rotator. Therefore, the way to construct a polarization-independent isolator is not to avoid manipulating the polarization. Because an optical wave of any polarization state can be decomposed into two orthogonal linearly polarized components, the basic idea behind any polarization-independent isolator is to separate these two components at the input, manipulate them separately through the nonreciprocal Faraday rotator, and then combine them at the output. This concept can be implemented with a variety of designs. Figure 7.9 shows one example. This device

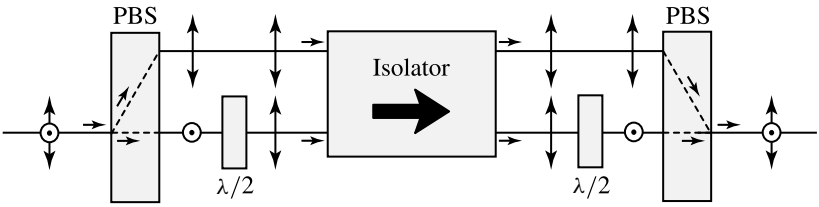


Figure 7.9 Polarization-independent optical isolator and its principle of operation. The polarization-dependent isolator used in this design maintains input polarization direction at the output. PBS indicates a polarizing beam splitter. $\lambda/2$ labels a half-wave plate.

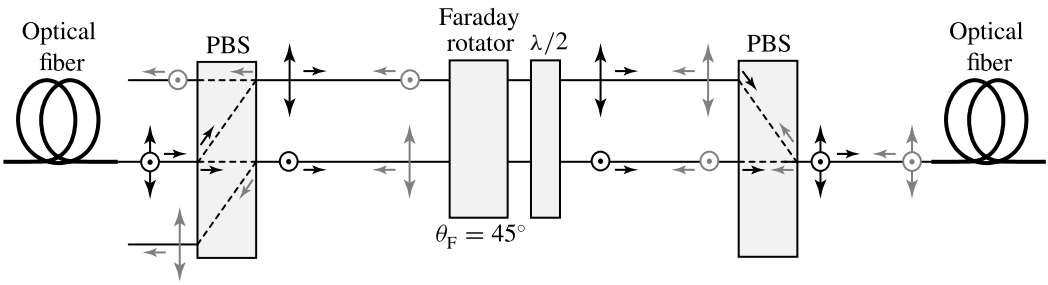


Figure 7.10 Polarization-independent optical isolator used in a fiber transmission line and its principle of operation. The coupling between the fiber tip and the isolator can be carried out in many different ways and is not specified here. PBS indicates a polarizing beam splitter. $\lambda/2$ labels a half-wave plate.

consists of two birefringent plates functioning as *polarizing beam splitters* based on the phenomenon of spatial beam walk-off discussed in Section 1.6, two half-wave plates for orienting the polarization in proper directions, and a polarization-dependent isolator that maintains the input polarization direction at the output, such as the one shown in Fig. 7.7(*b*) or that shown in Fig. 7.8. In the forward direction, an input wave of any polarization state is split by the input birefringent plate into two orthogonal linearly polarized components. The first half-wave plate rotates the polarization of the lower beam into the same direction as the upper beam, which is the proper polarization direction for transmission through the polarization-dependent isolator. At the output, the second half-wave plate again rotates the polarization of the lower beam by 90°, turning it back to its original input polarization direction. The output birefringent plate then combines the two beams into one of the same polarization state as that of the input wave. The isolation function of this device in the reverse direction is self-evident from the function of the polarization-dependent isolator inside the device.

For a polarization-independent isolator used in a fiber transmission line, the isolation function can be accomplished by sufficiently displacing the backward-propagating optical wave, instead of extinguishing it, so that it does not couple into the input fiber core in the reverse direction. Using this principle, the structure of a polarization-independent isolator can be substantially simplified. Figure 7.10 shows one design of

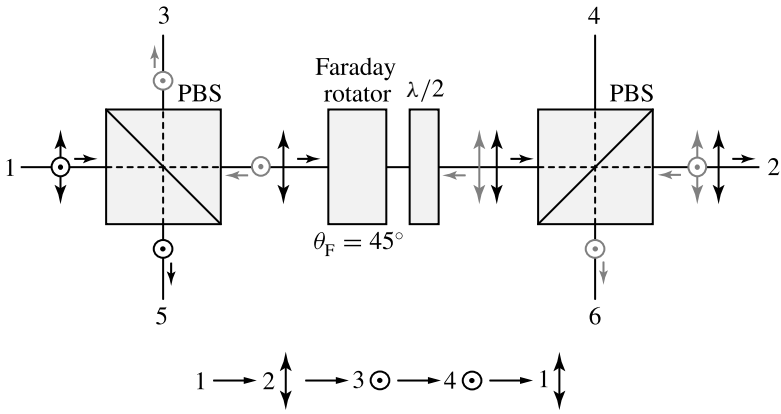


Figure 7.11 Polarization-dependent circulator. The circulator loops in the sequence $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$ among the four nonreciprocal ports. Ports 5 and 6 are reciprocal ports and are not part of the circulator loop. PBS indicates a polarizing beam splitter. $\lambda/2$ labels a half-wave plate.

such a device, which consists of two birefringent plates, one 45° Faraday rotator, and one half-wave plate. Also illustrated in Fig. 7.10 is the principle of operation of this device.

Polarization-dependent circulators

Figure 7.11 shows the structure of a polarization-dependent optical circulator. It consists of two polarizing beam splitters, one 45° Faraday rotator, and one half-wave plate. A polarizing beam splitter cube splits s- and p-polarized waves by transmitting and reflecting them, respectively, at the interface of the prisms that constitute the cube. The orientation of the half-wave plate is such that the net polarization rotation angle through the combination of the Faraday rotator and the half-wave plate is zero for a linearly polarized forward-propagating wave, but is 90° for a backward-propagating wave. Therefore, an s-polarized wave entering port 1 exits port 2 s polarized; an s-polarized wave entering port 2 exits port 3 p polarized; a p-polarized wave entering port 3 exits port 4 p polarized; finally, a p-polarized wave entering port 4 exits port 1 s polarized. It can be seen that these four ports are nonreciprocal because wave propagation in the reverse sequence is forbidden. For each of these four nonreciprocal ports, the input and output polarizations are the same. Ports 5 and 6 in this particular device are reciprocal ports and are not part of the circulator. If an optical wave of the wrong polarization direction enters a particular nonreciprocal port, it cannot enter the loop of the circulator but is lost through one of the reciprocal ports. For example, if a p-polarized wave enters port 1, it is lost through port 5. Consequently, the device is clearly a four-port polarization-dependent circulator. Other designs based on similar concepts are also possible.

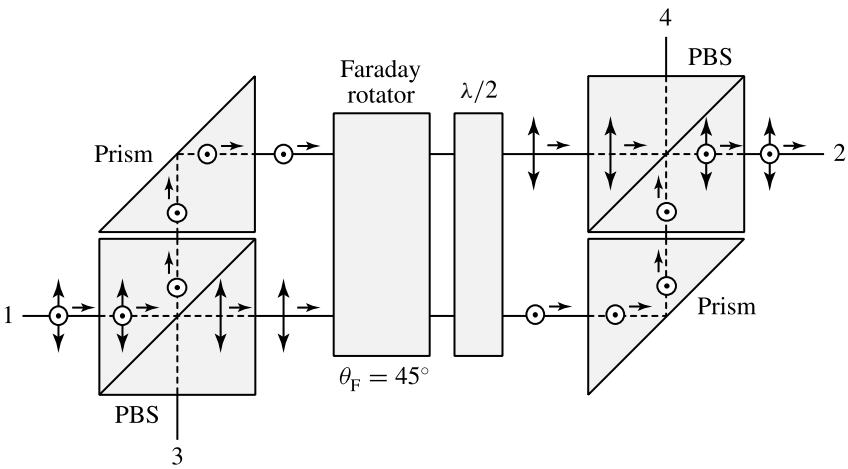


Figure 7.12 Four-port polarization-independent optical circulator. The looping sequence is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$. PBS indicates a polarizing beam splitter. $\lambda/2$ labels a half-wave plate.

Polarization-independent circulators

For the same reasons as those discussed above for the need of polarization-independent isolators, there is also a need for polarization-independent optical circulators in many applications. The basic idea for constructing a polarization-independent circulator is also the same as that discussed above for constructing a polarization-independent isolator. Following that concept, a polarization-independent circulator can be implemented by making slight modifications on the structure of the polarization-dependent circulator shown in Fig. 7.11. The resulting device is shown in Fig. 7.12. In this device, an optical wave of any polarization state entering any of the four nonreciprocal ports is split into two orthogonally polarized beams. Both are transmitted through the combination of a 45° Faraday rotator and a properly oriented half-wave plate. The two beams are combined afterwards at the succeeding port. The polarization state of the output beam is the same as that of the input beam. It can be easily verified that the device functions as a four-port circulator and that its function is independent of the polarization of the optical wave propagating through it.

Other designs are also possible for polarization-independent circulators. In particular, the polarizing beam splitter cubes can be replaced by birefringent plates to make very compact devices, but at the expense of increasing the complexity of the device.

7.5 Magneto-optic modulators and sensors

Polarization and amplitude modulators that are based on the Faraday effect and are driven by currents or magnetic fields can be easily realized. In comparison to the electro-optic polarization and amplitude modulators discussed in Chapter 6, these devices

have similar functions but quite different characteristics. The mechanism responsible for magneto-optic polarization modulators is circular birefringence, whereas that for electro-optic polarization modulators is linear birefringence. If the input optical wave is linearly polarized, the output of an ideal magneto-optic polarization modulator is linearly polarized, but that of an electro-optic polarization modulator is elliptically polarized in general and is linearly polarized only when the applied voltage is equal to an integral multiple of the half-wave voltage. The basic structure of both magneto-optic and electro-optic amplitude modulators consists of a polarization modulator and a polarizer–analyzer pair.

The basic configuration of a magneto-optic amplitude modulator is simply that of the polarization-dependent optical isolator shown in Fig. 7.7(a), except that θ_F for a modulator can have any value and the polarizer at the output is now referred to as the analyzer. Usually, there is absorption loss in the Faraday rotator, as well as in the polarizer and the analyzer. If the Faraday rotator has negligible magnetic circular dichroism, the transmitted optical wave remains linearly polarized, though attenuated. The intensity transmittance of the modulator is then given by

$$T = \frac{I_{\text{out}}}{I_{\text{in}}} = T_0 \mathrm{e}^{-\alpha l} \cos^2(\theta_F - \theta_p) = \frac{T_0}{2} \mathrm{e}^{-\alpha l}[1 + \cos 2(\theta_F - \theta_p)], \tag{7.45}$$

where α and l are the absorption coefficient and the length, respectively, of the Faraday rotator and T_0 accounts for losses in the polarizer, the analyzer, and other components such as the nonmagnetic substrate supporting a magnetic film. If the input optical wave is linearly polarized, $0 < T_0 \leq 1$. If it is unpolarized, $0 < T_0 \leq 1/2$.

If the absolute value of θ_F is small, θ_p is chosen to be 45° for a linear response. Then T varies linearly with θ_F:

$$T = \frac{T_0}{2} \mathrm{e}^{-\alpha l}(1 + \sin 2\theta_F) \approx T_0 \mathrm{e}^{-\alpha l} \left(\frac{1}{2} + \theta_F \right). \tag{7.46}$$

In this case, the transmittance T has the highest sensitivity in response to variations in the value of θ_F around the point $\theta_F = 0$. Because θ_F of a paramagnetic or diamagnetic Faraday rotator is linearly proportional to the magnetic field, and thus is also linearly proportional to the modulating current, a linear response that has a high sensitivity over a large dynamic range can be obtained for a modulator using such a Faraday rotator. In certain applications, however, a value of θ_p different from 45° is chosen for objectives other than a linear response.

To measure the value of the Faraday rotation angle θ_F independently of the fluctuations in the input optical intensity and the absolute calibration of the detection system, a dual-quadrature polarimetric configuration as shown in Fig. 7.13 can be employed. In this configuration, a polarizing beam splitter, such as a Glan prism, is used to divide the output beam into two orthogonal linearly polarized beams detected by two differential photodetectors of matched responsivity. The output readings from the two photodetectors are taken to compute a normalized difference signal

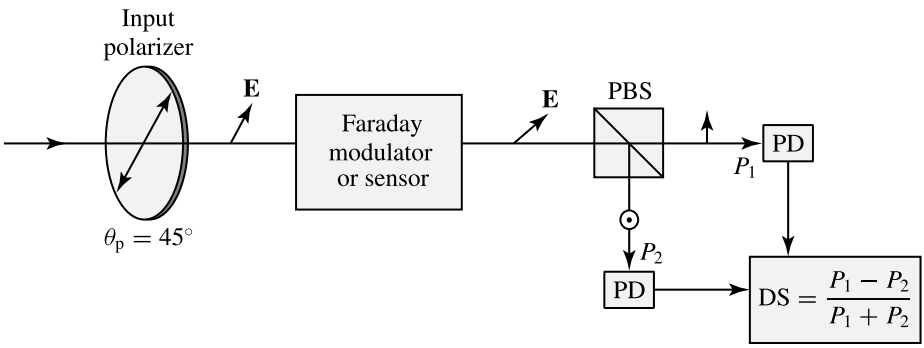


Figure 7.13 Dual-quadrature polarimetric detection scheme for the measurement of the Faraday rotation angle. PD indicates a photodetector. PBS indicates a polarizing beam splitter.

$$\mathrm{DS} = \frac{P_1 - P_2}{P_1 + P_2}, \tag{7.47}$$

where P_1 and P_2 are the optical powers of the two beams detected by the differential photodetectors. By properly orienting the principal axis of the polarizing beam splitter with respect to that of the input polarizer for $\theta_p = 45°$ so that DS $= 0$ when $\theta_F = 0$, the difference signal has the following dependence on the Faraday rotation angle:

$$\mathrm{DS} = \sin 2\theta_F. \tag{7.48}$$

Current and magnetic field sensors

A magneto-optic amplitude modulator can be used as a current or magnetic field sensor. For this kind of application, a linear response is desired. Therefore, the absolute value of θ_F is kept small within the range of operation, and the analyzer is carefully oriented at $\theta_p = 45°$ with respect to the polarizer so that (7.46) is valid. Paramagnetic or diamagnetic materials, such as silica glass, terbium-doped glasses, TGG, $Bi_{12}SiO_{20}$ (BSO), and $Bi_{12}GeO_{20}$ (BGO), are used.

There are two different types of current sensors, namely, *linked* and *unlinked*. In a *linked* sensor, the conductor carrying the current to be measured is fully enclosed by the magneto-optic medium. In an *unlinked* device, the magneto-optic medium does not fully enclose the conductor.

Figure 7.14 shows two examples of the linked type. In Fig. 7.14(*a*), the Faraday rotator is made of a monolithic magneto-optic material, such as a single piece of silica glass. The conductor passes through the central opening of this medium. A linearly polarized wave is guided by total internal reflection at the properly shaped corners to travel closely along the magnetic field line encircling the conductor. To multiply the Faraday rotation angle, sophisticated optical design for guiding the optical wave to encircle multiple turns around the conductor can be implemented over this basic structure. Alternatively,

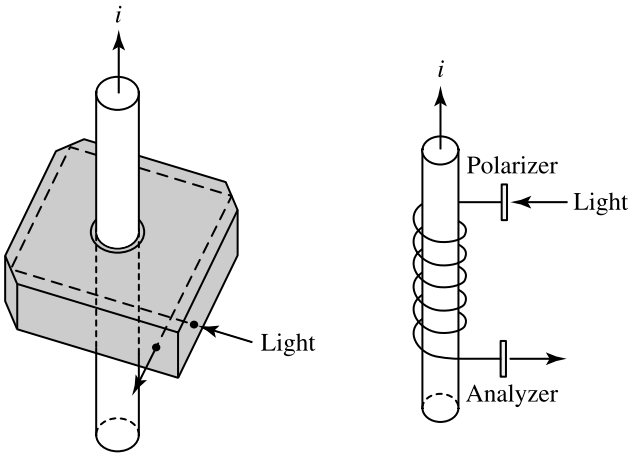


Figure 7.14 Magneto-optic current sensors of linked type that use (*a*) a monolithic Faraday rotator and (*b*) an optical fiber wound around the conductor as the Faraday rotator.

an optical fiber wound around the conductor, shown in Fig. 7.14(*b*), can be used. For a linked current sensor, the Faraday rotation angle given in (7.25) has to be modified because both the magnetic field and the optical path loop around the current. Using Ampere's law, we have

$$\theta_{\mathrm{F}} = V \oint \boldsymbol{H}_0 \cdot \mathrm{d}\mathbf{l} = VNi, \tag{7.49}$$

where N is the number of turns for which the optical path encircles the current i.

EXAMPLE 7.4 A fiber-optic current sensor of linked type as shown in Fig. 7.14(*b*) consists of 20 turns of silica fiber wound around the conductor. The detection scheme has the dual-quadrature polarimetric configuration shown in Fig. 7.13 with a polarized He–Ne laser at $\lambda = 632.8$ nm used as the light source. The sensor has a dynamic range from 1 A to 1 kA. What is the smallest Faraday rotation angle the sensor is required to measure? What is the largest linearity error in the measurement?

Solution At 632.8 nm wavelength, the Verdet constant of silica fiber is $V = 3.93 \times 10^{-6}$ rad A^{-1} from Table 7.1. To obtain a current reading of 1 A at the lower end of its dynamic range, the sensor, with $N = 20$, is required to be capable of measuring a Faraday rotation angle as small as

$$\theta_{\mathrm{F}} = 3.939 \times 10^{-6} \times 20 \times 1 \text{ rad} = 78.6 \text{ μrad}.$$

Linearity error of the measurement occurs because of the difference between the signal, $\mathrm{DS} = \sin 2\theta_{\mathrm{F}}$, that is obtained from the reading of the sensor according to (7.48) and the response, $2\theta_{\mathrm{F}}$, that is directly proportional to the current. It increases as the absolute

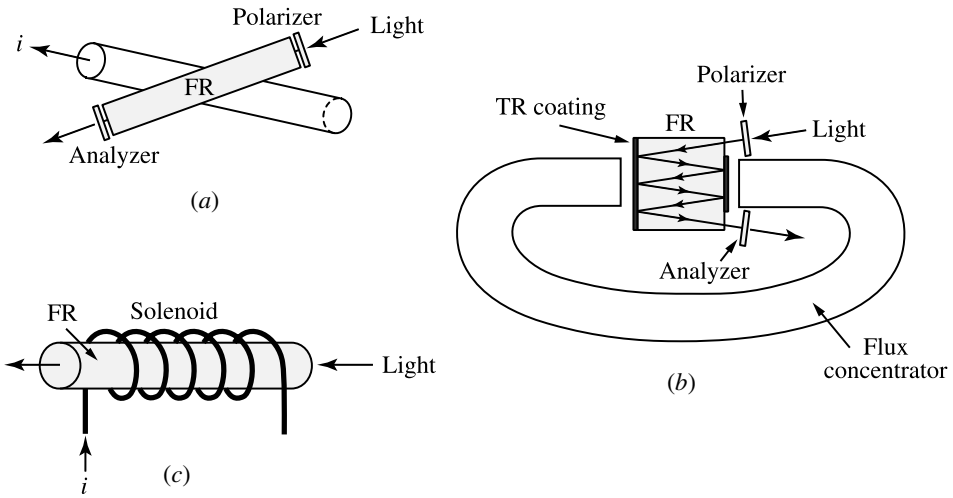


Figure 7.15 Magneto-optic current sensors of unlinked type that have (*a*) a simple structure, (*b*) a sophisticated structure with a flux concentrator and multiple optical passes through the rotator, and (*c*) a Faraday rotator in a solenoid. FR represents a Faraday rotator. TR means total reflection.

value of θ_F increases from zero toward $\pi/2$. Therefore, the largest linearity error occurs at the upper end of the dynamic range at $i = 1$ kA for which $\theta_F = 78.6$ mrad. It is found as

$$\text{Linearity error} = 1 - \frac{\sin 2\theta_F}{2\theta_F} = 1 - \frac{\sin(2 \times 0.0786)}{2 \times 0.0786} \approx 0.41\%.$$

Unlinked magneto-optic current sensors can also take a variety of different structures. Figure 7.15(*a*) shows a simple structure, whereas a more sophisticated structure is shown in Fig. 7.15(*b*). In the latter structure, the Faraday rotator is placed in the gap of a magnetic core that serves the purpose of a *flux concentrator* to enhance the sensitivity of the device by concentrating the magnetic flux through the rotator. Figure 7.15(*c*) shows yet another structure consisting of a Faraday rotator that is looped around by a current-carrying conductor in the form of a solenoid; the relation in (7.49) also applies to this structure with N being the number of turns of the conducting wire in the solenoid. One can also take advantage of the nonreciprocal nature of the Faraday effect to multiply the total Faraday rotation angle in an unlinked device, thus further enhancing the sensitivity of the device, by properly applying total-reflection coatings on the rotator surfaces for the optical wave to have multiple internal passes in the rotator, as also shown in Fig. 7.15(*b*).

There are advantages and disadvantages for both linked and unlinked types of devices. As can be seen from (7.49), a linked device measures the current directly and is virtually immune to interference from external stray magnetic fields. An unlinked device does not measure the current directly, but measures the magnetic field induced

by the current. It requires careful calibration for a correct reading of the current because it is susceptible to external interference and spatial variations of the magnetic field strength. One major problem of linked devices, however, is the maintenance of the correct polarization direction of the linearly polarized optical wave looping around in the rotator. All changes in the polarization direction have to be caused solely by the Faraday effect. Any other effects, such as improper internal reflection in a monolithic rotator and linear birefringence in a fiber caused by bending stress, that lead to polarization changes have to be eliminated in order to obtain a correct reading of the current being measured. The bandwidth of a magneto-optic current sensor of either type is ultimately limited by the optical transit time through the sensing element. Therefore, the device is capable of sensing AC currents at very high frequencies even when relatively long fibers are used. The light source used does not have to be polarized because it is polarized by the polarizer at the input end of the device before entering the Faraday rotator.

From the above discussions, it is clear that any unlinked current sensor can also be used as a magnetic field sensor.

Spatial light modulators

A magneto-optic spatial light modulator consists of a one- or two-dimensional spatial array of independently addressable Faraday rotators placed between a polarizer and an analyzer. A single-crystal magnetic thin film, commonly a bismuth-substituted iron garnet film of a high specific Faraday rotation coefficient, is grown on a lattice-matched, transparent, nonmagnetic garnet substrate, typically $Gd_3Ga_5O_{12}$ (GGG) or its derivatives such as one doped with Ca, Mg, and Zr. This substrate allows a large amount of Bi to be incorporated into the iron garnet film for a large specific Faraday rotation. The magnetic film is structured into a one- or two-dimensional array of isolated mesas using microprocessing technology. Each mesa defines a *pixel* (picture element) of the spatial light modulator. It is required that the film has a sufficiently large uniaxial magnetic anisotropy with a positive anisotropy constant in the direction normal to the surface, thus ensuring that the magnetization always points either up or down normal to the film surface.

The magnetization state of a pixel is controlled by two orthogonally running conductors that intersect at one corner of the mesa, as shown in Fig. 7.16. Switching of the magnetization state is accomplished by changing the magnetization direction. A uniform magnetic field stronger than the saturation field of the film material can be externally applied to the entire array to switch all of the pixels in the array to a given state, thus refreshing the array by erasing any existing pattern. The array can then be configured into any desired pattern by switching the magnetization state of selected pixels using the magnetic field generated by the currents flowing through the matrix conductors. The combined magnetic field generated by the currents flowing through

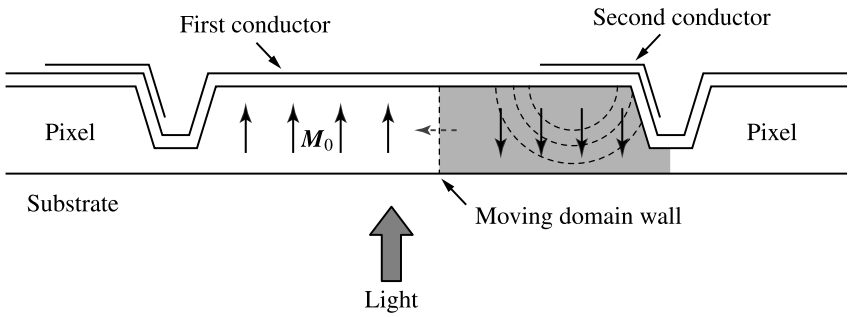


Figure 7.16 Pixel configuration and current-controlled switching process in a magneto-optic spatial light modulator.

the two conductors intersecting at a selected mesa is sufficient to initiate the switching, but not that generated by the current through either conductor alone. Thus, the entire array of pixels can be electrically addressed using orthogonally crossing matrix drive conductors.

In switching the magnetization state of a pixel, the magnetic field triggers movement of the magnetic domain wall across the mesa, as also illustrated in Fig. 7.16. If the magnetic field exceeds the saturation field and lasts long enough, the domain wall can sweep across the entire mesa, resulting in complete switching of the magnetization direction. If, instead, the currents generating the magnetic field are terminated at the moment when the domain wall reaches the bottom of the film but has not swept across the mesa, the mesa will be *nucleated*, containing equal areas magnetized in opposite directions. Consequently, there are three different magnetization states for a pixel: two uniformly magnetized states and the nucleated state. When a pixel is in one of the two uniformly magnetized states, a linearly polarized optical wave transmitted by the pixel experiences a Faraday rotation angle of either $\rho_F l$ or $-\rho_F l$, where ρ_F is the specific Faraday rotation and l is the thickness of the film. When it is in the nucleated state, the Faraday rotation for the optical wave transmitted by the pixel averages out to be zero.

A spatial light modulator can be used either in *binary operation*, by switching between the two uniformly magnetized states of opposite magnetization direction, or in *ternary operation*, by switching among all of the three different magnetization states.

The basic configuration of a *transmission-mode* magneto-optic spatial light modulator in binary operation is illustrated in Fig. 7.17. The input light, which can be either polarized or unpolarized to begin with, is polarized by the polarizer. Using (7.45), the transmittance for the two uniformly magnetized states can be expressed as

$$T = T_0 e^{-\alpha l} \cos^2(\pm\rho_F l - \theta_p). \tag{7.50}$$

In order for the device to have the highest possible contrast ratio, the surfaces of the Faraday rotator are antireflection coated to eliminate reflections that can introduce improper polarization changes to the transmitted light. In addition, the value of θ_p has

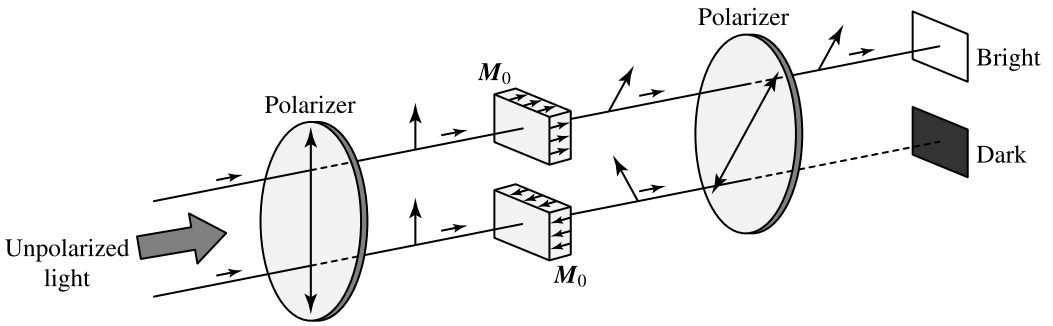


Figure 7.17 Illustration of a transmission-mode magneto-optic spatial light modulator in binary operation.

to be chosen such that $T_{\text{OFF}} = 0$ in the OFF state. One choice is $\theta_{\text{p}} = 90^\circ - \rho_{\text{F}} l$ for the magnetization state yielding a Faraday rotation angle of $-\rho_{\text{F}} l$ to represent the OFF state. Another choice is $\theta_{\text{p}} = 90^\circ + \rho_{\text{F}} l$ for the magnetization state corresponding to a Faraday rotation angle of $\rho_{\text{F}} l$ to represent the OFF state. In either case, the transmittance in the ON state is

$$T_{\text{ON}} = T_0 \text{e}^{-\alpha l} \sin^2(2\rho_{\text{F}} l). \tag{7.51}$$

Clearly, the optimum thickness of the magnetic film that yields the highest transmittance in the ON state depends on the values of both α and ρ_{F} (see Problem 7.5.6). In practice, T_{OFF} is never exactly zero because of residual reflections from the rotator surfaces and the circular dichroism in the film. Ignoring the effect of circular dichroism, the OFF-state transmittance is given by (see Problem 7.5.8)

$$T_{\text{OFF}} = T_0 R_1 R_2 \text{e}^{-3\alpha l} \sin^2(2\rho_{\text{F}} l), \tag{7.52}$$

where R_1 and R_2 are the reflectivities of the two surfaces of the Faraday rotator consisting of the magnetic film on a substrate. The contrast ratio of the device is given by (see Problem 7.5.8)

$$\text{Contrast ratio} = \frac{T_{\text{ON}}}{T_{\text{OFF}}} = \frac{\text{e}^{2\alpha l}}{R_1 R_2}. \tag{7.53}$$

Clearly, to maximize the contrast ratio, the residual reflections have to be minimized with high-quality antireflection coating while maximizing the ON-state transmittance.

EXAMPLE 7.5 A Bi : YIG film of 10 μm thickness on GGG substrate is used for a transmission-mode magneto-optic spatial light modulator operated at 632.8 nm wavelength. At this wavelength, the film has an absorption coefficient of $\alpha = 0.108\ \mu\text{m}^{-1}$ and a specific Faraday rotation of $\rho_{\text{F}} = 1.68 \times 10^{-2}$ rad μm^{-1}. The sample is antireflection coated on both surfaces with $T_0 = 0.9$ caused only by the absorption in the

polarizer and analyzer. What is the ON-state transmittance of the modulator? If the film thickness is increased to 16 μm, what is the ON-state transmittance?

Solution For $l = 10$ μm, we find with the given parameters that

$$T_{\text{ON}} = 0.9 \times \text{e}^{-0.108\times 10} \times \sin^2(2 \times 1.68 \times 10^{-2} \times 10) = 3.3\%.$$

For $l = 16$ μm, the transmittance is increased to

$$T_{\text{ON}} = 0.9 \times \text{e}^{-0.108\times 16} \times \sin^2(2 \times 1.68 \times 10^{-2} \times 16) = 4.2\%.$$

Further increase in the film thickness does not increase the ON-state transmittance but results in a decrease in the ON-state transmittance (see Problem 7.5.6).

If the analyzer is oriented such that $\theta_{\text{p}} = 90^\circ$, (7.50) yields the same transmittance, which is proportional to $\sin^2(\rho_{\text{F}} l)$, for the two uniformly magnetized states. These two states can still be distinguished because their transmitted fields have a π phase difference. Consequently, this arrangement leads to the *binary phase-only mode* of operation. In this configuration, the transmittance of a pixel in the nucleated state is zero. A *ternary phase-only mode* of operation consisting of the $+1$, 0, and -1 states is possible if the nucleated state is also included in the operation.

Using a *reflection-mode* spatial light modulator, the magnetic film can be halved in thickness while maintaining the same contrast ratio as that of a transmission-mode device. In a reflection-mode device, the back surface of the Faraday rotator is made totally reflective while the front surface is antireflection coated. Light entering the rotator from the front surface passes through the film twice before leaving the film, also from the front surface. Because of the nonreciprocal characteristic of the Faraday effect, the Faraday rotation is cumulative for both passes. The linear absorption is also cumulative. Therefore, if the film of a reflection-mode device is half as thick as that of a transmission-mode device, the reflectance of the reflection-mode device in any particular state is the same as the transmittance of the transmission-mode device in the corresponding state.

The magneto-optic spatial light modulator has several unique features that enable it to find many useful applications, such as parallel optical signal processing, optical pattern recognition, image coding, and reconfigurable optical interconnects. On the one hand, the device is electrically addressable and has a high frame rate because the magnetization state of a pixel can be switched in a time as short as 1 ns. In practical applications, typical current pulses used for the switching are on the order of 100 ns. On the other hand, the device has the nonvolatility to hold a pattern for a long time because the pixels do not spontaneously demagnetize without the externally applied magnetic field or the controlling current pulses. A very high contrast ratio can be obtained by optimizing the film thickness and by eliminating the absorption and stray reflections of the optical components as much as possible. The size of a pixel is typically on the

order of 10 μm × 10 μm to 100 μm × 100 μm. A very large number of pixels can be incorporated in a device for a high image resolution.

7.6 Magneto-optic recording

In magneto-optic recording, digitized information stored in a magnetic thin film is read using the magneto-optic Faraday or Kerr effect. There are certain similarities between the principle of magneto-optic recording and that of the magneto-optic spatial light modulator. Indeed, because of its nonvolatility, a magneto-optic spatial light modulator also has the ability to hold digitized information for later access. Reading of the recorded information is performed using the magneto-optic Faraday effect. However, while the application of a magneto-optic spatial light modulator is primarily dynamic information processing, the purpose of magneto-optic recording is data storage and retrieval. Therefore, there are many fundamental differences between them due to different practical considerations.

The media for magneto-optic recording are ferromagnetic or ferrimagnetic thin films supported by nonmagnetic substrates. The presence of a sufficiently large uniaxial magnetic anisotropy with a positive anisotropy constant in the direction normal to the film surface is required to ensure that the film has two clearly distinguishable, oppositely directed magnetization states, which represent the binary logical states. In amorphous magnetic films prepared by evaporation or sputtering, this condition can be achieved by properly choosing the deposition parameters to create an anisotropic atomic arrangement along the film normal.

The materials suitable for the application of magneto-optic recording include magnetic oxides, particularly the Bi-substituted garnets, metallic Pt–Co and Pd–Co multilayers, and magnetic alloys. The most popular magneto-optic recording materials today are amorphous ferrimagnetic rare-earth transition-metal (RE–TM) alloys containing one or more of the rare earths Gd, Tb, and Dy in addition to one or more of the transition metals Fe and Co. The most prominent examples are GdTbFe and TbFeCo alloys. The magnetizations of these rare-earth and transition-metal atoms vary differently with temperature and are antiferromagnetically coupled. The rare-earth magnetization is larger at low temperatures, whereas the transition-metal magnetization is larger at high temperatures. Consequently, an alloy of a proper RE–TM composition is a ferrimagnetic material that has a *compensation temperature*, T_{comp}, below its Curie temperature T_c. At T_{comp}, the magnetizations of the rare earth and the transition metal are equal and opposite, resulting in zero net magnetization. The coercive field, H_c, exhibits a singularity tending toward infinity at T_{comp}. Above the compensation point, H_c decreases as the temperature is increased toward T_c. The magnetization and coercivity of such an alloy as a function of temperature are shown in Figs. 7.18(*a*) and (*b*), respectively. The

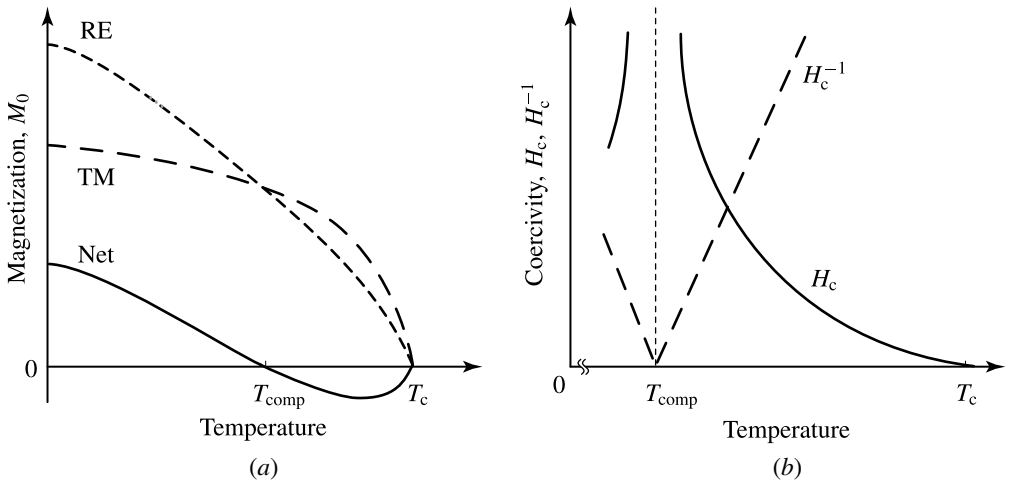


Figure 7.18 Temperature-dependent characteristics of (*a*) the magnetization and (*b*) the coercivity of a rare-earth transition-metal alloy.

compensation and Curie temperatures, as well as the temperature characteristics of H_c, of an alloy can be controlled by properly choosing the composition of the alloy.

The information is stored in the recording medium by means of magnetic domains. Writing, erasing, and rewriting are achieved by switching the magnetization direction through a thermomagnetic process with optical heating by a focused laser beam. The composition of the medium is chosen to have a compensation point close to room temperature and a Curie temperature between 400 and 600 K. On the one hand, this medium provides a high coercivity at room temperature to stabilize the information stored in the magnetic domains and to allow for a high storage density. On the other hand, the coercivity of the heated spot can be significantly lowered in the write process with a laser beam of a moderate power to raise the temperature of the heated spot near or above the Curie temperature.

The thermomagnetic switching process is based on a simple principle that the magnetization in a locally heated volume of the film can be oriented to the direction of an applied magnetic field when the coercivity is lowered at a high temperature to be less than the applied magnetic field. Writing is accomplished by focusing a laser beam of a moderately high power, typically in the range of 5–10 mW, to a diffraction-limited spot on the medium. The write process is performed either by modulating the laser power at a constant magnetic field or by modulating the magnetic field at a constant laser power during the pass of the laser beam over the medium along the data track. The latter permits direct overwrite, but its switching frequency is limited to less than 10 MHz due to the operation margins of the coil generating the magnetic field. Erasure of the written information is accomplished by heating the medium with a constant laser power at a constant magnetic field to revert the magnetization of the bit to be

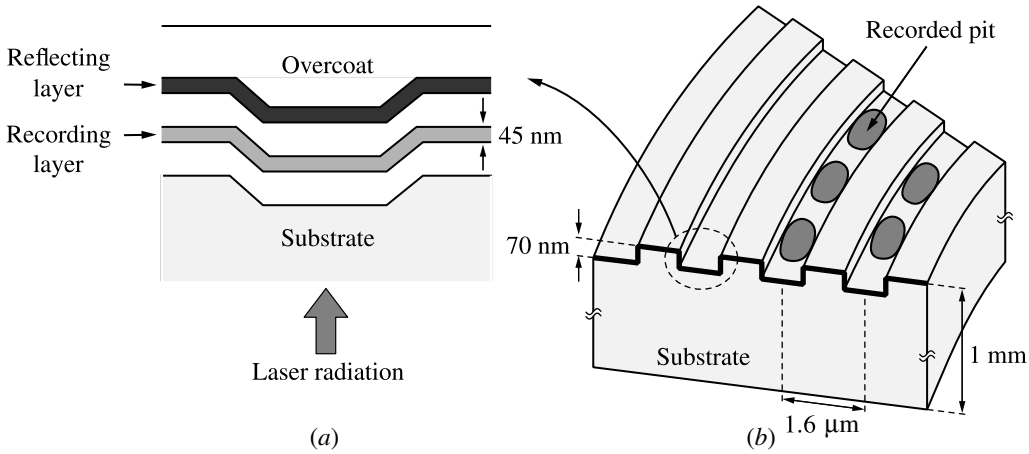


Figure 7.19 (*a*) Multilayer structure and (*b*) tracking pregrooves of a magneto-optic disk.

erased back to the preset direction. The simplest overwrite scheme uses two passes of the laser beam over the medium: the first pass for erasing the old information and the second for writing the new data. More sophisticated schemes using one or two laser beams in a single pass for direct overwrite are also developed. Magneto-optic recording is both erasable and rewritable because the thermomagnetic switching process is reversible.

The optimum magneto-optic disk is based on a multilayer structure to achieve long-term stability, high read-out efficiency, and high switching sensitivity. As shown in Fig. 7.19(*a*), a typical disk is composed of a pregrooved, transparent polyvinyl chloride or glass substrate of about 1 mm thickness, a precoated antireflection dielectric layer of 80 nm thickness, a magneto-optic medium layer of 45 nm thickness, a thin space layer, a reflective metallic layer of 30 nm thickness, and, finally, a protective polymer layer of a few micrometers thickness. The laser light used for writing, erasing, or reading is incident on the disk from the substrate side, as is also illustrated in Fig. 7.19(*a*). The composition and thickness of each layer, as well as the number of different layers, vary in disks designed for different applications. The pregrooved structure serves for tracking in the recording system. The depth of the pregrooves is in the range of 50–70 nm, and the spacing between neighboring tracks is designed to be twice the focused laser beam spot size. A typical example is shown in Fig. 7.19(*b*). By monitoring light reflected from the pregrooved structure, tracking servomechanisms maintain the position of the recording head accurately along the track and keep the focal point of the laser beam on the surface of the magneto-optic medium layer.

The written domain size is governed more by the magnetic properties of the medium and the domain nucleation and growth processes than by the diameter of the focused laser beam. Domain sizes smaller than the *Rayleigh resolution limit*, $d_R = 1.22\lambda/\text{NA}$, are quite easily written in a good magneto-optic medium. The practical spot size is

approximately given by

$$d_0 = \frac{0.5\lambda}{\text{NA}}, \tag{7.54}$$

where λ is the laser wavelength and NA is the numerical aperture of the objective lens. Taking into account the spacing between the pregrooved tracks in a disk, this yields an *areal bit density* of

$$\text{ABD} = \frac{1}{2d_0^2} = \frac{2(\text{NA})^2}{\lambda^2}. \tag{7.55}$$

Typically an objective lens with a numerical aperture of 0.5 or greater is used so that the spot has submicrometer dimensions.

EXAMPLE 7.6 A magneto-optic recording system uses a diode laser emitting at $\lambda = 800$ nm and an objective lens of NA $= 0.5$. For this system, the laser spot has a submicrometer spot size of 800 nm, according to (7.54). Using (7.55), it is found that the areal bit density is approximately 78 Mbit cm^{-2} (or 503 Mbit in^{-2}). For this bit density, the *mark area*, which is the area for each bit, is 1.28 μm^2.

To achieve a high recording density, one avenue is to reduce the wavelength of the laser light utilized. According to (7.55), a 40% increase in areal bit density can be realized by replacing the diode lasers at 800 nm with red diode lasers of wavelengths at around 670 nm. The areal bit density can be quadrupled by using an InGaN laser at 400 nm wavelength. Further increase of the bit density can be accomplished by using lenses of high numerical aperture. The bit density is also dependent upon the data coding scheme. Using efficient codes for the data, such as those employing the magnetic flux changes between domains rather than the domains themselves to represent the information, the bit density can be effectively doubled. An areal bit density as high as 100 Gbit in^{-2}, equivalent to 15.5 Gbit cm^{-2}, has been achieved (see Problem 7.6.1).

For effective optical heating, high optical absorption in the magnetic film is required. Consequently, the polar Kerr effect, rather than the Faraday effect, is most commonly used for the read process. From (7.38), it can be seen that the polar Kerr angle θ_{K} obeys the relation $\theta_{\text{K}}(\boldsymbol{M}_0) = -\theta_{\text{K}}(-\boldsymbol{M}_0)$. Therefore, the two possible directions of magnetization correspond to a positive and a negative Kerr signal, respectively. The magneto-optic recording system consists of a polarization-sensitive optical head, as is shown in Fig. 7.20(*a*). During readout, the laser power is typically reduced to about one-tenth of that used for writing, which is far below the threshold to write or erase. Upon reflection from the magnetic medium, the plane of polarization of the light is rotated by the polar Kerr effect. From (7.38), it can be seen that in the polar Kerr effect a Kerr rotation angle is always accompanied by a Kerr ellipticity. Therefore, the reflected light is passed through a wave plate to compensate for the ellipticity

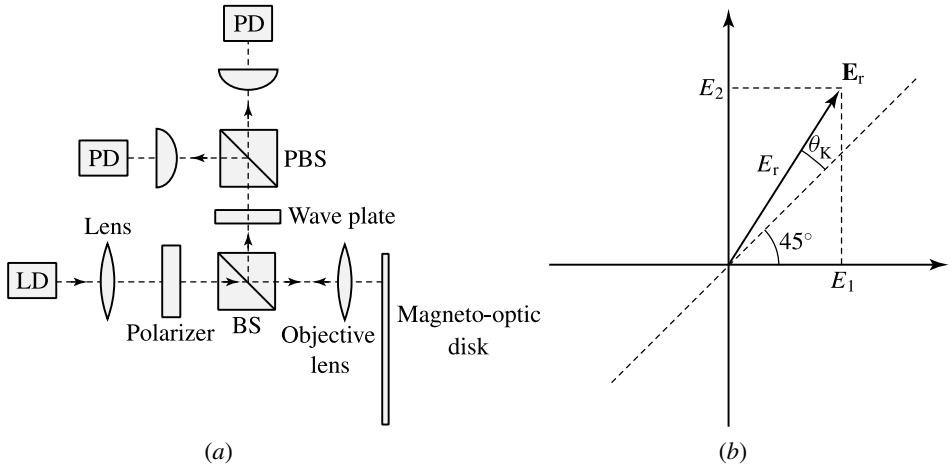


Figure 7.20 (*a*) Schematics of a magneto-optic recording head assembly. (*b*) Field decomposition, by a polarizing beam splitter, of the Kerr-rotated reflected light for the differential photodetectors. LD indicates a laser diode. PD indicates a photodetector. BS indicates a beam splitter. PBS indicates a polarizing beam splitter.

introduced by the polar Kerr effect. It is then decomposed into two components by a polarizing beam splitter that is set at 45° with respect to the polarization of the incident light, as shown in Fig. 7.20(*b*). The two components are directed to a set of differential photodetectors. With no magneto-optic rotation of the plane of polarization, the intensity of light at each photodetector is the same, yielding no difference signal between the two photodetectors. With a rotation of the plane of polarization, one photodetector receives more light than the other, resulting in a difference signal. As shown in Fig. 7.20(*b*), with a polarization rotation angle of θ_K, the optical fields of the reflected light passing through the polarizing beam splitter to reach the two photodetectors are $E_1 = E_r \cos(\pi/4 + \theta_K)$ and $E_2 = E_r \sin(\pi/4 + \theta_K)$, respectively. Consequently, the signal current produced by the differential photodetectors is given by

$$i_s = \mathcal{R} P_0 R \left[\sin^2 \left(\frac{\pi}{4} + \theta_K \right) - \cos^2 \left(\frac{\pi}{4} + \theta_K \right) \right] = \mathcal{R} P_0 R \sin 2\theta_K, \tag{7.56}$$

where $\mathcal{R}$ is the responsivity of the photodetectors, P_0 is the average laser power incident upon the medium surface, and R is the reflectivity of the medium surface. Clearly, the polarity of the difference signal indicates the direction of the magnetization because i_s changes sign with θ_K.

From (7.56), the difference-signal current can be increased by increasing the detector responsivity, the laser power, the medium reflectivity, or the Kerr rotation angle. However, the important parameter characterizing the medium and the recording system is the signal-to-noise ratio (SNR) rather than the difference signal alone. The SNR obtainable from a magneto-optic recording system is fundamentally limited by shot noise in the differential photodetectors. The shot-noise, which is determined by the

total detector current rather than by the difference signal current, is given by

$$\overline{i_{\mathrm{n}}^2} = 2eB\mathcal{R}P_0R, \tag{7.57}$$

where e is the electronic charge and B is the bandwidth of the detectors. Because the maximum values of $|\theta_{\mathrm{K}}|$ for magneto-optic recording media are on the order of 10 mrad, the shot-noise-limited SNR is given by

$$\mathrm{SNR} = 10\log\frac{\overline{i_{\mathrm{s}}^2}}{\overline{i_{\mathrm{n}}^2}} \approx 10\log\frac{2\mathcal{R}P_0R\theta_{\mathrm{K}}^2}{eB}, \tag{7.58}$$

where θ_{K} is in radians. The SNR can be increased by increasing the laser power or the value of $R\theta_{\mathrm{K}}^2$. The value of $R\theta_{\mathrm{K}}^2$ is purely determined by the recording medium and is regarded as the figure of merit of a magneto-optic recording medium. Because the laser power in the read process is limited by the threshold for writing or erasing, this figure of merit is a very important consideration in the choice of a medium. For RE–TM alloys, the maximum values of $|\theta_{\mathrm{K}}|$ are below 0.5°, and $R \approx 0.4$ at a laser wavelength around 800 nm. Other media have slightly higher values of θ_{K}, but have other disadvantages. For example, MnBi has a Kerr rotation angle of $\theta_{\mathrm{K}} = 0.7°$, but it is polycrystalline, resulting in a noise figure high above the shot-noise limit and a poor SNR in spite of the large Kerr rotation.

EXAMPLE 7.7 A representative set of parameters for a magneto-optic recording system is $\mathcal{R} = 0.4$ A W^{-1}, $P_0 = 1$ mW, $R = 0.4$, $\theta_{\mathrm{K}} = 0.4° = 7$ mrad, and $B = 10$ MHz, yielding a shot-noise-limited SNR of 40 dB.

In a magneto-optic recording system, there is no direct contact between the recording head and the medium. The spacing between the optical head and the disk can be of the order of millimeters. Therefore, head crashes are not a concern as they are in magnetic recording systems. Because the optical access is provided through a transparent substrate and the laser beam is highly focused on the surface of the magneto-optic film, small particles on the back side of the substrate are out of the focal plane of the laser beam and do not significantly affect recording or readback. These characteristics of the magneto-optic recording technology make it possible to have a removable disk.

7.7 Guided-wave magneto-optic devices

It is possible to implement various kinds of guided-wave magneto-optic devices for optical modulation, switching, and many other functions. Nevertheless, there has been very little interest in developing such devices because equal or better performance of the functions provided by such devices can be accomplished by their electro-optic or acousto-optic counterparts. Among devices of equal performance, the magneto-optic ones have

certain disadvantages. Magneto-optic waveguides are not compatible with the dielectric and semiconductor waveguides used in most photonic devices because they have to be fabricated with magnetic materials, most commonly magnetic garnets, on special substrates that can support such waveguides. Furthermore, magneto-optic modulators and switches compare less favorably to the voltage-controlled electro-optic modulators and switches of the same function, particularly those in waveguide structures, because they have to be controlled by currents. However, magneto-optic devices utilizing the linear magneto-optic effect have the unique advantage of nonreciprocity, which is not possible for devices utilizing electro-optic or acousto-optic effects. Therefore, the most important guided-wave magneto-optic devices are nonreciprocal devices, including guided-wave optical isolators and circulators. In such devices, the core of the waveguide consists of a material that has a spontaneous magnetization. No controlling current is needed. Most of them utilize YIG or Bi-substituted YIG waveguides on GGG substrates.

Nonreciprocal TE–TM mode converters

There are some fundamental differences between guided-wave devices and bulk devices. Light propagates in a waveguide in the form of waveguide modes. Polarization rotation in a waveguide is accomplished through coupling between orthogonally polarized modes, such as the TE and TM modes in a planar waveguide and the TE-like and TM-like modes in a three-dimensional waveguide. Figure 7.21 shows a YIG waveguide on a GGG substrate with a magnetization $\boldsymbol{M}_0 = M_{0z}\hat{z}$ and a permittivity tensor described by (7.16). From the expression in (7.7), this permittivity tensor can be divided into two parts by writing $\boldsymbol{\epsilon}(\boldsymbol{M}_0) = \boldsymbol{\epsilon}(0) + \Delta\boldsymbol{\epsilon}(\boldsymbol{M}_0)$. The waveguide modes are defined by the permittivity tensor:

$$\boldsymbol{\epsilon}(0) = \epsilon_0 \begin{bmatrix} n_\perp^2 & 0 & 0 \\ 0 & n_\perp^2 & 0 \\ 0 & 0 & n_\parallel^2 \end{bmatrix}, \tag{7.59}$$

which does not include the circular birefringence caused by the magnetization. These modes are coupled through the linear magneto-optic effect caused by the perturbing

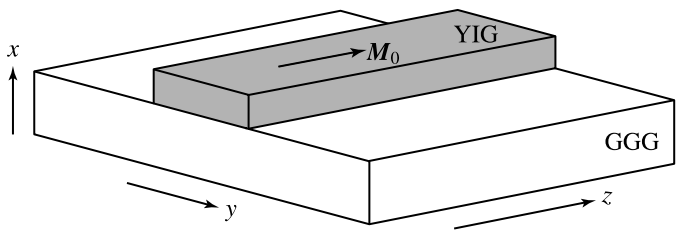


Figure 7.21 Nonreciprocal TE–TM-mode converter with a magnetic YIG waveguide on a GGG substrate. The magnetization of the waveguide is in the longitudinal direction.

permittivity tensor:

$$\Delta\boldsymbol{\epsilon}(\boldsymbol{M}_0) = \epsilon_0 \begin{bmatrix} 0 & \mathrm{i}\xi & 0 \\ -\mathrm{i}\xi & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{7.60}$$

which is responsible for circular birefringence. For a waveguide that supports only fundamental TE-like and TM-like modes, the coupling coefficients can be found using (4.36). With $\Delta\boldsymbol{\epsilon}$ given in (7.60) for $\boldsymbol{M}_0 = M_{0z}\hat{z}$, the self-coupling coefficients, κ_{EE} and κ_{MM} for TE-like and TM-like modes, respectively, are both found to be zero. Consequently, the propagation constants of both modes are not influenced by the perturbation caused by $\Delta\boldsymbol{\epsilon}$, and the phase mismatch between them is simply given by

$$2\delta = \Delta\beta = \beta_{\mathrm{TM}} - \beta_{\mathrm{TE}}, \tag{7.61}$$

where β_{TE} and β_{TM} are determined by taking the dielectric tensor of the waveguide core to be only $\boldsymbol{\epsilon}(0)$ given in (7.59). However, there is a nonvanishing coupling coefficient between these two modes given by (see Problem 7.7.3(a))

$$\begin{aligned} \kappa &= \kappa_{\mathrm{EM}} = \kappa_{\mathrm{ME}}^{*} \\ &= \omega \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y \hat{\boldsymbol{\mathcal{E}}}_{\mathrm{TE}}^{*}(x,y)\cdot\Delta\boldsymbol{\epsilon}(x,y)\cdot\hat{\boldsymbol{\mathcal{E}}}_{\mathrm{TM}}(x,y) \\ &\approx -\mathrm{i}\rho_{\mathrm{F}}\Gamma_{\mathrm{EM}}, \end{aligned} \tag{7.62}$$

where ρ_{F} is the specific Faraday rotation, given by (7.28), of the waveguide material and

$$\begin{aligned} \Gamma_{\mathrm{EM}} &= \frac{2\beta_{\mathrm{TE}}^{1/2}\beta_{\mathrm{TM}}^{1/2}}{\omega\mu_0}\frac{1}{M_{0z}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{d}x\mathrm{d}y M_{0z}(x,y)\hat{\mathcal{E}}_{\mathrm{TE},y}^{*}(x,y)\hat{\mathcal{E}}_{\mathrm{TM},x}(x,y) \\ &\approx \frac{1}{M_{0z}}\frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{d}x\mathrm{d}y M_{0z}(x,y)\hat{\mathcal{E}}_{\mathrm{TE},y}^{*}(x,y)\hat{\mathcal{E}}_{\mathrm{TM},x}(x,y)}{\left[\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{d}x\mathrm{d}y|\hat{\mathcal{E}}_{\mathrm{TE},y}|^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{d}x\mathrm{d}y|\hat{\mathcal{E}}_{\mathrm{TM},x}|^2\right]^{1/2}} \end{aligned} \tag{7.63}$$

is the overlap factor for the magneto-optic coupling of TE-like and TM-like modes through the linear magneto-optic effect. The distribution of the magnetization in the entire structure is described by $M_{0z}(x,y)$, which has a value of $M_{0z}(x,y) = M_{0z}$ in the magnetic core region and a value of $M_{0z}(x,y) = 0$ in the nonmagnetic regions.

The coupling efficiency from one mode to another in a nonreciprocal magneto-optic TE–TM mode converter follows that of the codirectional coupler discussed in Section 4.3. However, unlike the electro-optic TE–TM mode converter discussed

in Section 6.4, a magneto-optic TE–TM mode converter is a nonreciprocal device. If an optical wave is allowed to travel multiple passes back and forth in a nonreciprocal magneto-optic TE–TM mode converter, the net coupling efficiency from one mode to the other depends only on the total distance, but not on the direction, over which the wave has traveled within the waveguide. When the TE and TM modes are perfectly phase matched, the conversion efficiency is simply

$$\eta = \sin^2 |\kappa| l, \tag{7.64}$$

where l is the *cumulative distance traveled in both directions*. An optical wave that makes a round trip in a nonreciprocal magneto-optic TE–TM mode converter due to reflection at the far end of the waveguide does not return to the input end in its original polarization state unless the value of $|\kappa| l$, with l being twice the length of the waveguide for a round trip, happens to be an integral multiple of π. In contrast, in a reciprocal TE–TM mode converter, irrespective of how many round-trip passes an optical wave has traveled, the length l used in (7.64) for calculation of the coupling efficiency is the physical distance measured from the input end to the point where the conversion is being evaluated, not the cumulative distance traveled by the optical wave. Consequently, an optical wave that makes a round trip in a reciprocal TE–TM mode converter always returns to the input end in its original polarization state (see Problem 7.7.1).

According to the discussions in Section 7.1, the material used in a magneto-optic device must have no other birefringence that dominates the magneto-optic effect used for the operation of the device. The materials, such as YIG and GGG, used in magneto-optic waveguides are generally isotropic materials. The phase mismatch between TE-like and TM-like modes in such waveguides is caused by the structural birefringence due to the difference in the boundary conditions imposed by the waveguide structure on the TE-like and TM-like modes. This structural birefringence is not very large in an ordinary waveguide, but in a magneto-optic waveguide it can easily dominate the circular birefringence caused by the linear magneto-optic effect because the value of ρ_F is generally very small. Therefore, the phase mismatch caused by the structural birefringence, which always results in $\Delta\beta = \beta_{TM} - \beta_{TE} < 0$ according to (2.69), is generally too large for a magneto-optic TE–TM mode converter.

The general concept for reducing the phase mismatch and attaining phase matching in a nonreciprocal magneto-optic TE–TM mode converter is to introduce other birefringence of opposite sign in the waveguide to counterbalance the structural birefringence. There are several approaches to implementing this concept. The simplest is to incorporate in the magnetic waveguide core a proper amount of *stress-induced birefringence*, which is caused by stress in the waveguide due to a slight lattice mismatch between the waveguide core and the substrate, or *growth-induced birefringence*, which is not caused by lattice mismatch but by the dopants in the waveguide core such as the bismuth atoms in a Bi-substituted YIG layer. Another approach is to have a layer of anisotropic crystal, such as $LiIO_3$, grown on top of the magnetic waveguide core in such

a way that the fields of the orthogonally polarized TE-like and TM-like modes penetrating into this layer see a proper amount of difference in the refractive index to counterbalance the difference in their boundary conditions. The desired counterbalancing birefringence for the elimination of phase mismatch can also be introduced by artificial structures, such as periodic grooves, on the waveguide, as the discussions in Section 5.1 demonstrate.

Nonreciprocal phase shifters

A unique nonreciprocal waveguide device that has no counterpart among bulk devices is the nonreciprocal phase shifter. The function of this device depends on nonreciprocal coupling between the transverse and longitudinal electric field components of a waveguide mode through a magnetization that has a component perpendicular to both of them. For simplicity, we consider a planar magneto-optic waveguide whose magnetic core layer has a magnetization, $\boldsymbol{M}_0 = M_{0y}\hat{y}$, in a direction perpendicular to both longitudinal and transverse field components, $\hat{\mathcal{E}}_{\mathrm{TM},z}\hat{z}$ and $\hat{\mathcal{E}}_{\mathrm{TM},x}\hat{x}$, respectively, of the TM waveguide mode, as shown in Fig. 7.22. The permittivity tensor of this magnetic layer can be written as $\boldsymbol{\epsilon}(\boldsymbol{M}_0) = \boldsymbol{\epsilon}(0) + \Delta\boldsymbol{\epsilon}(\boldsymbol{M}_0)$ with

$$\boldsymbol{\epsilon}(0) = \epsilon_0 \begin{bmatrix} n_\perp^2 & 0 & 0 \\ 0 & n_\parallel^2 & 0 \\ 0 & 0 & n_\perp^2 \end{bmatrix} \tag{7.65}$$

and

$$\Delta\boldsymbol{\epsilon}(\boldsymbol{M}_0) = \epsilon_0 \begin{bmatrix} 0 & 0 & -\mathrm{i}\xi \\ 0 & 0 & 0 \\ \mathrm{i}\xi & 0 & 0 \end{bmatrix}, \tag{7.66}$$

where ξ is linearly proportional to M_{0y} and $\xi(-M_{0y}) = -\xi(M_{0y})$. In this case, the propagation constants of the waveguide modes without perturbation of the linear magneto-optic effect due to $\xi(M_{0y})$ are determined by taking $\boldsymbol{\epsilon}(0)$ given in (7.65) alone as

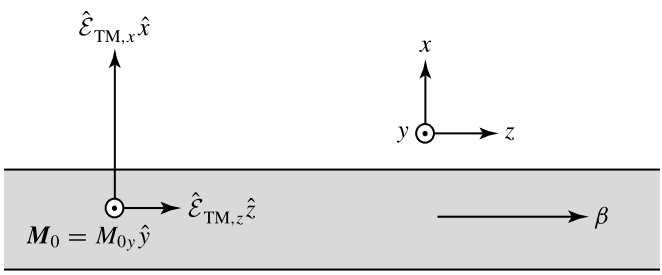


Figure 7.22 Nonreciprocal phase shifter for the TM mode in a planar magneto-optic waveguide. The magnetization direction is perpendicular to both electric field components of the TM mode. The self-coupling coefficient of the TM mode changes sign upon reversal of the propagation direction.

the permittivity tensor of the waveguide core. The effect of $\Delta\epsilon$ on the propagation characteristics of the waveguide modes is again evaluated using coupled-mode theory in a procedure similar to that used above in the treatment of the TE–TM mode converter. For the planar waveguide shown in Fig. 7.22, it is found that $\kappa_{\mathrm{EM}} = \kappa_{\mathrm{ME}} = \kappa_{\mathrm{EE}} = 0$ and (see Problem 7.7.3(b))

$$\kappa_{\mathrm{MM}} = \omega \int_{-\infty}^{\infty} \mathrm{d}x \hat{\boldsymbol{\mathcal{E}}}^{*}_{\mathrm{TM}}(x) \cdot \Delta\epsilon(x) \cdot \hat{\boldsymbol{\mathcal{E}}}_{\mathrm{TM}}(x)$$

$$= 2\omega\epsilon_0 \frac{\xi}{M_{0y}} \mathrm{Im}\left[\int_{-\infty}^{\infty} \mathrm{d}x M_{0y}(x) \hat{\mathcal{E}}^{*}_{\mathrm{TM},x}(x) \hat{\mathcal{E}}_{\mathrm{TM},z}(x)\right], \tag{7.67}$$

where $M_{0y}(x) = M_{0y}$ in the magnetic core layer and $M_{0y}(x) = 0$ outside the magnetic core. There is no TE–TM mode coupling in this waveguide, and the propagation constant of a TE mode is not perturbed by the linear magneto-optic effect. However, because $\hat{\mathcal{E}}_x$ and $\hat{\mathcal{E}}_z$ of a TM mode are 90° out-of-phase, as can be seen by comparing (2.35) with (2.36), κ_{MM} exists if the integral in (7.67) yields a nonzero value. It can be shown that this integral is zero if the waveguide is symmetric but is nonzero if the two boundaries of the magnetic core layer are different (see Problem 7.7.3(b)). Consequently, the linear magneto-optic effect can induce a change in the propagation constant of a TM mode in an asymmetric waveguide.

If the designation of the x, y, and z coordinate axes is fixed, the sign of ξ is fixed in a given waveguide with a fixed magnetization. However, the product $\hat{\mathcal{E}}^{*}_{\mathrm{TM},x}\hat{\mathcal{E}}_{\mathrm{TM},z}$ changes sign together with the propagation constant when the direction of propagation of a TM mode is reversed, as can be seen from Fig. 7.22. According to (7.67), this sign change leads to a corresponding sign change in κ_{MM} with the reversal of the propagation direction:

$$\kappa^{\mathrm{b}}_{\mathrm{MM}} = -\kappa^{\mathrm{f}}_{\mathrm{MM}}. \tag{7.68}$$

Taking into account these changes caused by the linear magneto-optic effect, the propagation constants of a TM mode in forward and backward directions of propagation are

$$\beta^{\mathrm{f}}_{\mathrm{TM}} = \beta_{\mathrm{TM}} + \kappa^{\mathrm{f}}_{\mathrm{MM}} \tag{7.69}$$

and

$$\beta^{\mathrm{b}}_{\mathrm{TM}} = -\beta_{\mathrm{TM}} - \kappa^{\mathrm{b}}_{\mathrm{MM}} = -\beta_{\mathrm{TM}} + \kappa^{\mathrm{f}}_{\mathrm{MM}}, \tag{7.70}$$

respectively, where β_{TM} represents the absolute value of the propagation constant of a TM mode in the absence of the linear magneto-optic effect. In (7.70), the value of the backward-propagation constant is chosen to be negative by following the convention defined in the coupled-mode theory of Sections 4.2 and 4.3. Because $\beta^{\mathrm{b}}_{\mathrm{TM}} \neq -\beta^{\mathrm{f}}_{\mathrm{TM}}$ in

an asymmetric waveguide where $\kappa_{MM} \neq 0$, the phase shift experienced by a TM mode propagating in such a waveguide is nonreciprocal. The nonreciprocal phase shift results from coupling between the longitudinal and transverse electric field components of a TM mode through the linear magneto-optic effect. Such a phenomenon does not exist for TE mode fields, nor does it exist for fields in bulk homogeneous media. It does not exist for TM modes of a symmetric waveguide, either.

Optical isolators

Guided-wave optical isolator can be implemented by using either a nonreciprocal TE–TM mode converter or a nonreciprocal phase shifter.

A waveguide isolator using a nonreciprocal TE–TM mode converter follows the same basic concept of a polarization-dependent optical isolator in bulk form. The key component is a 45° Faraday rotator, which has the function of turning the direction of polarization of a linearly polarized wave by 45° in a single pass and by 90° in double passes. One major difference between a waveguide Faraday rotator in the form of a nonreciprocal TE–TM mode converter and a bulk Faraday rotator has to be recognized, however.

An optical wave that is linearly polarized at the input remains linearly polarized along its path through a bulk Faraday rotator. Only its direction of polarization is rotated. A 45° Faraday rotator made in a bulk material is easily obtained by simply choosing the length l_F of the Faraday rotator properly so that $\theta_F = \rho_F l_F = \pi/4$. In comparison, implementation of a 45° Faraday rotator in a waveguide structure using a nonreciprocal magneto-optic TE–TM mode converter is not so straightforward if phase mismatch between the TE-like and TM-like modes is not completely eliminated. If perfect phase matching is accomplished so that $\Delta\beta = 0$, it can be shown from solution of the coupled-mode equations that the TE-like and TM-like modes excited by any linearly polarized input wave and coupled by the coefficient given in (7.62) remain in phase throughout the waveguide. In this situation, a 45° Faraday rotator in a waveguide structure is accomplished by choosing the length of the waveguide to be

$$l_F = \frac{\pi}{4|\rho_F|\Gamma_{EM}} \tag{7.71}$$

so that the coupling efficiency given in (7.64) has a value of $\eta(l = l_F) = 1/2$ for a single pass and a value of $\eta(l = 2l_F) = 1$ for a round-trip pass. A TE-polarized input wave will be completely converted to the orthogonal TM polarization after a round trip through such a waveguide, and vice versa, as shown in Fig. 7.23(*a*).

If any phase mismatch exists so that $\Delta\beta \neq 0$, the phase between the TE-like and TM-like components of a guided field varies along the waveguide for any input polarization. Although it is possible, in the case that $|\Delta\beta| \leq 2|\rho_F|\Gamma_{EM}$, to choose a length of the waveguide such that $\eta = 1/2$ for a single pass, it still does not make the waveguide

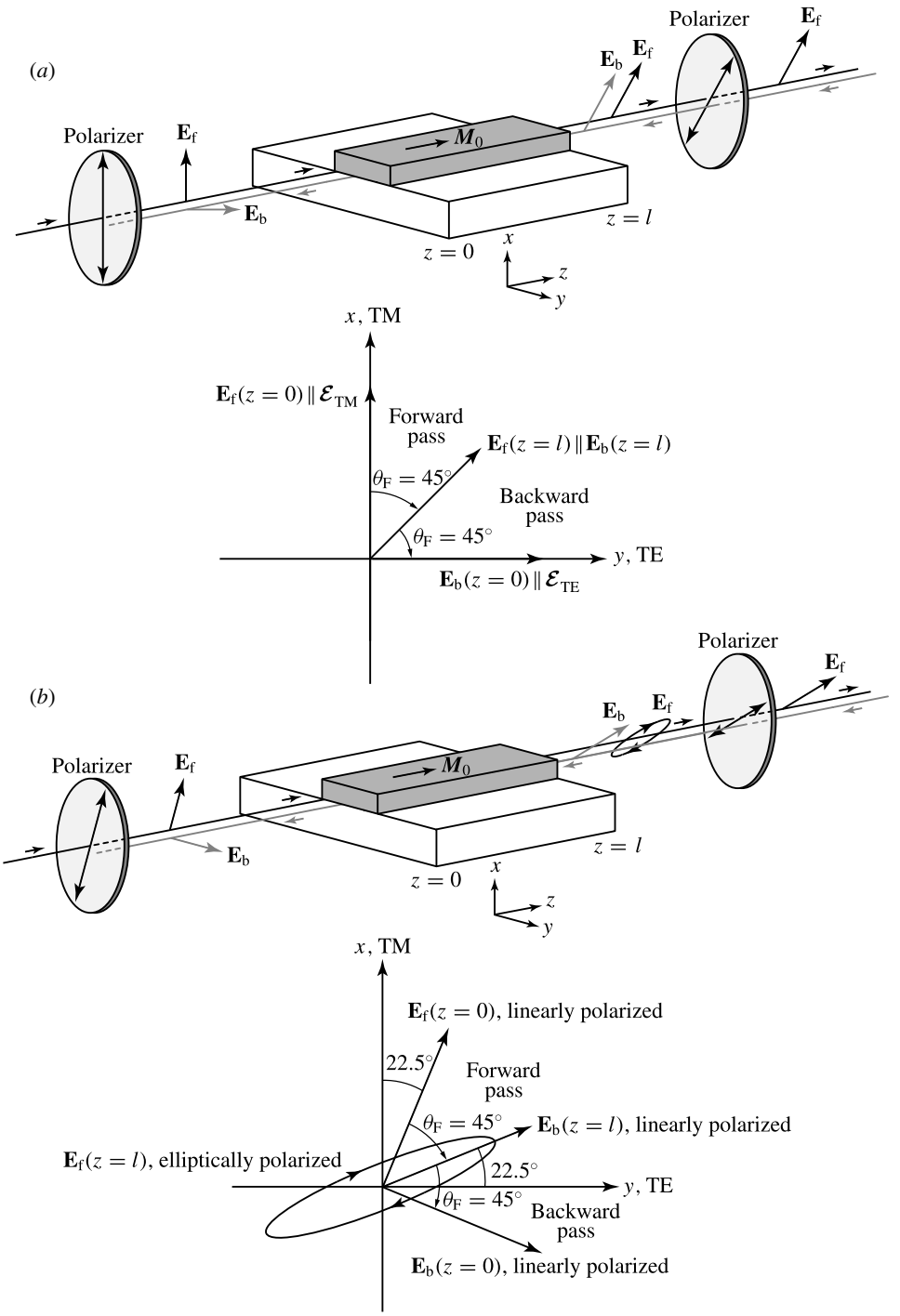


Figure 7.23 Implementation of 45° Faraday rotators using (*a*) a phase-matched nonreciprocal magneto-optic TE–TM mode converter and (*b*) a nonreciprocal magneto-optic TE–TM mode converter with a finite phase mismatch.

a 45° Faraday rotator due to the fact that its output in a single pass is elliptically polarized. Consequently, it is not possible to have a 45° Faraday rotator with a purely TE-polarized or purely TM-polarized input wave if phase mismatch exists (see Problem 7.7.4). To accomplish a high reverse isolation for an isolator, however, any light propagating in the reverse direction must be blocked by a polarizer at the input end of the device. Because an elliptically polarized wave cannot be completely blocked by a polarizer, it is necessary that a 45° Faraday rotator be used, at least for the reverse propagation direction.

A 45° Faraday rotator using a waveguide that has a phase mismatch of $\Delta\beta \neq 0$ is possible, however, if the length of the waveguide is chosen to be (see Problem 7.7.5(a))

$$l_F = \frac{1}{\left[\rho_F^2 \Gamma_{EM}^2 + (\Delta\beta)^2/4\right]^{1/2}} \tan^{-1}\left[1 + \frac{(\Delta\beta)^2}{4\rho_F^2 \Gamma_{EM}^2}\right]^{1/2}, \tag{7.72}$$

while the input polarization is chosen to be 22.5° off the TE- or TM-polarization direction on the proper side determined by the sign of ρ_F, as shown in Fig. 7.23(*b*). To make an isolator, the orientation of the output polarizer is chosen at 22.5° on the proper side so that any light that is back coupled from the output end returns to the input end linearly polarized. It can then be blocked by the input polarizer whose axis is chosen to be orthogonal to this polarization and at 45° with respect to that of the output polarizer. Therefore, perfect isolation can always be accomplished for any values of $\Delta\beta$ and ρ_F. However, because of the simultaneous presence of a nonreciprocal effect caused by ρ_F and a reciprocal effect caused by $\Delta\beta$, the waveguide does not function as a 45° Faraday rotator for a wave propagating in the forward direction with its input polarization defined by the input polarizer. As a result, the wave becomes elliptically polarized at the output, resulting in an insertion loss given by (see Problem 7.7.5(c))

$$\text{Insertion loss} = L_0 - 10 \log \frac{8\rho_F^2 \Gamma_{EM}^2}{8\rho_F^2 \Gamma_{EM}^2 + (\Delta\beta)^2}, \tag{7.73}$$

where L_0 accounts for all background insertion loss including coupling losses to the waveguide and absorption losses in the waveguide and polarizers.

EXAMPLE 7.8 A magneto-optic waveguide has the simple structure of a magnetic Bi : YIG film of 4 μm thickness on a nonmagnetic GGG substrate. The top of the Bi : YIG film is exposed to the air. The waveguide is used as a nonreciprocal TE–TM converter in a guided-wave optical isolator for 1.15 μm wavelength. At this wavelength, $n_1 = 2.178$ for the Bi : YIG film, $n_2 = 1.945$ for the GGG substrate, and $\rho_F = 280°\ \text{cm}^{-1}$, equivalent to 0.49 rad mm^{-1}. The device is operated in the fundamental TE_0 and TM_0 modes of the waveguide. It has a background insertion loss of $L_0 = 3$ dB. Find the required length of the waveguide and the total insertion loss of the isolator.

Solution First, we solve for the TE_0 and TM_0 mode parameters of the asymmetric slab waveguide with $n_1 = 2.178$, $n_2 = 1.945$, $n_3 = 1$, and $d = 4$ μm to find that $\beta_{TE} = 11.877\,17\ \mu m^{-1}$, $\beta_{TM} = 11.875\,93\ \mu m^{-1}$, and $\Gamma_{EM} = 0.999$. Therefore, $\Delta\beta = \beta_{TM} - \beta_{TE} = -1.24\ mm^{-1}$. Using these parameters, we find from (7.72) that the required length for the waveguide is

$$l_F = \frac{1}{(0.49^2 \times 0.999^2 + 1.24^2/4)^{1/2}} \tan^{-1}\left(1 + \frac{1.24^2}{4 \times 0.49^2 \times 0.999^2}\right)^{1/2} \text{mm} = 1.29 \text{ mm}.$$

Using (7.73), we find that

$$\text{Insertion loss} = 3 \text{ dB} - 10 \times \log \frac{8 \times 0.49^2 \times 0.999^2}{8 \times 0.49^2 \times 0.999^2 + 1.24^2} \text{ dB} = 5.6 \text{ dB}.$$

The phase mismatch in the waveguide contributes an additional 2.6 dB to the insertion loss.

A practical problem arises in the use of a 45° Faraday rotator in a waveguide. This rotation angle poses no problem for a bulk device, but it results in a mixture of orthogonally polarized modes at the input or the output, or both, of the waveguide that carries out this essential function for an isolator. This situation is not consistent with that in the applications of most guided-wave devices, which normally have a well-defined single TE-like or TM-like mode at both input and output ends. Some guided-wave devices are designed to function only properly for a particular mode of polarization. Therefore, a practical guided-wave optical isolator that can be integrated with other guided-wave devices must operate with single-polarization mode fields at both its input and output ends, meaning that the total amount of polarization rotation between its input and output ends has to be 0° or an integral multiple of 90°. Because nonreciprocity is required of an isolator, such a guided-wave optical isolator must be a *unidirectional TE–TM mode converter*, which converts a TE-like mode into a TM-like mode, and vice versa, in one direction of propagation but has no net polarization conversion in the opposite direction of propagation. It is not possible to construct such a unidirectional TE–TM mode converter using the linear magneto-optic effect alone, nor is it possible without using the linear magneto-optic effect. A similar function in bulk devices, shown in Fig. 7.7(*b*), is accomplished by combination of a 45° Faraday rotator and a quarter-wave plate. In guided-wave devices, such a function can be accomplished by combination of a nonreciprocal magneto-optic TE–TM mode converter functioning as a 45° Faraday rotator and a reciprocal TE–TM mode converter functioning as a 45° linear polarization rotator analogous to a quarter-wave plate. These two mode converters can be either placed in tandem or distributedly mixed.

A unidirectional TE–TM mode converter can be realized by using the reciprocal linear magnetic birefringence of the Cotton–Mouton effect for the reciprocal TE–TM mode converter. In this approach, shown in Fig. 7.24(*a*), both nonreciprocal and reciprocal

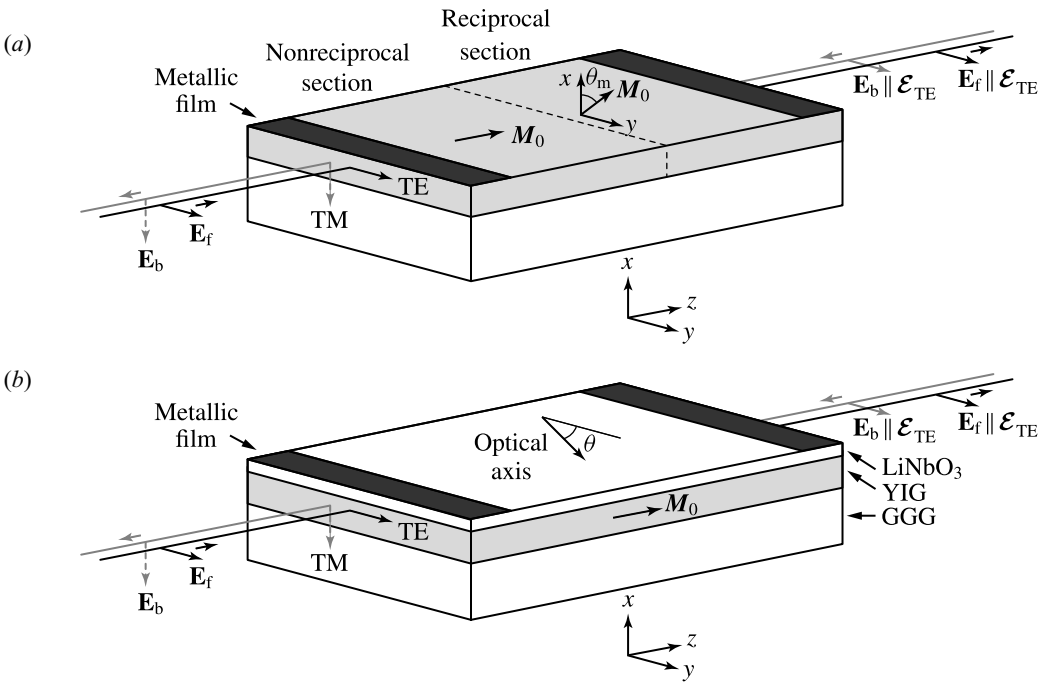


Figure 7.24 Optical isolators using unidirectional TE–TM mode converters. (*a*) The reciprocal Cotton–Mouton effect is utilized for the section of the reciprocal TE–TM mode converter placed in tandem with the nonreciprocal TE–TM mode converter. In the nonreciprocal section, $\boldsymbol{M}_0$ is in the longitudinal direction. In the reciprocal section, $\boldsymbol{M}_0$ is in the transverse xy plane and is tilted away from the mode-field polarization directions. (*b*) An anisotropic crystal with its optical axis properly tilted but lying in the yz plane is used as the top layer of the waveguide to function as a distributed reciprocal TE–TM mode converter. The metallic film sections in each device function as mode filters to transmit only TE-like modes.

TE–TM mode converters are realized using the same magnetic waveguide materials. In the nonreciprocal section the magnetization is parallel to the longitudinal direction of the waveguide, whereas in the reciprocal section it lies in the transverse plane and is tilted at an angle, θ_m, with respect to the transverse TM electric field polarization. By choosing a proper value of the angle θ_m for the magnetization direction and a proper length for the reciprocal section of the waveguide, the desired reciprocal 45° linear polarization rotation can be realized even when $\Delta\beta \neq 0$. A second approach uses an anisotropic crystal, such as $LiNbO_3$ or $LiIO_3$, for the reciprocal function. This approach can be carried out by using such an anisotropic crystal for a layer on top of the magnetic waveguide, as shown in Fig. 7.24(*b*). In this arrangement, the nonreciprocal and reciprocal TE–TM mode converters are distributedly mixed. Coupling between TE-like and TM-like modes for reciprocal conversion is caused by penetration of the mode fields into the nonmagnetic top layer. The optical axis of the anisotropic crystal has to be tilted at a proper angle away from the transverse TE and TM electric field polarization

directions in order for the TE-like and TM-like modes to have a nonzero coupling coefficient of the amount needed for the reciprocal 45° linear polarization rotation.

Other components needed in completing a guided-wave optical isolator using a unidirectional TE–TM mode converter are the input and output mode selectors analogous to the polarizers used in a bulk isolator. These components can be easily implemented with metal-loaded strips at both input and output ends of the device, as also shown in Fig. 7.24. With proper design, the metallic films on top of the waveguide can have strong attenuation for the TM-like mode but very little attenuation for the TE-like mode, based on the fact that the transverse electric field component of the TM-like mode is perpendicular to the metallic film surface but that of the TE-like mode is parallel to it. Thus the metallic film sections function as mode filters for transmitting only TE-like modes. When $LiNbO_3$ on YIG is used to implement the concept shown in Fig. 7.24(*b*), it is also possible to design the waveguide structure such that the TE-like mode is a guided mode while the TM-like mode is a leaky mode. There is no need for additional mode selectors in an isolator using such a semileaky waveguide, which already has a built-in mode-filtering function.

A very different type of guided-wave optical isolator that has no bulk counterpart uses a nonreciprocal phase shifter in an asymmetric Mach–Zehnder waveguide interferometer, as shown in Fig. 7.25. Because a nonreciprocal phase shifter is possible only for the TM-like mode, this optical isolator functions only in the TM-like mode. This device consists of two asymmetric waveguides. One waveguide has a properly oriented transverse magnetization and functions as a nonreciprocal phase shifter. The other is not magnetized and functions as a reciprocal phase shifter. In the absence of the perturbation from the linear magneto-optic effect, there is a reciprocal difference of $\Delta\beta_{TM}$ between the upper and lower waveguides in the propagation constants of the TM-like mode due to asymmetry between the two waveguides. The TM-like mode field propagating through the magnetized arm has a net reciprocal phase advance of $\Delta\varphi_{rec} = \Delta\beta_{TM} l = \pi/2$ over that propagating through the nonmagnetized arm, where l is the length of the

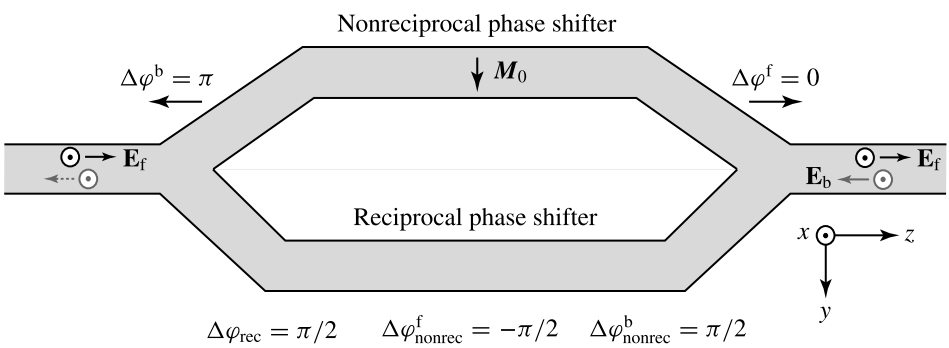


Figure 7.25 Optical isolator using a nonreciprocal phase shifter in an asymmetric Mach–Zehnder waveguide interferometer. This device allows transmission in the forward direction by constructive interference and blocks transmission in the backward direction by destructive interference.

phase-shifter section of the interferometer. The magnetized waveguide is further designed to have in the forward-propagation direction a net nonreciprocal phase shift of $\Delta\varphi^{\mathrm{f}}_{\mathrm{nonrec}} = \kappa^{\mathrm{f}}_{\mathrm{MM}} l = -\pi/2$ due to the linear magneto-optic effect and a corresponding nonreciprocal phase shift of $\Delta\varphi^{\mathrm{b}}_{\mathrm{nonrec}} = \kappa^{\mathrm{b}}_{\mathrm{MM}} l = \pi/2$ in the backward direction. The two waveguide arms are connected at both input and output ends with 3-dB Y-junctions. Consequently, in the forward direction, the combined reciprocal and nonreciprocal phase difference between the two arms is $\Delta\varphi^{\mathrm{f}} = \Delta\varphi_{\mathrm{rec}} + \Delta\varphi^{\mathrm{f}}_{\mathrm{nonrec}} = 0$, resulting in total transmission of the TM-like mode launched into the device. In the backward direction, a combined phase difference of $\Delta\varphi^{\mathrm{b}} = \Delta\varphi_{\mathrm{rec}} + \Delta\varphi^{\mathrm{b}}_{\mathrm{nonrec}} = \pi$ causes destructive interference to completely block the transmission of the TM-like mode.

Optical circulators

A guided-wave optical circulator can be realized with a nonreciprocal balanced-bridge interferometer by simply replacing the Y-junctions at the input and output ends of the nonreciprocal Mach–Zehnder interferometer shown in Fig. 7.25 with ordinary 3-dB directional couplers. The resulting device and its function are shown in Fig. 7.26. From the discussions regarding the operation of a balanced-bridge interferometer in Section 6.4, it can be easily seen by applying (6.79) and (6.80) that this nonreciprocal interferometer is in the cross state for forward propagation and is in the parallel state for backward propagation when it is designed to have $\Delta\varphi^{\mathrm{f}} = 0$ and $\Delta\varphi^{\mathrm{b}} = \pi$. Consequently, it functions as a four-port optical circulator with a $1 \to 4 \to 2 \to 3 \to 1$ looping sequence.

It is also possible to implement an optical circulator using a nonreciprocal directional coupler switch that consists of a nonreciprocal phase-shifter in one of its two coupling arms and a nonmagnetized reciprocal waveguide in another, as shown in Fig. 7.27. Because the cross state of a simple directional coupler switch is accessible only with

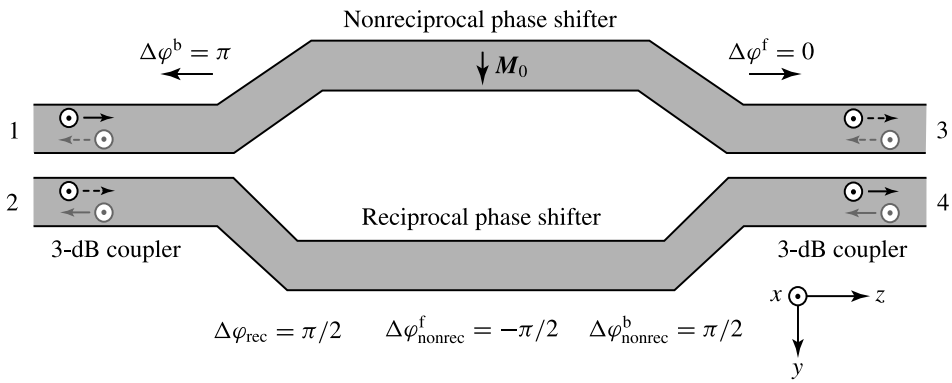


Figure 7.26 Optical circulator using a nonreciprocal phase shifter in a balanced-bridge interferometer. This device operates in the cross state in the forward direction and in the parallel state in the backward direction. The looping sequence is $1 \to 4 \to 2 \to 3 \to 1$.

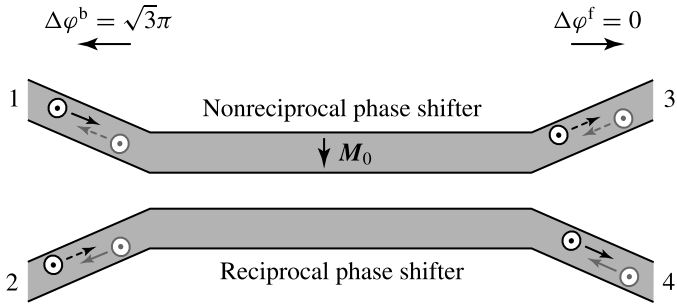


Figure 7.27 Optical circulator using a nonreciprocal phase shifter in a directional coupler switch. This device operates in the cross state in the forward direction and in the parallel state in the backward direction. The looping sequence is $1 \to 4 \to 2 \to 3 \to 1$.

a total phase difference of $\Delta\varphi = 0$ between its two arms, the length of the device has to be chosen according to (6.83), and the phase difference for reaching the parallel state is determined by (6.84). Consequently, the reciprocal and the nonreciprocal phase shifts needed for proper operation of this device are different from those needed for the nonreciprocal Mach–Zehnder and balanced-bridge interferometers. The shortest length that can be chosen for the phase-shifter section is $l = l_{\mathrm{c}}^{\mathrm{PM}}$ with a corresponding phase difference of $\Delta\varphi = \sqrt{3}\pi$ between the two coupling arms for the parallel state. With these parameters, this device should be designed to have $\Delta\varphi_{\mathrm{rec}} = \sqrt{3}\pi/2$ and $\Delta\varphi^{\mathrm{f}}_{\mathrm{nonrec}} = -\Delta\varphi^{\mathrm{b}}_{\mathrm{nonrec}} = -\sqrt{3}\pi/2$ so that $\Delta\varphi^{\mathrm{f}} = 0$ and $\Delta\varphi^{\mathrm{b}} = \sqrt{3}\pi$ for it to operate in the cross state in the forward direction and in the parallel state in the backward direction, as also shown in Fig. 7.27.

Both devices shown in Figs. 7.26 and 7.27 function only with the TM-like mode because of the use of the nonreciprocal phase shifter.

PROBLEMS

7.2.1 Diagonalize the matrix representing ϵ in (7.16) to verify that the normal modes of wave propagation in a magneto-optic medium are the three eigenvectors given in (7.18) with corresponding principal indices of refraction given in (7.20). Show that $\hat{e}_+$ and $\hat{e}_-$ form a set of orthonormal eigenmode unit vectors for the propagation of an electromagnetic wave in the z direction.

7.2.2 Linear birefringence often coexists with circular birefringence in a magneto-optic medium. When linear birefringence appears in the xy plane perpendicular to the magnetic field in the z direction, the dielectric permittivity tensor of the medium can be expressed as

$$\epsilon = \epsilon_0 \begin{bmatrix} n^2 & \mathrm{i}\xi & 0 \\ -\mathrm{i}\xi & n^2 + \zeta & 0 \\ 0 & 0 & n_z^2 \end{bmatrix}, \tag{7.74}$$

where ζ characterizes the linear birefringence that can be either positive or negative. Consider the simple case when n, ζ, and ξ are all real parameters so that both linear dichroism and circular dichroism are absent in the medium.

a. Show that the normal modes of wave propagation in this medium are

$$\hat{e}_1 = \frac{1}{\sqrt{2}}\left[\left(1+\frac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\right)^{1/2}\hat{x} + \mathrm{i}\left(1-\frac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\right)^{1/2}\hat{y}\right], \quad (7.75)$$

$$\hat{e}_2 = \frac{1}{\sqrt{2}}\left[\left(1-\frac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\right)^{1/2}\hat{x} - \mathrm{i}\left(1+\frac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\right)^{1/2}\hat{y}\right], \quad (7.76)$$

and $\hat{z}$, with corresponding principal indices of refraction given respectively by

$$n_1 = \left(n^2+\frac{\zeta-\sqrt{\zeta^2+4\xi^2}}{2}\right)^{1/2} \approx n+\frac{\zeta-\sqrt{\zeta^2+4\xi^2}}{4n}, \quad (7.77)$$

$$n_2 = \left(n^2+\frac{\zeta+\sqrt{\zeta^2+4\xi^2}}{2}\right)^{1/2} \approx n+\frac{\zeta+\sqrt{\zeta^2+4\xi^2}}{4n}, \quad (7.78)$$

and n_z.

b. Show that $\hat{e}_1$, $\hat{e}_2$, and $\hat{z}$ form a set of orthonormal normal modes for optical waves in free space.

c. Show that in the absence of linear birefringence, $\hat{e}_1$ and $\hat{e}_2$ reduce to the circularly polarized normal modes $\hat{e}_+$ and $\hat{e}_-$, respectively, while n_1 and n_2 reduce to n_+ and n_-, respectively.

d. Show that in the absence of circular birefringence, $\hat{e}_1$ and $\hat{e}_2$ reduce to the linearly polarized normal modes $\hat{x}$ and $\hat{y}$, respectively, and n_1 and n_2 reduce to n_x and n_y, respectively.

e. Represent $\hat{x}$ and $\hat{y}$ in terms of $\hat{e}_1$ and $\hat{e}_2$.

7.2.3 Show, by using the results in Problem 7.2.2, that the simultaneous effects of both linear and circular birefringence in a Faraday rotator of a length l on an optical wave of a wavelength λ traveling through the rotator is to transform an input field of $\mathbf{E}_{\text{in}} = (\mathcal{E}_x\hat{x}+\mathcal{E}_y\hat{y})_{\text{in}}\mathrm{e}^{-\mathrm{i}\omega t}$ to an output field of $\mathbf{E}_{\text{out}} = (\mathcal{E}_x\hat{x}+\mathcal{E}_y\hat{y})_{\text{out}}\mathrm{e}^{\mathrm{i}kl-\mathrm{i}\omega t}$ by the following Jones matrix:

$$\begin{bmatrix}\mathcal{E}_x\\ \mathcal{E}_y\end{bmatrix}_{\text{out}} = \begin{bmatrix}\cos\theta - \mathrm{i}\dfrac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\sin\theta & -\dfrac{2\xi}{\sqrt{\zeta^2+4\xi^2}}\sin\theta\\ \dfrac{2\xi}{\sqrt{\zeta^2+4\xi^2}}\sin\theta & \cos\theta + \mathrm{i}\dfrac{\zeta}{\sqrt{\zeta^2+4\xi^2}}\sin\theta\end{bmatrix}\begin{bmatrix}\mathcal{E}_x\\ \mathcal{E}_y\end{bmatrix}_{\text{in}}, \quad (7.79)$$

where $k = \pi(n_1 + n_2)/\lambda$ and

$$\theta = \frac{\pi\sqrt{\zeta^2 + 4\xi^2}}{2\lambda n} l. \tag{7.80}$$

7.2.4 Show that in an isotropic medium subject to $\boldsymbol{H}_0$ if an optical wave propagates along an arbitrary direction with respect to the direction of $\boldsymbol{H}_0$, the angle of the Faraday rotation is that of (7.25) but with H_{0z} in the formula replaced by the projection of $\boldsymbol{H}_0$ on the path of propagation. How should (7.25) be modified if the isotropic medium is ferromagnetic and the optical wave does not propagate in a direction parallel or antiparallel to $\boldsymbol{M}_0$?

7.2.5 If a linearly polarized optical wave at 1.064 μm is sent through the TGG Faraday rotator described in Example 7.1, what is the Faraday rotation angle in a single pass? With the given crystal length, if a Faraday rotation angle of 45° in a single pass is desired for the wave, what should the magnetic flux be? Answer the same questions for a linearly polarized wave at 750 nm wavelength.

7.2.6 Fully magnetized YIG, YbBi : YIG, Bi : YIG, and Ce : YIG listed in Table 7.2 are used to make Faraday rotators at 1.55 μm wavelength. Compare the performance characteristics of these rotators by finding the length of each crystal needed for a Faraday rotation angle of 45° and the corresponding loss due to absorption in each rotator. If the crystals are antireflection coated so that there are no reflection losses at the input and output surfaces, what is the transmittance of each Faraday rotator?

7.2.7 Iron has a very large specific Faraday rotation, but it cannot be used to make useful Faraday rotators. Use the data given in Table 7.2 to give a numerical explanation for the possible difficulty. Are the other ferromagnets listed in Table 7.2 any better than iron?

7.2.8 Show, using (7.23) and the relations defined in (1.66) and (1.67), that an initially linearly polarized wave entering a medium of circular dichroism, where n_+ and n_- are complex, becomes elliptically polarized with a Faraday rotation angle of θ_F and a Faraday ellipticity of ε_F given in (7.29) and (7.30), respectively, after the optical wave propagates a distance l. Show that the Faraday ellipticity can be approximated as

$$\varepsilon_F \approx \text{Im}\left[\frac{\pi\xi}{\lambda n_\perp} l\right] \tag{7.81}$$

when $|\varepsilon_F| \ll 1$. Compare the Faraday rotation angle and the Faraday ellipticity to the Kerr rotation angle and the Kerr ellipticity given in (7.38).

7.2.9 How should the expressions for the Verdet constant and the specific Faraday rotation given in (7.26) and (7.28), respectively, be modified for a material that has circular dichroism?

7.2.10 At 546 nm wavelength, iron has a complex index of refraction of $n = 2.73 + i3.3$ and, when fully magnetized, a complex linear magneto-optic constant of $\xi = -0.18 + i0.74$ at room temperature.

a. Find the absorption coefficient, measured per meter and in decibels per centimeter, of iron at 546 nm. Compare the result to the data given in Table 7.2.

b. Find the specific Faraday rotation of fully magnetized iron at 546 nm. Compare the result to the data given in Table 7.2.

c. A linearly polarized optical wave at 546 nm is normally incident on the surface of fully magnetized iron and propagates in the direction of the magnetization. What are the Faraday rotation angle θ_F and the Faraday ellipticity ε_F of the transmitted wave after it travels a distance of the skin depth, $d_{skin} = 1/\alpha$?

d. Compare the values of θ_F and ε_F to those of θ_K and ε_K found in Example 7.2.

7.2.11 In this problem, we study the difference between the Faraday effect and natural optical activity. A linearly polarized plane wave propagating along a fixed z axis in the medium is considered. Circular polarizations, $\hat{e}_+$ and $\hat{e}_-$, are the normal modes of the wave propagation.

a. Show that the sense of Faraday rotation is independent of the forward or backward direction of wave propagation. (As a result, the amount of the Faraday rotation is doubled instead of being canceled when an optical wave passing through a magneto-optic material is reflected to traverse its original path in the opposite direction back to the input end.)

b. Show that the sense of polarization rotation caused by natural optical activity is reversed when the direction of wave propagation is reversed. (Therefore, the net amount of the polarization rotation is zero when an optical wave passing through a naturally optically active material is reflected to traverse its original path in the opposite direction back to the input end.)

7.2.12 Name two physical effects that can make a linearly polarized wave rotate polarization while always retaining linearly polarized as the wave propagates through a medium. Experimentally, how do you tell the difference between these two effects?

7.3.1 At 1 μm wavelength, nickel has a complex index of refraction of $n = 2.1 + i5.1$ and, when fully magnetized, a complex linear magneto-optic constant of $\xi = 0.23 + i0.4$ at room temperature. A linearly polarized optical wave at 1 μm is normally incident on a fully magnetized nickel surface in the polar configuration shown in Fig. 7.3.

a. Find the absorption coefficient, measured per meter and in decibels per centimeter, of nickel at 1 μm.

b. Find the specific Faraday rotation of fully magnetized nickel at 1 μm. What is the figure of merit in degrees per decibel.

c. Find the Faraday rotation angle θ_F and the Faraday ellipticity ε_F of the transmitted wave after it travels a distance of the skin depth, $d_{skin} = 1/\alpha$.
d. Find the Kerr rotation angle θ_K and the Kerr ellipticity ε_K of the reflected wave.

7.3.2 Verify the relations in (7.33) using the boundary conditions for the electromagnetic fields at the interface between free space and a magnetic medium.

7.3.3 The approximation used in (7.20) for n_+ and n_- and the conditions $|n_\perp - 1| > 1$ and $|n_\perp + 1| > 1$ are valid for magneto-optic Kerr materials of interest.
a. Show that r_+ and r_- given in (7.33) can be expressed in terms of the Kerr rotation angle and the Kerr ellipticity as follows:

$$r_+ \approx (1 + \varepsilon_K - i\theta_K)r, \tag{7.82}$$

$$r_- \approx (1 - \varepsilon_K + i\theta_K)r, \tag{7.83}$$

where

$$r = \frac{1 - n_\perp}{1 + n_\perp}. \tag{7.84}$$

b. Show that the reflectivities, $R_+ = |r_+|^2$ and $R_- = |r_-|^2$, for the two circularly polarized modes have the following relation:

$$R_+ - R_- = 4\varepsilon_K R, \tag{7.85}$$

where $R = |r|^2$.

7.3.4 Use the parameters of iron at 546 nm from Problem 7.2.10 to find the value of $R_+ - R_-$ for circularly polarized waves at 546 nm that are normally incident on a fully magnetized iron surface in the polar configuration. Use the parameters of nickel at 1 μm from Problem 7.3.1 to find the value of $R_+ - R_-$ for circularly polarized waves at 1 μm that are normally incident on a fully magnetized nickel surface in the polar configuration.

7.4.1 Show that the insertion loss and the reverse isolation of a polarization-dependent isolator of the structure shown in Fig. 7.7(*a*) can be expressed as (7.43) and (7.44), respectively. Show that it is possible to minimize the insertion loss while maximizing the reverse isolation at the same time by setting $\theta_F = \theta_p = 45^\circ$. What are the expressions for the minimum insertion loss and the maximum reverse isolation?

7.4.2 An optical isolator of the structure shown in Fig. 7.7(*a*) is desired to have an insertion loss less than 1 dB and a reverse isolation larger than 30 dB. The input and output polarizers have the same extinction ratio. The Faraday rotator generates both a rotation angle θ_F and an ellipticity ε_F for a linearly polarized input optical wave in a single pass. The device has a background optical loss of L_0.

a. If $\theta_F = 45^\circ$, what are the required values of the background loss, the polarizer extinction ratio, and the Faraday ellipticity that allow the device to satisfy the stated performance specifications?
b. If $\theta_F = 40^\circ$ instead, how can the stated performance specifications be met?

7.4.3 An optical isolator of the structure shown in Fig. 7.7(*a*) has a specified maximum insertion loss of L (dB) and a specified minimum reverse isolation of I (dB), where $I \gg L$. The Faraday rotator generates a single-pass rotation angle of $\theta_F = 45^\circ + \Delta\theta_F$, which varies with the operating condition. The Faraday ellipticity is within the limit that allows the device to meet its performance specifications. Show that to meet the specifications, the following condition has to be satisfied:

$$|\Delta\theta_F| < \frac{1}{2}\cos^{-1}\left[10^{(L_0-L)/20}\right], \tag{7.86}$$

where L_0 is the background optical loss defined in (7.43) and (7.44).

7.4.4 The wavelength dependence of the Faraday rotation angle of YbBi : YIG can be approximated by

$$\frac{1}{\theta_F}\frac{\partial\theta_F}{\partial\lambda} = -3.6\ \mu\text{m}^{-1}. \tag{7.87}$$

If an optical isolator consisting of a YbBi : YIG Faraday rotator has a background optical loss of $L_0 = 0.5$ dB and a high reverse isolation of 60 dB, find the bandwidth of its operation for an insertion loss that is kept below 1 dB. If the device is designed for a center wavelength of 1.3 μm, what is its useful wavelength range of application? To maintain maximum reverse isolation, how much should the output polarizer angle θ_p be varied as the wavelength is varied over this wavelength range? Assume that the background loss is not wavelength dependent within the wavelength range of interest.

7.4.5 An optical isolator consisting of a YbBi : YIG Faraday rotator as described in Problem 7.4.4 has an insertion loss of 0.5 dB and a reverse isolation of 60 dB with $\theta_F = 45^\circ$ at 1.3 μm wavelength at 300 K. The output polarizer angle of the isolator is fixed at $\theta_p = 45^\circ$ and cannot be adjusted. Assume that the background loss remains constant within this wavelength range and that ε_F is negligibly small.

a. What are the variations in the insertion loss and in the reverse isolation over a wavelength range between 1.28 and 1.32 μm?
b. The temperature dependence of the Faraday rotation angle of YbBi : YIG can be approximated by

$$\frac{1}{\theta_F}\frac{\partial\theta_F}{\partial T} = -4.2 \times 10^{-4}\ \text{K}^{-1}. \tag{7.88}$$

What are the variations in the insertion loss and in the reverse isolation for a temperature variation of ±20 K?

7.4.6 Verify the function of the device shown in Fig. 7.11 as a polarization-dependent optical circulator and the function of the device shown in Fig. 7.12 as a polarization-independent optical circulator.

7.5.1 For a magneto-optic amplitude modulator of the basic configuration shown in Fig. 7.7(*a*), find an expression for the transmittance to replace (7.45) in the case when linear birefringence coexists with magneto-optic circular birefringence in the Faraday rotator as described in Problem 7.2.2. The birefringent phase factor and the Faraday rotation angle resulting from the linear birefringence and the circular birefringence, respectively, are

$$\varphi = \frac{\omega\zeta}{2nc}l = \frac{\pi\zeta}{\lambda n}l, \qquad \theta_{\mathrm{F}} = \frac{\omega\xi}{2nc}l = \frac{\pi\xi}{\lambda n}l. \tag{7.89}$$

Show that when θ_{p} is chosen to be 45° in this situation, we have, in place of (7.46), the following transmittance:

$$T = \frac{T_0}{2}\mathrm{e}^{-\alpha l}\left(1 + \frac{2\theta_{\mathrm{F}}}{\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}}\sin\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}\right). \tag{7.90}$$

7.5.2 The Faraday rotator of a length l in a magneto-optic amplitude modulator of the dual-quadrature polarimetric configuration shown in Fig. 7.13 has both linear birefringence and magneto-optic circular birefringence, as described in Problem 7.2.2. For an optical wave of wavelength λ, the birefringent phase factor and the Faraday rotation angle resulting from the linear birefringence and the circular birefringence, respectively, are

$$\varphi = \frac{\omega\zeta}{2nc}l = \frac{\pi\zeta}{\lambda n}l, \qquad \theta_{\mathrm{F}} = \frac{\omega\xi}{2nc}l = \frac{\pi\xi}{\lambda n}l. \tag{7.91}$$

a. In the case when a linearly polarized beam is launched through the input polarizer and θ_{p} is chosen to be 45°, show that

$$\mathrm{DS} = \frac{2\theta_{\mathrm{F}}}{\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}}\sin\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}. \tag{7.92}$$

b. In the case when a circularly polarized beam is launched in the absence of the input polarizer, show that

$$\mathrm{DS} = \frac{\varphi}{\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}}\sin\sqrt{\varphi^2 + 4\theta_{\mathrm{F}}^2}. \tag{7.93}$$

7.5.3 The dynamic range of the fiber-optic current sensor described in Example 7.4 is expanded to 3 kA at the upper end. If the number of turns of fiber is kept at 20, what is the largest linearity error? In what current range is the linearity error below 1%? If the linearity error is to be kept below 1% over the entire

dynamic range from 1 A to 3 kA by reducing the number of fiber turns, what is the maximum number of turns allowed? What is the smallest Faraday rotation angle to be measured in this situation?

7.5.4 A linked fiber-optic current sensor of the dual-quadrature polarimetric configuration shown in Fig. 7.13 consists of a 2-m-long silica fiber wound around the conductor. At the 632.8 nm wavelength of the light source, the refractive index of silica is $n = 1.457$. The largest Faraday rotation angle corresponding to the maximum current to be measured is 100 mrad. If there is linear birefringence in the fiber, find the largest value of the birefringence ζ that is allowed for a reduction of the difference signal DS by less than 10%. What is the corresponding birefringence measured as the difference in the refractive indices between the two birefringent axes? What are the value of ζ and the corresponding refractive index birefringence for a 50% reduction of the difference signal DS?

7.5.5 An unlinked current sensor for a dynamic range from 10 μA to 1 A consists of a TGG Faraday rotator in a 1000-turn solenoid in a configuration shown in Fig. 7.15(*c*). A semiconductor laser at 750 nm wavelength is used as the light source. What is the range of Faraday rotation angle that the detection system has to be designed to measure in order to cover the dynamic range of this device?

7.5.6 Show, using (7.51), that the magnetic film of a magneto-optic spatial light modulator has the following optimum thickness for the highest ON-state transmittance:

$$l_{\text{opt}} = \frac{1}{2\rho_{\text{F}}} \tan^{-1} \frac{4\rho_{\text{F}}}{\alpha}, \tag{7.94}$$

where ρ_{F} and α are, respectively, the specific Faraday rotation and the absorption coefficient of the magnetic film.

7.5.7 A magneto-optic spatial light modulator is constructed with a Bi : YIG film on LLC substrate between a polarizer and an analyzer. The magnetic film has $\rho_{\text{F}} = 2.58 \times 10^{-2}$ rad μm^{-1} and $\alpha = 0.086$ μm^{-1} at 632.8 nm wavelength and $\rho_{\text{F}} = 1.17 \times 10^{-2}$ rad μm^{-1} and $\alpha = 0.013$ μm^{-1} at 788 nm wavelength. All surfaces are broadband antireflection coated so that $T_0 = 0.95$ at both wavelengths due to the absorption of the polarizer and analyzer.

a. Find the optimum thickness of the film for the highest ON-state transmittance at 632.8 nm. With this film thickness, what are the values of T_{ON} at 632.8 and 788 nm, respectively?

b. Find the optimum thickness of the film for the highest ON-state transmittance at 788 nm. With this film thickness, what are the values of T_{ON} at 632.8 and 788 nm, respectively?

7.5.8 Show that the OFF-state transmittance of a magneto-optic spatial light modulator is that given in (7.52) when the reflectivities of the rotator surfaces

are considered but the circular dichroism is ignored. With the result, show that the contrast ratio of this device is that given in (7.53).

7.5.9 The rotator surfaces of the magneto-optic spatial light modulator described in Problem 7.5.7 are antireflection coated to have residual reflectivities of $R_1 = R_2 = 1\%$ at 632.8 nm and $R_1 = R_2 = 0.5\%$ at 788 nm.

a. If the film thickness is optimized for 632.8 nm wavelength, what are the OFF-state transmittance and the contrast ratio of the device at 632.8 nm? What are they at 788 nm?

b. If the film thickness is optimized for 788 nm wavelength, what are the OFF-state transmittance and the contrast ratio of the device at 632.8 nm? What are they at 788 nm?

7.6.1 Magneto-optic recording of very high areal bit densities at 15.5 Gbit cm^{-2} (100 Gbit in^{-2}) and beyond is possible. It is accomplished using a blue laser at 488 nm wavelength and a specially designed super-Solid Immersion Lens that has a numerical aperture of 1.8.

a. What are the theoretical laser spot size on the medium and the areal bit density according to (7.55)? The measured spot size is actually 182 nm. Based on this spot size, what is the areal bit density according to (7.55)?

b. For an areal bit density of 15.5 Gbit cm^{-2} (100 Gbit in^{-2}), what is the required mark area?

c. To accomplish the high bit density of 15.5 Gbit cm^{-2} (100 Gbit in^{-2}), a center aperture detection magnetic super-recording (CAD-MSR) medium, which allows optical reading of a very small magnetic domain, is used. A portion of the focused laser beam is optically masked to create an effective aperture smaller than the diffraction-limited laser spot size. The bit spacing is then reduced to 40 nm. What track pitch, which is the spacing between neighboring tracks, should be chosen to accomplished a bit density of 15.5 Gbit cm^{-2} (100 Gbit in^{-2})?

7.6.2 A magneto-optic recording system uses a medium that has a reflectance of $R = 0.6$ and a Kerr rotation angle of $\theta_K = 0.3^\circ$. The read laser power is 1 mW, and the differential photodetectors have a responsivity of $\mathcal{R} = 0.35$ A W^{-1}.

a. What is the shot-noise-limited SNR of the system if the system has a bandwidth of 10 MHz?

b. A large bandwidth of 100 MHz is desired. If the read and detection system is kept unchanged but a different medium can be used, what should the Kerr rotation angle of the medium be in order to maintain a shot-noise-limited SNR of at least 30 dB?

7.7.1 An optical wave is launched into a guided-wave TE–TM mode converter as a TE-like mode at the input end and is allowed to make round-trip passes in the waveguide.

a. Show that if the device is a nonreciprocal Faraday TE–TM mode converter, the coupling efficiency to the TM-like mode is given by (7.64) with l being the cumulative length traveled by the optical wave in both directions in the waveguide.
b. Show that if the device is a reciprocal TE–TM mode converter, the coupling efficiency to the TM-like mode is also given by (7.64) but with l being the distance from the input end to the point where the conversion efficiency is being evaluated rather than the cumulative traveling length over multiple passes in the waveguide.
c. Identify the root of the difference between the nonreciprocal and the reciprocal TE–TM mode converters.

7.7.2 Find the conversion efficiency of a nonreciprocal Faraday TE–TM mode converter in a single pass through a waveguide of length l when a finite phase mismatch exists between the TE and TM modes. Find the conversion efficiency in a round trip over such a waveguide. Show that both single-pass and round-trip efficiencies reduce to the expression given in (7.64) in the situation of perfect phase matching. Can the round-trip efficiency be obtained by simply doubling the length in the single-pass efficiency in the presence of a phase mismatch? Can this be done in the situation of perfect phase matching? Give a physical explanation to the answers.

7.7.3 In the absence of the linear magneto-optic effect, the mode characteristics of the planar waveguide shown in Fig. 7.22 are simply those discussed in Section 2.5. We have seen that the waveguide functions as a nonreciprocal TE–TM mode converter if the waveguide core has a magnetization $\boldsymbol{M}_0 = M_{0z}\hat{z}$ and that it functions as a nonreciprocal phase shifter for the TM mode if $\boldsymbol{M}_0 = M_{0y}\hat{y}$.
a. Verify that $\kappa_{\mathrm{EE}} = \kappa_{\mathrm{MM}} = 0$ and $\kappa_{\mathrm{EM}} = \kappa_{\mathrm{ME}}^{*} \approx -\mathrm{i}\rho_{\mathrm{F}}\Gamma_{\mathrm{EM}}$, as given in (7.62), if $\boldsymbol{M}_0 = M_{0z}\hat{z}$.
b. With $\boldsymbol{M}_0 = M_{0y}\hat{y}$ as shown in Fig. 7.22, show that $\kappa_{\mathrm{EM}} = \kappa_{\mathrm{ME}} = \kappa_{\mathrm{EE}} = 0$ and k_{MM} given by (7.67) has a nonzero value only for asymmetric magnetic waveguides.
c. What are the coupling coefficients if the magnetization in the waveguide core is perpendicular to the boundary surfaces of the planar waveguide, i.e., $\boldsymbol{M}_0 = M_{0x}\hat{x}$? Can you give a physical argument to explain the results obtained in this case?

7.7.4 In this problem, we consider the effect of phase mismatch between the TE and TM modes in a nonreciprocal magneto-optic TE–TM mode converter that is used as a guided-wave Faraday rotator.
a. Show that if a finite phase mismatch exists, so that $\Delta\beta \neq 0$, an optical wave launched at one end of the waveguide as a linearly polarized wave in the purely TE or purely TM mode does not appear at the other end linearly

polarized even when the length of the waveguide is chosen such that the power conversion efficiency is $\eta = 1/2$. Therefore, the waveguide does not function as a 45° Faraday rotator. What is the phase difference between the TE and TM mode fields at the waveguide output?

b. If the wave makes a round trip in the waveguide after being launched as a linearly polarized wave in a purely TE or purely TM mode, is it linearly polarized after it returns to the input end in the case when $\Delta\beta \neq 0$ but l is chosen so that $\eta = 1/2$?

c. Does the mode converter function as a 90° rotator in a round trip?

7.7.5 A magneto-optic TE–TM mode converter can be used to construct a guided-wave optical isolator even when the TE and TM modes are not phase matched. The consequence is an increased insertion loss caused by the phase mismatch.

a. Show that a guided-wave 45° Faraday rotator can be constructed if the length of the waveguide is chosen to be that given in (7.72) while a linearly polarized wave is launched at one end with its polarization being 22.5° off the TE- or TM-polarization direction on one side, which is determined by the sign of ρ_F.

b. Does this rotator function as a 90° rotator in a round trip?

c. Show that if the input and output polarizers of the isolators are arranged to maximize the reverse isolation, then the insertion loss of the device is that given in (7.73).

7.7.6 A guided-wave optical isolator for 1.15 μm wavelength is based on a magneto-optic waveguide that has the structure of the one described in Example 7.8 but has a magnetic Bi : YIG film of 2 μm thickness. Solving for the waveguide structure results in the following parameters for the TE_0 and TM_0 modes: $\beta_{TE} = 11.819\,98\ \mu m^{-1}$, $\beta_{TM} = 11.811\,73\ \mu m^{-1}$, and $\Gamma_{EM} = 0.994$. This device is operated in the fundamental TE_0 and TM_0 modes of the waveguide and has a background insertion loss of $L_0 = 3$ dB. Find the required length of the waveguide and the total insertion loss of the isolator. Compare the characteristics of this device with those of the one described in Example 7.8.

SELECT BIBLIOGRAPHY

Altman, C. and Suchy, K., *Reciprocity, Spatial Mapping and Time Reversal in Electromagnetics*. Dordrecht, The Netherlands: Kluwer Academic, 1991.

Bradley, A., *Optical Storage for Computers: Technology and Applications*. Chichester: Ellis Horwood, 1989.

Eremenko, V. V., Kharchenko, N. F., Litvinenko, Yu. G., and Naumenko, V. M., *Magneto-Optics and Spectroscopy of Antiferromagnets*. New York: Springer-Verlag, 1992.

Iizuka, K., *Elements of Photonics in Free Space and Special Media*, Vol. I. New York: Wiley, 2002.

Landau, L. D. and Lifshitz, E. M., *Electrodynamics of Continuous Media*. Oxford: Pergamon Press, 1960.

Levi, L., *Applied Optics: A Guide to Optical System Design*. Vols. 1 and 2. New York: Wiley, 1980.

Nishihara, H., Haruna, M., and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Nye, J. F., *Physical Properties of Crystals*. London: Oxford University Press, 1957.

Paoletti, A., ed., *Physics of Magnetic Garnets*. Amsterdam: North-Holland, 1978.

Parker, S. P., *Optical Source Book*. New York: McGraw-Hill, 1987.

Post, E. J., *Formal Structure of Electromagnetics*. Amsterdam: North-Holland, 1962.

Schwartz, K., *The Physics of Optical Recording*. New York: Springer-Verlag, 1993.

Sirotin, Yu. I. and Shaskolskaya, M. P., *Fundamentals of Crystal Physics*. Moscow: Mir Publishers, 1982.

Smit, J., ed., *Magnetic Properties of Materials*. New York: McGraw-Hill, 1971.

Sugano, S. and Kojima, N., eds., *Magneto-Optics*. Berlin: Springer, 2000.

Winkler, G., ed., *Magnetic Garnet*. Braunschweig: Vieweg und Sohn, 1981.

ADVANCED READING LIST

Ando, K., "Nonreciprocal devices for integrated optics," *Proceedings of the International Society for Optical Engineering* **1126**: 58–65, 1989.

Blendelli, G. and Donati, S., "Optical isolators for telecommunications: review and current trends," *European Transactions on Telecommunications and Related Technologies* **3**(4): 373–380, July–Aug. 1992.

Carey, R., Newman, D. M., and Thomas, B. W. J., "Magneto-optic recording," *Journal of Physics D* **28**(11): 2207–2227, Nov. 1995.

Castera, J. P. and Meunier, P. L., "Nonreciprocal devices in integrated optics," *Fiber and Integrated Optics* **8**(1): 71–85, 1989.

Chang, K. W., Schmidt, S., Sorin, W. V., Yarnell, J. L., Chou, H., and Newton, S. A., "A high-performance optical isolator for lightwave systems," *Hewlett–Packard Journal*: 45–50, Feb. 1991.

Davis, J. A., and Waas, J. M., "Current status of the magneto-optic spatial light modulator," *Proceedings of the International Society for Optical Engineering* **1150**: 27–43, 1990.

Freiser, M. J., "A survey of magnetooptic effects," *IEEE Transactions on Magnetics* **MAG-4**(2): 152–161, June 1968.

Hansen, P., "Magneto-optic recording materials and technologies," *Journal of Magnetism and Magnetic Materials* **83**: 6–12, 1990.

Ishikawa, H., Nakajima, K., Machida, K., and Tanii, A., "Optical isolators using Bi-substituted rare-earth iron garnet films," *Optical and Quantum Electronics* **22**: 517–528, 1990.

Koga, M. and Matsumoto, T., "High-isolation polarization-insensitive optical circulator for advanced optical communication systems," *Journal of Lightwave Technology* **10**(9): 1210–1217, Sep. 1992.

Kryder, M. H., "Magneto-optic recording technology," *Journal of Applied Physics* **57**(1): 3913–3918, Apr. 1985.

"Advances in magneto-optic recording technology," *Journal of Magnetism and Magnetic Materials* **83**: 1–5, 1990.

Ning, Y. N., Wang, Z. P., Palmer, A. W., Grattan, K. T. V., and Jackson, D. A., "Recent progress in optical current sensing techniques," *Review of Scientific Instruments* **66**(5): 3097–3111, May 1995.

Pershan, P. S., "Magneto-optical effects," *Journal of Applied Physics* **38**(3): 1482–1490, Mar. 1967.

Qiu, Z. Q. and Bader, S. D., "Surface magneto-optic Kerr effect," *Review of Scientific Instruments* **71**(3): 1243–1255, Mar. 2000.

Robinson, C. C., "Electromagnetic theory of the Kerr and the Faraday effects for oblique incidence," *Journal of the Optical Society of America* **54**(10): 1220–1224, Oct. 1964.

Schmitt, H. J., "Magneto-optic devices," *Proceedings of the International Society for Optical Engineering* **1274**: 208–219, 1990.

Zak, J., Moog, E. R., Liu, C., and Bader, S. D., "Universal approach to magneto-optics," *Journal of Magnetism and Magnetic Materials* **89**: 107–123, 1990.

8 Acousto-optic devices

Scattering of light by acoustic waves was first investigated by Brillouin. The acoustic frequencies involved in Brillouin scattering fall in the ultrasonic and hypersonic regions. Hypersonic waves in a medium are caused by thermal excitation, whereas ultrasonic waves can be excited electronically using piezoelectric transducers. The acoustic waves used in acousto-optics are generally ultrasonic waves that have frequencies in the range between about 100 kHz and a few gigahertz. The basic principles of acousto-optic devices are based on the scattering of light by the periodic index variations generated by an acoustic wave in the supporting medium. These periodic index variations form a moving index grating, generated by a traveling acoustic wave, or a standing index grating, generated by a standing acoustic wave. Because the speed of sound in dielectric media that are used for device applications typically falls in the range between 1 and 10 km s^{-1}, the index gratings generated by ultrasonic acoustic waves have grating periods ranging from the order of 1 μm to a few centimeters. A unique property of the index grating created by an acoustic wave is that its period and modulation depth can be varied by varying the frequency and amplitude, respectively, of the acoustic wave through variation in the electronic signal applied to the transducer. Therefore, the operating parameters of an acousto-optic device can be controlled electronically. Practical acousto-optic devices include modulators, beam deflectors, frequency shifters, couplers, switches, and spectrum analyzers.

8.1 Elastic waves

An acoustic wave in a medium is an elastic wave of space- and time-dependent periodic deformation in the medium. A *traveling* plane acoustic wave can be expressed as

$$\mathbf{u}(\mathbf{r}, t) = \mathcal{U} \cos(\mathbf{K} \cdot \mathbf{r} - \Omega t). \tag{8.1}$$

A *standing* plane acoustic wave is a combination of two contrapropagating traveling waves of equal amplitude, wavelength, and frequency. It can be described by

$$\mathbf{u}(\mathbf{r}, t) = \mathcal{U} \cos(\mathbf{K} \cdot \mathbf{r}) \cos \Omega t. \tag{8.2}$$

In (8.1) and (8.2), $\mathbf{u}(\mathbf{r}, t)$ represents the time-dependent displacement of the point at $\mathbf{r}$ in the medium subject to deformation, $\mathcal{U}$ is the amplitude of the elastic wave, and $\mathbf{K}$ and $\Omega = 2\pi f$ are its wavevector and angular frequency, respectively. The physical meaning of the displacement vector $\mathbf{u}(\mathbf{r}, t)$ is that in a fixed coordinate system, an atom, or an ion, in the medium at location $\mathbf{r}$ before deformation is moved to location $\mathbf{r} + \mathbf{u}$ under deformation caused by the elastic wave. Therefore, $\mathbf{u}(\mathbf{r}, t)$ describes the motion of the particles in the medium supporting the acoustic wave. The direction of the amplitude vector $\mathcal{U}$ defines the polarization of the acoustic wave, while that of the wavevector $\mathbf{K}$ describes the direction of propagation of the wave. The two contrapropagating traveling waves that form the standing wave expressed in (8.2) propagate in $\mathbf{K}$ and $-\mathbf{K}$ directions, respectively. The magnitude of the wavevector is

$$K = \frac{2\pi}{\Lambda} = \frac{\Omega}{v_a} = \frac{2\pi f}{v_a}, \tag{8.3}$$

where Λ and v_a are the wavelength and phase velocity, respectively, of the acoustic wave.

For any given direction of propagation of an acoustic wave in any medium, there are *three* orthogonal normal modes of polarization. If one mode is polarized along the direction of $\mathbf{K}$, the directions of polarization of the other two modes are perpendicular to $\mathbf{K}$. An acoustic wave polarized in the direction of $\mathbf{K}$ is known as a *longitudinal wave*, while one with a polarization perpendicular to $\mathbf{K}$ is called a *transverse wave* or a *shear wave*. In isotropic media and cubic crystals, the three normal modes are always one purely longitudinal and two purely transverse for any direction of acoustic wave propagation. In anisotropic crystals other than those in the triclinic system, the normal modes again consist of one purely longitudinal wave and two purely transverse waves if the acoustic wave propagates along a crystal axis of two-, three-, four-, or six-fold symmetry. In general, however, the polarization directions of the normal modes of an acoustic wave in an anisotropic crystal are not necessarily parallel or perpendicular to the direction of wave propagation. Then, a mode that has its polarization close to the direction of $\mathbf{K}$ is called *quasi-longitudinal*, and one whose polarization is close to being perpendicular to $\mathbf{K}$ is called *quasi-transverse*. Figure 8.1 illustrates the characteristics of different modes of acoustic waves. At a given acoustic frequency in a given medium, the acoustic velocity v_a, and, consequently, the wavelength Λ, and the value of K all depend on both the direction of propagation and that of the polarization of an acoustic wave. In an isotropic medium, the two transverse modes are degenerate, meaning that they have the same acoustic velocity, but they are generally not degenerate with the longitudinal mode. In a cubic crystal, the two transverse modes are degenerate only for waves propagating along certain directions, such as the [100] and [111] directions of the crystal. In anisotropic crystals, all three normal modes are generally nondegenerate.

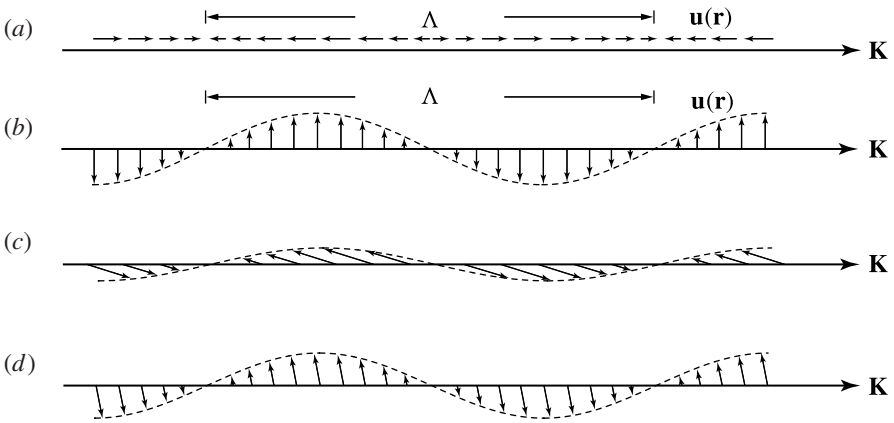


Figure 8.1 Spatial variations of displacement vectors for (*a*) longitudinal acoustic wave, (*b*) transverse acoustic wave, (*c*) quasi-longitudinal acoustic wave, and (*d*) quasi-transverse acoustic wave.

Deformation of a medium can be characterized by a second-rank *displacement gradient tensor* defined by

$$\frac{\partial u_i}{\partial x_j}, \tag{8.4}$$

where the indices $i, j = 1, 2, 3$ represent the coordinates x, y, z. The mechanical strains associated with deformation are described by a symmetric *strain tensor*, $\mathbf{S} = [S_{ij}]$, defined by

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{8.5}$$

The three tensor elements S_{xx}, S_{yy}, and S_{zz} are *tensile strains*, while the other elements $S_{yz} = S_{zy}$, $S_{zx} = S_{xz}$, and $S_{xy} = S_{yx}$ are *shear strains*. In addition, there is an antisymmetric *rotation tensor*, $\mathbf{R} = [R_{ij}]$, defined by

$$R_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right). \tag{8.6}$$

Clearly, $R_{xx} = R_{yy} = R_{zz} = 0$, while $R_{yz} = -R_{zy}$, $R_{zx} = -R_{xz}$, and $R_{xy} = -R_{yx}$. Note that the elements of the strain and rotation tensors, as well as those of the displacement gradient tensor, are dimensionless. For elastic deformation caused by an acoustic wave such as that described by (8.1), all of these tensor elements are space- and time-dependent quantities.

If the reference coordinate system is chosen such that one of its axes lines up with $\mathbf{K}$, a longitudinal acoustic wave generates only one tensile strain component and no rotation while a transverse acoustic wave generates only shear strains and rotation. For example, if we take $\mathbf{K} = K\hat{x}$, a longitudinal wave has $\boldsymbol{\mathcal{U}} = \mathcal{U}\hat{x}$. Then, the only nonzero element of the strain tensor is S_{xx}, and all elements of the rotation tensor vanish. For

a transverse wave, $\mathcal{U} = \mathcal{U}_y\hat{y} + \mathcal{U}_z\hat{z}$. If both $\mathcal{U}_y$ and $\mathcal{U}_z$ are nonzero, the only nonzero elements of the strain tensor are $S_{xy} = S_{yx}$ and $S_{zx} = S_{xz}$, while the nonzero elements of the rotation tensor are $R_{xy} = -R_{yx}$ and $R_{zx} = -R_{xz}$. If an acoustic wave is neither purely longitudinal nor purely transverse, it can generate both tensile and shear strains as well as many elements of the rotation tensor.

8.2 Photoelastic effect

Mechanical strain in a medium causes changes in the optical property of the medium due to the *photoelastic effect*. The basis of acousto-optic interaction is the *dynamic photoelastic effect* in which the periodic time-dependent mechanical strain caused by an acoustic wave induces periodic time-dependent variations in the optical properties of the medium.

The photoelastic effect is traditionally defined in terms of changes in the elements of the relative impermeability tensor caused by strain:

$$\eta_{ij}(\mathbf{S}) = \eta_{ij} + \Delta\eta_{ij}(\mathbf{S}) = \eta_{ij} + \sum_{k,l} p_{ijkl}S_{kl}, \tag{8.7}$$

where p_{ijkl} are dimensionless *elasto-optic coefficients*, also called *strain-optic coefficients* or *photoelastic coefficients*, and they form a fourth-rank tensor. Because $\eta_{ij} = \eta_{ji}$ and $S_{kl} = S_{lk}$, the elasto-optic tensor $[p_{ijkl}]$ is symmetric in i and j and in k and l. The rules of index contraction defined in (1.115) can be used to reduce the double indices ij and kl to single indices α and β:

$$p_{ijkl} = p_{jikl} = p_{ijlk} = p_{jilk} = p_{\alpha\beta}, \qquad \text{where } \alpha, \beta = 1, 2, \ldots, 6. \tag{8.8}$$

In general, $p_{\alpha\beta} \neq p_{\beta\alpha}$. Then, (8.7) can be expressed as

$$\eta_\alpha(\mathbf{S}) = \eta_\alpha + \Delta\eta_\alpha(\mathbf{S}) = \eta_\alpha + \sum_\beta p_{\alpha\beta}S_\beta, \tag{8.9}$$

where the elements S_β are defined by the following rules:

$$\begin{aligned} &S_1 = S_{xx}, \quad S_2 = S_{yy}, \quad S_3 = S_{zz}, \\ &S_4 = 2S_{yz}, \; S_5 = 2S_{zx}, \; S_6 = 2S_{xy}. \end{aligned} \tag{8.10}$$

Note that the factor 2 in the definitions of S_4, S_5, and S_6 deviates from the standard rules used in index contraction. *Similarly to the electro-optic Kerr effect, the photoelastic effect exists in all matters, including centrosymmetric crystals and isotropic media. Acousto-optic interactions are not precluded by any symmetry property of a medium.* Table 8.1 lists the matrix form of $p_{\alpha\beta}$ for various point groups. Also included in the table is the $p_{\alpha\beta}$ matrix for isotropic media, which has only two independent elements, p_{11} and p_{12}. The elasto-optic coefficients are dimensionless. The largest of them for a particular material typically has a value on the order of 0.1–0.5.

Table 8.1 *Matrix form of elasto-optic coefficients for all point groups*[a]

Triclinic	1 $\overline{1}$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} \\ p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} \end{bmatrix}$
Monoclinic	2 m $2/m$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & p_{15} & 0 \\ p_{21} & p_{22} & p_{23} & 0 & p_{25} & 0 \\ p_{31} & p_{32} & p_{33} & 0 & p_{35} & 0 \\ 0 & 0 & 0 & p_{44} & 0 & p_{46} \\ p_{51} & p_{52} & p_{53} & 0 & p_{55} & 0 \\ 0 & 0 & 0 & p_{64} & 0 & p_{66} \end{bmatrix}$
Orthorhombic	222 $mm2$ mmm	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & 0 \\ p_{21} & p_{22} & p_{23} & 0 & 0 & 0 \\ p_{31} & p_{32} & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{66} \end{bmatrix}$
Tetragonal	4 $\overline{4}$ $4/m$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & p_{16} \\ p_{12} & p_{11} & p_{13} & 0 & 0 & -p_{16} \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & p_{45} & 0 \\ 0 & 0 & 0 & -p_{45} & p_{44} & 0 \\ p_{61} & -p_{61} & 0 & 0 & 0 & p_{66} \end{bmatrix}$
	422 $\overline{4}2m$ $4mm$ $4/mmm$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & 0 \\ p_{12} & p_{11} & p_{13} & 0 & 0 & 0 \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{66} \end{bmatrix}$
Trigonal	3 $\overline{3}$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \\ p_{12} & p_{11} & p_{13} & -p_{14} & -p_{15} & -p_{16} \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ p_{41} & -p_{41} & 0 & p_{44} & p_{45} & p_{46} \\ -p_{46} & p_{46} & 0 & -p_{45} & p_{44} & p_{41} \\ -p_{16} & p_{16} & 0 & -p_{15} & p_{14} & \frac{1}{2}(p_{11} - p_{12}) \end{bmatrix}$
	32 $3m$ $\overline{3}m$	$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & 0 & 0 \\ p_{12} & p_{11} & p_{13} & -p_{14} & 0 & 0 \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ p_{41} & -p_{41} & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{44} & p_{41} \\ 0 & 0 & 0 & 0 & p_{14} & \frac{1}{2}(p_{11} - p_{12}) \end{bmatrix}$

(*continued*)

Table 8.1 (*cont.*)

Hexagonal

$$\begin{matrix}6\\ \overline{6}\\ 6/m\end{matrix}\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & p_{16} \\ p_{12} & p_{11} & p_{13} & 0 & 0 & -p_{16} \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & p_{45} & 0 \\ 0 & 0 & 0 & -p_{45} & p_{44} & 0 \\ -p_{16} & p_{16} & 0 & 0 & 0 & \frac{1}{2}(p_{11}-p_{12}) \end{bmatrix}$$

$$\begin{matrix}622\\ 6mm\\ \overline{6}m2\\ 6/mmm\end{matrix}\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & 0 \\ p_{12} & p_{11} & p_{13} & 0 & 0 & 0 \\ p_{31} & p_{31} & p_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}(p_{11}-p_{12}) \end{bmatrix}$$

Cubic

$$\begin{matrix}23\\ m3\end{matrix}\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 & 0 & 0 \\ p_{13} & p_{11} & p_{12} & 0 & 0 & 0 \\ p_{12} & p_{13} & p_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{44} \end{bmatrix} \quad \begin{matrix}432\\ \overline{4}3m\\ m3m\end{matrix}\begin{bmatrix} p_{11} & p_{12} & p_{12} & 0 & 0 & 0 \\ p_{12} & p_{11} & p_{12} & 0 & 0 & 0 \\ p_{12} & p_{12} & p_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{44} \end{bmatrix}$$

Isotropic

$$\begin{bmatrix} p_{11} & p_{12} & p_{12} & 0 & 0 & 0 \\ p_{12} & p_{11} & p_{12} & 0 & 0 & 0 \\ p_{12} & p_{12} & p_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}(p_{11}-p_{12}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}(p_{11}-p_{12}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2}(p_{11}-p_{12}) \end{bmatrix}$$

[a] Nye, J. F., *Physical Properties of Crystals*. London: Oxford University Press, 1957, pp. 250–251.

Many crystals, for example $LiNbO_3$ and TeO_2, that are of interest in acousto-optic applications are piezoelectric. The elasto-optic coefficients of such crystals are modified by the piezoelectric effect due to the fact that the strain produced by an acoustic wave generates an electric field in a crystal, which also causes index changes in the crystal through the electro-optic effect. The modifications due to this secondary effect can be quite significant in certain crystals. These modifications depend on the direction of propagation of the acoustic wave. Furthermore, they can rotate the index ellipsoid and induce birefringence, as can be imagined from our experience with the Pockels effect discussed in Chapter 6. The modifications to p_{ijkl} due to this piezoelectric effect still maintain the symmetry described in (8.8), but they do not follow the matrix form listed in Table 8.1. Their matrix form depends on the combination of the wave propagation direction and crystal symmetry.

The photoelastic effect associated with the rotational deformation characterized by the rotation tensor was not included in the definition of the elasto-optic coefficients p_{ijkl} expressed in (8.7). A complete description of the photoelastic effect has to include

the contributions from both strain and rotation as

$$\Delta\eta_{ij}(\mathbf{S},\mathbf{R}) = \sum_{k,l}\left(p_{ijkl}S_{kl} + p'_{ijkl}R_{kl}\right). \tag{8.11}$$

In this relation, p_{ijkl} can be considered to include modifications due to the piezoelectric effect, with the understanding that they are not exactly the same as those listed in Table 8.1 and that they depend on the direction of propagation of the acoustic wave. If the indices i, j, k, and l are referenced to the principal axes of a crystal, we have

$$p'_{ijkl} = -\frac{1}{2}\left(\frac{1}{n_i^2} - \frac{1}{n_j^2}\right)\left(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}\right), \tag{8.12}$$

where n_i and n_j represent the principal indices of refraction of the crystal. It can be seen that p'_{ijkl} is symmetric in i and j but is antisymmetric in k and l. Therefore, index contraction cannot be applied to the indices k and l in (8.11) and (8.12). Nevertheless, in (8.11) the indices i and j can still be contracted by following the rules in (1.115) into a single index α to write $\Delta\eta_{ij} = \Delta\eta_\alpha$ for it to be used to express the index ellipsoid in the form of (6.19). From (8.12), we find that p'_{ijkl} vanishes for isotropic media and cubic crystals and that *the rotational effect is significant only in strongly birefringent crystals*.

In the treatment of acousto-optic diffraction using coupled-wave theory, it is desirable to express the photoelastic effect caused by strain and rotation in a medium formally in terms of a change in the permittivity of the medium as

$$\boldsymbol{\epsilon}(\omega,\mathbf{S},\mathbf{R}) = \boldsymbol{\epsilon}(\omega) + \Delta\boldsymbol{\epsilon}(\omega,\mathbf{S},\mathbf{R}) = \boldsymbol{\epsilon}(\omega) + \epsilon_0\Delta\boldsymbol{\chi}(\omega,\mathbf{S},\mathbf{R}), \tag{8.13}$$

where $\boldsymbol{\epsilon}(\omega)$ is the dielectric permittivity tensor of the medium in the absence of strain and rotation fields. The effect of strain and rotation on an optical field $\mathbf{E}(\omega)$ propagating in a medium is characterized by a polarization of

$$\Delta\mathbf{P}(\omega,\mathbf{S},\mathbf{R}) = \Delta\boldsymbol{\epsilon}(\omega,\mathbf{S},\mathbf{R})\cdot\mathbf{E}(\omega). \tag{8.14}$$

Once the elements of $\Delta\boldsymbol{\eta}$ caused by strain and rotation are found through the procedure described above, the elements of $\Delta\boldsymbol{\epsilon}$ can be found using the relation in (6.17) or that in (6.18).

Acousto-optic figure of merit

In the case when $\Delta\eta_{ij}$ is independent of rotation tensor elements but is a function of strain tensor elements only, we can use (8.7) and (6.18) to express the photoelastic changes in the permittivity tensor as

$$\Delta\epsilon_{ij} = -\epsilon_0 n_i^2 n_j^2 \Delta\eta_{ij} = -\epsilon_0 n_i^2 n_j^2 \sum_{k,l} p_{ijkl}S_{kl}, \qquad i,j,k,l = 1,2,3. \tag{8.15}$$

The strain tensor elements depend on the propagation direction and the polarization mode of the acoustic wave. For an acoustic wave that has a wavevector $\mathbf{K}$ and an angular

frequency Ω, the strain tensor elements vary with space and time as

$$S_{kl} = \mathcal{S}_{kl} \sin(\mathbf{K} \cdot \mathbf{r} - \Omega t), \tag{8.16}$$

where $\mathcal{S}_{kl}$ is the amplitude of the strain. The magnitude of vector $\mathbf{K}$ depends on the polarization of the acoustic wave because longitudinal and transverse acoustic modes propagating in the same direction generally have different velocities. Then the photoelastic permittivity tensor is a function of space and time:

$$\Delta\boldsymbol{\epsilon} = \Delta\tilde{\boldsymbol{\epsilon}} \sin(\mathbf{K} \cdot \mathbf{r} - \Omega t), \tag{8.17}$$

where $\Delta\tilde{\boldsymbol{\epsilon}}$ is the amplitude of $\Delta\boldsymbol{\epsilon}$, and its elements are

$$\Delta\tilde{\epsilon}_{ij} = -\epsilon_0 n_i^2 n_j^2 \sum_{k,l} p_{ijkl} \mathcal{S}_{kl}. \tag{8.18}$$

The intensity in watts per square meter of an acoustic wave that has a strain amplitude $\mathcal{S}$ is given by

$$I_\mathrm{a} = \frac{1}{2} \mathcal{S}^2 \rho v_\mathrm{a}^3, \tag{8.19}$$

where ρ is the density of the medium and v_a is the velocity of the specific acoustic mode under consideration. Therefore, the strain amplitude in (8.18) can be properly calculated from the acoustic intensity using the relation:

$$\mathcal{S} = \left(\frac{2 I_\mathrm{a}}{\rho v_\mathrm{a}^3} \right)^{1/2}. \tag{8.20}$$

In acousto-optic diffraction, as we shall see in the following section, the coupling coefficient between an incident optical wave of a unit polarization vector $\hat{e}_\mathrm{i}$ and a diffracted optical wave of a unit polarization vector $\hat{e}_\mathrm{d}$ is determined by the following effective permittivity:

$$\Delta\tilde{\epsilon}_\mathrm{id} = \hat{e}_\mathrm{i}^* \cdot \Delta\tilde{\boldsymbol{\epsilon}} \cdot \hat{e}_\mathrm{d}. \tag{8.21}$$

The diffraction efficiency, however, is proportional to the square of the coupling coefficient in the low-efficiency limit. In practical acousto-optic applications, it is usually convenient to use an *acousto-optic figure of merit* that is defined as

$$M_2 = \frac{|\Delta\tilde{\epsilon}_\mathrm{id}|^2}{2\epsilon_0^2 n_\mathrm{i} n_\mathrm{d} I_\mathrm{a}} = \frac{n_\mathrm{i}^3 n_\mathrm{d}^3 p^2}{\rho v_\mathrm{a}^3}, \tag{8.22}$$

where n_i and n_d are the refractive indices seen by the incident and diffracted optical waves, respectively, p is an effective elasto-optic coefficient properly characterizing the interaction, and v_a is the velocity of the acoustic mode involved in the interaction.

Note that the parameters in the definition of M_2 have to be chosen properly according to the mode of operation. Therefore, the figure of merit is specific to the mode of operation. It depends on both the propagation direction and the polarization mode of the acoustic wave, as well as on the polarizations of the optical waves. The figure

of merit M_2 has the unit of cubic seconds per kilogram, which is equivalent to square meters per watt. The properties of some representative acousto-optic materials are listed in Table 8.2. The values of M_2 and v_a listed in this table are measured under specific experimental conditions. They are subject to changes under different conditions.

Another factor to be considered in the practical applications of an acousto-optic material is the acoustic attenuation due to acoustic absorption of the medium, which generally increases with the square of the acoustic frequency. As a consequence, popular materials such as silica glass, TeO_2, $PbMoO_4$, and Ge are limited to applications at acoustic frequencies well below 1 GHz and are often used in an acoustic frequency range between 10 and 500 MHz. GaP can be used in the acoustic frequency range between 500 MHz and 1 GHz. Because of its relatively low acoustic attenuation, $LiNbO_3$ is suitable for applications at high acoustic frequencies up to 5 GHz.

Isotropic medium

We consider, for simplicity, the propagation of a plane acoustic wave in an isotropic medium. From the discussions in the preceding section, there are one longitudinal and two transverse modes for an acoustic wave in an isotropic medium. We can define the x direction of the coordinate system to be the direction of propagation of the acoustic wave so that $\mathbf{K} = K\hat{x}$. Then a longitudinal wave can be expressed as

$$\mathbf{u}(x,t) = \hat{x}\mathcal{U}\cos(K_L x - \Omega t). \tag{8.23}$$

The two orthogonal transverse modes are degenerate. A transverse wave polarized in the y direction can be written as

$$\mathbf{u}(x,t) = \hat{y}\mathcal{U}\cos(K_T x - \Omega t), \tag{8.24}$$

and that polarized in the z direction as

$$\mathbf{u}(x,t) = \hat{z}\mathcal{U}\cos(K_T x - \Omega t). \tag{8.25}$$

Because $K_L \neq K_T$ in general, the longitudinal and transverse waves have different acoustic velocities, $v_{a,L} = \Omega/K_L$ and $v_{a,T} = \Omega/K_T$, respectively.

For the longitudinal wave given in (8.23), the only nonvanishing strain tensor element is the tensile strain

$$S_1 = S_{xx} = \mathcal{S}_1 \sin(K_L x - \Omega t), \tag{8.26}$$

where $\mathcal{S}_1 = \mathcal{S}_{xx} = -K_L\mathcal{U}$ is the amplitude of the space- and time-dependent periodic strain wave. Using (8.15) and the matrix form of $p_{\alpha\beta}$ for isotropic media in Table 8.1, we find that

$$\Delta\epsilon = \Delta\tilde{\epsilon}\sin(K_L x - \Omega t) = -\epsilon_0 n^4 \begin{bmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{12} \end{bmatrix} \mathcal{S}_1 \sin(K_L x - \Omega t). \tag{8.27}$$

Table 8.2 *Properties of representative acousto-optic materials*

Material	Point group	Density, ρ (10^3 kg m^{-3})	Acoustic wave			Optical wave polarization[c]	Index[d], n	Figure of merit M_2 (10^{-15} m^2 W^{-1})
			Direction[a]	Mode[b]	v_a (km s^{-1})			
Silica glass	Isotropic	2.20	Any	L	5.97	$\perp$	1.457	1.50
				L	5.97	$\parallel$	1.457	0.30
				T	3.76	Any	1.457	0.46
Water	Isotropic	1.00	Any	L	1.50	$\perp$	1.333	160
SF-4 glass	Isotropic	3.59	Any	L	3.63	$\perp$	1.616	4.51
TF-7 glass	Isotropic	4.59	Any	L	3.63	$\perp$	1.728	5.00
$LiNbO_3$	$3m$	4.64	[100]	L	6.57	$\perp$	2.201	7.00
			[001]	T	3.59	$\perp$	2.291	5.60
TeO_2	422	6.00	[001]	L	4.20	$\perp$	2.260	34.5
			[001]	L	4.20	$\parallel$	2.412	25.6
			[110]	T	0.616	Circular	2.260	1200
$PbMoO_4$	$4/m$	6.95	[001]	L	3.63	$\perp$	2.386	36.1
			[001]	L	3.63	$\parallel$	2.262	36.3
GaP	$\bar{4}3m$	4.13	[110]	L	6.32	$\parallel$	3.31	44.8
			[100]	L	5.85	$\perp$	3.31	10.7
			[100]	L	5.85	$\parallel$	3.31	36.3
			[100]	T	4.13	Any	3.31	23.4

[a] Acoustic wave propagation direction **K** with respect to crystal axes.

[b] L, longitudinal acoustic mode; T, transverse acoustic mode.

[c] Optical field polarization **E** with respect to acoustic wave propagation direction **K**. $\perp$, $\mathbf{E} \perp \mathbf{K}$; $\parallel$, $\mathbf{E} \parallel \mathbf{K}$.

[d] Refractive index at 632.8 nm optical wavelength. In the case of an anisotropic crystal, the relevant index used is a function of the optical wave polarization with respect to the principal axes of the crystal.

For the y-polarized transverse wave given in (8.24), the only nonvanishing strain tensor elements are the shear strains $S_{xy} = S_{yx}$; thus

$$S_6 = 2S_{xy} = \mathcal{S}_6 \sin(K_T x - \Omega t), \tag{8.28}$$

where $\mathcal{S}_6 = 2\mathcal{S}_{xy} = -K_T \mathcal{U}$. Because $p_{66} = (p_{11} - p_{12})/2$ in an isotropic medium, we have

$$\Delta\boldsymbol{\epsilon} = \Delta\tilde{\boldsymbol{\epsilon}} \sin(K_T x - \Omega t) = -\epsilon_0 n^4 \begin{bmatrix} 0 & \frac{1}{2}(p_{11} - p_{12}) & 0 \\ \frac{1}{2}(p_{11} - p_{12}) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathcal{S}_6 \sin(K_T x - \Omega t). \tag{8.29}$$

Similarly, for the z-polarized transverse wave given in (8.25), we have

$$S_5 = 2S_{zx} = \mathcal{S}_5 \sin(K_T x - \Omega t) \tag{8.30}$$

and

$$\Delta\boldsymbol{\epsilon} = \Delta\tilde{\boldsymbol{\epsilon}} \sin(K_T x - \Omega t) = -\epsilon_0 n^4 \begin{bmatrix} 0 & 0 & \frac{1}{2}(p_{11} - p_{12}) \\ 0 & 0 & 0 \\ \frac{1}{2}(p_{11} - p_{12}) & 0 & 0 \end{bmatrix} \mathcal{S}_5 \sin(K_T x - \Omega t), \tag{8.31}$$

where $\mathcal{S}_5 = 2\mathcal{S}_{zx} = -K_T \mathcal{U}$.

We see from the above examples that the elements of the $\Delta\tilde{\boldsymbol{\epsilon}}$ tensor have the form

$$\Delta\tilde{\epsilon}_{ij} = -\epsilon_0 n^4 p \mathcal{S}, \tag{8.32}$$

where p is the appropriate elasto-optic coefficient and $\mathcal{S}$ is the amplitude of the appropriate strain tensor element representing the traveling acoustic wave under consideration. In an isotropic medium, the acousto-optic figure of merit defined in (8.22) is simplified as

$$M_2 = \frac{n^6 p^2}{\rho v_a^3}. \tag{8.33}$$

Note that even in an isotropic medium, M_2 still depends on the mode of the acoustic wave and the polarizations of the optical waves involved in the interaction.

EXAMPLE 8.1 Fused silica glass is an isotropic material that has only two independent elasto-optic coefficients, $p_{11} = 0.121$ and $p_{12} = 0.271$. A longitudinal acoustic wave at a frequency of 500 MHz is generated to propagate in the x direction. Use the data listed in Table 8.2 to find the wavelength of the acoustic wave and the figure of merit at 632.8 nm optical wavelength for optical waves of different polarizations. If the acoustic wave has an intensity of 10 W cm^{-2}, what are the photoelastic index changes?

Solution From Table 8.2, we find that $v_{a,L} = 5.97$ km s^{-1} for a longitudinal acoustic wave in silica glass. At an acoustic frequency of $f = 500$ MHz, the wavelength of the longitudinal acoustic wave is found using (8.3) to be

$$\Lambda_L = \frac{v_{a,L}}{f} = 11.92\ \mu\text{m}.$$

The longitudinal acoustic wave has a wavevector $\mathbf{K} = K_L\hat{x}$. Therefore, the acousto-optic permittivity change $\Delta\boldsymbol{\epsilon}$ has the form of that given in (8.27) for isotropic silica glass. For an optical wave that is linearly polarized in the x direction, parallel to the acoustic wavevector $\mathbf{K}$, we have $p = p_{11}$ and the following figure of merit:

$$M_2^{\parallel} = \frac{n^6 p_{11}^2}{\rho v_{a,L}^3} = \frac{1.457^6 \times 0.121^2}{2.2 \times 10^3 \times (5.97 \times 10^3)^3}\ \text{s}^3\ \text{kg}^{-1} = 3.0 \times 10^{-16}\ \text{s}^3\ \text{kg}^{-1}$$
$$= 3.0 \times 10^{-16}\ \text{m}^2\ \text{W}^{-1}.$$

For an optical wave that is polarized in any direction in the yz plane, perpendicular to the acoustic wavevector $\mathbf{K}$, we have $p = p_{12}$ and the following figure of merit:

$$M_2^{\perp} = \frac{n^6 p_{12}^2}{\rho v_{a,L}^3} = \frac{1.457^6 \times 0.271^2}{2.2 \times 10^3 \times (5.97 \times 10^3)^3}\ \text{s}^3\ \text{kg}^{-1} = 1.5 \times 10^{-15}\ \text{s}^3\ \text{kg}^{-1}$$
$$= 1.5 \times 10^{-15}\ \text{m}^2\ \text{W}^{-1}.$$

With $I_a = 10$ W cm$^{-2} = 1 \times 10^5$ W m^{-2}, we have the following strain amplitude:

$$\mathcal{S} = \left(\frac{2I_a}{\rho v_{a,L}^3}\right)^{1/2} = \left[\frac{2 \times 10^5}{2.2 \times 10^3 \times (5.97 \times 10^3)^3}\right]^{1/2} = 2.07 \times 10^{-5}.$$

Therefore, $\Delta\tilde{\epsilon}_{xx} = -\epsilon_0 n^4 p_{11}\mathcal{S} = -1.13 \times 10^{-5}\epsilon_0$ and $\Delta\tilde{\epsilon}_{yy} = \Delta\tilde{\epsilon}_{zz} = -\epsilon_0 n^4 p_{12}\mathcal{S} = -2.53 \times 10^{-5}\epsilon_0$. Because $\Delta\boldsymbol{\epsilon}$ is diagonal and because $|\Delta\tilde{\epsilon}_{ii}/\epsilon_0| \ll n^2$, we have

$$\Delta n_x(x,t) = \frac{\Delta\tilde{\epsilon}_{xx}}{2n\epsilon_0}\sin(K_L x - \Omega t) = -3.38 \times 10^{-6}\sin(K_L x - \Omega t),$$
$$\Delta n_y(x,t) = \frac{\Delta\tilde{\epsilon}_{yy}}{2n\epsilon_0}\sin(K_L x - \Omega t) = -8.68 \times 10^{-6}\sin(K_L x - \Omega t),$$
$$\Delta n_z(x,t) = \frac{\Delta\tilde{\epsilon}_{zz}}{2n\epsilon_0}\sin(K_L x - \Omega t) = -8.68 \times 10^{-6}\sin(K_L x - \Omega t),$$

where $K_L = 2\pi/\Lambda_L$ and $\Omega = 2\pi f$.

8.3 Acousto-optic diffraction

We see from the preceding two sections that the space- and time-dependent periodic permittivity changes induced by a traveling plane acoustic wave of the form given in (8.1) can be generally expressed as

$$\Delta\epsilon = \Delta\tilde{\epsilon}\sin(\mathbf{K}\cdot\mathbf{r} - \Omega t), \tag{8.34}$$

where $\mathbf{K}$ depends on both the polarization and the propagation direction of the acoustic wave. In general, $\Delta\tilde{\epsilon}$ is a function of the strain and the rotation generated by the acoustic wave in the medium, the elasto-optic coefficients of the medium, the mode and direction of the acoustic wave, and the frequency and polarization of the optical wave, but it is independent of the values of K and Ω. When an optical wave at a frequency ω is incident on this medium, the interaction between the optical wave and the periodic modulation described by (8.34) can generate diffracted optical waves at frequencies $\omega \pm \Omega$. The diffracted waves at $\omega \pm \Omega$ can be diffracted once more to generate waves at frequencies $\omega \pm 2\Omega$. If this process is allowed to cascade, we will end up with a series of diffracted optical waves at frequencies $\omega + q\Omega$, where q admits both positive and negative integers and is the *order of acousto-optic diffraction*.

For acousto-optic diffraction from a traveling acoustic wave, each diffraction order has a unique propagation direction. Therefore, there is a single wavevector $\mathbf{k}_q$ associated with the optical wave component at the frequency $\omega_q = \omega + q\Omega$. Following the formulation of the coupled-wave theory discussed in Section 4.1, the total optical field consisting of all interacting components can then be expressed in the form of (4.5):

$$\mathbf{E}(\mathbf{r}, t) = \sum_q \boldsymbol{\mathcal{E}}_q(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}-\mathrm{i}\omega_q t} = \sum_q \hat{e}_q\mathcal{E}_q(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}-\mathrm{i}\omega_q t}, \tag{8.35}$$

where $\hat{e}_q$ is the unit vector defining the polarization of $\boldsymbol{\mathcal{E}}_q$. According to (8.14), the polarization induced by interaction of the acoustic wave with the optical field component of the frequency ω_q is $\Delta\mathbf{P}_q(\mathbf{r}) = \Delta\epsilon(\omega_q)\cdot\mathbf{E}_q(\mathbf{r}) = \Delta\epsilon(\omega_q)\cdot\boldsymbol{\mathcal{E}}_q(\mathbf{r})\exp(\mathrm{i}\mathbf{k}_q\cdot\mathbf{r})$. Because $\omega \gg \Omega$, dispersion of the medium within the frequency range of interaction can be ignored to take $\Delta\epsilon(\omega_q) = \Delta\epsilon$. According to (4.6), the total induced polarization is

$$\begin{aligned}
\Delta\mathbf{P}(\mathbf{r}, t) &= \sum_q \Delta\epsilon\cdot\boldsymbol{\mathcal{E}}_q(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}-\mathrm{i}\omega_q t} \\
&= \sum_q \Delta\tilde{\epsilon}\cdot\boldsymbol{\mathcal{E}}_q(\mathbf{r})\sin(\mathbf{K}\cdot\mathbf{r} - \Omega t)\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}-\mathrm{i}\omega_q t} \\
&= \frac{1}{2\mathrm{i}}\sum_q \Delta\tilde{\epsilon}\cdot\boldsymbol{\mathcal{E}}_q(\mathbf{r})\left[\mathrm{e}^{\mathrm{i}(\mathbf{k}_q+\mathbf{K})\cdot\mathbf{r}-\mathrm{i}\omega_{q+1}t} - \mathrm{e}^{\mathrm{i}(\mathbf{k}_q-\mathbf{K})\cdot\mathbf{r}-\mathrm{i}\omega_{q-1}t}\right] \\
&= \frac{1}{2\mathrm{i}}\sum_q \Delta\tilde{\epsilon}\cdot\left[\boldsymbol{\mathcal{E}}_{q-1}(\mathbf{r})\mathrm{e}^{\mathrm{i}(\mathbf{k}_{q-1}+\mathbf{K})\cdot\mathbf{r}} - \boldsymbol{\mathcal{E}}_{q+1}(\mathbf{r})\mathrm{e}^{\mathrm{i}(\mathbf{k}_{q+1}-\mathbf{K})\cdot\mathbf{r}}\right]\mathrm{e}^{-\mathrm{i}\omega_q t}.
\end{aligned} \tag{8.36}$$

Comparing (8.36) with (4.6), we find that

$$\Delta \mathbf{P}_q(\mathbf{r}) = \frac{1}{2\mathrm{i}} \Delta\tilde{\epsilon} \cdot \left[\boldsymbol{\mathcal{E}}_{q-1}(\mathbf{r}) \mathrm{e}^{\mathrm{i}(\mathbf{k}_{q-1}+\mathbf{K})\cdot\mathbf{r}} - \boldsymbol{\mathcal{E}}_{q+1}(\mathbf{r}) \mathrm{e}^{\mathrm{i}(\mathbf{k}_{q+1}-\mathbf{K})\cdot\mathbf{r}} \right]. \tag{8.37}$$

Using (4.11), we have the following equation for coupling of optical waves through their interaction with a traveling acoustic wave in an isotropic medium:

$$(\mathbf{k}_q \cdot \boldsymbol{\nabla}) \boldsymbol{\mathcal{E}}_q = \frac{\omega_q^2 \mu_0}{4} \Delta\tilde{\epsilon} \cdot \left[\boldsymbol{\mathcal{E}}_{q-1} \mathrm{e}^{\mathrm{i}(\mathbf{k}_{q-1}+\mathbf{K}-\mathbf{k}_q)\cdot\mathbf{r}} - \boldsymbol{\mathcal{E}}_{q+1} \mathrm{e}^{\mathrm{i}(\mathbf{k}_{q+1}-\mathbf{K}-\mathbf{k}_q)\cdot\mathbf{r}} \right]. \tag{8.38}$$

In an anisotropic medium, the valid coupled-wave equation can be obtained by taking the transverse component on both sides of (8.38) according to (4.18). From this coupled-wave equation, the following general observations can be made.

1. Each optical frequency is directly coupled only to its neighboring frequencies shifted by Ω or $-\Omega$.
2. Coupling between optical waves of different polarizations is possible in both isotropic and anisotropic media because $\Delta\tilde{\epsilon}$ is generally anisotropic. In an isotropic medium, this coupling is possible when $p_{11} \neq p_{12}$, as can be seen from the demonstration in the preceding section.
3. The coupling efficiency depends on the polarization and the propagation direction of the optical waves being coupled, as well as on the polarization and the propagation direction of the acoustic wave.
4. Coupling between $\boldsymbol{\mathcal{E}}_q$ and $\boldsymbol{\mathcal{E}}_{q-1}$ is phase matched when $\mathbf{k}_{q-1} = \mathbf{k}_q - \mathbf{K}$, whereas the phase-matching condition for coupling between $\boldsymbol{\mathcal{E}}_q$ and $\boldsymbol{\mathcal{E}}_{q+1}$ is $\mathbf{k}_{q+1} = \mathbf{k}_q + \mathbf{K}$. The efficiency for the coupling between two specific wave components of different frequencies depends critically on the amount of phase mismatch in the coupling process.

Consequently, acousto-optic diffraction displays many different phenomena under different experimental conditions. Each phenomenon is useful for certain applications.

Raman–Nath diffraction

We consider the diffraction, in an isotropic medium, of a plane optical wave at a frequency ω by a column of plane acoustic wave in a geometry shown in Fig. 8.2(a). The acoustic wave propagates in the x direction so that $\mathbf{K} = K\hat{x}$. The value of K depends on the polarization of the acoustic wave, as is demonstrated in the preceding section. The incident optical wave propagates in a direction close to normal to the acoustic column so that its wavevector $\mathbf{k}_\mathrm{i}$ makes a small angle θ_i with respect to the z-coordinate axis, as also shown in Fig. 8.2(a). The acoustic wave column has a finite width in the z direction, but it extends in the x direction far beyond the region of interaction. We also assume that the interaction is two-dimensional so that there are no variations in the

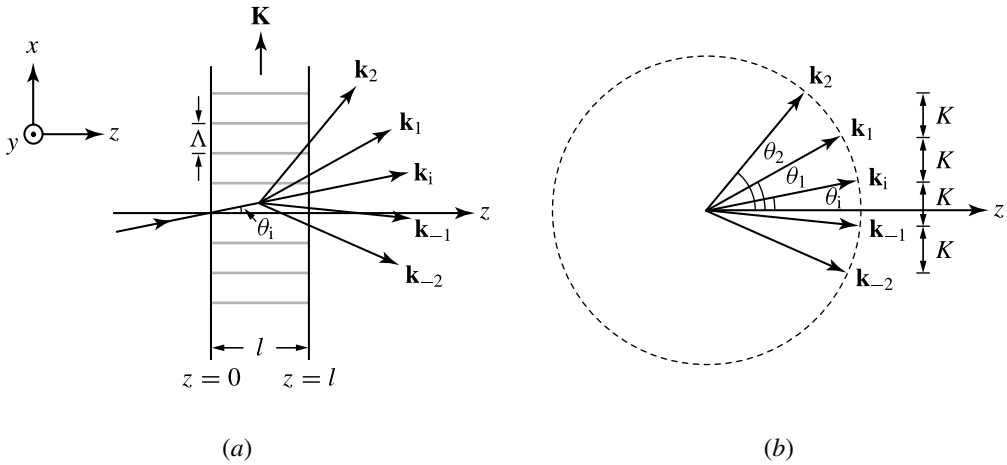


Figure 8.2 (*a*) Configuration and (*b*) wavevector diagram for Raman–Nath diffraction in an isotropic medium. Phase matching in the x direction determines the propagation angles of the diffracted waves. Phase mismatch exists only in the z direction.

y direction for both the optical and the acoustic waves. With these assumptions, any changes in the amplitude of the optical field caused by its interaction with the acoustic wave column occur only along the z direction even though the propagation direction of the optical wave might not be parallel to the z-coordinate axis. Consequently, we have $\boldsymbol{\mathcal{E}}_q(\mathbf{r}) = \boldsymbol{\mathcal{E}}_q(z)$ though $\mathbf{k}_q = k_{q,x}\hat{x} + k_{q,z}\hat{z}$ and $k_{q,x} \neq 0$ in general.

Under the conditions discussed above, the coupled-wave equation in (8.38) can be written as

$$\frac{\mathrm{d}\boldsymbol{\mathcal{E}}_q}{\mathrm{d}z} = \frac{\omega_q^2 \mu_0}{4k_{q,z}} \Delta\tilde{\epsilon} \cdot \left\{ \boldsymbol{\mathcal{E}}_{q-1} \exp[\mathrm{i}\left(k_{q-1,x} + K - k_{q,x}\right) x + \mathrm{i}\left(k_{q-1,z} - k_{q,z}\right) z] \right.$$
$$\left. - \boldsymbol{\mathcal{E}}_{q+1} \exp[\mathrm{i}\left(k_{q+1,x} - K - k_{q,x}\right) x + \mathrm{i}\left(k_{q+1,z} - k_{q,z}\right) z] \right\}, \quad (8.39)$$

according to (4.12). Because the field amplitudes in this equation vary with z only, the x-dependent phases on the right-hand side of this equation must vanish, resulting in the following phase-matching condition:

$$K = k_{q,x} - k_{q-1,x} = k_{q+1,x} - k_{q,x}. \quad (8.40)$$

This phase-matching condition determines the propagation direction of each diffracted wave component. Because $\omega \gg \Omega$, we can take the approximation that $k_q = n(\omega + q\Omega)/c \approx n\omega/c = k$. Then, (8.40) can be written as

$$K = k\left(\sin\theta_q - \sin\theta_{q-1}\right) = k\left(\sin\theta_{q+1} - \sin\theta_q\right), \quad (8.41)$$

where θ_q is the directional angle of $\mathbf{k}_q$ with respect to the z axis, as shown in Fig. 8.2(*b*). The zeroth order, $q = 0$, represents the undiffracted component with $\mathbf{k}_0 = \mathbf{k}_\mathrm{i}$ and

$\theta_0 = \theta_\mathrm{i}$ at the original frequency $\omega_0 = \omega$. The recursion relation in (8.41) can then be reduced to

$$\sin\theta_q = \sin\theta_\mathrm{i} + \frac{qK}{k}. \tag{8.42}$$

For small angles of incidence and diffraction, (8.42) can be written as

$$\theta_q \approx \theta_\mathrm{i} + \frac{qK}{k} = \theta_\mathrm{i} + q\frac{\lambda}{n\Lambda} = \theta_\mathrm{i} + q\frac{\lambda f}{nv_\mathrm{a}}, \tag{8.43}$$

where n is the refractive index of the medium. By expressing the phase mismatch in the z direction in terms of the angle of propagation for each wave component, (8.39) becomes

$$\frac{\mathrm{d}\mathcal{E}_q}{\mathrm{d}z} = \frac{\omega_q^2\mu_0}{4k\cos\theta_q}\left[\Delta\tilde{\epsilon}_{q,q-1}\mathcal{E}_{q-1}\mathrm{e}^{\mathrm{i}kz\left(\cos\theta_{q-1}-\cos\theta_q\right)} - \Delta\tilde{\epsilon}_{q,q+1}\mathcal{E}_{q+1}\mathrm{e}^{\mathrm{i}kz\left(\cos\theta_{q+1}-\cos\theta_q\right)}\right], \tag{8.44}$$

where $\Delta\tilde{\epsilon}_{q,q-1} = \hat{e}_q^* \cdot \Delta\tilde{\boldsymbol{\epsilon}} \cdot \hat{e}_{q-1}$ and $\Delta\tilde{\epsilon}_{q,q+1} = \hat{e}_q^* \cdot \Delta\tilde{\boldsymbol{\epsilon}} \cdot \hat{e}_{q+1}$. The coupled equations represented by (8.44) are known as the *Raman–Nath equations*.

The solution of (8.44) depends on many experimental parameters. In the special case when the direction of propagation of the incident optical wave is normal to the direction of propagation of the acoustic wave so that $\theta_\mathrm{i} = 0$, the phase-mismatch parameters in (8.44) can be approximated as

$$kz(\cos\theta_{q-1} - \cos\theta_q) \approx kz\left(q - \frac{1}{2}\right)\frac{K^2}{k^2}, \tag{8.45}$$

$$kz(\cos\theta_{q+1} - \cos\theta_q) \approx -kz\left(q + \frac{1}{2}\right)\frac{K^2}{k^2}, \tag{8.46}$$

using (8.43) to expand $\cos\theta_q$. If the interaction length l along the z direction is small so that

$$q\frac{K^2 l}{k} \ll 1, \tag{8.47}$$

the cumulative phase mismatch over the interaction length can be neglected. Then (8.44) can be approximated as

$$\frac{\mathrm{d}\mathcal{E}_q}{\mathrm{d}z} \approx \frac{\omega_q^2\mu_0}{4k}\left(\Delta\tilde{\epsilon}_{q,q-1}\mathcal{E}_{q-1} - \Delta\tilde{\epsilon}_{q,q+1}\mathcal{E}_{q+1}\right). \tag{8.48}$$

In this situation, acousto-optic coupling allows many diffraction orders to be observed. This is the regime of *Raman–Nath diffraction*. The condition for Raman–Nath diffraction is usually stated as

$$Q = \frac{K^2 l}{k} = 2\pi\frac{\lambda l}{n\Lambda^2} = 2\pi\frac{\lambda f^2 l}{nv_\mathrm{a}^2} \ll 1, \tag{8.49}$$

although the condition in (8.47) is more precise when high diffraction orders are considered. Typically, one chooses $Q \leq 0.3$ for Raman–Nath diffraction. In addition to this condition, *Raman–Nath diffraction occurs only when the optical wave propagates in a direction normal, or nearly normal, to the direction of propagation of the acoustic wave.*

Because $\Delta\tilde{\epsilon}_{q,q-1}$ and $\Delta\tilde{\epsilon}_{q,q+1}$ have different values and both vary with q, there is no general solution to the coupled equations of (8.48). Nevertheless, with the incident optical wave propagating in the z direction and the acoustic wave propagating in the x direction, we do have $\Delta\tilde{\epsilon}_{q,q-1} = \Delta\tilde{\epsilon}_{q,q+1} = \Delta\tilde{\epsilon}_{\text{id}}$, which is independent of q in the following special situations (see Problem 8.3.1).

1. If the acoustic wave is longitudinally polarized and the incident optical wave is linearly polarized either in the x or y direction, then $\hat{e}_q = \hat{e}_{\text{i}}$ for all q.
2. If the acoustic wave is a y-polarized transverse wave, then $\hat{e}_q = \hat{e}_{\text{i}}$ for all even values of q and $\hat{e}_q = \hat{e}_1$ for all odd values of q. This statement is true irrespective of the polarization state of the incident optical wave, which can be linear, circular, or elliptical.

In these special cases, (8.48) can be written as

$$\frac{\mathrm{d}\mathcal{E}_q}{\mathrm{d}z} = \frac{\omega^2\mu_0\Delta\tilde{\epsilon}_{\text{id}}}{4k}\left(\mathcal{E}_{q-1} - \mathcal{E}_{q+1}\right), \tag{8.50}$$

where the approximation of $\omega_q \approx \omega$ is taken. We can cast (8.50) in the form

$$\frac{\mathrm{d}\mathcal{E}_q}{\mathrm{d}\zeta} = \frac{1}{2}\left(\mathcal{E}_{q-1} - \mathcal{E}_{q+1}\right) \tag{8.51}$$

by taking

$$\zeta = \frac{\omega^2\mu_0\Delta\tilde{\epsilon}_{\text{id}}}{2k}z. \tag{8.52}$$

The recursion relation in (8.51) is that of the Bessel functions given in (3.21). Therefore, its solutions are the Bessel functions:

$$\mathcal{E}_q(z) = \mathcal{E}_0(0)J_q\left(\frac{\omega^2\mu_0\Delta\tilde{\epsilon}_{\text{id}}}{2k}z\right), \tag{8.53}$$

where $\mathcal{E}_0(0)$ is the amplitude of the incident optical wave, which is the zeroth order with $q = 0$, at the input plane $z = 0$. We can define a coupling coefficient

$$|\kappa| = \frac{\pi}{\lambda}\left(\frac{M_2 I_{\text{a}}}{2}\right)^{1/2}, \tag{8.54}$$

where λ is the optical wavelength in free space and M_2 is related to $\tilde{\epsilon}_{\text{id}}$ according to the relation in (8.22). Then, for an interaction length l, we have

$$\mathcal{E}_q(l) = \mathcal{E}_0(0)J_q(-2|\kappa|l) = \mathcal{E}_0(0)J_{-q}(2|\kappa|l), \tag{8.55}$$

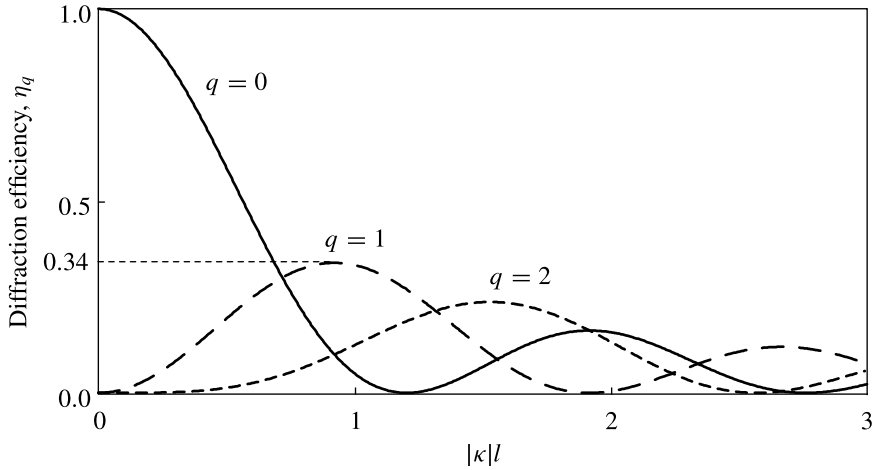


Figure 8.3 Raman–Nath diffraction efficiencies of a few leading diffraction orders.

where we have used the identity $J_q(-x) = J_{-q}(x)$ for the Bessel functions. The leading orders of the Bessel functions have been plotted in Fig. 3.2. The *Raman–Nath diffraction efficiency* for order q over an interaction length l is

$$\eta_q = \frac{I_q(l)}{I_0(0)} = J_{-q}^2(2|\kappa|l) = J_q^2(2|\kappa|l) = \eta_{-q}. \tag{8.56}$$

Because $J_{-q}^2 = J_q^2$ according to (3.20), the diffraction orders q and $-q$ have the same diffraction efficiency. Figure 8.3 shows the Raman–Nath diffraction efficiencies of a few leading orders.

In the above, we have considered Raman–Nath diffraction in an isotropic medium. Raman–Nath diffraction in an anisotropic crystal that involves a polarization change between successive orders would require successive anisotropic phase matching and is generally not possible.

EXAMPLE 8.2 A longitudinal acoustic wave propagates in a piece of fused silica glass in the x direction. An optical wave at 632.8 nm wavelength propagating in the z direction is diffracted by this acoustic wave in the Raman–Nath regime. (a) If the acoustic frequency is kept at $f = 100$ MHz, what is the limit on the interaction length l? (b) If the acousto-optic interaction length is $l = 1$ cm, what is the requirement on the acoustic frequency f? (c) Find the first-order diffraction efficiency for an interaction length of $l = 1$ cm and an acoustic intensity of $I_a = 1$ W cm^{-2}.

Solution For a longitudinal acoustic wave in fused silica, we have $v_a = v_{a,L} = 5.97$ km s^{-1} from Table 8.2. We also have $n = 1.457$ for fused silica at $\lambda = 632.8$ nm.

(a) For $f = 100$ MHz, we have the following limit for the interaction length:

$$l \ll \frac{n v_a^2}{2\pi \lambda f^2} = 1.3 \text{ mm}.$$

From these results, we see clearly that Raman–Nath diffraction takes place only at low acoustic frequencies or short interaction lengths.

(b) For $l = 1$ cm, the requirement that $Q \ll 1$ sets the following limit on the acoustic frequency:

$$f \ll \left(\frac{n v_a^2}{2\pi \lambda l}\right)^{1/2} = 36 \text{ MHz}.$$

(c) Because $M_2 = 1.5 \times 10^{-15}$ m^2 W^{-1}, we find, using (8.54), that $|\kappa| = 13.64$ m^{-1} for $I_a = 1$ W cm^{-2}. Therefore, $2|\kappa|l = 0.2728$ for $l = 1$ cm, and the first-order diffraction efficiency for $q = 1$ and $q = -1$ alike is

$$\eta_1 = \eta_{-1} = J_1^2(0.2728) \approx 0.018.$$

Bragg diffraction

When the interaction length l is sufficiently large so that

$$Q = \frac{K^2 l}{k} = 2\pi \frac{\lambda l}{n\Lambda^2} = 2\pi \frac{\lambda f^2 l}{n v_a} \gg 1, \tag{8.57}$$

the cumulative phase mismatch for each pair of coupled wave components cannot be neglected when solving (8.44). Consequently, it is necessary to have perfect, or nearly perfect, phase matching between two coupled wave components in order to have a significant diffraction efficiency from one of them to the other. This condition defines the regime of *Bragg diffraction*. In practice, $Q \geq 4\pi$ is often chosen for Bragg diffraction.

The incident wave, being the zeroth order with a wavevector $\mathbf{k}_i$ and a frequency ω, is directly coupled only to the diffraction orders $q = 1$ and $q = -1$. It can be seen by taking $q = 0$ in (8.38) that the phase-matching condition for generation of the diffraction order $q = 1$ at the *up-shifted frequency* $\omega_1 = \omega + \Omega$ is

$$\mathbf{k}_d = \mathbf{k}_1 = \mathbf{k}_i + \mathbf{K}, \tag{8.58}$$

whereas that for generation of the diffraction order $q = -1$ at the *down-shifted frequency* $\omega_{-1} = \omega - \Omega$ is

$$\mathbf{k}_d = \mathbf{k}_{-1} = \mathbf{k}_i - \mathbf{K}. \tag{8.59}$$

For any diffraction order q to be generated, it is found by reduction that the phase-matching condition of $\mathbf{k}_q = \mathbf{k}_i + q\mathbf{K}$ has to be satisfied for $\omega_q = \omega + q\Omega$. In addition, because each diffraction order is directly coupled only to its neighboring orders, Bragg

diffraction at a high order requires successive generation of low diffraction orders, thus simultaneous satisfaction of corresponding phase-matching conditions. With the exception of some very special cases, these requirements cannot be fulfilled. Consequently, only one diffraction order, either $q = 1$ or $q = -1$, is usually generated in Bragg diffraction from a traveling acoustic wave.

Bragg diffraction occurs in both isotropic and anisotropic media when the phase-matching condition in (8.58) or that in (8.59) is satisfied. The polarization of the diffracted wave is determined by the property of the $\Delta\tilde{\epsilon}$ tensor and the polarization of the incident optical wave in much the same manner as that discussed above for Raman–Nath diffraction. Therefore, in both isotropic and anisotropic media, acousto-optic Bragg diffraction can be accompanied by a change of polarization between the incident and diffracted waves. In an isotropic medium, $k_\mathrm{d} \approx k_\mathrm{i} = k$ no matter whether there is a change of polarization in the process or not. In an anisotropic medium, $k_\mathrm{d} = k_\mathrm{i}$ when the incident and the diffracted waves have the same polarization, but $k_\mathrm{d} \neq k_\mathrm{i}$ in general when they have different polarizations. The type of acousto-optic diffraction in which $k_\mathrm{d} \neq k_\mathrm{i}$ is called *birefringent diffraction*, whereas that in which $k_\mathrm{d} = k_\mathrm{i}$ is called *nonbirefringent diffraction*.

With the incident wave propagating at a directional angle θ_i and the diffracted wave propagating at a directional angle θ_d, the phase-matching conditions in (8.58) and (8.59) can be expressed as

$$k_\mathrm{d}\cos\theta_\mathrm{d} = k_\mathrm{i}\cos\theta_\mathrm{i}, \qquad k_\mathrm{d}\sin\theta_\mathrm{d} = k_\mathrm{i}\sin\theta_\mathrm{i} \pm K, \tag{8.60}$$

where the plus sign is for *up-shifted diffraction* and the minus sign is for *down-shifted diffraction*. Therefore, for phase-matched, up-shifted Bragg diffraction, the angles of incidence and diffraction are

$$\theta_\mathrm{i} = -\sin^{-1}\frac{K^2 + k_\mathrm{i}^2 - k_\mathrm{d}^2}{2k_\mathrm{i}K} = -\sin^{-1}\frac{\lambda f}{2n_\mathrm{i}v_\mathrm{a}}\left[1 + \frac{v_\mathrm{a}^2}{\lambda^2 f^2}(n_\mathrm{i}^2 - n_\mathrm{d}^2)\right], \tag{8.61}$$

$$\theta_\mathrm{d} = \sin^{-1}\frac{K^2 + k_\mathrm{d}^2 - k_\mathrm{i}^2}{2k_\mathrm{d}K} = \sin^{-1}\frac{\lambda f}{2n_\mathrm{d}v_\mathrm{a}}\left[1 + \frac{v_\mathrm{a}^2}{\lambda^2 f^2}(n_\mathrm{d}^2 - n_\mathrm{i}^2)\right], \tag{8.62}$$

respectively.

For phase-matched, down-shifted Bragg diffraction, the signs of both θ_i and θ_d change from those given above for up-shifted diffraction. Note that when $k_\mathrm{i} \neq k_\mathrm{d}$, either the incident or the diffracted wave has to be an extraordinary wave; sometimes both of them are. Therefore, in the above equations, the value of k_i depends on θ_i and that of k_d depends on θ_d, in general. In Bragg diffraction, the angles of incidence and diffraction are not limited to small values as is the case in Raman–Nath diffraction. As can be seen from (8.61) and (8.62), depending on the values of k_i, k_d, and K involved in the process, the values of θ_i and θ_d dictated by the phase-matching conditions can be anywhere between $-\pi/2$ and $\pi/2$, if they exist. For certain combinations of k_i, k_d, and

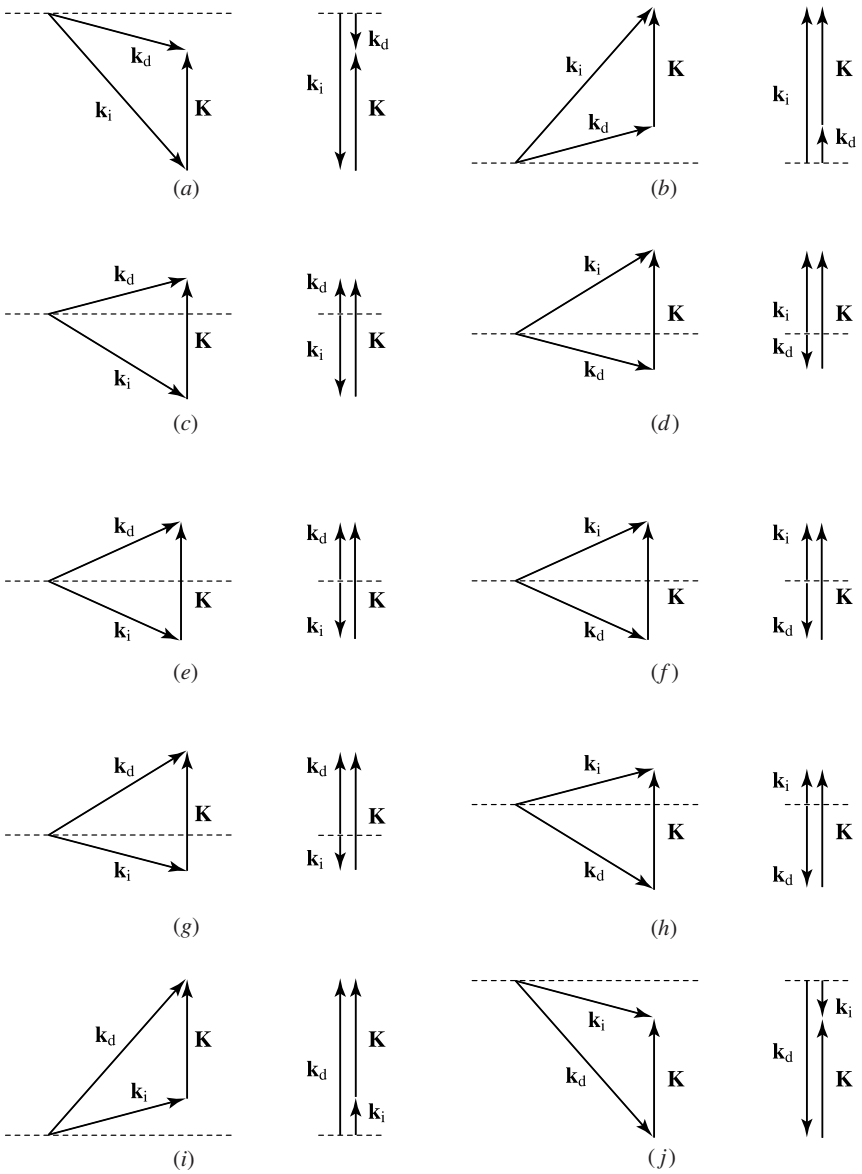


Figure 8.4 Phase-matching configurations for Bragg diffraction from a traveling acoustic wave of a wavevector $\mathbf{K} = K\hat{x}$ under various situations: (*a*)–(*d*) $k_d < k_i$, (*e*) and (*f*) $k_d = k_i$, (*g*)–(*j*) $k_d > k_i$. Configurations for up-shifted diffraction are shown in (*a*), (*c*), (*e*), (*g*), and (*i*) on the left. Those for down-shifted diffraction are shown in (*b*), (*d*), (*f*), (*h*), and (*j*) on the right. Both noncollinear and collinear phase-matching configurations are shown in each case.

K, there are no solutions for θ_i and θ_d. In such cases, Bragg diffraction cannot occur because the required phase-matching condition cannot be satisfied.

Figure 8.4 shows various phase-matching configurations for up-shifted and down-shifted Bragg diffraction from a traveling acoustic wave of a given wavevector $\mathbf{K}$. Several remarks on phase-matched Bragg diffraction can be made:

1. Bragg diffraction in an isotropic medium, with or without a change of polarization, and that in an anisotropic medium without a change of polarization between the incident and the diffracted waves are nonbirefringent because $k_d \approx k_i = k$ in such cases. The angles of incidence and diffraction in nonbirefringent Bragg diffraction are of equal magnitude but opposite signs:

$$\theta_i = -\theta_B, \qquad \theta_d = \theta_B, \tag{8.63}$$

for up-shifted diffraction, and

$$\theta_i = \theta_B, \qquad \theta_d = -\theta_B, \tag{8.64}$$

for down-shifted diffraction, where

$$\theta_B = \sin^{-1}\frac{K}{2k} = \sin^{-1}\frac{\lambda f}{2nv_a} \tag{8.65}$$

is known as the *Bragg angle*. For a monochromatic optical wave incident at any θ_i, there is always one and only one value of K that can satisfy the phase-matching condition for either up-shifted or down-shifted Bragg diffraction, depending on the sign of θ_i, as demonstrated in Figs. 8.4(e) and (f).

2. For Bragg diffraction in an isotropic medium, the values of K and f that allow phase-matched interaction fall in the following range:

$$0 \le K \le 2k, \quad \text{or} \quad 0 \le f \le \frac{2nv_a}{\lambda}. \tag{8.66}$$

For birefringent Bragg diffraction in an anisotropic medium in general, the range of K and f for phase-matched interaction is

$$|k_i - k_d| \le K \le |k_i + k_d|, \quad \text{or} \quad \frac{|n_i - n_d|v_a}{\lambda} \le f \le \frac{|n_i + n_d|v_a}{\lambda}. \tag{8.67}$$

3. For birefringent Bragg diffraction in an anisotropic medium, there is a change of polarization between the incident and the diffracted waves, and k_d can be either smaller or larger than k_i. In the case when $k_d < k_i$, Bragg diffraction occurs only if the incident angle satisfies $\pi/2 \ge |\theta_i| > \cos^{-1}(k_d/k_i)$. For each acceptable value of θ_i, there may exist two values or only one value of K for phase matching if the diffracted wave is extraordinary so that the value of k_d depends on θ_d, but two values of K always exist for each acceptable θ_i if the diffracted wave is ordinary. Up-shifted diffraction occurs when θ_i has a negative value, as demonstrated in Figs. 8.4(a) and (c). Down-shifted diffraction occurs when θ_i has a positive value, as demonstrated in Figs. 8.4(b) and (d). In the case when $k_d > k_i$, for any incident angle except $\theta_i = 0$, one value of K exists for up-shifted diffraction and another for down-shifted diffraction, if the values of both k_i and k_d are fixed. For given $\mathbf{k}_i$ and $\mathbf{K}$, this implies that $\mathbf{k}_d$ for up-shifted diffraction and that for down-shifted diffraction have different directions and different magnitudes, as demonstrated in Figs. 8.4(g)–(j).

4. When $\theta_i = \pi/2$ or $-\pi/2$, the phase-matching configurations are collinear. In these configurations, also shown in Fig. 8.4, $\mathbf{k}_i$ and $\mathbf{k}_d$ are both collinear with $\mathbf{K}$ but can be either parallel or antiparallel to each other.

EXAMPLE 8.3 The angles of incidence and diffraction for Bragg diffraction are determined by the parameters of both the optical and the acoustic waves in the medium, including λ, n_i, and n_d for the optical waves and f and v_a for the acoustic wave. To see the dependences of θ_i and θ_d on these parameters clearly, it is convenient to define a dimensionless normalized acoustic frequency:

$$\hat{f} = \frac{\lambda f}{(n_i + n_d)v_a}. \tag{8.68}$$

Then θ_i and θ_d given in (8.61) and (8.62) for up-shifted Bragg diffraction can be expressed as

$$\theta_i = -\sin^{-1}\frac{\hat{f} + \hat{f}_{min}/\hat{f}}{1 + \hat{f}_{min}}, \quad \theta_d = \sin^{-1}\frac{\hat{f} - \hat{f}_{min}/\hat{f}}{1 - \hat{f}_{min}}, \tag{8.69}$$

where

$$\hat{f}_{min} = \frac{n_i - n_d}{n_i + n_d}. \tag{8.70}$$

Note that $\hat{f}_{min}$ can be either positive or negative depending on whether $n_i > n_d$ or $n_d > n_i$. For down-shifted Bragg diffraction, the signs of both θ_i and θ_d change from those seen in (8.69). For nonbirefringent Bragg diffraction, we simply set $n_i = n_d = n$ so that $\hat{f}_{min} = 0$ and $\theta_i = -\theta_d = -\theta_B$ with

$$\theta_B = \sin^{-1}\hat{f}. \tag{8.71}$$

We find from (8.69) that θ_i and θ_d have solutions only if the frequency $\hat{f}$ falls within the range:

$$|\hat{f}_{min}| \le \hat{f} \le 1. \tag{8.72}$$

This condition is the same as that in (8.67) for birefringent Bragg diffraction and that in (8.66) for nonbirefringent Bragg diffraction. Therefore, $|\hat{f}_{min}|$ is the minimum normalized acoustic frequency that allows a phase-matched birefringent Bragg interaction, whereas there is no minimum acoustic frequency for phase-matched nonbirefringent Bragg diffraction. In the case when $n_i > n_d$ so that $\hat{f}_{min} > 0$, we also find from (8.69) that $\theta_d = 0$ while $|\theta_i|$ has a minimum value of $|\theta_i^{min}|$ for an incident angle of

$$\theta_i^{min} = -\sin^{-1}\frac{2|\hat{f}_{min}|^{1/2}}{1 + |\hat{f}_{min}|} = -\cos^{-1}\frac{n_d}{n_i} \tag{8.73}$$

at the frequency of

$$\hat{f}_t = |\hat{f}_{min}|^{1/2}. \tag{8.74}$$

In the case when $n_d > n_i$, we find $\theta_i = 0$ and $\theta_d = \theta_d^{min} = \cos^{-1} n_i/n_d$ at the frequency

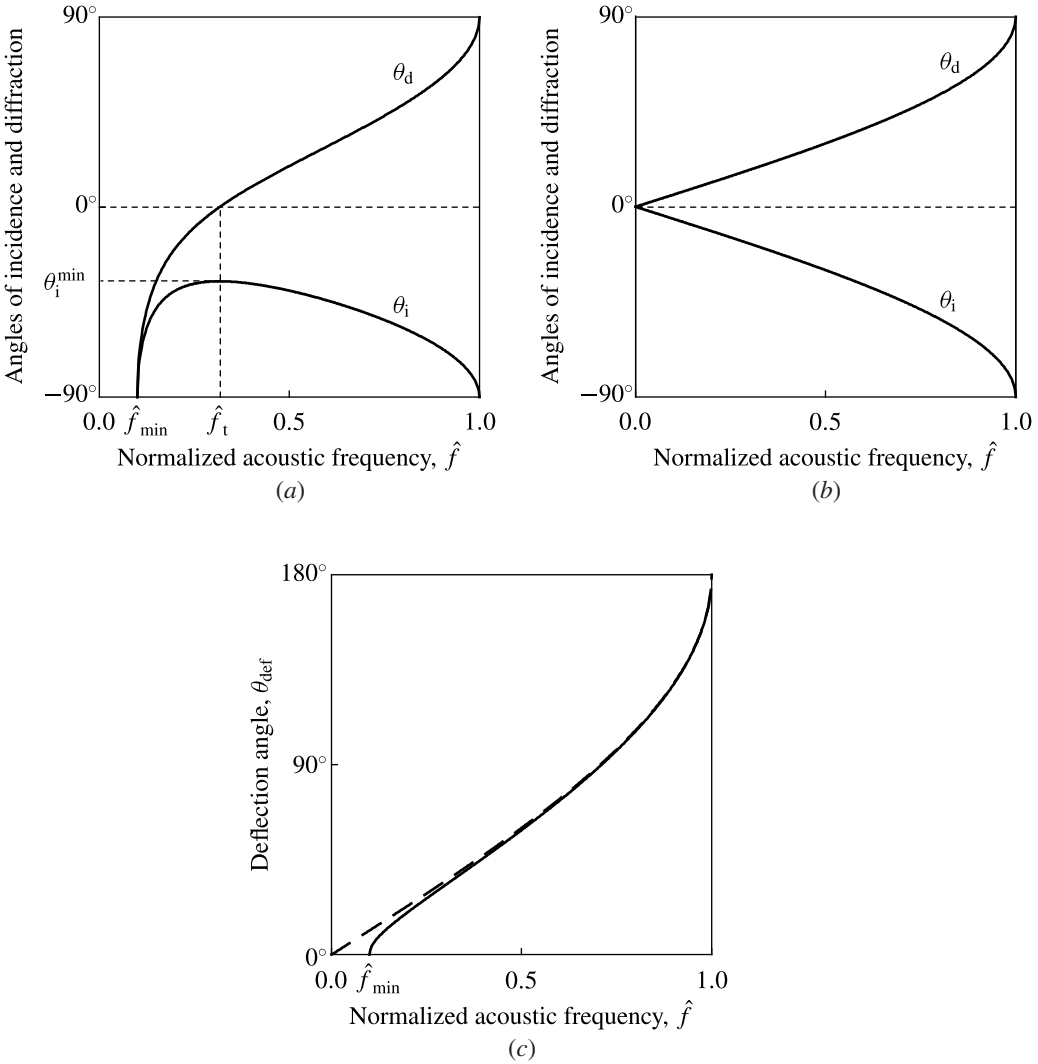


Figure 8.5 Angles of incidence and diffraction as a function of the dimensionless normalized acoustic frequency, $\hat{f}$, for (*a*) up-shifted birefringent Bragg diffraction with $n_i = 2.2$ and $n_d = 1.8$ and (*b*) up-shifted nonbirefringent Bragg diffraction with $n = 2$. (*c*) Deflection angle as a function of the dimensionless normalized acoustic frequency for both birefringent (solid curve) and nonbirefringent (dashed curve) cases. Tangential phase matching for birefringent diffraction occurs at the frequency $\hat{f}_t$.

$\hat{f}_t$. Phase matching for birefringent Bragg diffraction at $\hat{f}_t$ so that one angle is zero and another has a minimum absolute value is known as *tangential phase matching* or $90°$*phase matching* because either $\mathbf{k}_d$ (in the case when $n_i > n_d$) or $\mathbf{k}_i$ (in the case when $n_d > n_i$) is perpendicular to $\mathbf{K}$ in this situation.

To illustrate the dependences of θ_i and θ_d on the value of $\hat{f}$ numerically and to compare birefringent and nonbirefringent interactions, we consider a birefringent case

with $n_\mathrm{i} = 2.2$ and $n_\mathrm{d} = 1.8$ and a nonbirefringent case with $n = 2$ so that $n_\mathrm{i} > n_\mathrm{d}$ and $n_\mathrm{i} + n_\mathrm{d} = 2n = 4$. For birefringent diffraction, we find that $\hat{f}_{\min} = 0.1$ and that $\theta_\mathrm{i} = \theta_\mathrm{i}^{\min} = -35.1^\circ$ and $\theta_\mathrm{d} = 0$ at $\hat{f}_\mathrm{t} = 0.316$. The values of θ_i and θ_d as a function of $\hat{f}$ are shown in Fig. 8.5(a) for birefringent Bragg diffraction and in Fig. 8.5(b) for nonbirefringent Bragg diffraction. The *deflection angle* $\theta_\mathrm{def} = \theta_\mathrm{d} - \theta_\mathrm{i}$ measured from the direction of the incident wave to that of the diffracted wave is shown in Fig. 8.5(c) for both cases. As can be seen, the dependences of θ_i and θ_d on $\hat{f}$ differ significantly between birefringent and nonbirefringent cases, particularly at frequencies near and below $\hat{f}_\mathrm{t}$. However, there is very little difference in the deflection angle θ_def between the two cases except at frequencies near and below $|\hat{f}_{\min}|$ where birefringent diffraction has a cutoff but nonbirefringent diffraction does not.

Taking the approximation that $\omega_1 \approx \omega \approx \omega_{-1}$ and considering the general situation that the phase-matching condition may not be perfectly satisfied, we find from (8.38) the following coupled equations for Bragg diffraction:

$$\cos\theta_\mathrm{i}\frac{\partial \mathcal{E}_\mathrm{i}}{\partial z} + \sin\theta_\mathrm{i}\frac{\partial \mathcal{E}_\mathrm{i}}{\partial x} = \mp\frac{\omega^2\mu_0\Delta\tilde{\epsilon}_\mathrm{id}}{4k_\mathrm{i}}\mathcal{E}_\mathrm{d}\mathrm{e}^{\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}}, \tag{8.75}$$

$$\cos\theta_\mathrm{d}\frac{\partial \mathcal{E}_\mathrm{d}}{\partial z} + \sin\theta_\mathrm{d}\frac{\partial \mathcal{E}_\mathrm{d}}{\partial x} = \pm\frac{\omega^2\mu_0\Delta\tilde{\epsilon}_\mathrm{di}}{4k_\mathrm{d}}\mathcal{E}_\mathrm{i}\mathrm{e}^{-\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}}, \tag{8.76}$$

where $\Delta\tilde{\epsilon}_\mathrm{id} = \hat{e}_\mathrm{i}^* \cdot \Delta\tilde{\boldsymbol{\epsilon}} \cdot \hat{e}_\mathrm{d} = \Delta\tilde{\epsilon}_\mathrm{di}^*$. For up-shifted diffraction, $\Delta\mathbf{k} = \mathbf{k}_\mathrm{d} - \mathbf{k}_\mathrm{i} - \mathbf{K}$, and the top signs are used on the right-hand side of both (8.75) and (8.76). For down-shifted diffraction, $\Delta\mathbf{k} = \mathbf{k}_\mathrm{d} - \mathbf{k}_\mathrm{i} + \mathbf{K}$, and the bottom signs are used. According to the discussions in Section 4.1, it is also understood that in the case of diffraction in an anisotropic medium, (8.75) and (8.76) represent the transverse components of the fields being coupled. We can define the following normalized amplitude for an optical field:

$$A = \left(\frac{2k}{\omega\mu_0}\right)^{1/2}\mathcal{E} \tag{8.77}$$

so that, according to (1.98), the intensity of the field is simply

$$I = |A|^2. \tag{8.78}$$

Then, the coupled equations for Bragg diffraction can be written as

$$\cos\theta_\mathrm{i}\frac{\partial A_\mathrm{i}}{\partial z} + \sin\theta_\mathrm{i}\frac{\partial A_\mathrm{i}}{\partial x} = \mathrm{i}\kappa A_\mathrm{d}\mathrm{e}^{\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}}, \tag{8.79}$$

$$\cos\theta_\mathrm{d}\frac{\partial A_\mathrm{d}}{\partial z} + \sin\theta_\mathrm{d}\frac{\partial A_\mathrm{d}}{\partial x} = \mathrm{i}\kappa^* A_\mathrm{i}\mathrm{e}^{-\mathrm{i}\Delta\mathbf{k}\cdot\mathbf{r}}, \tag{8.80}$$

where, for up-shifted or down-shifted diffraction, respectively,

$$\kappa = \mathrm{i}\frac{\omega^2\mu_0\Delta\tilde{\epsilon}_\mathrm{id}}{4k_\mathrm{i}^{1/2}k_\mathrm{d}^{1/2}} \quad \text{or} \quad \kappa = -\mathrm{i}\frac{\omega^2\mu_0\Delta\tilde{\epsilon}_\mathrm{id}}{4k_\mathrm{i}^{1/2}k_\mathrm{d}^{1/2}}. \tag{8.81}$$

Using (8.22), we find that the Bragg coupling coefficient can be expressed in terms of the acousto-optic figure of merit as

$$|\kappa| = \frac{\pi}{\lambda}\left(\frac{M_2 I_a}{2}\right)^{1/2}, \tag{8.82}$$

which has the same form as that of the coupling coefficient for Raman–Nath diffraction defined in (8.54). In general, θ_i and θ_d can have any values between $-\pi/2$ and $\pi/2$, and A_i and A_d can vary with both x and z. However, two extreme cases that have much simplified solutions for the coupled equations in (8.79) and (8.80) are of particular interest and are discussed in the following.

Small-angle Bragg diffraction: $\theta_i \approx 0$

In case the angles of incidence and diffraction are both very small, $\cos\theta_i$ and $\cos\theta_d$ can both be approximated by unity, and $\sin\theta_i$ and $\sin\theta_d$ are both approximately zero. This is the situation in which the optical waves propagate almost perpendicularly to the acoustic wave. It normally occurs when the acoustic wavelength is much larger than the optical wavelength but the interaction length is large so that (8.57) is satisfied. In this case, the field amplitudes, A_i and A_d, vary primarily with z only, and the phase mismatch is $\Delta\mathbf{k} = \Delta k\hat{z}$. The coupled equations in (8.79) and (8.80) then reduce to

$$\frac{dA_i}{dz} = i\kappa A_d e^{i\Delta kz}, \tag{8.83}$$

$$\frac{dA_d}{dz} = i\kappa^* A_i e^{-i\Delta kz}. \tag{8.84}$$

The boundary conditions are $A_i(0) \neq 0$ and $A_d(0) = 0$. These coupled equations describe *codirectional coupling* of the incident and the diffracted waves with symmetric coupling coefficients. They have the solutions obtained in Section 4.3 for codirectionally coupled modes when we identify Δk with 2δ. Therefore, the *codirectional Bragg diffraction efficiency* over an interaction length l is

$$\eta = \frac{I_d(l)}{I_i(0)} = \frac{|A_d(l)|^2}{|A_i(0)|^2} = \frac{1}{1 + \Delta k^2/4|\kappa|^2}\sin^2\left(|\kappa| l\sqrt{1 + \Delta k^2/4|\kappa|^2}\right). \tag{8.85}$$

The diffraction efficiency in the case of perfect phase matching is

$$\eta_{PM} = \sin^2|\kappa| l. \tag{8.86}$$

Using (8.82), it is found that the acoustic intensity needed for 100% Bragg diffraction efficiency is

$$I_a = \frac{\lambda^2}{2M_2 l^2}. \tag{8.87}$$

Collinear Bragg diffraction: $\theta_i = \pm\pi/2$

Another special case of particular interest is collinear Bragg diffraction, which occurs when the incident optical wave propagates collinearly with the acoustic wave. The wavevector $\mathbf{k}_i$ can be either parallel or antiparallel to $\mathbf{K}$, corresponding to $\theta_i = \pi/2$ and $-\pi/2$, respectively. The phase-matching condition in (8.60) then requires that θ_d be either $\pi/2$ or $-\pi/2$. Therefore, $\mathbf{k}_d$ is collinear with both $\mathbf{k}_i$ and $\mathbf{K}$. As can be seen from the collinear phase-matching configurations shown in Fig. 8.4, the diffracted optical wave can propagate either codirectionally or contradirectionally with respect to the incident optical wave. *Collinear Bragg diffraction in an isotropic medium is always contradirectional* because $k_i = k_d = K/2$ in this situation. However, *collinear Bragg diffraction in an anisotropic medium can be codirectional or contradirectional* because two values of K exist for phase-matching in each case when $k_i \neq k_d$. Codirectional Bragg diffraction occurs in an anisotropic medium with an acoustic wave at a low frequency corresponding to a small K value, whereas contradirectional Bragg diffraction occurs with an acoustic wave at a high frequency corresponding to a large K value.

From the above discussions, we find that the coupled equations describing collinear Bragg diffraction can be either those of codirectional coupling or those of contradirectional coupling, depending on whether the propagation directions of the incident and the diffracted waves are codirectional or contradirectional. In either case, with $\theta_i = \pm\pi/2$, $\theta_d = \pm\pi/2$, and $\mathbf{K} = K\hat{x}$, both A_i and A_d vary only with x and $\Delta\mathbf{k} = \Delta k\hat{x}$. The coupled equations in (8.79) and (8.80) then reduce to

$$\pm\frac{dA_i}{dx} = i\kappa A_d e^{i\Delta kx}, \tag{8.88}$$

$$\pm\frac{dA_d}{dx} = i\kappa^* A_i e^{-i\Delta kx}. \tag{8.89}$$

In (8.88), the plus or minus sign on the left-hand side is chosen respectively according to whether $\mathbf{k}_i$ points in the direction of $\hat{x}$ or $-\hat{x}$. Similarly, in (8.89), the plus or minus sign is chosen respectively according to whether $\mathbf{k}_d$ points in the direction of $\hat{x}$ or $-\hat{x}$.

1. When $\mathbf{k}_i$ and $\mathbf{k}_d$ point in the same direction, the same sign is chosen in both (8.88) and (8.89) for codirectional coupling between the incident and the diffracted waves. The boundary conditions for codirectional Bragg diffraction are $A_i(0) \neq 0$ and $A_d(0) = 0$, and the diffraction efficiency is that given in (8.85) in general and that in (8.86) for perfect phase matching. *Collinear, codirectional Bragg diffraction is birefringent and is possible only in anisotropic media. It is always accompanied by a change of polarization between the incident and the diffracted waves.*
2. When $\mathbf{k}_i$ and $\mathbf{k}_d$ point in opposite directions, different signs are chosen in (8.88) and (8.89) for contradirectional coupling between the incident and the diffracted waves. For contradirectional Bragg diffraction over an interaction length l, the boundary

conditions are $A_i(0) \neq 0$ and $A_d(l) = 0$ if $\mathbf{k}_d = -k_d\hat{x}$ and are $A_i(0) \neq 0$ and $A_d(-l) = 0$ if $\mathbf{k}_d = k_d\hat{x}$. The solutions obtained in Section 4.3 for contradirectionally coupled modes can be used when we identify Δk with 2δ. Therefore, the *contradirectional Bragg diffraction efficiency* is

$$\eta = \frac{I_d(0)}{I_i(0)} = \frac{|A_d(0)|^2}{|A_i(0)|^2} = \frac{\sinh^2\left(|\kappa| l\sqrt{1 - \Delta k^2/4|\kappa|^2}\right)}{\cosh^2\left(|\kappa| l\sqrt{1 - \Delta k^2/4|\kappa|^2}\right) - \Delta k^2/4|\kappa|^2}. \tag{8.90}$$

The diffraction efficiency in the case of perfect phase matching is

$$\eta_{PM} = \tanh^2 |\kappa| l. \tag{8.91}$$

Collinear, contradirectional Bragg diffraction is possible in both isotropic and anisotropic media. A change of polarization between the incident and the diffracted waves may or may not occur in this process.

EXAMPLE 8.4 An optical wave at 632.8 nm wavelength interacts with a longitudinal acoustic wave of a frequency $f = 100$ MHz in a piece of fused silica glass over an interaction length of $l = 4$ cm. The incident optical wave is polarized in a direction $\hat{e}_i$ that is perpendicular to vector $\mathbf{K}$ of the acoustic wave. Is this interaction in the Bragg regime? What incident angle θ_i should be chosen for phase matching? What is the deflection angle θ_{def}? Find the acoustic intensity that is required for a 100% diffraction efficiency if that is possible.

Solution From Table 8.2, we find that $v_a = v_{a,L} = 5.97$ km s^{-1} for a longitudinal acoustic wave in silica glass and $n = 1.457$ at 632.8 nm. At the acoustic frequency of $f = 100$ MHz, we find that

$$Q = 2\pi\frac{\lambda f^2 l}{n v_a^2} = 2\pi \times \frac{632.8 \times 10^{-9} \times (100 \times 10^6)^2 \times 4 \times 10^{-2}}{1.457 \times (5.97 \times 10^3)^2} = 30.6 \gg 1.$$

Therefore, the interaction is in the Bragg regime, and phase matching is required. Because this is nonbirefringent phase matching, the angles of incidence and diffraction are both defined by the following Bragg angle:

$$\theta_B = \sin^{-1}\frac{\lambda f}{2n v_a} = \sin^{-1}\frac{632.8 \times 10^{-9} \times 100 \times 10^6}{2 \times 1.457 \times 5.97 \times 10^3} = 0.21^\circ.$$

We then have $\theta_i = -\theta_B = -0.21^\circ$ and $\theta_{def} = 2\theta_B = 0.42^\circ$ for up-shifted diffraction and $\theta_i = \theta_B = 0.21^\circ$ and $\theta_{def} = -2\theta_B = -0.42^\circ$ for down-shifted diffraction.

Because $\theta_i \approx 0$ in this problem, this is a case of small-angle Bragg diffraction. It is therefore possible to accomplish a 100% diffraction efficiency. Because $\hat{e}_i \perp \mathbf{K}$, the relevant figure of merit for this interaction is $M_2^{\perp} = 1.5 \times 10^{-15}$ m^2 W^{-1} found in

Example 8.1. We then find, using (8.87), that the required acoustic intensity is

$$I_a = \frac{\lambda^2}{2M_2 l^2} = \frac{(632.8 \times 10^{-9})^2}{2 \times 1.5 \times 10^{-15} \times (4 \times 10^{-2})^2} \text{ W m}^{-2} = 8.29 \text{ W cm}^{-2}.$$

Diffraction from a standing acoustic wave

So far only diffraction from a traveling acoustic wave has been considered. We have seen that each spatial diffraction order defined by a wavevector $\mathbf{k}_q$ at a diffraction angle θ_q contains a single, uniquely defined frequency of $\omega_q = \omega + q\Omega$. This is not the case, however, for diffraction from a standing acoustic wave.

A standing acoustic wave can be considered as a linear superposition of two contrapropagating traveling waves with both $\mathbf{K}$ and $-\mathbf{K}$ existing simultaneously for phase matching. The implication of this situation is two-fold: (1) both up-shifted and down-shifted frequencies are simultaneously generated in each phase-matched direction of diffraction, and (2) each shifted optical frequency generated by diffraction can be diffracted back to the direction of the incident wave with a further shift in frequency. This process cascades. Figure 8.6 shows the cascading process in the case of Raman–Nath diffraction from a standing acoustic wave. At the output, each of the even spatial orders, including the undiffracted zeroth order, consists of all of the frequencies up- or down-shifted by even multiples of Ω, whereas each of the odd spatial orders consists of all of the frequencies up- or down-shifted by odd multiples of Ω.

For Bragg diffraction from a standing acoustic wave, the incident angle can be either θ_i or $-\theta_i$ of the form given in (8.61) because $\mathbf{K}$ and $-\mathbf{K}$ exist simultaneously. In either case, both up-shifted and down-shifted frequencies are generated in the direction of the

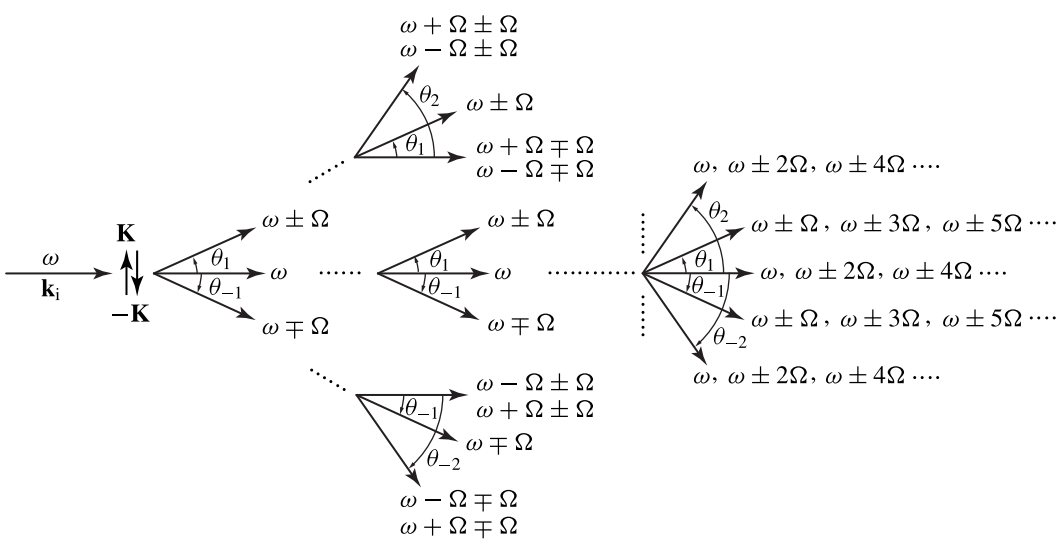


Figure 8.6 Cascading process in Raman–Nath diffraction from a standing acoustic wave.

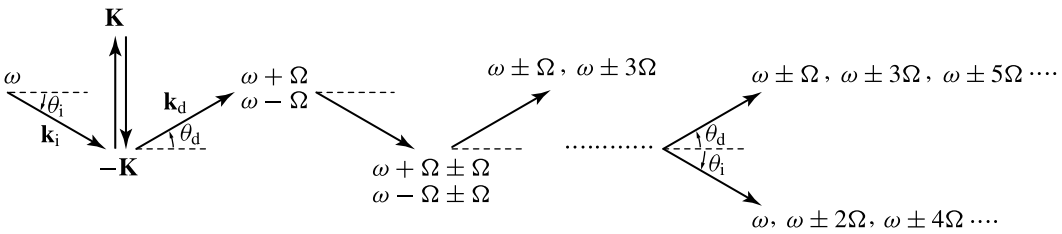


Figure 8.7 Cascading process in Bragg diffraction from a standing acoustic wave.

corresponding $\mathbf{k}_d$. This process cascades. Consequently, the undiffracted beam in the $\mathbf{k}_i$ direction contains a series of even side bands at $\omega \pm 2m\Omega$, and the diffracted beam in the $\mathbf{k}_d$ direction contains the odd side bands at $\omega \pm (2m+1)\Omega$. Figure 8.7 shows the cascading process in Bragg diffraction.

A formal analysis of diffraction from a standing acoustic wave using coupled-wave theory can be carried out, but it is very complicated because each spatial order consists of many different frequency components. Closed-form analytical solutions can be obtained, however, without solving coupled-wave equations but by extending the results obtained from the analysis of diffraction from a traveling wave. For a standing acoustic wave of the form given in (8.2), we have

$$\Delta\epsilon = \Delta\tilde{\epsilon}\sin(\mathbf{K}\cdot\mathbf{r})\cos\Omega t = \Delta\tilde{\epsilon}(t)\sin(\mathbf{K}\cdot\mathbf{r}). \tag{8.92}$$

Because $\omega \gg \Omega$, the temporal variation in $\Delta\tilde{\epsilon}(t) = \Delta\tilde{\epsilon}\cos\Omega t$ is very slow compared to the optical cycles. Consequently, in place of (8.35), we can expand the field as

$$\mathbf{E}(\mathbf{r}, t) = \sum_q \boldsymbol{\mathcal{E}}_q(\mathbf{r}, t)e^{i\mathbf{k}_q\cdot\mathbf{r}-i\omega t} = \sum_q \hat{e}_q\mathcal{E}_q(\mathbf{r}, t)e^{i\mathbf{k}_q\cdot\mathbf{r}-i\omega t}, \tag{8.93}$$

where q represents the spatial diffraction order and $\mathbf{k}_q = \mathbf{k}_i + q\mathbf{K}$.

For Raman–Nath diffraction, it follows from the analysis leading to (8.55) that for the spatial diffraction order $q \geq 0$, we have

$$\begin{aligned}\frac{\mathcal{E}_q(l, t)}{\mathcal{E}_0(0, 0)} &= J_q(-2|\kappa|l\cos\Omega t)\\ &= \sum_{n=0}^{q}\sum_{m=0}^{\infty}\sum_{p=-m}^{m}\frac{(-1)^{q+p}q!(2m)!(|\kappa|l)^m}{2^{q+2m}m!n!(q-n)!(m-p)!(m+p)!}J_{q+m}(2|\kappa|l)e^{i(q-2n-2p)\Omega t},\end{aligned} \tag{8.94}$$

and $\mathcal{E}_{-q}(l, t) = (-1)^q\mathcal{E}_q(l, t)$ for the negative spatial orders. In the expansion of (8.94), we have used the following multiplication theorem for the Bessel functions:

$$J_q(x\cos\phi) = \cos^q\phi\sum_{m=0}^{\infty}\frac{\sin^{2m}\phi}{m!}\left(\frac{x}{2}\right)^m J_{q+m}(x), \qquad \text{for } q \geq 0. \tag{8.95}$$

As can be seen from (8.94), the temporal dependence of $\mathcal{E}_q(l, t)$ contains all of the positive and negative even harmonics of Ω if q is an even integer, and it contains all

of the positive and negative odd harmonics of Ω if q is an odd integer. Therefore, each even spatial order consists of a series of frequency components at $\omega \pm 2m\Omega$, while each odd spatial order consists of a series of frequency components at $\omega \pm (2m+1)\Omega$.

A similar analysis can be carried out for Bragg diffraction in different situations. For small-angle Bragg diffraction with perfect phase matching, we have

$$\begin{aligned} A_{\mathrm{i}}(l,t) &= A_{\mathrm{i}}(0,0)\cos(|\kappa|l\cos\Omega t) \\ &= A_{\mathrm{i}}(0,0)\sum_{m=-\infty}^{\infty}(-1)^m J_{2m}(|\kappa|l)\mathrm{e}^{\mathrm{i}2m\Omega t}, \end{aligned} \tag{8.96}$$

$$\begin{aligned} A_{\mathrm{d}}(l,t) &= \mathrm{i}A_{\mathrm{i}}(0,0)\sin(|\kappa|l\cos\Omega t) \\ &= \mathrm{i}A_{\mathrm{i}}(0,0)\sum_{m=-\infty}^{\infty}(-1)^m J_{2m+1}(|\kappa|l)\mathrm{e}^{\mathrm{i}(2m+1)\Omega t}, \end{aligned} \tag{8.97}$$

where, for the expansion, we have used the following identities:

$$\cos(x\cos\phi) = \sum_{m=-\infty}^{\infty}(-1)^m J_{2m}(x)\mathrm{e}^{\mathrm{i}2m\phi}, \tag{8.98}$$

$$\sin(x\cos\phi) = \sum_{m=-\infty}^{\infty}(-1)^m J_{2m+1}(x)\mathrm{e}^{\mathrm{i}(2m+1)\phi}. \tag{8.99}$$

It can be seen clearly from (8.96) and (8.97) that the undiffracted beam consists of the frequency component at ω and all of the even side bands at $\omega \pm 2m\Omega$, while the diffracted beam consists of all of the odd side bands at $\omega \pm (2m+1)\Omega$, as discussed above.

The concept of intensity refers to the flow of energy through a unit area. It can be clearly defined for a traveling wave but is not applicable to a standing wave. However, a standing wave can be considered to be the linear superposition of two contrapropagating traveling waves of equal amplitude and, therefore, of equal intensity. For a standing acoustic wave described by (8.2), we find the amplitude of the strain tensor element representing the two contrapropagating acoustic waves to be $\mathcal{S}^{\mathrm{f}} = \mathcal{S}^{\mathrm{b}} = \mathcal{S}/2$, where $\mathcal{S}$ is the same as the tensor element $\mathcal{S}_{kl}$ defined in (8.18). Therefore, the intensities of the two contrapropagating traveling acoustic waves are given by

$$I_{\mathrm{a}}^{\mathrm{f}} = I_{\mathrm{a}}^{\mathrm{b}} = \frac{1}{2}\left(\frac{\mathcal{S}}{2}\right)^2 \rho v_{\mathrm{a}}^3. \tag{8.100}$$

Using (8.20) for the relation between I_{a} and $\mathcal{S}$ used in defining M_2 in (8.22), we find from the above that $I_{\mathrm{a}} = 4I_{\mathrm{a}}^{\mathrm{f}} = 4I_{\mathrm{a}}^{\mathrm{b}}$. We then find, using (8.22), that

$$|\Delta\tilde{\epsilon}_{\mathrm{id}}|^2 = 8\epsilon_0^2 n_{\mathrm{i}} n_{\mathrm{d}} M_2 I_{\mathrm{a}}^{\mathrm{f}} = 8\epsilon_0^2 n_{\mathrm{i}} n_{\mathrm{d}} M_2 I_{\mathrm{a}}^{\mathrm{b}}. \tag{8.101}$$

Using (8.101) in (8.53) and (8.81) for the cases of Raman–Nath and Bragg diffraction, respectively, we find that the coupling coefficient for both cases is

$$|\kappa| = \frac{2\pi}{\lambda}\left(\frac{M_2 I_{\mathrm{a}}^{\mathrm{f}}}{2}\right)^{1/2} = \frac{2\pi}{\lambda}\left(\frac{M_2 I_{\mathrm{a}}^{\mathrm{b}}}{2}\right)^{1/2}. \tag{8.102}$$

This is the coupling coefficient that appears in (8.94) and in (8.96) and (8.97) for diffraction from a standing acoustic wave. It has a form different from that of (8.54) and (8.82) for diffraction from a traveling acoustic wave.

In both Raman–Nath and Bragg diffraction from a standing acoustic wave, the number of frequencies that appear in each spatially separated, diffracted or undiffracted, beam is determined by the dispersion and the bandwidth of the medium, as well as by the total length and the strength of interaction.

8.4 Acousto-optic modulators

The acoustic wave of an acousto-optic modulator is *amplitude modulated*. The operation of an acousto-optic modulator is based on the dependence of the acousto-optic diffraction efficiency on the intensity of the acoustic wave. The acoustic intensity can be controlled by an electrical signal that generates the acoustic wave in a modulator. An acousto-optic modulator is an electronically addressed *amplitude modulator* that accepts an electrical modulation signal to vary the intensity of an optical beam accordingly.

Acousto-optic modulators have been put to many different applications. The straightforward application is amplitude modulation of an optical beam, thus encoding a modulation signal on an optical carrier or providing loss modulation to an optical system such as a Q-switched or mode-locked laser. Sophisticated applications include time-domain convolution and correlation of wide-band RF signals in signal processing systems.

Generally speaking, an acousto-optic modulator can operate either in the Bragg regime or in the Raman–Nath regime. In the low-efficiency limit, the efficiency of the first diffraction order of a Raman–Nath-type modulator is similar to the diffraction efficiency of a Bragg-type modulator. For low-efficiency diffraction from a traveling wave with $|\kappa| l \ll 1$, the efficiency of the first diffraction order of a Raman–Nath-type modulator is, according to (8.56),

$$\eta_1 = J_1^2(2|\kappa| l) \approx |\kappa|^2 l^2 = \frac{\pi^2 M_2 l^2}{2\lambda^2} I_{\mathrm{a}}, \tag{8.103}$$

while the diffraction efficiency of a Bragg-type modulator with perfect phase matching is, according to (8.86),

$$\eta_{\mathrm{PM}} = \sin^2 |\kappa| l \approx |\kappa|^2 l^2 = \frac{\pi^2 M_2 l^2}{2\lambda^2} I_{\mathrm{a}}. \tag{8.104}$$

A similar comparison can be made in the case of diffraction from a standing wave using (8.94) and (8.97). However, notwithstanding this similarity, an acousto-optic modulator operating in the Raman–Nath regime has a few disadvantages in comparison to one operating in the Bragg regime. For this reason, most acousto-optic modulators designed for practical applications are of Bragg type.

One obvious disadvantage of a Raman–Nath-type modulator easily seen from Fig. 8.3 is that none of the diffraction orders can reach a diffraction efficiency higher than 34%. In comparison, a phase-matched Bragg-type modulator has a maximum diffraction efficiency of 100%. Other disadvantages stem from the condition in (8.49), which is required for Raman–Nath diffraction. Under the constraint of this condition, the acceptable interaction length l decreases quadratically as a function of acoustic frequency and linearly as a function of optical wavelength. At high acoustic frequencies and/or for long optical wavelengths, the interaction length becomes impracticably small so that a very large acoustic intensity is required in order to have a significant diffraction efficiency. A modulator of Raman–Nath type is thus limited to applications with low acoustic frequencies and, consequently, small bandwidths. Such limitations do not apply to a modulator of Bragg type.

Traveling-wave modulators

Traveling acoustic waves are used in the majority of acousto-optic modulators. The most important performance characteristics to be considered for a traveling-wave modulator are its diffraction efficiency η, its bandwidth, measured by a 3-dB modulation bandwidth $f_{\mathrm{m}}^{\mathrm{3dB}}$, and its speed, measured by a modulation response risetime, t_{r}. A traveling acousto-optic modulator is normally operated in the Bragg regime with a focused optical beam of small spot size to reduce the acoustic transit time across the optical beam and thus increase its modulation bandwidth and speed.

Figure 8.8 shows the diagram of a typical solid-state acousto-optic modulator operating with a traveling acoustic wave in the Bragg regime. The acousto-optic cell, which can be a crystal or a noncrystalline glass, is attached to a piezoelectric transducer at one end and is terminated by an angled surface at the other end. The piezoelectric transducer consists of a metallic electrode, a piezoelectric crystal such as $LiNbO_3$, and one or more metallic bonding layers for the attachment of the piezoelectric crystal to the acousto-optic cell. It converts the applied RF electrical signal into an acoustic signal and couples the acoustic power to the acousto-optic cell to generate the traveling acoustic wave. The angled termination surface reflects the acoustic wave away from the incident direction so as to prevent the reflected acoustic wave from interacting with the optical beam. This back surface is also often loaded with an acoustically absorbing material to reduce acoustic reflection.

The cross-sectional area of the acoustic beam in the modulator shown in Fig. 8.8 is HL, which is defined by the length L and height H of the transducer. The acoustic

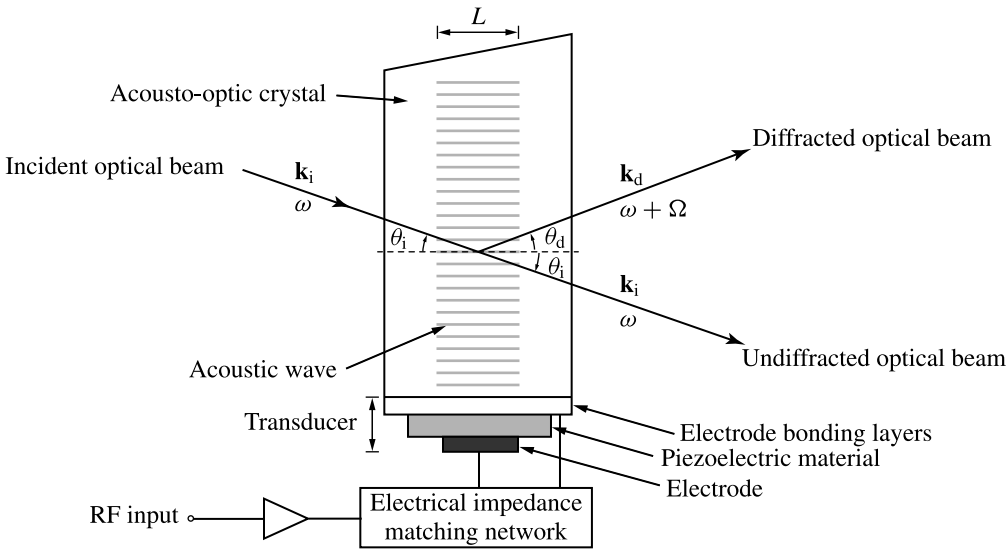


Figure 8.8 Typical solid-state acousto-optic modulator operating with a traveling acoustic wave in the Bragg regime. Up-shifted diffraction is illustrated here. For an anisotropic acousto-optic modulator, $|\theta_i| \neq |\theta_d|$. For an isotropic acousto-optic modulator, $|\theta_i| = |\theta_d| = \theta_B$.

intensity is

$$I_a = \frac{P_a}{HL} = \frac{\eta_t P_e}{HL}, \tag{8.105}$$

where P_a is the acoustic power delivered by the transducer to the acoustic medium, P_e is the power of the electrical signal driving the transducer, and η_t is the conversion efficiency of the transducer from electric to acoustic power. Using (8.82) and (8.86), we find that the diffraction efficiency of a Bragg-type traveling-wave modulator with perfect phase matching can be expressed as

$$\eta_{PM} = \sin^2\left[\frac{\pi}{\lambda}\left(\frac{M_2}{2HL}P_a\right)^{1/2} l\right] = \sin^2\left[\frac{\pi}{\lambda}\left(\frac{M_2}{2HL}\eta_t P_e\right)^{1/2} l\right]. \tag{8.106}$$

In the low-efficiency limit, the diffraction efficiency is linearly proportional to the modulation power:

$$\eta_{PM} \approx \frac{\pi^2 M_2 l^2}{2\lambda^2 HL} P_a = \frac{\pi^2 M_2 l^2}{2\lambda^2 HL}\eta_t P_e, \qquad \text{if } \eta_{PM} \ll 1. \tag{8.107}$$

For a modulator using a traveling acoustic wave, a time-dependent $P_e(t)$ that carries the modulation signal is applied to the device so that the diffraction efficiency varies with time. When an acousto-optic modulator is used as a loss modulator, the transmittance from the incident optical beam to the undiffracted optical beam is $T = 1 - \eta_{PM}$ in the situation of perfect phase matching.

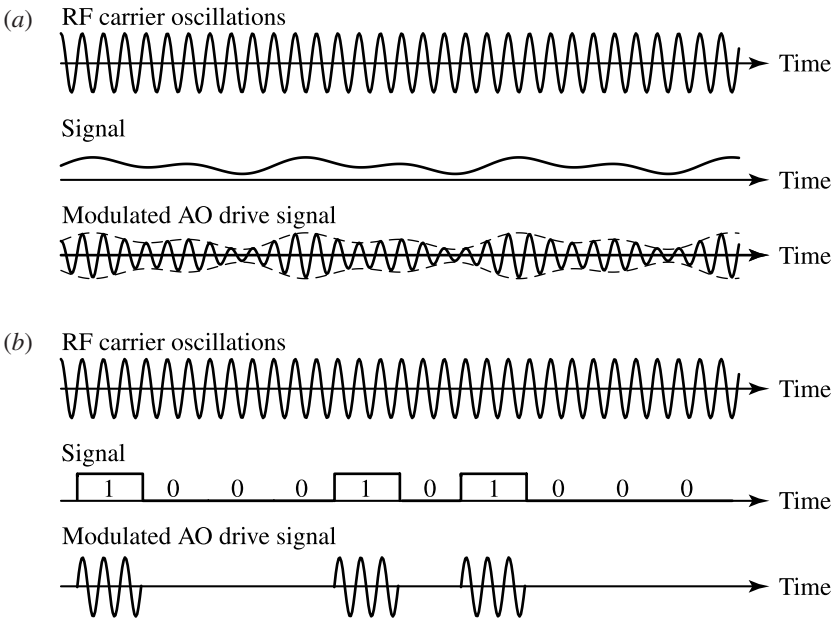


Figure 8.9 (*a*) Continuously varying amplitude modulation signal and (*b*) pulsed digital amplitude modulation signal carried by a traveling acoustic wave.

Most acousto-optic modulators take the configuration of small-angle Bragg interaction shown in Fig. 8.8. In this configuration, acousto-optic interaction takes place when the optical beam passes through the width of the acoustic beam. The interaction length is then defined by the length of the transducer: $l = L$.

The amplitude modulation signal that is applied to a traveling-wave acousto-optic modulator can be either a continuously varying signal, as shown in Fig. 8.9(*a*), or a pulsed digital signal, as shown in Fig. 8.9(*b*). In either case, the modulation signal appears as a modulation on the amplitude, thus on the intensity, of the acoustic wave at a carrier frequency of $f_0 = \Omega_0/2\pi$. The frequency, f_m, of a sinusoidal modulation signal that is imposed on an acoustic carrier wave of a frequency f_0 must satisfy the condition that $f_\mathrm{m} < f_0$. This modulation generates two side-band acoustic frequencies at $f_0 - f_\mathrm{m}$ and $f_0 + f_\mathrm{m}$. Through the dependence of the acousto-optic diffraction efficiency on the intensity of the acoustic wave, the intensities of the diffracted and undiffracted optical beams vary with the amplitude modulation signal accordingly.

In a practical modulator, both the incident optical beam and the acoustic beam have a certain degree of beam divergence because of the finite dimensions of their cross-sectional sizes. The performance characteristics of a traveling-wave acousto-optic modulator are determined by three basic parameters: (1) the factor Q; (2) the *beam divergence ratio*

$$a = \frac{\Delta\theta_\mathrm{o}}{\Delta\theta_\mathrm{a}} \tag{8.108}$$

of the optical beam divergence $\Delta\theta_o$ to the acoustic beam divergence $\Delta\theta_a$; and (3) the *acoustic transit time*

$$\tau_a = \frac{d}{v_a}, \tag{8.109}$$

defined as the time it takes for an acoustic wavefront to travel across a light beam that has a beam diameter d at the interaction point.

To ensure that the modulator operates in the Bragg regime, the value of Q has to satisfy (8.57). More precisely, to eliminate parasitic coupling of optical intensity to high diffraction orders sufficiently, it is necessary to have

$$Q = 2\pi \frac{\lambda l}{n\Lambda^2} \geq 4\pi. \tag{8.110}$$

The divergence of an optical beam is inversely proportional to its beam waist with a proportionality factor determined by the shape of the beam profile. For an optical beam of the fundamental Gaussian spatial profile discussed in Section 1.7, the divergence is given by (1.137) as

$$\Delta\theta_o = \frac{2\lambda}{\pi n w_0}, \tag{8.111}$$

where w_0 is the minimum Gaussian beam spot size in the acoustic medium. The acoustic beam can be considered to have a rectangular transverse spatial distribution with a beam width L determined by the length of the transducer, as shown in Fig. 8.8. Its divergence is given by

$$\Delta\theta_a = \frac{\Lambda}{L}. \tag{8.112}$$

Assuming that acousto-optic interaction takes place at the optical beam waist, as in the most favorable situation, the acoustic transit time is given by $\tau_a = d_0/v_a = 2w_0/v_a$. *The most important characteristics to be considered for a traveling-wave acousto-optic modulator are the diffraction efficiency and the modulator response.* Optimization of these characteristics dictates the choice of the values of a and τ_a in designing a practical acousto-optic modulator.

In the analysis of acousto-optic diffraction presented in Section 8.3, we have considered only the interaction between plane optical and plane acoustic waves, both of which were assumed to have zero divergence. In this ideal situation, perfect phase matching can be accomplished over the entire optical wavefront when the conditions in (8.60) are satisfied. Then the formulas in (8.106) and (8.107) are accurate for calculation of the diffraction efficiency. In a realistic situation where the optical beam has a nonzero divergence angle $\Delta\theta_o$, the optical wavefront covers a range of directions from $-\Delta\theta_o/2$ to $\Delta\theta_o/2$ with respect to its central direction of propagation. If the acoustic beam has an infinite plane wavefront with a single, well-defined direction of propagation, the phase-matching conditions in (8.60) cannot be simultaneously satisfied for the entire range of

optical wavefront directions. Consequently, the overall diffraction efficiency of the optical beam is degraded. In this situation, the formulas for the diffraction efficiency given in (8.106) and (8.107) become inaccurate but tend to overestimate the real efficiency of the device. In reality, however, the acoustic beam also has a nonzero divergence, and its wavefront covers a range of directions from $-\Delta\theta_a/2$ to $\Delta\theta_a/2$ with respect to its central direction of propagation. If $\Delta\theta_a \geq \Delta\theta_o$, it is possible for different optical wavefront directions to be phase matched by different acoustic wavefront directions. Therefore, to lessen the degradation in diffraction efficiency caused by divergence of the optical beam, a small value of a is required, meaning that *the optical beam has to be more collimated than the acoustic beam for maximizing the diffraction efficiency.*

The response of a traveling-wave acousto-optic modulator is primarily determined by the value of τ_a. If the modulation signal varies substantially within a time interval of τ_a, the acoustic wave intensity varies spatially across the width of the optical beam. This spatial nonuniformity of acoustic intensity across the cross section of the optical beam leads to nonuniform diffraction of the optical beam. As a result, the modulation signal carried by the acoustic wave is not faithfully converted to the modulation of the optical beam. For a given value of τ_a, this effect becomes more significant at higher modulation frequencies, thus degrading the response of a modulator to high-speed or high-frequency modulation signals. Quantitatively, the modulator response is measured by the *modulation bandwidth* in the case of a continuously varying modulation signal, or by the *modulation speed* in the case of a pulsed modulation signal. The modulation bandwidth is characterized by a 3-dB modulation frequency, f_m^{3dB}, at which point the frequency response of a modulator rolls off to 50% of its maximum response. The modulation speed is characterized by a *risetime*, t_r, which is defined as the time interval needed for the modulated optical intensity of the diffracted beam to rise from 10 to 90% of its steady-state value in response to a step modulation signal.

Detailed analysis of the response of a traveling-wave acousto-optic modulator involves convolution of the spatial intensity profile of the optical beam with the propagation of the acoustic wave carrying the modulation signal across the optical beam. The results depend on the spatial profile of the optical beam as well as on the value of a. For an optical beam of the fundamental Gaussian spatial profile, we have

$$f_m^{3dB} \approx \begin{cases} \dfrac{0.75}{\tau_a}, & a \ll 1, \\ \dfrac{0.86 - 0.13a}{\tau_a}, & a \gg 1, \end{cases} \tag{8.113}$$

and

$$t_r \approx \begin{cases} 0.65\tau_a, & a \ll 1, \\ (0.45 + 0.25a)\tau_a, & a \gg 1. \end{cases} \tag{8.114}$$

It can be seen that, besides degrading the diffraction efficiency, a large value of a also degrades the modulator response.

From the discussions presented so far, it seems that for the best performance of a modulator, the value of a should be made as small as possible. This is not true, however, because a and τ_a are both functions of the optical beam waist diameter. For a given modulator with a fixed acoustic beam divergence, the value of a can be reduced only by reducing the value of $\Delta\theta_o$ through collimation of the optical beam. The consequence is an increase in the optical beam waist diameter and a corresponding increase in the value of τ_a, thus degrading the bandwidth and the speed of the modulator. From (8.113) and (8.114), it is clear that a small value of τ_a is required for a large modulation bandwidth and, correspondingly, a high modulation speed. Indeed, to obtain a large modulation bandwidth and a high modulation speed, the optical beam has to be focused to a small beam waist located in the interaction region. These conflicting requirements lead to the need for properly choosing an optimum value of a depending on the requirements of a particular application. Once this choice is made, the value of τ_a and, consequently, the characteristics of the modulator response are basically determined.

Two additional issues regarding the functional reality of a traveling-wave acousto-optic modulator have to be considered. First, the amplitude modulation signal carried by the acoustic wave generates side-band frequencies on both high- and low-frequency sides of the carrier frequency f_0. For a modulator with a modulation bandwidth of f_m^{3dB}, the side-band frequencies cover the range from $f_0 - f_m^{3dB}$ to $f_0 + f_m^{3dB}$. Clearly, the lowest side-band frequency has to be a positive frequency in order for the modulation signal not to be distorted. In practical situations, it is often necessary to avoid nonlinear distortion of the modulation signal by requiring that the highest side-band frequency be smaller than the second harmonic of the lowest side-band frequency: $f_0 + f_m^{3dB} < 2(f_0 - f_m^{3dB})$. This requirement leads to the following condition for the carrier frequency:

$$f_0 \geq 3 f_m^{3dB}. \tag{8.115}$$

Another realistic issue is the need to separate the diffracted and the undiffracted optical beams cleanly at the output of the modulator. The clean separation between these two beams can be ensured by requiring that the deflection angle be larger than twice the beam divergence:

$$|\theta_{def}| = |\theta_d - \theta_i| > 2\Delta\theta_o. \tag{8.116}$$

In a modulator where the acousto-optic diffraction is nonbirefringent, $|\theta_{def}| = 2\theta_B$, and the above condition becomes

$$\theta_B > \Delta\theta_o. \tag{8.117}$$

The conditions discussed above set some constraints on the physical parameters of a traveling-wave acousto-optic modulator:

1. **Condition for Bragg diffraction.** The condition in (8.110) that $Q \geq 4\pi$ for Bragg diffraction sets the following minimum interaction length for a given acoustic carrier frequency:

$$L = l \geq \frac{2n\Lambda_0^2}{\lambda} = \frac{2nv_a^2}{\lambda f_0^2}, \tag{8.118}$$

where Λ_0 is the acoustic carrier wavelength corresponding to the acoustic carrier frequency f_0 and n is the refractive index of the medium. For a given value of a, we find, using (8.111) and (8.112), that this condition requires the Gaussian beam spot size located at the interaction region to be subject to the following condition:

$$d_0 = 2w_0 \geq \frac{8}{a\pi}\Lambda_0 = \frac{8v_a}{a\pi f_0}. \tag{8.119}$$

For a given acoustic beam width L and a given optical beam spot size w_0, the condition for Bragg diffraction sets the limit for the lowest acceptable carrier frequency:

$$f_0 \geq \left(\frac{2nv_a^2}{\lambda L}\right)^{1/2} = \frac{4v_a}{a\pi w_0} = \frac{8}{a\pi\tau_a}, \tag{8.120}$$

where we have used the relation $\tau_a = d_0/v_a = 2w_0/v_a$ by taking d in (8.109) to be the beam waist diameter d_0.

2. **Condition for beam separation and side-band limitation.** Using the definitions for θ_B, $\Delta\theta_o$, and $\Delta\theta_a$ in (8.65), (8.111), and (8.112), respectively, it can be shown that the condition that $\theta_B > \Delta\theta_o = a\Delta\theta_a$ given in (8.117) for clean separation between the diffracted and the undiffracted beams sets the following lower limit for the acoustic beam width:

$$L = \frac{a\pi n w_0 \Lambda}{2\lambda} \geq \frac{2anv_a^2}{\lambda f_0^2}, \tag{8.121}$$

and the following lower limit for the optical beam spot size:

$$d_0 = 2w_0 \geq \frac{8}{\pi}\Lambda_0 = \frac{8v_a}{\pi f_0}. \tag{8.122}$$

This latter constraint sets the following lower limit for the acoustic carrier frequency:

$$f_0 \geq \frac{4v_a}{\pi w_0} = \frac{8}{\pi\tau_a}. \tag{8.123}$$

This condition guarantees that $f_0 \geq 3.4 f_m^{3dB}$ because $f_m^{3dB} \leq 0.75/\tau_a$ for any value of parameter a, according to (8.113). Therefore, the condition that $f_0 \geq 3 f_m^{3dB}$ given in (8.115) imposed by side-band consideration is automatically satisfied as long as the condition in (8.122) for clean beam separation is satisfied.

From these discussions, we see that the value of the parameter a determines whether the physical parameters of a traveling-wave acousto-optic modulator are dictated by

the condition for Bragg diffraction or by the condition for clean beam separation. In the case when $a < 1$, the limits set by the condition for Bragg diffraction determine the physical parameters of the device because they are more stringent than those required by the condition for clean beam separation. In the case when $a > 1$, the limits set by the condition for clean beam separation are more stringent and thus define the physical parameters of the device.

In applications where the modulation bandwidth and speed have to be maximized, the optimum choice for the value of a is

$$a = 1.5, \tag{8.124}$$

for $f_{\mathrm{m}}^{3\mathrm{dB}} = 0.65/\tau_{\mathrm{a}}$ and $t_{\mathrm{r}} = 0.85\tau_{\mathrm{a}}$ with a small value of τ_{a} due to a focused beam waist in this situation. Then the acoustic beam width and the optical beam spot size are limited by the constraints set by (8.121) and (8.122), respectively, while the acoustic carrier frequency is subject to the condition in (8.123). In applications where the diffraction efficiency and the collimation of the optical beam have to be maximized at the expense of modulation speed, $a \ll 1$ is chosen so that $f_{\mathrm{m}}^{3\mathrm{dB}} = 0.75/\tau_{\mathrm{a}}$ and $t_{\mathrm{r}} = 0.65\tau_{\mathrm{a}}$ with a large value of τ_{a} due to an unfocused beam waist. Then the acoustic beam width and the optical beam spot size are limited by the constraints set by (8.118) and (8.119), respectively, while the acoustic carrier frequency is subject to the condition in (8.120). In all applications, the height of the transducer, however, only has to be $H \leq \sqrt{2}d_0 = 2\sqrt{2}w_0$ to cover the spot size of the optical beam at the interaction point.

The modulation bandwidth and modulation speed discussed above take into consideration only the interaction of the acoustic wave with the optical beam. Clearly, the overall response of a modulator to an electrical modulation signal is also subject to the bandwidth of the piezoelectric transducer and its supporting electronic circuitry. This *transducer bandwidth* is characterized by the frequency dependence of the conversion efficiency η_{t} defined in (8.105).

EXAMPLE 8.5 A fused silica traveling-wave acousto-optic modulator using the longitudinal acoustic mode at an acoustic carrier frequency of $f_0 = 100$ MHz is designed for the optical wavelength at $\lambda = 1.064$ μm. The optical wave is polarized in a direction perpendicular to the propagation directions of both the optical and the acoustic waves. The physical length L and height H of the transducer are chosen so that the device can be used for both high-speed application with a focused optical beam and low-speed application with a collimated optical beam. (a) Find the optimum optical beam spot size and the values of $f_{\mathrm{m}}^{3\mathrm{dB}}$ and t_{r} for the high-speed application. (b) Find the values of $f_{\mathrm{m}}^{3\mathrm{dB}}$ and t_{r} for the low-speed application with a collimated optical beam waist diameter of $d_0 = 1$ mm. (c) If the transducer efficiency is $\eta_{\mathrm{t}} = 60\%$, what is the electrical modulation power needed to obtain a modulation loss of 10% for the device?

Solution From Table 8.2, we find that $v_a = v_{a,L} = 5.97$ km s^{-1} for a longitudinal acoustic wave in silica glass. At $\lambda = 1.064$ μm, $n = 1.45$ for pure silica, as can be calculated by using (3.96).

(a) For the high-speed application, we choose $a = 1.5$. Then the beam spot size is limited by (8.122) to be

$$d_0 = 2w_0 \geq \frac{8v_a}{\pi f_0} = \frac{8 \times 5.97 \times 10^3}{\pi \times 100 \times 10^6} \text{ m} = 152 \text{ μm}.$$

Therefore, the beam waist can be focused to a minimum of $w_0 = 76$ μm to obtain a minimum acoustic transit time of $\tau_a = 8/\pi f_0 = 25.5$ ns using (8.123). We then have $f_m^{3dB} = 0.65/\tau_a = 25.5$ MHz and $t_r = 0.85\tau_a = 21.7$ ns. The width of the acoustic beam is required by (8.121) to be

$$L \geq \frac{2anv_a^2}{\lambda f_0^2} = \frac{2 \times 1.5 \times 1.45 \times (5.97 \times 10^3)^2}{1.064 \times 10^{-6} \times (100 \times 10^6)^2} \text{ m} = 1.46 \text{ cm}.$$

We can then choose a transducer length of $L = 1.5$ cm.

(b) With $L = 1.5$ cm and $w_0 = d_0/2 = 500$ μm for the low-speed application, we then find that

$$a = \frac{\Delta\theta_o}{\Delta\theta_a} = \frac{2\lambda f_0 L}{\pi n w_0 v_a} = \frac{2 \times 1.064 \times 10^{-6} \times 100 \times 10^6 \times 1.5 \times 10^{-2}}{\pi \times 1.45 \times 500 \times 10^{-6} \times 5.97 \times 10^3} = 0.235.$$

Because $a = 0.235 \ll 1$, we have to use (8.120) to find that $\tau_a \geq 8/a\pi f_0 = 108$ ns. We then have $f_m^{3dB} = 0.75/\tau_a \leq 6.9$ MHz and $t_r = 0.65\tau_a \geq 70$ ns for the low-speed application with a collimated beam spot size of $d_0 = 1$ mm.

Because this device is to be used for both high-speed and low-speed applications, the height of the transducer is dictated by the larger beam spot size in the low-speed application to be $H \geq \sqrt{2}d_0 = 1.4$ mm. Therefore, we can choose a transducer height of $H = 1.5$ mm.

(c) Because $M_2 \propto n^6$, we can use the value of $M_2 = 1.5 \times 10^{-15}$ m^2 W^{-1} for $n = 1.457$ at 632.8 nm to find that $M_2 = (1.45/1.457)^6 \times 1.5 \times 10^{-15}$ m^2 W^{-1} $= 1.46 \times 10^{-15}$ m^2 W^{-1} for $n = 1.45$ at 1.064 μm. Because both θ_i and θ_d are very small in the operation of this device, we have $l = L$. For a modulation loss of 10%, we need $\eta_{PM} = 0.1$. With $\eta_t = 60\%$, the required electrical power can be found by using (8.107) to be

$$P_e = \frac{2\lambda^2 H}{\pi^2 M_2 L} \frac{\eta_{PM}}{\eta_t} = \frac{2 \times (1.064 \times 10^{-6})^2 \times 1.5 \times 10^{-3} \times 0.1}{\pi^2 \times 1.46 \times 10^{-15} \times 1.5 \times 10^{-2} \times 0.6} \text{ W} = 2.6 \text{ W}.$$

This is the required power for low-speed modulation. For high-speed modulation, the required power would be somewhat higher for the same modulation loss of 10% because of the degradation in efficiency at a high modulation speed with a focused optical beam.

Standing-wave modulators

Acousto-optic modulators that utilize standing acoustic waves are used in some special applications such as laser mode locking. A standing-wave modulator provides sinusoidal amplitude modulation at a very high frequency. It differs from a traveling-wave modulator in many important aspects, from the device structure to the performance characteristics. The most important performance characteristics of a standing-wave acousto-optic modulator are its diffraction efficiency η and its loss modulation frequency, $f_m = 2f$, at twice the acoustic frequency in the low-efficiency limit. It is always operated in the Bragg regime with a well-collimated optical beam.

In order to create a standing acoustic wave, the acousto-optic cell is made to be a resonant acoustic cavity. Instead of the angled surface of the acousto-optic cell of a traveling-wave device, the surface at the far end across the cell width is made parallel to the near end that is attached to the piezoelectric transducer, as shown in Fig. 8.10. With a given cell width W measured in the direction of the acoustic wave, a standing acoustic wave is formed only when the acoustic wavelength satisfies the condition:

$$W = m\frac{\Lambda}{2}, \qquad m = \text{integer}. \tag{8.125}$$

Therefore, the device functions only at the following discrete acoustic resonance frequencies:

$$f = m\frac{v_a}{2W}, \qquad m = \text{integer}, \tag{8.126}$$

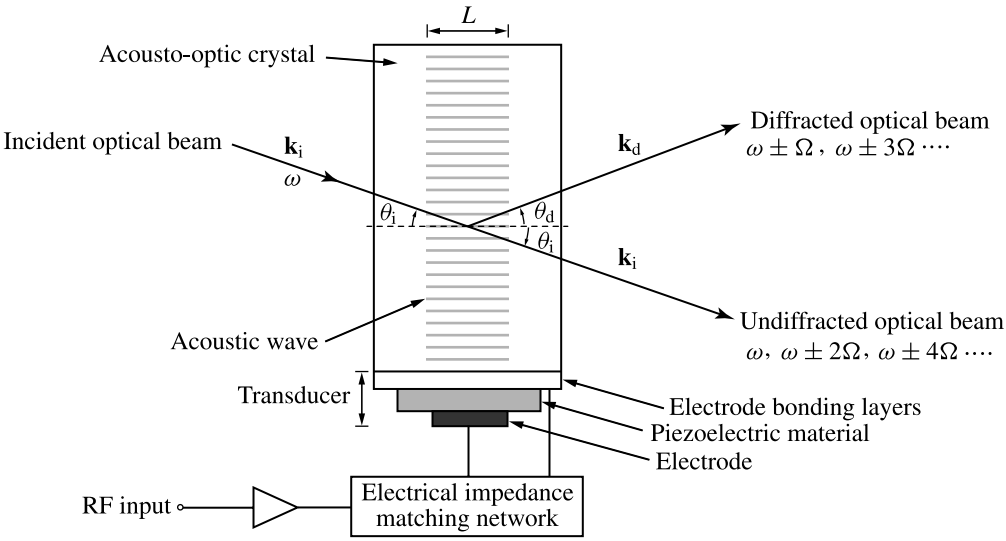


Figure 8.10 Typical solid-state acousto-optic modulator operating with a standing acoustic wave in the Bragg regime. For an anisotropic acousto-optic modulator, $|\theta_i| \neq |\theta_d|$. For an isotropic acousto-optic modulator, $|\theta_i| = |\theta_d| = \theta_B$.

which are determined by the cell width and the acoustic velocity. The resonance frequencies are sensitive to variations in the value of v_a caused by temperature fluctuations. In many applications of a standing-wave modulator, the temperature of the acousto-optic cell has to be carefully stabilized to maintain stable and efficient operation.

The acoustic power that has to be delivered by the transducer to the acoustic resonator is equal to the product of the acoustic energy stored in the resonator and the decay rate, γ_a, of that energy:

$$P_a = \left(\frac{I_a^f + I_a^b}{v_a} HLW \right) \gamma_a \approx \frac{2I_a^f}{v_a} HLW\gamma_a, \tag{8.127}$$

where we have taken $I_a^f \approx I_a^b$ for an efficient resonator. The acoustic power, P_a, is converted from an electrical power, P_e, by a transducer, which has a conversion efficiency of η_t: $P_a = \eta_t P_e$. Using (8.97) and (8.102), we find that the diffraction efficiency of a standing-wave acousto-optic modulator with perfect phase matching can be expressed as

$$\eta_{PM} = \sin^2 \left[\frac{\pi}{\lambda} \left(\frac{M_2 v_a}{HLW\gamma_a} P_a \right)^{1/2} l \cos \Omega t \right] = \sin^2 \left[\frac{\pi}{\lambda} \left(\frac{M_2 v_a}{HLW\gamma_a} \eta_t P_e \right)^{1/2} l \cos \Omega t \right]. \tag{8.128}$$

In the low efficiency limit, we have

$$\eta_{PM} \approx \frac{\pi^2 M_2 l^2 v_a}{\lambda^2 HLW\gamma_a} \eta_t P_e \cos^2 \Omega t = \frac{\pi^2 M_2 l^2 v_a}{2\lambda^2 HLW\gamma_a} \eta_t P_e (1 + \cos 2\Omega t), \qquad \text{if } \eta_{PM} \ll 1. \tag{8.129}$$

Again, $l = L$ in the configuration of small-angle Bragg diffraction. We see that, when a standing-wave acousto-optic modulator is operated in the low-efficiency limit, the intensity of the diffracted beam at its output is sinusoidally modulated at *twice* the acoustic carrier frequency with a *modulation depth* that is linearly proportional to the driving power. Unlike the situation in a traveling-wave modulator, there is no need to impose an additional modulation signal on the carrier. Therefore, P_e in the above equations is a constant. The transducer is driven by an unmodulated RF electrical signal at the desired acoustic frequency. A standing-wave modulator is capable of modulating an optical beam at very high frequencies, but the allowed modulation frequencies cannot be tuned continuously because they are discretely defined by the resonance frequencies of the acousto-optic cell.

A standing-wave acousto-optic modulator is often used as a loss modulator, such as in its use as a mode locker for a mode-locked laser. The transmittance from the incident optical beam to the undiffracted optical beam in a loss modulator is $T = 1 - \eta_{PM}$ in the case of perfect phase matching.

In the consideration of the performance characteristics of a standing-wave modulator, the only relevant parameters are Q and a. The acoustic transit time τ_a is irrelevant because the two contrapropagating acoustic waves that form the standing wave are not amplitude modulated. To ensure operation in the Bragg regime, the requirement that $Q \geq 4\pi$ given in (8.110) still has to be satisfied, leading to the same minimum length given in (8.118) for the transducer. Because the transit time is no longer relevant, the value of a is not subject to the conflicting requirements faced by a traveling-wave device. Therefore, the optical beam in a standing-wave modulator can be well collimated so that $a \ll 1$ to avoid degradation of the diffraction efficiency caused by divergence of the optical beam.

EXAMPLE 8.6 A fused silica standing-wave acousto-optic modulator using the longitudinal acoustic mode at the acoustic frequency of $f = 100$ MHz is designed for the optical wavelength at $\lambda = 1.064$ μm with specifications similar to those of the traveling-wave modulator described in Example 8.5. The optical wave is polarized in a direction perpendicular to the propagation directions of both the optical and the acoustic waves. It has a collimated optical beam waist diameter of $d_0 = 1$ mm. The transducer efficiency is also $\eta_t = 60\%$, and the length L and height H of the transducer are to be chosen properly for this device. The resonant acousto-optic cell of this standing-wave modulator has a width of $W = 3$ cm, resulting in a decay rate of $\gamma_a = 4 \times 10^4$ s^{-1}. Find the modulation frequency and the electrical modulation power needed to obtain a peak modulation loss of 10% for the device.

Solution Because the optical beam is collimated to have a beam waist diameter of $d_0 = 1$ mm, we know from Example 8.5 that $a \ll 1$ in this situation. The length L is thus subject to the condition in (8.118):

$$L \geq \frac{2nv_a^2}{\lambda f^2} = \frac{2 \times 1.45 \times (5.97 \times 10^3)^2}{1.064 \times 10^{-6} \times (100 \times 10^6)^2} \text{ m} = 9.7 \text{ mm},$$

which can be smaller than that chosen in Example 8.5. For easy comparison to Example 8.5, however, we choose the same transducer length of $L = 1.5$ cm and the same transducer height of $H = 1.5$ mm.

With an acoustic carrier frequency of $f = 100$ MHz, the loss modulation frequency is $f_m = 2f = 200$ MHz according to the discussions following (8.129). The operation of this device is small-angle Bragg diffraction with $l = L$. From (8.129) the required electrical power for a peak modulation loss of $\eta_{PM}^{max} = 10\%$ can be found to be

$$P_e = \frac{\lambda^2 H W \gamma_a}{\pi^2 M_2 L v_a} \frac{\eta_{PM}^{max}}{\eta_t}$$

$$= \frac{(1.064 \times 10^{-6})^2 \times 1.5 \times 10^{-3} \times 3 \times 10^{-2} \times 4 \times 10^4 \times 0.1}{\pi^2 \times 1.46 \times 10^{-15} \times 1.5 \times 10^{-2} \times 5.97 \times 10^3 \times 0.6} \text{ W} = 263 \text{ mW}.$$

In comparison to the traveling-wave modulator described in Example 8.5, we find that the standing-wave modulator has a much higher modulation frequency at a much reduced modulation power for a given modulation loss. However, a standing-wave modulator is not as versatile as a traveling-wave modulator because it only allows periodic sinusoidal modulation at twice its resonance frequencies.

8.5 Acousto-optic deflectors

The acoustic wave of an acousto-optic deflector is *frequency modulated*. Unlike an acousto-optic modulator, which is an amplitude modulator, an acousto-optic deflector is a *frequency modulator*, which allows its acoustic frequency to be varied electronically.

Acousto-optic deflectors have many applications. A *frequency shifter*, which has the sole purpose of generating a diffracted optical beam at an optical frequency shifted by the amount of the acoustic frequency from the input optical frequency, can be considered as the simplest form of an acousto-optic deflector. Acousto-optic deflectors are also used in such diverse applications as optical scanners, spatial light modulators, RF pulse compressors, and programmable optical interconnecters.

An acousto-optic deflector generally functions in the Bragg regime with a traveling wave. Its most important performance characteristics are its diffraction efficiency η and its number of resolvable spots N. An acousto-optic deflector has a structure similar to that of a traveling-wave acousto-optic modulator shown in Fig. 8.8, but it is always operated with a highly collimated optical beam with the parameter $a \ll 1$ to increase the value of N.

The basic principle of acousto-optic deflectors is simple: the acousto-optic deflection angle, $\theta_{\text{def}} = \theta_{\text{d}} - \theta_{\text{i}}$, which has an absolute value of $2\theta_{\text{B}}$ in the case of nonbirefringent Bragg diffraction, is determined by the phase-matching condition, which can be varied by varying the value of the propagation constant K of the acoustic wave. Because $K = 2\pi f / v_{\text{a}}$, the deflection angle can be varied by varying the acoustic frequency.

The efficiency of a deflector is also given by (8.106), or (8.107) in the low-efficiency limit. Because an acousto-optic deflector always uses the configuration of small-angle diffraction, its interaction length is defined by the length of its transducer: $l = L$. Certain requirements, such as the condition that $Q \geq 4\pi$ for the device to function in the Bragg regime, apply to both modulators and deflectors, but many key parameters of a deflector are subject to considerations different from those for determining the parameters of a modulator. While the modulation signal applied to a modulator is the amplitude variations on a constant acoustic carrier frequency, the signal applied to a deflector has a constant amplitude but a varying acoustic frequency. The overall bandwidth of an acousto-optic deflector is subject to both the Bragg bandwidth of the acousto-optic interaction and the transducer bandwidth determined by the frequency

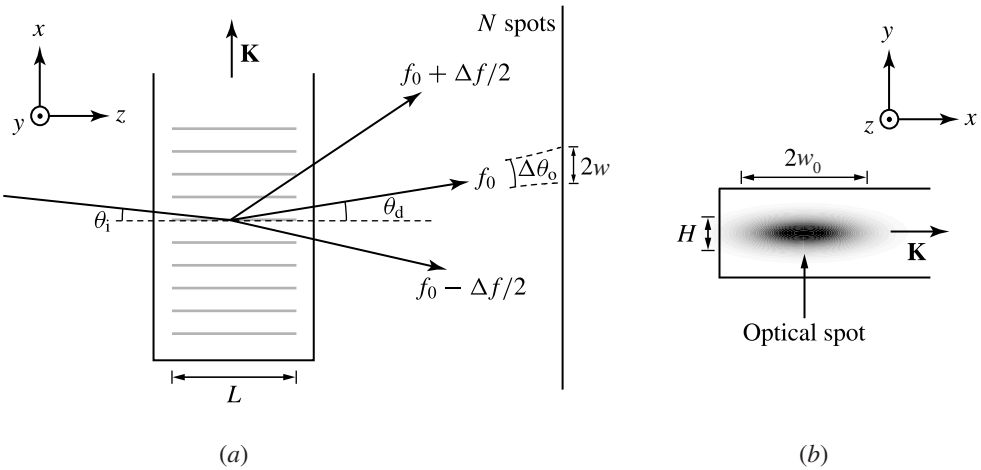


Figure 8.11 (*a*) Basic principle of an acousto-optic deflector illustrating the deflection range and the number of resolvable spots. (*b*) Side view of the device.

dependence of the conversion efficiency η_t. The bandwidth considered in the following refers to the Bragg bandwidth alone.

In an acousto-optic deflector, both the propagation direction of the incident optical beam and that of the acoustic wave are usually fixed when the value of K is varied by varying the acoustic frequency. In this situation, the angle of incidence, θ_i, is fixed, and the variation in the deflection angle is simply determined by the change in the angle of diffraction: $\delta\theta_{def} = \delta\theta_d$. Because the value of θ_i is fixed, perfect phase matching in all directions is not possible if the direction of vector **K** is fixed when the value of K is varied. As shown in Fig. 8.11, the angle of diffraction, θ_d, is now determined by the requirement for phase matching in the x direction parallel to **K**. For small variations of K, the value of θ_d varies linearly with K. If the acoustic frequency is varied over a range Δf from $f_0 - \Delta f/2$ to $f_0 + \Delta f/2$, where f_0 is the center frequency, the propagation constant of the acoustic wave varies over a range $\Delta K = 2\pi\Delta f/v_a$. Correspondingly, the deflection angle varies over a range of (see Problem 8.5.1(c))

$$\Delta\theta_d = \frac{\Delta K}{k_d} = \frac{\lambda}{n_d v_a}\Delta f, \tag{8.130}$$

where n_d is the index of refraction seen by the diffracted optical beam.

In practical applications, an acousto-optic deflector is controlled by varying the acoustic frequency for the diffracted optical beam to address different spatial locations. In the *random access* mode of operation, the acoustic frequency is changed discretely from one to another to access random positions. In the *continuous scan* mode of operation, the acoustic frequency is varied continuously so that the deflection angle changes continuously. An important parameter for an acousto-optic deflector is the number N of resolvable spots. Over a given deflection range $\Delta\theta_d$, this parameter

is determined by the divergence of the diffracted optical beam through the following relation:

$$N = \frac{\Delta\theta_{\mathrm{d}}}{\Delta\theta_{\mathrm{o}}}, \tag{8.131}$$

as illustrated in Fig. 8.11. If the optical beam has a Gaussian spatial profile, $\Delta\theta_{\mathrm{o}}$ has the form given in (8.111). Substituting (8.111) and (8.130) in (8.131), we have

$$N = \frac{\pi}{4}\tau_{\mathrm{a}}\Delta f, \tag{8.132}$$

where τ_{a} is the acoustic transit time defined in (8.109). For a deflector, τ_{a} is also called the *time aperture* of the device. The constant multiplication factor $\pi/4$ in (8.132) is specific to a Gaussian beam. This factor is different for different optical beam profiles but is generally on the order of unity. Thus the number of resolvable spots is given by the time–bandwidth product $\tau_{\mathrm{a}}\Delta f$.

For a given value of Δf, the value of N is solely determined by the value of τ_{a}. To increase the number of resolvable spots within a given frequency bandwidth, it is necessary to increase the acoustic time aperture by collimating the optical beam or by choosing an acousto-optic medium that has a low acoustic velocity. However, the acoustic transit time determines the response time of a deflector. In a deflector operating in the random access mode, it takes a temporal delay of τ_{a} for the deflector to address a new spatial position when the acoustic frequency is changed from one to another. Therefore, the *scan rate*, defined as the number of spots addressed per second, is $1/\tau_{\mathrm{a}}$ in the random access mode. The scan rate in the continuous scan mode is generally much higher than $1/\tau_{\mathrm{a}}$. Usually the speed requirement of a deflector is not very demanding and can be easily satisfied. Therefore, the choice of the parameter τ_{a} for a deflector is primarily determined by the desired number of resolvable spots within a given frequency bandwidth.

We have seen in the preceding section that the minimum height H of the transducer for an acousto-optic modulator is determined by the optical spot size. This is because the optical beam is normally focused to a small round spot. For a deflector, however, the optical spot size, $w_0 = v_{\mathrm{a}}\tau_{\mathrm{a}}/2$, in the acoustic wave propagation direction parallel to $\mathbf{K}$ as determined by (8.109) is usually quite large because a deflector normally requires a relatively large value of τ_{a} for a large value of N. In this situation, it is not necessary to maintain a round spot shape because the optical spot size in the direction perpendicular to $\mathbf{K}$ is irrelevant to the transit time τ_{a}. The optical beam can then take an elliptic spot shape as seen in Fig. 8.11 for a small value of H to increase the diffraction efficiency at a given acoustic power level. In this case, the height of the transducer is limited by the divergence of the acoustic beam to the order of

$$H \approx v_{\mathrm{a}}\left(\frac{\tau_{\mathrm{a}}}{f_0}\right)^{1/2}. \tag{8.133}$$

To prevent the second harmonics of the frequencies within the operating bandwidth from interfering with the function of a deflector, the highest frequency within the bandwidth has to be smaller than the second harmonic of the lowest frequency: $f_0 + \Delta f/2 < 2(f_0 - \Delta f/2)$. For a given center frequency, this requirement sets the upper limit for the bandwidth:

$$\Delta f \leq \frac{2}{3} f_0. \tag{8.134}$$

The maximum value of the *fractional bandwidth*, defined as $\Delta f/f_0$, of an acousto-optic deflector is set by this condition at 0.67. This condition has the same form as that of (8.115) if we identify the carrier frequency of a modulator with the center frequency of a deflector and Δf with $2f_{\mathrm{m}}^{3\mathrm{dB}}$. Note, however, that the bandwidth Δf and the modulation bandwidth $f_{\mathrm{m}}^{3\mathrm{dB}}$ are completely different in terms of their physical originals and their implications for device performance. As we have seen in the preceeding section, the modulation bandwidth $f_{\mathrm{m}}^{3\mathrm{dB}}$ is determined by the acoustic transit time τ_{a}; it is related to the response speed of a device. In contrast, the bandwidth Δf is determined by the maximum acceptable phase mismatch of a deflector; it is irrelevant to the speed of the device but determines the largest deflection angle and thus the number of resolvable spots. For a deflector, the response speed is still related to its modulation bandwidth as $f_{\mathrm{m}}^{3\mathrm{dB}} \approx 0.75/\tau_{\mathrm{a}}$ for $a \ll 1$, but $\Delta f \neq 2f_{\mathrm{m}}^{3\mathrm{dB}}$. Indeed, Δf and $f_{\mathrm{m}}^{3\mathrm{dB}}$ are not directly related. Therefore, τ_{a} and Δf can be simultaneously optimized to maximize the value of N in (8.132).

Nonbirefringent deflectors

When the value of K is varied, both the angle of incidence and the angle of diffraction have to change accordingly in order to maintain perfect phase matching. When the incident angle θ_{i} is fixed, a change in the value of K without a corresponding change in the direction of vector $\mathbf{K}$ results in a phase mismatch. Though phase matching in the x direction parallel to $\mathbf{K}$ is maintained through a change in the value of θ_{d}, a phase mismatch along the z direction perpendicular to vector $\mathbf{K}$ cannot be avoided. On either side of the acoustic frequency range, the maximum deviation from the center frequency is $\Delta f/2$. The corresponding maximum deviation in the propagation constant is $\Delta K/2$ from the center value K_0, as illustrated in Fig. 8.12. If θ_{i} is chosen to be the angle $\theta_{\mathrm{i0}} = \pm\theta_{\mathrm{B0}}$ for perfect phase matching at f_0, the phase mismatch that appears at either edge of the bandwidth is (see Problem 8.5.1(d))

$$|\Delta k| = \frac{K_0}{4k}\Delta K = \frac{\pi\lambda f_0}{2nv_{\mathrm{a}}^2}\Delta f \tag{8.135}$$

to first order. Instead of ensuring perfect phase matching at f_0, the incident angle can be chosen as

$$\theta_{\mathrm{i}} = \theta_{\mathrm{i0}}\left[1 + \left(\frac{\Delta f}{2f_0}\right)^2\right], \tag{8.136}$$

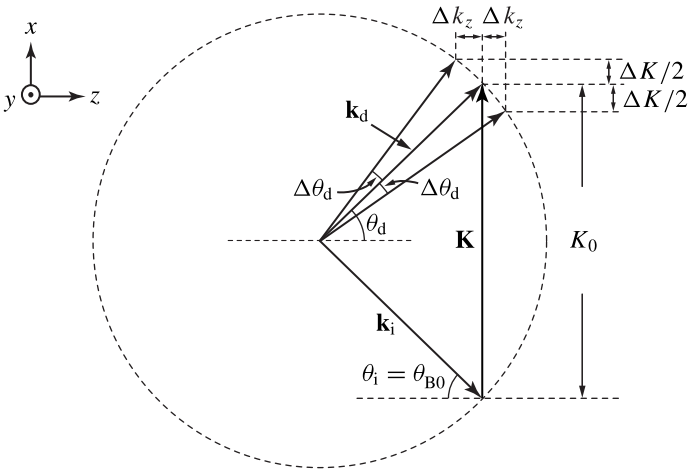


Figure 8.12 Phase-matching diagram of a nonbirefringent acousto-optic deflector showing phase mismatch at the edges of the deflection range.

to minimize the largest phase mismatch for a given bandwidth Δf. Then the phase mismatch can be minimized to the following value:

$$|\Delta k| = \frac{\pi \lambda f_0}{2 n v_\mathrm{a}^2} \Delta f \left[1 - \left(\frac{\Delta f}{2 f_0} \right)^2 \right]. \tag{8.137}$$

This phase mismatch results in a reduction in the diffraction efficiency. The maximum acceptable phase mismatch depends on the interaction length l and the acceptable reduction of efficiency at the edges of the bandwidth. A convenient general criterion is to limit the value of $|\Delta k| l$ to less than 0.9π:

$$|\Delta k| l = \frac{\pi \lambda f_0 l}{2 n v_\mathrm{a}^2} \Delta f \left[1 - \left(\frac{\Delta f}{2 f_0} \right)^2 \right] \approx \frac{\pi}{2} \frac{\Delta\theta_\mathrm{d}}{\Delta\theta_\mathrm{a}} \leq 0.9\pi, \tag{8.138}$$

where $\Delta\theta_\mathrm{a}$ is the divergence of the acoustic beam defined in (8.112). In the limit of low diffraction efficiency where $|\Delta k| > \kappa$, the constraint in (8.138) ensures that the diffraction efficiency at the edges of the bandwidth is not reduced by more than 3 dB below that at the center frequency (see Problem 8.3.16). Other criteria can be chosen based on the bandwidth and efficiency specifications of a device.

The condition in (8.138) can be appreciated from two different, but equivalent, viewpoints. From one viewpoint, the relation $|\Delta k| l \leq 0.9\pi$ indicates that in order to limit the degradation of the diffraction efficiency caused by phase mismatch to an acceptable level, the interaction length has to be kept below an upper limit set by the amount of phase mismatch at the edges of a given bandwidth. From another viewpoint, the relation $\Delta\theta_\mathrm{d} \leq 1.8 \Delta\theta_\mathrm{a}$ implied by (8.138) leads to a picture similar to that used in the preceding section in the discussion of the relation between $\Delta\theta_\mathrm{a}$ and $\Delta\theta_\mathrm{o}$ for a traveling-wave acousto-optic modulator. For efficient interaction over the entire deflection range, the divergence $\Delta\theta_\mathrm{a}$ of the acoustic beam has to be large enough that the acoustic wavefront

covers a sufficiently wide range of directions for all of the different diffraction directions within the range of $\Delta\theta_d$ to be phase matched by different acoustic wavefront directions. These two viewpoints are equivalent because the interaction length is defined by the width of the acoustic beam and a small acoustic beam width results in a large acoustic beam divergence.

Combining (8.131) and (8.138), we find that

$$a = \frac{\Delta\theta_o}{\Delta\theta_a} \leq \frac{1.8}{N} \ll 1, \tag{8.139}$$

where N for a practical deflector is generally a large number. The small value of a required by (8.139) can be appreciated by considering the fact that for an acousto-optic deflector, $\Delta\theta_o$ has to be much smaller than $\Delta\theta_d$ in order to have a large number of resolvable spots but $\Delta\theta_a$ has to be on the order of $\Delta\theta_d$ in order to prevent severe degradation in the diffraction efficiency due to phase mismatch at the edges of the range of deflection.

The lower limit for the length L of the piezoelectric transducer given by (8.118) is still valid due to the requirement for a deflector to operate in the Bragg regime throughout the entire bandwidth. Because a lower frequency sets a more stringent lower limit for L, we have to use the lowest frequency $f_0 - \Delta f/2$ for the Bragg condition. In addition, the relation in (8.138) and the fact that $l = L$ together set an upper limit for L. Combining these two limits, we find the following constraints for the length of the transducer in a nonbirefringent acousto-optic deflector:

$$\frac{1.8nv_a^2}{\lambda f_0 \Delta f}\left[1 - \left(\frac{\Delta f}{2f_0}\right)^2\right]^{-1} \geq L \geq \frac{2nv_a^2}{\lambda f_0^2}\left(1 - \frac{\Delta f}{2f_0}\right)^{-2}. \tag{8.140}$$

For optimum performance of a deflector, both a large Δf, for a large number of resolvable spots, and a large L, for a high efficiency, are desired. The optimum values of Δf and L for a nonbirefringent deflector can be found by solving (8.140) with equals sign for both places of the $\geq$ sign to be (see Problem 8.5.3)

$$\Delta f = 0.525 f_0 \quad \text{and} \quad L = 3.68\frac{nv_a^2}{\lambda f_0^2}. \tag{8.141}$$

EXAMPLE 8.7 A $LiNbO_3$ acousto-optic deflector for 1.3 μm optical wavelength is desired to have a 1 GHz bandwidth with 100 resolvable spots. The optical axis $\hat{z}$ of the crystal is parallel to the [001] direction, and the y-coordinate axis is taken to be in the [010] direction. The acoustic wave is a transverse mode propagating in the [001] z direction and polarized in the y direction. Its acoustic velocity is 3.59 km s^{-1}. Optical deflection takes place in the yz plane, and the incident optical wave is a Gaussian beam of an elliptical spot shape polarized in the x direction. The ordinary and extraordinary indices of refraction for $LiNbO_3$ at 1.3 μm are $n_o = 2.222$ and $n_e = 2.145$, respectively.

(a) Find the required acoustic time aperture τ_a and the optical spot size w_0 in the z direction. (b) Find the optimum center acoustic frequency f_0. (c) Find the optimum dimensions L and H for the transducer to maximize the efficiency. (d) What is the peak diffraction efficiency for 1 W of acoustic power?

Solution (a) With $\Delta f = 1$ GHz and $N = 100$, we find by using (8.132) that $\tau_a = 4N/\pi\Delta f = 127.3$ ns. Then the spot size in the z direction is $w_0 = v_a\tau_a/2 = 228.5$ μm. This is not the required spot size in the x direction, however, as the beam is elliptical.

(b) For interaction of the x-polarized optical wave with the z-propagating, y-polarized transverse acoustic mode in $LiNbO_3$, the only coupling is through the $\Delta\epsilon_{xx}$ element of the acousto-optic permittivity tensor with a figure of merit of $M_2 = 3.15 \times 10^{-15}$ m^2 W^{-1} at $\lambda = 1.3$ μm (see Problem 8.3.6). Therefore, this deflector is nonbirefringent with both incident and diffracted waves polarized in the x direction. We can then use (8.141) to find that the optimum center acoustic frequency is $f_0 = \Delta f/0.525 = 1.9$ GHz for $\Delta f = 1$ GHz.

(c) From (8.141), the optimum length of the transducer is

$$L = 3.68\frac{nv_a^2}{\lambda f_0^2} = \frac{3.68 \times 2.222 \times (3.59 \times 10^3)^2}{1.3 \times 10^{-6} \times (1.9 \times 10^9)^2}\ \text{m} = 22.5\ \mu\text{m}$$

and, from (8.133), the optimum height is

$$H = v_a\left(\frac{\tau_a}{f_0}\right)^{1/2} = 3.59 \times 10^3 \times \left(\frac{127.3 \times 10^{-9}}{1.9 \times 10^9}\right)^{1/2}\ \text{m} = 29.4\ \mu\text{m}.$$

(d) The peak diffraction efficiency at $f_0 = 1.9$ GHz for $P_a = 1$ W is, for $l = L$,

$$\eta_{PM} = \frac{\pi^2 M_2 L}{2\lambda^2 H}P_a = \frac{\pi^2 \times 3.15 \times 10^{-15} \times 22.5 \times 10^{-6}}{2 \times (1.3 \times 10^{-6})^2 \times 29.4 \times 10^{-6}} \times 1 = 0.7\%.$$

Birefringent deflectors

For a nonbirefringent deflector, the constraints in (8.140) dictate that increasing the bandwidth Δf leads to a reduction in the length L, and vice versa. It is therefore not possible to increase both Δf and L simultaneously above their respective optimum values given in (8.141).

Using birefringent Bragg diffraction under the special condition of *tangential phase matching*, also known as 90° *phase matching*, as discussed in Example 8.3, it is possible to increase the values of both Δf and L for a birefringent deflector beyond their optimum values for a nonbirefringent deflector, thus increasing the number of resolvable spots and the diffraction efficiency simultaneously. In the application of a deflector, it is required that θ_d varies sensitively to the acoustic frequency f while θ_i is fixed. Therefore, the conditions for birefringent tangential phase matching in a deflector are (1) $n_i > n_d$ so

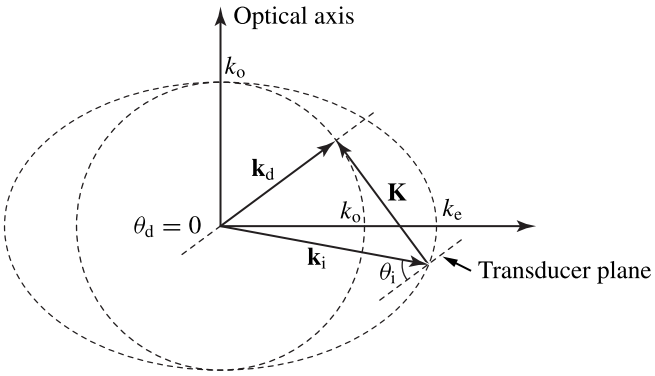


Figure 8.13 Tangential phase-matching scheme for a birefringent acousto-optic deflector.

that $k_i > k_d$ and (2) $K_0^2 = k_i^2 - k_d^2$ so that perfect phase matching occurs at the center frequency f_0 with $\theta_d = 0$, which are shown in Fig. 8.13. This frequency f_0 is determined by the normalized frequency $\hat{f}_t$ found in (8.74) as $f_0 = (n_i + n_d)v_a\hat{f}_t/\lambda$. Under this phase-matching condition, the maximum phase mismatch that appears at the edges of the bandwidth is (see Problem 8.5.4)

$$|\Delta k| \approx \frac{(\Delta K)^2}{8k_d} = \frac{\pi\lambda}{4n_d v_a^2}(\Delta f)^2. \tag{8.142}$$

Applying the criterion of $|\Delta k|L = |\Delta k|l \leq 0.9\pi$, we find that the bandwidth and the transducer length of a birefringent deflector under the tangential phase-matching condition are subject to the following constraints:

$$\frac{3.6n_d v_a^2}{\lambda(\Delta f)^2} \geq L \geq \frac{2n_d v_a^2}{\lambda f_0^2}\left(1 - \frac{\Delta f}{2f_0}\right)^{-2}. \tag{8.143}$$

This condition can be satisfied for the largest bandwidth allowed by the condition in (8.134). Therefore, it leads to the following optimum values for the bandwidth and the interaction length, respectively (see Problem 8.5.6):

$$\Delta f = \frac{2}{3}f_0 \quad \text{and} \quad L = 8.1\frac{n_d v_a^2}{\lambda f_0^2}. \tag{8.144}$$

In comparison to (8.141), we see that tangential phase matching for a birefringent deflector allows an interaction length that is more than twice that of the optimum length for a nonbirefringent deflector while having a 27% increase in its bandwidth to reach its allowable maximum.

The upper limit of length L can be further doubled if we move the phase-matching point slightly away from the tangential point by choosing $K_0^2 = k_i^2 - k_d^2 + (\Delta K)^2/8$, as shown in Fig. 8.14. Under this arrangement, perfect phase matching occurs at the two frequencies of $f_0 \pm \Delta f/2\sqrt{2}$, and the maximum phase mismatch appears at the center

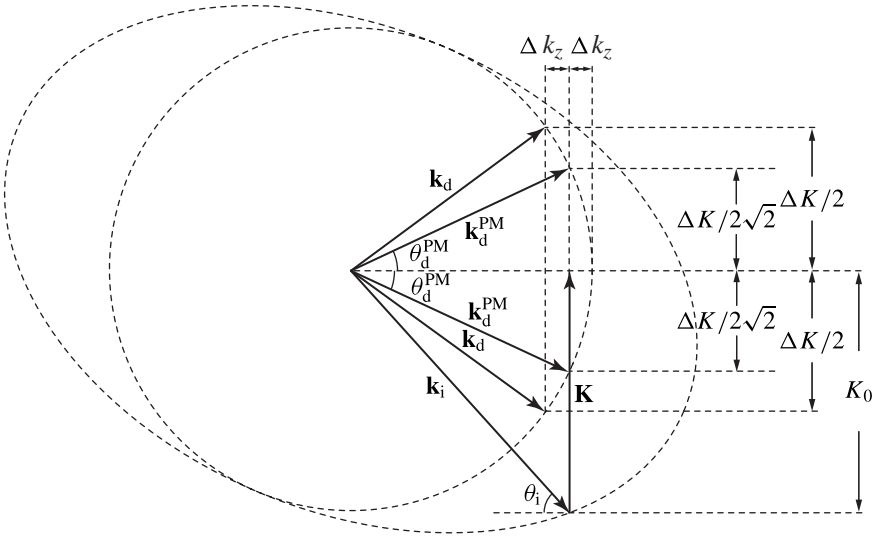


Figure 8.14 Optimum phase-matching scheme for a birefringent acousto-optic deflector of a large bandwidth.

frequency as well as at both edges of the bandwidth. This maximum phase mismatch is (see Problem 8.5.5(c))

$$|\Delta k| \approx \frac{(\Delta K)^2}{16k_\mathrm{d}} = \frac{\pi\lambda}{8n_\mathrm{d}v_\mathrm{a}^2}(\Delta f)^2, \tag{8.145}$$

which is only half the value of that in (8.142). Consequently, instead of (8.143), we have

$$\frac{7.2n_\mathrm{d}v_\mathrm{a}^2}{\lambda(\Delta f)^2} \geq L \geq \frac{2n_\mathrm{d}v_\mathrm{a}^2}{\lambda f_0^2}\left(1-\frac{\Delta f}{2f_0}\right)^{-2}. \tag{8.146}$$

Therefore, while the bandwidth remains at its allowable maximum, the interaction length is doubled (see Problem 8.5.6):

$$\Delta f = \frac{2}{3}f_0 \quad \text{and} \quad L = 16.2\frac{n_\mathrm{d}v_\mathrm{a}^2}{\lambda f_0^2}. \tag{8.147}$$

In the above discussions, we have assumed that n_d is a constant that does not vary with acoustic frequency. In the situation when the diffracted beam is an extraordinary wave, however, n_d can be a function of the diffraction angle and thus a function of the acoustic frequency. Then the optimization process is more complicated than what is discussed above, but the general concepts are still valid.

EXAMPLE 8.8 A $LiNbO_3$ acousto-optic deflector for 1.3 μm optical wavelength is desired to have a 1 GHz bandwidth with 100 resolvable spots such as the one described in Example 8.7. The acoustic wave is still a transverse mode propagating in the [001] z

direction, but is now polarized in the x direction. Its acoustic velocity is also 3.59 km s^{-1}. Optical deflection still takes place in the yz plane, and the incident optical wave is still a Gaussian beam of an elliptical spot shape polarized in the x direction. (a) Find the required acoustic time aperture τ_a and the optical spot size w_0 in the z direction. (b) Find the optimum center acoustic frequency f_0. (c) Find the optimum dimensions L and H for the transducer to maximize the efficiency. (d) What is the peak diffraction efficiency for 1 W of acoustic power?

Solution (a) With $\Delta f = 1$ GHz and $N = 100$, we still find, using (8.132), that $\tau_a = 4N/\pi\Delta f = 127.3$ ns. Then the spot size in the z direction is $w_0 = v_a\tau_a/2 = 228.5$ μm. These two parameters are the same as those found in Example 8.7.

(b) For interaction of the x-polarized optical wave with the z-propagating, x-polarized transverse acoustic mode in $LiNbO_3$, the only coupling is through the $\Delta\epsilon_{xy} = \Delta\epsilon_{yx}$ and $\Delta\epsilon_{zx} = \Delta\epsilon_{xz}$ elements of the acousto-optic permittivity tensor with a figure of merit of $M_2 = 1.075 \times 10^{-14}$ m^2 W^{-1} at $\lambda = 1.3$ μm (see Problem 8.3.6). Therefore, this deflector is birefringent with both an ordinary incident wave polarized in the x direction and an extraordinary diffracted wave polarized in the yz plane. We find that $n_i > n_d$ because $LiNbO_3$ is negative uniaxial. Therefore, tangential phase matching so that $\theta_d = 0$ for the optimum performance of the deflector is possible. This occurs at $\theta_i = -15.13^\circ$ for up-shifted diffraction, or $\theta_i = 15.13^\circ$ for down-shifted diffraction, at an acoustic frequency of $f_t = 1.6$ GHz for tangential phase matching (see Problem 8.3.6). Because $\theta_d = 0$, the diffracted wave is polarized in the z direction with $n_d = n_e = 2.145$. For this device, the center acoustic frequency is dictated by the tangential phase-matching condition to be $f_0 = f_t = 1.6$ GHz rather than by the desired Δf and the optimum condition in (8.144). With $\Delta f = 1$ GHz, we find that $\Delta f/f_0 = 0.625 < 2/3$.

(c) Because the value of $\Delta f/f_0$ is smaller than its optimum value allowed by (8.144), the value of L can be larger than that given in (8.144). It is determined by its upper bound in (8.143) to be

$$L = \frac{3.6 n_d v_a^2}{\lambda(\Delta f)^2} = \frac{3.6 \times 2.145 \times (3.59 \times 10^3)^2}{1.3 \times 10^{-6} \times (1 \times 10^9)^2}\ \text{m} = 76.6\ \mu\text{m}.$$

The height of the transducer can be chosen to be approximately

$$H = v_a\left(\frac{\tau_a}{f_0}\right)^{1/2} = 3.59 \times 10^3 \times \left(\frac{127.3 \times 10^{-9}}{1.6 \times 10^9}\right)^{1/2}\ \text{m} = 32\ \mu\text{m}.$$

(d) The peak diffraction efficiency at $f_0 = 1.6$ GHz for $P_a = 1$ W is, for $l = L$,

$$\eta_{PM} = \frac{\pi^2 M_2 L}{2\lambda^2 H} P_a = \frac{\pi^2 \times 1.075 \times 10^{-14} \times 76.6 \times 10^{-6}}{2 \times (1.3 \times 10^{-6})^2 \times 32 \times 10^{-6}} \times 1 = 7.5\%,$$

which is more than ten times that of the nonbirefringent deflector described in Example 8.7. By further using the optimum phase-matching scheme illustrated in Fig. 8.14, the length L can be doubled to $L = 153.2$ μm, thus doubling the peak efficiency to

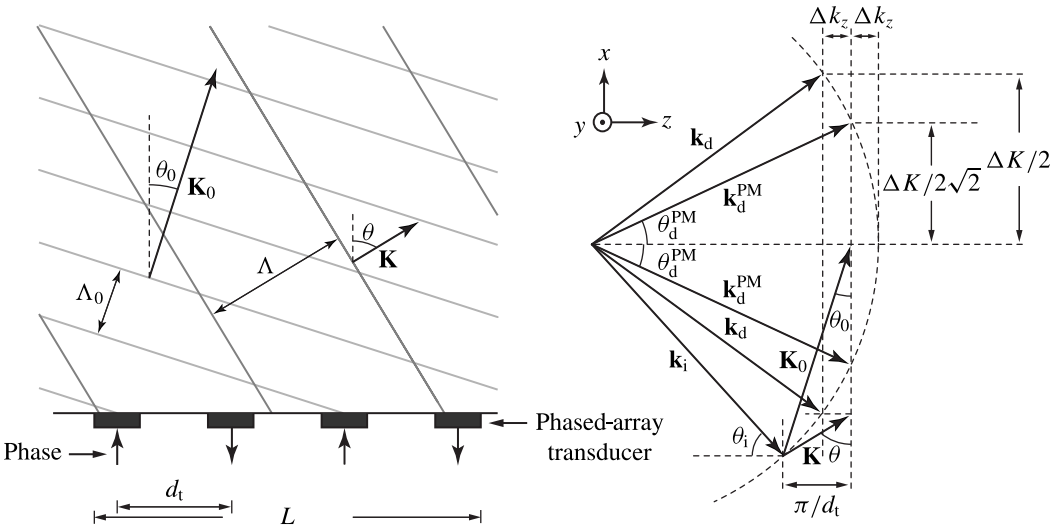


Figure 8.15 Phased-array transducer for acoustic beam steering in an acousto-optic deflector. As the acoustic frequency is tuned, both the acoustic wavelength and the acoustic wavefront direction change accordingly.

15%. However, with this optimum phase-matching scheme, the peak efficiency does not occur at $f_0 = 1.6$ GHz, but occurs at two frequencies: $f_0 + \Delta f/2\sqrt{2} = 1.954$ GHz and $f_0 - \Delta f/2\sqrt{2} = 1.246$ GHz.

Deflectors using phased-array transducers

The phase mismatch that exists over the bandwidth of a deflector is caused by the fact that the propagation directions of both the incident optical wave and the acoustic wave are fixed while the acoustic frequency is varied. In the operation of an acousto-optic deflector, it is clearly not practical to vary the direction of the incident optical wave in response to variations in the acoustic frequency in order to maintain perfect phase matching. However, using a multiple-element *phased-array transducer* shown in Fig. 8.15, the direction of the acoustic **K** vector can be steered to satisfy the phase-matching condition better as the acoustic frequency is varied.

The simplest phased-array transducer consists of equally spaced elements of the same width, but a phase shift of π between each pair of adjacent elements is introduced. The interference among the acoustic waves generated by the phase-shifted elements causes the acoustic power to concentrate in the direction of constructive interference. This direction shifts from the normal to the transducer by an angle θ given by (see Problem 8.5.7(a))

$$\theta \approx \sin\theta = \frac{\pi}{d_t K} = \frac{\Lambda}{2d_t} = \frac{v_a}{2d_t f}, \tag{8.148}$$

where d_t is the center-to-center distance between adjacent elements in the transducer array, as shown in Fig. 8.15. Therefore, as the acoustic frequency varies, the direction

of the acoustic beam is steered accordingly. In a nonbirefringent deflector using this simple phased-array transducer, the maximum phase mismatch over a given bandwidth can be reduced to the value given in (8.145) when the parameters of the transducer and the incident direction of the optical beam are chosen properly (see Problem 8.5.7). Therefore, the bandwidth Δf and the interaction length L of a nonbirefringent deflector using this phased-array transducer under optimum conditions are subject to the same constraints as given in (8.146). Their optimum values are thus those given in (8.147).

The simple phased-array transducer with a constant phase shift of π between adjacent elements does not completely eliminate phase mismatch within a given bandwidth, but it allows the performance of a nonbirefringent deflector to match the best performance of a birefringent deflector. Further improvement to obtain closer phase matching over a large bandwidth is possible with a more sophisticated phased-array transducer to fine tune the acoustic beam direction. For most applications, however, a complicated transducer is not practical because the effort is not justified by the benefit gained.

8.6 Acousto-optic tunable filters

An optical grating can be used for the separation or filtering of optical frequencies, as is seen in any grating spectrometer and in a distributed Bragg reflector as discussed in Section 5.1. It is also possible to use the index grating generated by an acoustic wave in a medium for such purposes. One advantage of such an acousto-optic filter, or acousto-optic spectrometer, is that it is electronically tunable because the period of the acousto-optic grating can be varied by altering the acoustic frequency. This electronic tunability allows an acousto-optic filter to have many sophisticated functions that are not possible with an ordinary spectroscopic device. For example, the acoustic frequency can be rapidly scanned or quickly switched from one to another, thus allowing very rapid optical spectrum analysis. The acoustic signal driving the filter can also be amplitude or frequency modulated, thus imposing a desired amplitude or frequency modulation on the optical signal at the selected optical frequency. As a result, the applications of acousto-optic tunable filters cover a very wide range from wavelength tuning of a laser to wavelength-division multiplexing and demultiplexing in optical communication systems, as well as various spectroscopic analyses.

An acousto-optic tunable filter functions exclusively in the Bragg regime because its tunability and frequency selectivity rely entirely on the phase-matching condition. It generally uses a traveling acoustic wave because the acoustic frequency cannot be tuned continuously in a standing-wave device. Practical acousto-optic filters, except for those in waveguide structures discussed in the following section, function exclusively with birefringent Bragg diffraction in anisotropic crystals. The most important characteristic parameters of an acousto-optic tunable filter are its chromatic resolving power, its

angular aperture, and its efficiency. Because high spectral resolution is one of its merits, it generally operates with the parameter $a \gg 1$ with a large angular aperture.

The selectivity and resolution of the optical frequency in the interaction of an optical wave with a grating increase linearly with interaction length. To have a long interaction length with an acousto-optic grating while avoiding the need to expand the acoustic beam width and the consequential reduction of the acoustic intensity at a given acoustic power, it is necessary to use collinear, or nearly collinear, Bragg diffraction. For collinear, *codirectional* Bragg diffraction, the phase-matching condition is $k_\mathrm{d} = k_\mathrm{i} \pm K$. The optical wavelength selected under this phase-matching scheme in an acousto-optic filter driven by an acoustic wave at frequency f is

$$\lambda = \frac{\Delta n v_\mathrm{a}}{f}, \tag{8.149}$$

where $\Delta n = |n_\mathrm{d} - n_\mathrm{i}|$. For collinear, *contradirectional* Bragg diffraction, the phase-matching condition is $k_\mathrm{d} = -k_\mathrm{i} \pm K$, and the relation between the selected optical wavelength and the acoustic frequency becomes $\lambda = (n_\mathrm{d} + n_\mathrm{i})v_\mathrm{a}/f$. The contradirectional scheme is not practical for an ordinary acousto-optic tunable filter using a typical acousto-optic material because it requires an acoustic frequency on the order of 10 GHz for a typical optical wavelength, say $\lambda = 1$ μm. In contrast, over the entire optical spectral range, the acoustic frequencies required for an acousto-optic tunable filter functioning in the codirectional scheme are, according to (8.149), generally in the practical range of 10 MHz to 1 GHz for typical acousto-optic materials, such as $LiNbO_3$, TeO_2, and $CaMoO_4$, that are used for this purpose. Therefore, the ideal phase-matching scheme for a practical acousto-optic filter is collinear, codirectional Bragg interaction. In certain applications, exact collinear interaction is not possible due to the limitation of the parameters of the available acousto-optic materials. Then noncollinear birefringent Bragg interaction under a special tangential phase-matching condition can also be used to design practical noncollinear acousto-optic tunable filters. *For an acousto-optic tunable filter that uses either collinear or noncollinear interaction, an anisotropic crystal is needed and there is always a polarization change between the incident and diffracted optical waves.* This statement, however, is not true for guided-wave acousto-optic tunable filters, as we shall see in the next section. In this section, we consider only collinear filters for simplicity, but the general concepts apply to noncollinear filters as well.

Figure 8.16 shows two configurations for collinear acousto-optic tunable filters. The interaction length l is measured along the *longitudinal direction*, rather than the transverse direction, of the acoustic beam. Therefore, the interaction length in an acousto-optic tunable filter is not determined by the transducer length: $l \neq L$. Because of birefringent Bragg interaction, the filtered optical beam at the selected wavelength has a polarization different from that of the unfiltered beam. In general, a very high extinction ratio can be easily achieved in separating orthogonally polarized optical waves. By

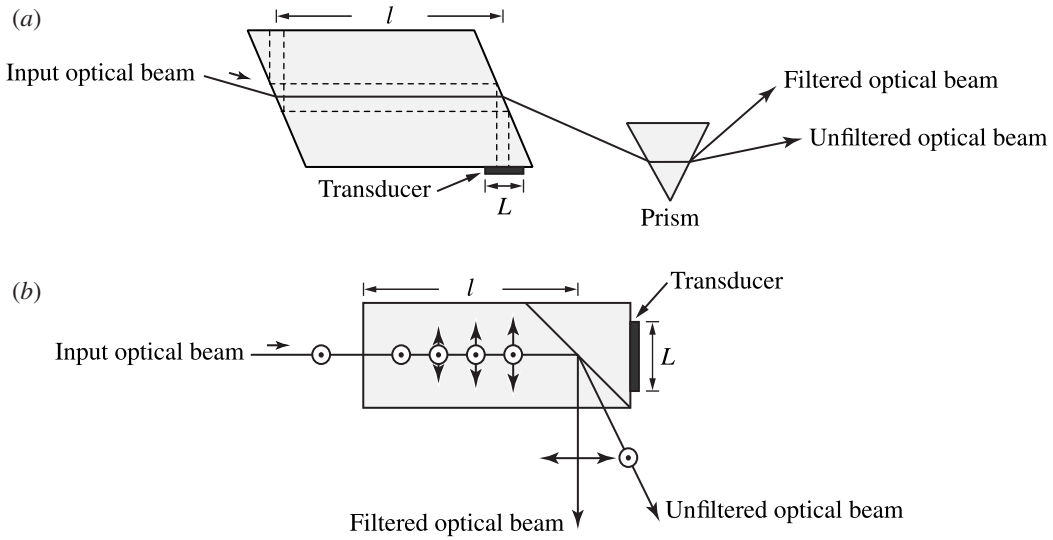


Figure 8.16 Configurations for collinear acousto-optic tunable filters.

using proper polarizing optical components to separate the filtered beam from the unfiltered beam, a very high signal-to-noise ratio can be obtained. This is another advantage of acousto-optic tunable filters.

In collinear, codirectional Bragg diffraction, if both the frequency and the propagation direction of the acoustic wave are fixed, the phase mismatch caused by a wavelength deviation of $\delta\lambda$ and an angular deviation of $\delta\theta_i$ in the incident optical wave is (see Problem 8.6.1)

$$\Delta k \approx \frac{2\pi(n_d - n_i)}{\lambda}\left[-\frac{\delta\lambda}{\lambda} + \frac{(\delta\theta_i)^2}{2}\right]. \tag{8.150}$$

A very important characteristic of an acousto-optic tunable filter is its optical spectral resolution. Over a spectral width of $\Delta\lambda$ centered at the wavelength λ that is selected by a given acoustic frequency f according to (8.149), the largest wavelength deviations away from λ are $\delta\lambda = \pm\Delta\lambda/2$ at the two edges of the spectral width. For a perfectly collimated optical beam with $\delta\theta_i = 0$, the general criterion that $|\Delta k|l \leq 0.9\pi$ leads to the following *spectral width passed by the filter*:

$$\Delta\lambda = \frac{0.9\lambda^2}{\Delta n l}. \tag{8.151}$$

The *chromatic resolving power* of the acousto-optic filter is then given by

$$R = \frac{\lambda}{\Delta\lambda} = \frac{\Delta n l}{0.9\lambda} = \frac{f l}{0.9 v_a} = \frac{l}{0.9\Lambda}. \tag{8.152}$$

Hence the resolving power is simply given by the number of acoustic wavelengths covered by the interaction length and is linearly proportional to the interaction length.

Another important characteristic of an acousto-optic tunable filter is its *angular aperture*. As is the case in the application of any spectroscopic instrument, the input optical beam incident on the acousto-optic tunable filter is usually not well collimated. A large angular aperture is thus needed for a high optical collection efficiency at the input, which then leads to a high overall efficiency for the device. The angular aperture of a given acousto-optic tunable filter is determined by the maximum angular divergence $\Delta\theta_o$ allowed for the incident optical beam. An angular divergence of $\Delta\theta_o$ in the incident optical beam corresponds to maximum angular deviations of $\delta\theta_i = \pm\Delta\theta_o/2$ with respect to the center direction of the beam. For any given optical spectral component, we find that

$$\Delta\theta_o = 2\left(\frac{0.9\lambda}{\Delta n l}\right)^{1/2} = \frac{2}{R^{1/2}} \tag{8.153}$$

by applying the criterion $|\Delta k| l \leq 0.9\pi$. Note that $\Delta\theta_o$ is the maximum allowable optical beam divergence *inside* the acousto-optic medium. The difference in the refractive index between the medium and free space causes a change in the divergence of an optical beam between free space and the acousto-optic medium. Therefore, the angular apertures for the incident and the diffracted waves in free space outside the medium are, respectively,

$$\Delta\theta_i = n_i\Delta\theta_o = \frac{2n_i}{R^{1/2}} \quad \text{and} \quad \Delta\theta_d = n_d\Delta\theta_o = \frac{2n_d}{R^{1/2}}. \tag{8.154}$$

For birefringent interaction in a collinear filter, one of the two optical waves has to be an extraordinary wave but both optical waves have to be polarized in a plane that is perpendicular to the acoustic $\mathbf{K}$ direction because $\mathbf{k}_i \parallel \mathbf{k}_d \parallel \mathbf{K}$. Therefore, interaction is generally restricted to the plane that is perpendicular to the optical axis of the crystal with the extraordinary optical wave polarized in the direction along the optical axis. Because the interaction length is not determined by the transducer length in a collinear configuration, it is possible to increase the diffraction efficiency and the resolving power of a collinear acousto-optic tunable filter simultaneously by choosing $l > L$. The length l is normally limited by the attenuation of the acoustic wave and by the availability of long acousto-optic crystals.

EXAMPLE 8.9 It is possible to make a collinear $LiNbO_3$ acousto-optic tunable filter utilizing a transverse acoustic mode that propagates in the [010] y direction and is polarized in the [100] x direction. The acoustic velocity of this mode is $v_a = 4.08$ km s^{-1}. The birefringent collinear interaction with $\mathbf{k}_i \parallel \mathbf{k}_d \parallel \mathbf{K} \parallel \hat{y}$ then couples the ordinary wave polarized with $\hat{e}_o = \hat{x}$ and the extraordinary wave polarized with $\hat{e}_e = \hat{z}$ through the acousto-optic tensor elements $\Delta\epsilon_{xz} = \Delta\epsilon_{zx}$ (see Problem 8.2.5(c)). This filter is used to select an ordinary incident wave spectrally at $\lambda = 1.3$ μm with a chromatic resolving power of $R = 10^3$. The ordinary and extraordinary refractive indices of $LiNbO_3$

at 1.3 μm are $n_o = 2.222$ and $n_e = 2.145$, respectively. Find the required acoustic frequency f and the required interaction length l. What are the spectral linewidth $\Delta\lambda$ and the angular apertures $\Delta\theta_i$ and $\Delta\theta_d$?

Solution The required acoustic frequency is

$$f = \frac{\Delta n v_a}{\lambda} = \frac{|2.222 - 2.145| \times 4.08 \times 10^3}{1.3 \times 10^{-6}} \text{ Hz} = 241.66 \text{ MHz}.$$

For $R = 10^3$, the needed interaction length is

$$l = \frac{0.9\lambda R}{\Delta n} = \frac{0.9 \times 1.3 \times 10^{-6} \times 10^3}{|2.222 - 2.145|} \text{ m} = 1.52 \text{ cm}.$$

The spectral linewidth is simply $\Delta\lambda = \lambda/R = 1.3$ nm. The angular apertures are $\Delta\theta_i = 2n_o/R^{1/2} = 0.141$ rad $= 8.05^\circ$ and $\Delta\theta_d = 2n_e/R^{1/2} = 0.136$ rad $= 7.77^\circ$.

8.7 Guided-wave acousto-optic devices

Guided-wave acousto-optic devices are developed along the same principles discussed in preceding sections for bulk devices. For the same reasons as those that require practical bulk acousto-optic devices to be of Bragg type, all practical guided-wave acousto-optic devices also function in the Bragg regime. With only a few exceptions, a bulk counterpart can be identified for each guided-wave acousto-optic device, both of which have the same basic operation principles and are subject to the same considerations in terms of performance characteristics.

Guided-wave acousto-optic devices differ from bulk acousto-optic devices in two major aspects. First, the acoustic wave used in a guided-wave device is a *surface acoustic wave* (SAW) that is excited by a transducer fabricated on the top of the device rather than the volume acoustic wave that is used in a bulk device and is excited by a transducer attached to one end of the device. Furthermore, as is true for any guided-wave device, the optical waves involved in the function of a guided-wave acousto-optic device are in the form of waveguide modes.

A SAW needed in a guided-wave acousto-optic device is excited using an interdigital transducer (IDT) that consists of an array of electrodes on the piezoelectric surface of the device, as shown in Fig. 8.17. Many anisotropic crystals, such as $LiNbO_3$, are piezoelectric. Both the IDT for generating the SAW and the optical waveguide can be fabricated directly on such a material. Isotropic materials, such as fused silica, are not piezoelectric. Many useful crystals, such as GaAs and quartz, are also nonpiezoelectric. When a nonpiezoelectric material is used as the waveguide substrate, a layer of piezoelectric material, such as ZnO, has to be placed on top for generation of a SAW through an IDT.

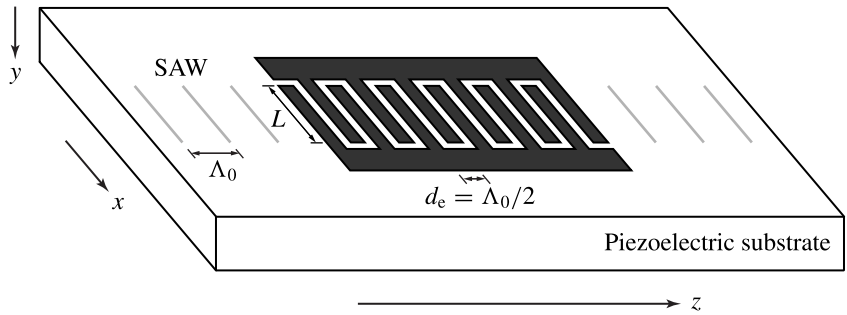


Figure 8.17 Generation of a surface acoustic wave by an interdigital transducer.

A SAW differs in many significant aspects from a bulk acoustic wave. These differences account for some qualitative and quantitative differences between a guided-wave acousto-optic device and its bulk counterpart. They also make the analysis of a guided-wave acousto-optic device complicated. Without getting into detailed analysis of SAWs, which in many cases of anisotropic media has no analytical solution, we consider here only some key features of SAWs. To facilitate the discussions, we consider a SAW propagating in the z direction on the surface of a material normal to the y direction, as shown in Fig. 8.17.

1. A SAW cannot be considered to be either longitudinal or transverse as a bulk acoustic wave is. Its mechanical displacement vector $\mathbf{u}$ has at least two components, and sometimes all three components, coupled together. Therefore, a SAW is always elliptically polarized. Moreover, the ellipticity of a SAW is a function of the distance y into the substrate.
2. In the case of an isotropic medium and also in the case of an anisotropic medium with yz being a crystal symmetry plane, $\mathbf{u}$ has only y and z components. We then have $\mathbf{u}(\mathbf{r}, t) = \hat{y}u_y(y, z, t) + \hat{z}u_z(y, z, t) = (\hat{y}\mathcal{U}_y(y) + \hat{z}\mathcal{U}_z(y))\cos(Kz - \Omega t)$ for a traveling SAW propagating in the z direction, where both $\mathcal{U}_y(y)$ and $\mathcal{U}_z(y)$ decay in the y direction in a distance on the order of the acoustic wavelength Λ. In this situation, we find that the nonvanishing strain tensor elements are $S_2 = S_{yy}$, $S_3 = S_{zz}$, and $S_4 = 2S_{yz}$.
3. In the case of an anisotropic medium with yz not coinciding with any crystal symmetry plane, $\mathbf{u}$ has all three x, y, and z components. All three components are functions of y and z but do not vary with x. Then, $\mathbf{u}(\mathbf{r}, t) = \hat{x}u_x(y, z, t) + \hat{y}u_y(y, z, t) + \hat{z}u_z(y, z, t) = (\hat{x}\mathcal{U}_x(y) + \hat{y}\mathcal{U}_y(y) + \hat{z}\mathcal{U}_z(y))\cos(Kz - \Omega t)$. We find in this situation that only $S_1 = S_{xx}$ vanishes. All other strain tensor elements exist at the same time.
4. While any strain tensor element $S(y, z, t)$ is a function of y, z, and t, its amplitude $\mathcal{S}(y)$ is a function of y only because the amplitude $\mathcal{U}(y)$ of the displacement is a function of y only. The strain field of a SAW is localized in a surface layer of

a thickness on the order of the acoustic wavelength, but it might oscillate as it tapers off. The amplitudes of different strain tensor elements normally have different magnitudes and phases as a SAW is elliptically polarized.

5. The total acoustic power of a SAW is then

$$P_{\mathrm{a}} = \frac{L}{2}\rho v_{\mathrm{a}}^3 \int_0^\infty \mathcal{S}^2(y)\mathrm{d}y, \tag{8.155}$$

where L is the width of the SAW beam and v_{a} is the velocity of the SAW. In a uniform IDT like the one shown in Fig. 8.17, L is defined by the overlap between adjacent electrodes rather than by the length of the electrodes.

6. In a piezoelectric medium such as $LiNbO_3$, acousto-optic interaction between a SAW and an optical wave generally involves the electro-optic effect caused by the electric field generated by the SAW. This phenomenon can have a significant effect on the figure of merit of the acousto-optic interaction.
7. The velocity of a SAW is smaller than the velocities of the longitudinal and transverse modes of the bulk acoustic wave in the same medium.

The polarity of the bias voltage alternates periodically in an IDT, resulting in a spatial periodicity that is characterized by a pair of electrodes of opposite polarity in each period. For an IDT that is designed to function around a center acoustic frequency of f_0, this periodicity has to match the wavelength Λ_0 of the SAW at this frequency. Thus the center-to-center distance between two adjacent electrodes of opposite polarity is $d_{\mathrm{e}} = \Lambda_0/2$. The corresponding SAW frequency

$$f_0 = \frac{v_{\mathrm{a}}}{2d_{\mathrm{e}}} \tag{8.156}$$

is called the *synchronous frequency* of the IDT. To generate a SAW at a frequency f, which can be different from f_0, an RF signal at this frequency f is applied to the electrodes of the IDT. As illustrated in Fig. 8.17, an IDT radiates bidirectionally two SAWs of equal power at both ends. Only one of these two SAWs is useful at a time for any given acousto-optic interaction in a device.

The conversion efficiency of an IDT is determined by the matching circuit that drives the IDT and the radiation impedance of the electrode array in the IDT. The frequency response of an IDT has a bandwidth that is determined by the number N_{e} of the electrodes or, more precisely, the number $N_{\mathrm{g}} = N_{\mathrm{e}} - 1$ of the gaps in the electrode array. Therefore, the useful acoustic power, which accounts for the acoustic output from *only one end* of the IDT, is

$$P_{\mathrm{a}}(f) = \eta_{\mathrm{t}}(f, N_{\mathrm{g}})P_{\mathrm{e}}(f) = \eta_{\mathrm{ckt}}(f, N_{\mathrm{g}})\eta_{\mathrm{a}}(f, N_{\mathrm{g}})P_{\mathrm{e}}(f), \tag{8.157}$$

where $P_{\mathrm{e}}(f)$ is the electrical power at the RF frequency f, $\eta_{\mathrm{ckt}}(f, N_{\mathrm{g}})$ accounts for the

efficiency of the matching circuit, and

$$\eta_a(f, N_g) = \frac{\sin^2[N_g\pi(f - f_0)/2f_0]}{N_g^2 \sin^2[\pi(f - f_0)/2f_0]} \approx \frac{\sin^2[N_g\pi(f - f_0)/2f_0]}{[N_g\pi(f - f_0)/2f_0]^2} \tag{8.158}$$

is the normalized *acoustic radiation efficiency* of the IDT. This normalized efficiency has a unity peak value at the center frequency: $\eta_a(f_0, N_g) = 1$. The bandwidth of η_a depends on the number of electrodes. From (8.158), we find that the 3-dB acoustic radiation bandwidth, f_{3dB}, of the IDT for $\eta_a(f_0 \pm f_{3dB}/2, N_g) = 1/2$ is

$$f_{3dB} = \frac{1.772}{N_g} f_0. \tag{8.159}$$

The specific functional form of $\eta_{ckt}(f, N_g)$ depends on the type of circuit used to drive the IDT. Both its peak value and its bandwidth depend on the number of electrodes and the circuit design. For a given type of circuit, the peak efficiency increases but the bandwidth decreases as the number of electrodes in an IDT is increased. The circuit is usually designed to have a bandwidth much larger than the acoustic radiation bandwidth so that η_{ckt} is practically a constant at its peak value in the frequency range within $f_0 \pm f_{3dB}/2$. The theoretical maximum value of η_{ckt} is 1/2 because an IDT radiates bidirectionally. Therefore, in the most ideal situation with an optimized broadband circuit of maximum efficiency, we have $\eta_t(f, N_g) \leq \eta_a(f, N_g)/2$ with a bandwidth limited by f_{3dB} given in (8.159).

EXAMPLE 8.10 A z-propagating SAW in y-cut $LiNbO_3$ has a phase velocity $v_a = 3.488$ km s^{-1}. An IDT is designed to generate this SAW at a center frequency of 500 MHz and a 3-dB bandwidth of at least 100 MHz. Find the center-to-center separation between the electrodes and the number of electrodes for the IDT.

Solution The synchronous frequency of the IDT has to be the center acoustic frequency $f_0 = 500$ MHz. Therefore, we find from (8.156) that the center-to-center electrode separation needs to be

$$d_e = \frac{v_a}{2f_0} = \frac{3.488 \times 10^3}{2 \times 500 \times 10^6} \text{ m} = 3.488 \text{ μm}.$$

For $f_{3dB} \geq 100$ MHz, we find from (8.159) that

$$N_g = 1.772\frac{f_0}{f_{3dB}} \leq \frac{1.772 \times 500 \times 10^6}{100 \times 10^6} = 8.86.$$

Therefore, the maximum value for N_g is 8, and the maximum number of electrodes is $N_e = 9$.

A guided-wave acousto-optic device can be either coplanar or collinear. Similarly to their bulk counterparts, guided-wave acousto-optic modulators and deflectors generally use small-angle Bragg diffraction in a coplanar configuration while guided-wave

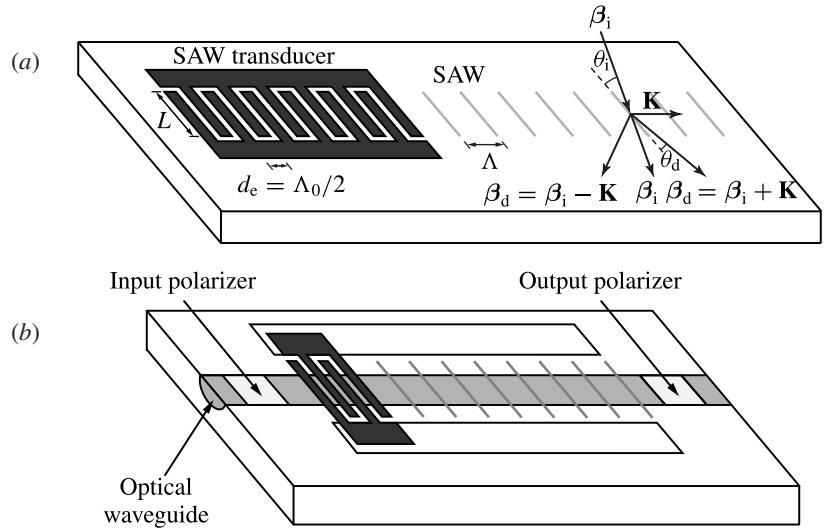


Figure 8.18 Basic configurations for (*a*) coplanar and (*b*) collinear guided-wave acousto-optic devices.

acousto-optic tunable filters and mode converters have the configuration of collinear Bragg diffraction. The basic configurations for coplanar and collinear guided-wave acousto-optic devices are shown in Figs. 8.18(*a*) and (*b*), respectively. In either case, the phase-matching condition is dictated by the wavevectors of the participating normal modes of the optical waveguide:

$$\boldsymbol{\beta}_d = \boldsymbol{\beta}_i \pm \mathbf{K}, \tag{8.160}$$

where $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_d$ are the normal-mode wavevectors of the incident and diffracted optical beams, respectively, $\mathbf{K}$ is the wavevector of the SAW, and the plus and minus signs are for up- and down-shifted diffraction, respectively. Most of the results obtained in Section 8.3 for Bragg diffraction in a bulk medium can be applied directly to Bragg diffraction in an optical waveguide by (1) replacing $\mathbf{k}_i$ and $\mathbf{k}_d$ by $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_d$, respectively, and (2) interpreting n_i and n_d as $n_i = \lambda\beta_i/2\pi$ and $n_d = \lambda\beta_d/2\pi$ to represent the *effective refractive indices* seen by relevant waveguide modes. The phase mismatch, if it exists, becomes $\Delta\mathbf{k} = \boldsymbol{\beta}_d - \boldsymbol{\beta}_i - \mathbf{K}$ for up-shifted diffraction and $\Delta\mathbf{k} = \boldsymbol{\beta}_d - \boldsymbol{\beta}_i + \mathbf{K}$ for down-shifted diffraction. After making these modifications, the results obtained in the preceding three sections for bulk acousto-optic devices are also valid for their guided-wave counterparts, with the exception of the diffraction efficiency and some other characteristics discussed later in this section.

The coupled equations for Bragg diffraction in an optical waveguide have to be obtained by employing a coupled-mode analysis. They have the same form as (8.79) and (8.80), with A_i and A_d now representing the amplitudes of waveguide modes that characterize the incident and the diffracted optical beams, respectively. The Bragg coupling coefficient obtained from the coupled-mode analysis can be expressed in a

form similar to that in (8.81):

$$\kappa = \pm \mathrm{i} \frac{\omega^2 \mu_0}{4 \beta_\mathrm{i}^{1/2} \beta_\mathrm{d}^{1/2}} \Delta \tilde{\epsilon}_\mathrm{id}^\mathrm{eff}, \tag{8.161}$$

where the plus and the minus signs are for up- and down-shifted diffraction, respectively, and

$$\Delta \tilde{\epsilon}_\mathrm{id}^\mathrm{eff} = \frac{2 \beta_\mathrm{i}^{1/2} \beta_\mathrm{d}^{1/2}}{\omega \mu_0} \int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\mathrm{i}^* \cdot \Delta \tilde{\boldsymbol{\epsilon}} \cdot \hat{\boldsymbol{\mathcal{E}}}_\mathrm{d} \mathrm{d}y. \tag{8.162}$$

If the photoelastic effect is the sole contribution to the changes in the dielectric properties of the medium in a guided-wave device, we have

$$\Delta \tilde{\epsilon}_\mathrm{id}^\mathrm{eff} \approx -\epsilon_0 \left(\frac{2 n_\mathrm{i} n_\mathrm{d} M_2 P_\mathrm{a}}{L \Lambda} \right)^{1/2} \Gamma_\mathrm{id}, \tag{8.163}$$

where n_i and n_d are the refractive indices seen by the modes of the incident and the diffracted optical beams, respectively, and

$$\Gamma_\mathrm{id} = \frac{\displaystyle\int_0^\infty \mathcal{S}(y) \hat{\mathcal{E}}_\mathrm{i}^*(y) \hat{\mathcal{E}}_\mathrm{d}(y) \mathrm{d}y}{\left[\dfrac{1}{\Lambda} \displaystyle\int_0^\infty \mathcal{S}^2(y) \mathrm{d}y \int_{-\infty}^{\infty} |\hat{\mathcal{E}}_\mathrm{i}(y)|^2 \mathrm{d}y \int_{-\infty}^{\infty} |\hat{\mathcal{E}}_\mathrm{d}(y)|^2 \mathrm{d}y \right]^{1/2}} \tag{8.164}$$

is a *dimensionless* overlap factor for the optical modes and the strain field of the SAW. The overlap factor Γ_id accounts for the different spatial distributions of the optical modes and the SAW. It is a function of the acoustic frequency f because the strain field distribution of a SAW depends on the acoustic frequency. The Bragg coupling coefficient for a guided-wave acousto-optic device can then be expressed as

$$|\kappa| = \frac{\pi}{\lambda} \left(\frac{M_2 P_\mathrm{a}}{2 L \Lambda} \right)^{1/2} \Gamma_\mathrm{id} = \frac{\pi}{\lambda} \left(\frac{M_2 f P_\mathrm{a}}{2 L v_\mathrm{a}} \right)^{1/2} \Gamma_\mathrm{id}. \tag{8.165}$$

Consequently, for a coplanar or a codirectional, collinear guided-wave acousto-optic device that uses a traveling acoustic wave, the diffraction efficiency for an interaction length l in the case of perfect phase matching is

$$\eta_\mathrm{PM} = \sin^2 \left[\frac{\pi}{\lambda} \left(\frac{M_2 f}{2 L v_\mathrm{a}} P_\mathrm{a}(f) \right)^{1/2} \Gamma_\mathrm{id}(f) l \right], \tag{8.166}$$

or, in the low-efficiency limit,

$$\eta_\mathrm{PM} \approx \frac{\pi^2 M_2 l^2 f}{2 \lambda^2 L v_\mathrm{a}} \Gamma_\mathrm{id}^2(f) P_\mathrm{a}(f), \qquad \text{if } \eta_\mathrm{PM} \ll 1. \tag{8.167}$$

In expressing $\Delta\tilde{\epsilon}_{id}^{eff}$ in the form of (8.163), we have assumed that only the photoelastic effect contributes to the changes in the dielectric properties of the medium. The discussions and conclusions that follow (8.163) are subject to this assumption. Many practical acousto-optic materials, such as $LiNbO_3$ for example, are piezoelectric and have large Pockels coefficients. Through the piezoelectric effect, the SAW can generate a significant electric field, which can then induce a significant change in the dielectric properties of the medium through the Pockels effect. In this situation, $\Delta\tilde{\epsilon}$ has contributions from both photoelastic and electro-optic effects. Then, the parameters M_2 and Γ_{id} that appear in (8.165)–(8.167) have contributions from both effects. In certain situations, such as in the interaction of a TE mode with a z-propagating SAW in a y-cut $LiNbO_3$ waveguide, the contribution from the electro-optic effect can even dominate that from the photoelastic effect.

Modulators

A guided-wave acousto-optic modulator generally has a coplanar configuration like that shown in Fig. 8.18(a). The general characteristics of acousto-optic modulators discussed in Section 8.4 remain true for guided-wave modulators as well. The only characteristic that is different between a guided-wave acousto-optic modulator and its bulk counterpart is the functional dependence of the diffraction efficiency on various device parameters, as is discussed above and can be seen by comparing (8.166) with (8.106). Although a modulator is driven by a time-dependent signal at a fixed carrier frequency, the diffraction efficiency of each frequency component in the modulation signal varies because of its frequency dependence, which is expressed in (8.166). As a result, the modulation bandwidth of a guided-wave acousto-optic modulator is not only determined by the acoustic transit time τ_a, but is also subject to the bandwidth of the IDT determined by (8.158) and that of the overlap factor $\Gamma_{id}(f)$.

Deflectors

Guided-wave acousto-optic deflectors are coplanar devices. They have a basic configuration like that shown in Fig. 8.18(a) though more sophisticated transducers are often used for broadband deflectors. The discussions in Section 8.5 are generally applicable to guided-wave acousto-optic deflectors.

The most important parameter that characterizes the performance of an acousto-optic deflector is its bandwidth Δf because, as the relation in (8.132) shows, the number of spatially resolvable spots that can be addressed by a deflector is linearly proportional to its bandwidth. The bandwidth of a guided-wave acousto-optic deflector is limited by the phase-matching bandwidth of the acousto-optic Bragg diffraction, the IDT bandwidth, and the bandwidth of the overlap factor.

The characteristics of the Bragg phase-matching bandwidth have been fully discussed in Section 8.5. The Bragg bandwidth can be increased either by using birefringent Bragg

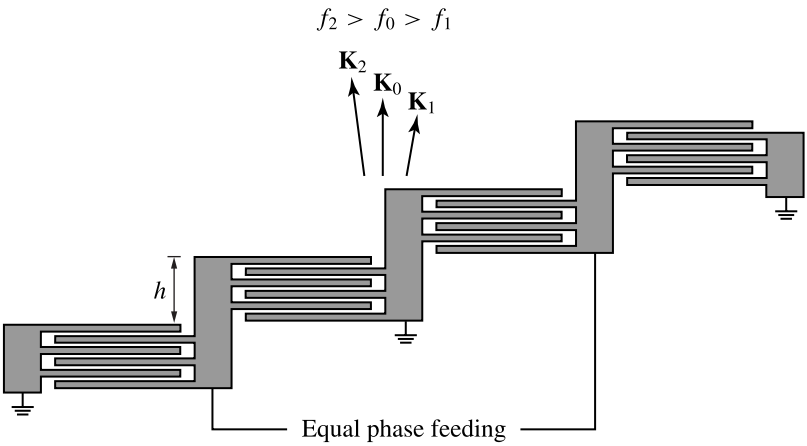


Figure 8.19 Phased-array interdigital transducer for the generation of a surface acoustic wave whose propagation direction is steered by the acoustic frequency to track the Bragg condition.

diffraction under the tangential phase-matching condition or by using a phased-array transducer in a nonbirefringent deflector. These approaches are applicable to guided-wave deflectors as well. In a birefringent guided-wave Bragg deflector that uses the tangential phase-matching condition, the incident and the diffracted optical beams are waveguide modes of orthogonal polarizations. In a planar waveguide, this kind of interaction involves coupling between TE and TM modes. A bandwidth comparable to that of an optimized birefringent deflector is possible for a nonbirefringent deflector that uses a phased-array transducer.

In a guided-wave device, a phased-array transducer consists of multiple elements of identical IDTs that are equally spaced but are successively displaced at a constant step h, as shown in Fig. 8.19. The step height is chosen to be an integral multiple of the center-to-center distance between adjacent electrodes:

$$h = md_e = m\frac{\Lambda_0}{2}, \qquad m = \text{integer}. \tag{8.168}$$

At the center frequency f_0, the SAWs generated by the IDT elements are all in phase, resulting in a combined SAW propagating in the normal direction. At any frequency f shifted away from f_0, the regular spatial displacement between successive IDT elements causes a frequency-dependent regular phase shift between adjacent individual SAWs, thereby steering the direction of the combined SAW wavefront according to the variation in the acoustic frequency.

As can be seen from (8.158) and subsequent discussions, the bandwidth of a regular IDT is quite limited. A phased-array IDT as shown in Fig. 8.19 increases the Bragg bandwidth of the acousto-optic interaction, but it does not improve the bandwidth of the IDT itself. Consequently, the bandwidth of a deflector using such a phased-array IDT is still subject to the limitation of the same IDT bandwidth. This limitation

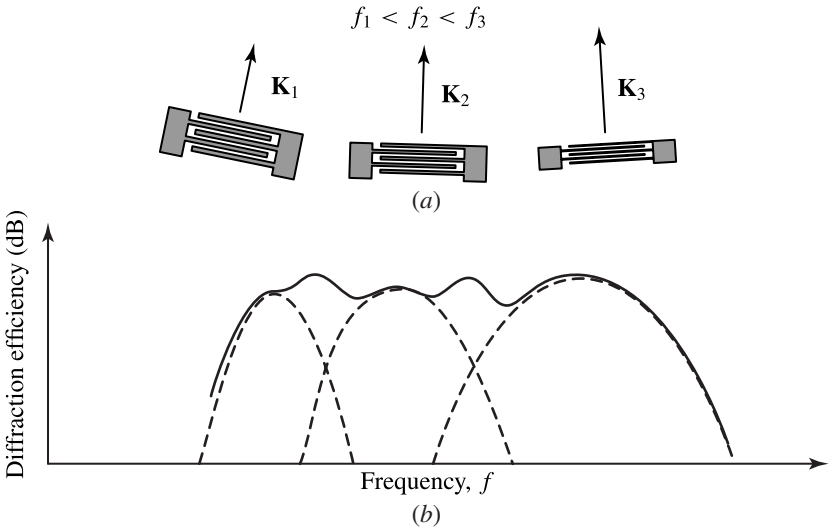


Figure 8.20 (*a*) Multiple tilted interdigital transducers of staggered center frequencies. (*b*) Individual frequency responses and composite frequency response of the transducers.

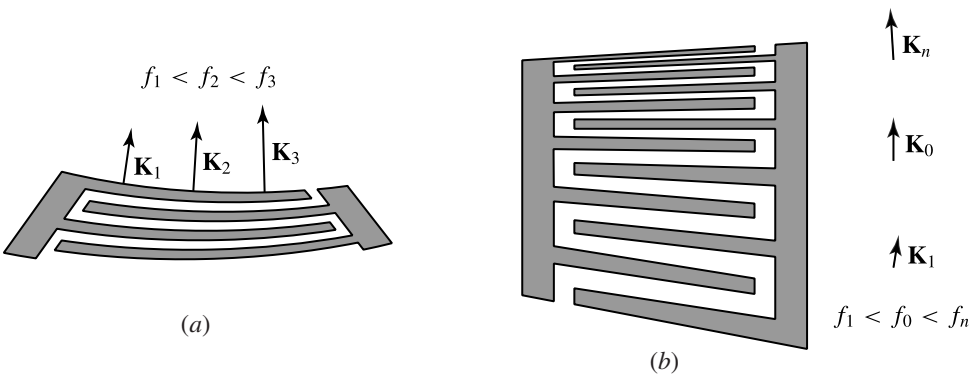


Figure 8.21 (*a*) Curved interdigital transducer with tapered electrodes. (*b*) Tilted-finger chirped interdigital transducer.

can be overcome by using an array of *multiple tilted IDTs* of staggered synchronous frequencies, as shown in Fig. 8.20(*a*). The individual IDTs are tilted in order for the Bragg phase-matching condition to be satisfied at each synchronous frequency. The difference in the tilt angles between two adjacent IDTs is equal to the difference between the Bragg angles at the corresponding synchronous frequencies of the two IDTs. The frequency response of each individual IDT is centered at the corresponding synchronous frequency. The synchronous frequencies of the individual IDTs are staggered in such a way that the individual frequency responses of adjacent IDTs intersect at -6 dB down from the peak response in order to obtain a large composite overall frequency response, as illustrated in Fig. 8.20(*b*). Two broadband IDT structures that evolve from the same concept are shown in Fig. 8.21. The *curved IDT* with tapered electrodes shown

in Fig. 8.21(*a*) results from parallel connection of multiple tilted IDTs of continuously varying parameters, whereas the *tilted-finger chirped IDT* shown in Fig. 8.21(*b*) results from serial stacking of multiple tilted IDTs of continuously varying parameters.

An array of multiple tilted IDTs has a broad composite transducer bandwidth. It also provides a broad Bragg bandwidth due to tracking of the Bragg phase-matching condition by the tilted angles of the individual IDTs. The curved IDT and the tilted-finger chirped IDT have similar characteristics. In contrast, a phased-array IDT increases only the Bragg bandwidth but not the transducer bandwidth. However, at any given acoustic frequency, all of the elements in a phased-array IDT are simultaneously excited, whereas in any of the broadband IDTs only part of the structure is effectively excited at a given acoustic frequency due to the design of staggered synchronous frequencies. Therefore, a deflector that uses a phased-array IDT can have a larger diffraction efficiency than one that uses one of the broadband IDTs. In order to optimize the bandwidth and the efficiency at the same time, a combination of different schemes is often used. For example, an array of multiple tilted phased-array IDTs of staggered synchronous frequencies can be designed for a nonbirefringent deflector of high efficiency and large bandwidth. In a birefringent deflector that has a large Bragg bandwidth, the Bragg angle varies little with varying acoustic frequencies due to the tangential phase-matching scheme. For such a device to have a large overall bandwidth, its Bragg bandwidth can be matched with a large IDT bandwidth by using an array of multiple *untilted* IDTs of staggered synchronous frequencies or a *parallel*-finger chirped IDT.

To utilize the Bragg bandwidth and the IDT bandwidth fully, the bandwidth of the overlap factor is generally made large enough by ensuring, through the design of the waveguide structure, sufficient overlap of the interacting optical modes with the SAW field distribution.

Tunable filters

Guided-wave acousto-optic tunable filters are typically collinear devices though, as is mentioned in Section 8.6, noncollinear configurations are also possible. The general discussions presented in Section 8.6 apply to guided-wave tunable filters as well. However, guided-wave tunable filters have some unique characteristics due to modal dispersion that does not exist in bulk devices. Because modes of the same polarization but different orders, such as TE_0 and TE_1, have different propagation constants, a collinear guided-wave acousto-optic tunable filter that is based on a change in the mode order without a change in the polarization is possible. However, guided-wave acousto-optic tunable filters that involve orthogonally polarized modes cannot be constructed with isotropic materials though the orthogonally polarized TE and TM modes always have different propagation constants even when the waveguide structure is made of layers of isotropic materials (see Problem 8.7.1). It is possible to construct collinear acousto-optic filters with TE–TM mode conversion using waveguides in properly oriented anisotropic substrates (see Problem 8.7.4).

Mode converters

The guided-wave acousto-optic filters discussed above can be used as guided-wave acousto-optic mode converters. The only difference between these two types of devices is the purpose of their applications. Any guided-wave acousto-optic tunable filter cannot operate without converting one waveguide mode to another, but its purpose is to select a desired optical wavelength out of a possibly broad spectrum. In contrast, the functional purpose of a guided-wave acousto-optic mode converter is to convert a waveguide mode from one polarization to another or from one mode order to another. A high mode-conversion efficiency is the most desirable characteristic of a mode converter. A mode converter is generally designed for the phase-matching condition to be satisfied at a desired optical wavelength of application. Furthermore, the input to the device is a coherent optical wave in the designated incident waveguide mode at this particular wavelength so that a high mode-conversion efficiency can be obtained.

PROBLEMS

8.2.1 Fused silica glass is an isotropic material that has only two independent elasto-optic coefficients, $p_{11} = 0.121$ and $p_{12} = 0.271$. A transverse acoustic wave at a frequency of 500 MHz that is polarized in the y direction is generated to propagate in the x direction. Use the data listed in Table 8.2 to answer the following questions.

a. Find the wavelength of the transverse acoustic wave and compare it to that of the longitudinal acoustic wave found in Example 8.1.

b. Find the figures of merit at 632.8 nm optical wavelength for acousto-optic interaction with optical waves of linear polarization that are parallel and perpendicular to $\mathbf{K}$, respectively.

c. Find the change $\Delta\epsilon(x, t)$ in the permittivity tensor of this material caused by such an acoustic wave at an intensity of 10 W cm^{-2}.

8.2.2 Consider an acoustic wave propagating along the [100] axis of a cubic crystal of $\bar{4}3m$ symmetry such as GaAs or GaP. Take $\hat{x}$ to be along the crystal [100] axis so that $\mathbf{K} = K\hat{x}$.

a. Find the tensor $\Delta\epsilon$ induced by each normal mode of the acoustic wave in terms of elasto-optic coefficients and strain tensor elements.

b. From the answers in (a), find the conditions under which acousto-optic diffraction in GaP does not change the polarization of the optical waves.

8.2.3 The three independent elasto-optic coefficients for a GaP crystal of $\bar{4}3m$ symmetry are $p_{11} = -0.151$, $p_{12} = -0.082$, and $p_{44} = -0.072$. The density of the crystal is $\rho = 4.13 \times 10^3$ kg m^{-3}. The refractive index of GaP at 632.8 nm wavelength is 3.31. For an acoustic wave propagating along one of the crystal axes, say [100], the velocity of its longitudinal mode is $v_{a,L} = 5.85$ km s^{-1} and

that of its transverse modes is $v_{a,T} = 4.13$ km s^{-1}. Take $\hat{x}$ to be the direction of the [100] crystal axis. The acoustic wave has an intensity of 10 W cm^{-2}.

a. Find the figures of merit for acousto-optic interactions of a longitudinal acoustic wave with optical waves at 632.8 nm of linear polarizations that are parallel and perpendicular to **K**, respectively.
b. Find the index changes caused by this longitudinal acoustic wave.
c. Find the figures of merit for acousto-optic interactions of a transverse acoustic wave with optical waves of linear polarizations that are parallel and perpendicular to **K**, respectively.
d. Find the changes in the permittivity tensor caused by this transverse acoustic wave.

8.2.4 Though the linear optical property of a cubic crystal is similar to that of an isotropic material in the sense that a cubic crystal has only one refractive index independent of the optical polarization, the acousto-optic properties of a cubic crystal such as GaP are different from those of an isotropic material such as silica glass. We consider in this problem a longitudinal acoustic wave propagating along the [110] direction of a GaP crystal. The velocity of this acoustic wave is $v_{a,L} = 6.32$ km s^{-1}, which is different from that of the longitudinal wave propagating along one of the crystal axes described in Problem 8.2.3. To describe the acousto-optic interaction in this situation, we take the new coordinates $\hat{x}' = (\hat{x} + \hat{y})/\sqrt{2}$, $\hat{y}' = (\hat{x} - \hat{y})/\sqrt{2}$, and $\hat{z}' = \hat{z}$, where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are the coordinate axes along the crystal axes. Then $\hat{x}'$ is along the [110] direction of the crystal. The elasto-optic coefficients transformed to this new coordinate system are $p'_{11} = (p_{11} + p_{12} + 2p_{44})/2$, $p'_{12} = (p_{11} + p_{12} - 2p_{44})/2$, and $p'_{44} = (p_{11} - p_{12})/2$.

a. Use the data given in Problem 8.2.3 for GaP to find the figures of merit for acousto-optic interactions of this [110] longitudinal acoustic wave with optical waves at 632.8 nm of linear polarizations that are parallel and perpendicular to **K**, respectively.
b. Find the index changes caused by such a longitudinal acoustic wave of an intensity of 10 W cm^{-2}.

8.2.5 $LiNbO_3$ is a trigonal crystal of $3m$ symmetry. It is uniaxial with an optical axis parallel to the crystal [001] direction that is designated the z axis. The crystal [100] and [010] directions are designated as the x and y axes, respectively.

a. Consider an acoustic wave that propagates in the crystal along the [001] direction with $\mathbf{K} = K\hat{z}$. Find the tensor $\Delta\boldsymbol{\epsilon}$ induced by each normal mode of this acoustic wave in terms of elasto-optic coefficients and strain tensor elements.
b. Find the tensor $\Delta\boldsymbol{\epsilon}$ induced by each normal mode of an acoustic wave that propagates along the [100] direction with $\mathbf{K} = K\hat{x}$.
c. Find the tensor $\Delta\boldsymbol{\epsilon}$ induced by each normal mode of an acoustic wave that propagates along the [010] direction with $\mathbf{K} = K\hat{y}$.

d. From the answers in (a)–(c), find the conditions under which acousto-optic diffraction in $LiNbO_3$ does not change the polarization of the optical waves.
e. From the answers in (a)–(c), find the conditions under which the incident and diffracted optical waves of acousto-optic Bragg diffraction in $LiNbO_3$ are orthogonally polarized.

8.2.6 The tetragonal crystal $PbMoO_4$ has $4/m$ symmetry. It is negative uniaxial with the z axis designated as its optical axis. Its density is $\rho = 6.95 \times 10^3$ kg m^{-3}, and its ordinary and extraordinary refractive indices at $\lambda = 632.8$ nm are $n_o = 2.386$ and $n_e = 2.262$, respectively. The velocity of its longitudinal acoustic mode propagating along the crystal [001] direction with $\mathbf{K} = k\hat{z}$ is $v_a = 3.63$ km s^{-1}. Its nonvanishing independent elasto-optic coefficients are $p_{11} = 0.24$, $p_{12} = 0.24$, $p_{13} = 0.255$, $p_{16} = 0.017$, $p_{31} = 0.175$, $p_{33} = 0.300$, $p_{44} = 0.067$, $p_{45} = -0.01$, $p_{61} = 0.013$, and $p_{66} = 0.025$. Find the acoustic figures of merit at 632.8 nm for optical polarizations that are perpendicular and parallel to $\mathbf{K}$, respectively.

8.3.1 Consider Raman–Nath diffraction, in an isotropic material, of an optical wave propagating in the z direction by an acoustic wave propagating in the x direction as shown in Fig. 8.2.
a. Show that in the diffraction by a longitudinal acoustic wave the polarizations of the diffracted waves of all orders are the same as that of the incident wave if the incident optical wave is linearly polarized in either the x or y direction.
b. Show that in the diffraction by a y-polarized transverse acoustic wave the polarizations of all even-order diffracted waves are the same as that of the incident wave and those of all odd-order diffracted waves are the same as that of the first-order diffracted wave for any polarization state of the incident optical wave.
c. What happens in the interaction with a z-polarized transverse acoustic wave?

8.3.2 An optical wave at 632.8 nm wavelength interacts at normal incidence with a longitudinal acoustic wave of 10 MHz frequency in a piece of TF-7 flint glass over an interaction length of 5 mm. The optical wave is linearly polarized in a direction perpendicular to the $\mathbf{K}$ vector of the acoustic wave. Use the data given in Table 8.2 for TF-7 glass to answer the following questions: Is the interaction in the Raman–Nath or Bragg regime? What is the maximum value for the first-order diffraction efficiency η_1? Find the acoustic intensity needed to reach this maximum diffraction efficiency. Find the first-order diffraction efficiency for an acoustic intensity of 5 W cm^{-2}. What is the first-order diffraction angle?

8.3.3 The nonvanishing independent elasto-optic coefficients of $LiNbO_3$, which has $3m$ symmetry, are $p_{11} = -0.026$, $p_{12} = 0.090$, $p_{13} = 0.133$, $p_{14} = -0.075$, $p_{31} = 0.179$, $p_{33} = 0.071$, $p_{41} = -0.151$, and $p_{44} = 0.146$. For acoustic

waves propagating in the [100] direction with $\mathbf{K} = K\hat{x}$, the acoustic velocities are $v_{a,L} = 6.57$ km s^{-1} for the longitudinal mode polarized in the x direction, $v_{a,T} = 4.08$ km s^{-1} for the transverse mode polarized in the y direction, and $v_{a,T} = 3.59$ km s^{-1} for the transverse mode polarized in the z direction. The density of $LiNbO_3$ is 4.64×10^3 kg m^{-3}. The optical axis is the z axis. The ordinary and extraordinary indices of $LiNbO_3$ at 1.3 μm optical wavelength are $n_o = 2.222$ and $n_e = 2.145$, respectively. We are interested in up-shifted Bragg diffraction at $\lambda = 1.3$ μm by a z-polarized transverse acoustic mode. Both $\mathbf{k}_i$ and $\mathbf{k}_d$ of the incident and diffracted optical waves lie in the xy plane. The incident optical wave is a z-polarized extraordinary wave so that $\hat{e}_i = \hat{z}$. Use the results obtained in Problem 8.2.5 to answer the following questions.

a. Is the diffracted wave an ordinary or extraordinary wave? Is this birefringent or nonbirefringent diffraction?
b. What are the minimum and maximum acoustic frequencies that define the frequency range for phase-matched interaction?
c. Is tangential phase matching possible? If it is possible, what is the required acoustic frequency f_0? What are the values of θ_i and θ_d?
d. What is the figure of merit M_2 at the tangential phase-matching point if that is possible?

8.3.4 Answer the questions given in Problem 8.3.3 in the case when the incident optical wave is an ordinary wave polarized in the xy plane so that $\hat{e}_i \perp \hat{z}$.

8.3.5 Answer the questions given in Problem 8.3.3 in the case when the acoustic wave is a y-polarized transverse mode and the incident optical wave is a z-polarized extraordinary wave so that $\hat{e}_i = \hat{z}$.

8.3.6 The velocities of acoustic waves propagating in the [001] direction of $LiNbO_3$ with $\mathbf{K} = K\hat{z}$ are $v_{a,L} = 7.27$ km s^{-1} for the longitudinal mode polarized in the z direction and $v_{a,T} = 3.59$ km s^{-1} for both of the transverse modes polarized in the x and y directions, respectively. We consider up-shifted Bragg diffraction at $\lambda = 1.3$ μm with both $\mathbf{k}_i$ and $\mathbf{k}_d$ of the incident and diffracted optical waves in the yz plane. The acoustic wave is the x-polarized transverse mode. The incident optical wave is an x-polarized ordinary wave so that $\hat{e}_i = \hat{x}$. Answer the questions given in Problem 8.3.3. What difference does it make if the acoustic wave is y polarized?

8.3.7 Show that, for birefringent Bragg diffraction in an anisotropic crystal in the case when $n_i > n_d$, phase-matched diffraction is not possible if the incident angle of the optical wave falls in the range of $|\theta_i| < \cos^{-1}(n_d/n_i)$. Show also that there is no such limitation if $n_i < n_d$.

8.3.8 Show that for given values of incident and diffracted optical wavenumbers, k_i and k_d, respectively, in an anisotropic medium, the value of the acoustic wavenumber, K, has to be in the range given in (8.67) in order for phase-matched Bragg diffraction to be possible.

8.3.9 Consider phase-matched Bragg diffraction in which the acoustic wave has fixed polarization and direction of propagation but its frequency can be varied. Assume for simplicity that the values of n_i and n_d for the incident and diffracted waves do not depend on the angles of incidence and diffraction, θ_i and θ_d, respectively. Plot θ_i and θ_d as a function of the varying acoustic frequency for the following cases.
   a. Up-shifted diffraction in the case when $n_i > n_d$.
   b. Down-shifted diffraction in the case when $n_i > n_d$.
   c. Up-shifted diffraction in the case when $n_i < n_d$.
   d. Down-shifted diffraction in the case when $n_i < n_d$.
   e. Up-shifted diffraction in an isotropic medium.
   f. Down-shifted diffraction in an isotropic medium.

8.3.10 In some special situations, more than one diffraction order can be simultaneously phase matched in Bragg diffraction from a traveling acoustic wave in an anisotropic crystal. Consider for simplicity the interaction in a uniaxial crystal with a configuration in which $\mathbf{k}_i$, $\mathbf{k}_d$, and $\mathbf{K}$ all lie on the plane normal to the optical axis of the crystal. Then, for both ordinary and extraordinary waves, the magnitudes of the wavevectors $\mathbf{k}_i$ and $\mathbf{k}_d$ are independent of the angles θ_i and θ_d. Take the two characteristic refractive indices of the crystal to be n_1 and n_2 with $n_1 > n_2$. The incident optical wave has a wavelength λ and an angular frequency ω. For phase matching, we vary the acoustic frequency f to vary the value of K. Consider both birefringent and nonbirefringent phase-matching possibilities in answering the following questions.
   a. If $n_i = n_1 > n_2$, is it possible to generate two up-shifted beams at $\omega + \Omega$ and $\omega + 2\Omega$ simultaneously through phase-matched Bragg diffraction? If it is possible, how many different scenarios can be found? What are the required conditions for each case? What are the phase-matching acoustic frequency, in terms of $\hat{f}$ defined in (8.68), and the angles of incidence and diffraction for the beams?
   b. Answer the questions in (a) for $n_i = n_2 < n_1$.
   c. Answer the questions in (a) for the simultaneous generation of two down-shifted beams at $\omega - \Omega$ and $\omega - 2\Omega$ in the case when $n_i = n_1 > n_2$.
   d. Answer the questions in (c) for $n_i = n_2 < n_1$.
   e. It is also possible to generate one up-shifted beam at $\omega + \Omega$ and one down-shifted beam at $\omega - \Omega$ simultaneously. Find the condition, the phase-matching acoustic frequency, and the angles of incidence and diffraction for each possible case.

8.3.11 Give an example for the situations described in Problem 8.3.10 by considering Bragg diffraction in $LiNbO_3$ with $n_o = 2.222$ and $n_e = 2.145$ at the 1.3 μm optical wavelength.

8.3.12 In the absence of an actively generated acoustic wave, thermal vibrations in a medium can serve as the required acoustic waves to cause frequency-shifted

diffraction of an optical wave. This phenomenon is the well-known Brillouin scattering, which is essentially phase-matched Bragg diffraction by the acoustic modes of a medium. If an optical wave is launched into a silica optical fiber, in which direction or directions do the frequency-shifted optical waves generated by spontaneous Brillouin scattering propagate? What frequency components can be found in the scattered waves? Use the Sellmeier equation given in (3.96) and the acoustic velocity of $v_a = 5.97$ km s^{-1} to find the frequency shift, known as the Brillouin frequency, f_B, at the following optical wavelengths: 632.8 and 850 nm, and 1.06, 1.3, and 1.55 μm.

8.3.13 A z-polarized extraordinary optical wave at 1.3 μm is collinearly diffracted by a longitudinal acoustic wave propagating in the [100] x direction of $LiNbO_3$ to generate a y-polarized ordinary optical wave. The velocity of this longitudinal acoustic wave is $v_a = 6.57$ km s^{-1}. The density of $LiNbO_3$ is 4.64×10^3 kg m^{-3}. The acousto-optic coupling of these two optical waves in this configuration is through the elasto-optic coefficient $p_{41} = -0.151$ of $LiNbO_3$, and the ordinary and extraordinary refractive indices are $n_o = 2.222$ and $n_e = 2.145$, respectively. The acoustic intensity is 10 W cm^{-2}, and the interaction length is 1 cm. If the acoustic frequency has to be kept below 5 GHz for practical reasons, in which directions do the optical waves propagate? Find the phase-matching frequency and the diffraction efficiency.

8.3.14 Compare Raman–Nath diffraction and Bragg diffraction regarding their basic principles, requirements, and characteristics. Explain why almost all practical acousto-optic devices operate in the Bragg regime.

8.3.15 Discuss the primary differences between nonbirefringent acousto-optic Bragg diffraction and birefringent acousto-optic Bragg diffraction.

8.3.16 In this problem, we consider Bragg diffraction in the limit of low diffraction efficiency.

a. Show that for both small-angle Bragg diffraction and collinear Bragg diffraction, the diffraction efficiency in the case of perfect phase matching is

$$\eta_{PM} \approx |\kappa|^2 l^2 = \frac{\pi^2 M_2 I_a}{2\lambda^2} l^2 \tag{8.169}$$

in the limit that $\eta_{PM} \ll 1$.

b. Show that in the situation where (8.169) is valid, the dependence of the diffraction efficiency on the phase-mismatch parameter for $|\Delta k| > |\kappa|$ can be described as

$$\eta \approx \eta_{PM} \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2} \tag{8.170}$$

for both small-angle and collinear diffraction.

c. What are the values of Δk for which $\eta = \eta_{PM}/2$?

8.4.1 Show that if the length L of an acousto-optic modulator is chosen to be the minimum limited by (8.118) when $a < 1$ or that limited by (8.121) when $a > 1$, the Rayleigh range z_R of the Gaussian optical beam is always larger than L so that the optical beam remains well collimated throughout the length of the modulator no matter whether the beam is focused or not. Therefore, the choice of a modulator height of $H = 2\sqrt{2}w_0$ is sufficient to cover the beam spot.

8.4.2 A fused silica traveling-wave acousto-optic modulator similar to the one described in Example 8.5 is desired for high-speed application with a risetime of $t_r = 10$ ns at $\lambda = 1.064$ μm optical wavelength. The acoustic wave is the longitudinal mode, and the optical wave is polarized in a direction perpendicular to the propagation directions of both optical and acoustic waves. Find the 3-dB modulation bandwidth f_m^{3dB}, the required carrier frequency f_0, and the optimum optical beam spot size w_0 for this device. Practical considerations for the construction and operation of this device require that the physical length L and height H of the transducer be larger than 0.5 mm. The transducer efficiency is $\eta_t = 80\%$. Design the device by properly choosing the values of L and H so that the device has a modulation loss of at least 5% for an electrical modulation power of 1 W.

8.4.3 Answer the questions in Example 8.5 for a fused silica traveling-wave acousto-optic modulator designed for an optical wavelength at $\lambda = 632.8$ nm instead of 1.064 μm. Compare the results with those found in Example 8.5 for 1.064 μm wavelength.

8.4.4 Use the data listed in Table 8.2 to answer the questions in Example 8.5 for a $PbMoO_4$ traveling-wave acousto-optic modulator that is designed for an optical wavelength at $\lambda = 1.064$ μm. The acoustic wave is a longitudinal mode propagating in the [001] direction, and the optical wave is an ordinary wave polarized in a direction perpendicular to the propagation directions of the optical and acoustic waves. The ordinary and extraordinary refractive indices of $PbMoO_4$ at 1.064 μm are $n_o = 2.298$ and $n_e = 2.200$, respectively. Compare the results with those found in Example 8.5 for the fused silica modulator.

8.4.5 A silica standing-wave acousto-optic modulator is used as a mode locker for a Nd : YAG laser at 1.064 μm wavelength to provide a periodic loss modulation at 100 MHz in tune with the 100 MHz longitudinal mode spacing of the laser so that the longitudinal modes of the laser are locked together in phase. At one end, the acousto-optic cell of the modulator is attached to a matched transducer that has dimensions $L = 4$ cm and $H = 1$ cm. The width of the cell is $W = 4$ cm. For fused silica, $n = 1.45$ and $M_2 = 1.46 \times 10^{-15}$ m^2 W^{-1} at $\lambda = 1.064$ μm, and $v_a = 5.97$ km s^{-1}.

a. What should the acoustic frequency f used to drive this modulator be?

b. Does the device function in the Bragg regime with clean separation between diffracted and undiffracted beams?

c. The acoustic decay rate γ_a of the resonant cell is contributed by the losses associated with absorption inside the cell and acoustic reflection at the far end of the cell. The absorption loss rate is quadratically proportional to the acoustic frequency and is given by $\gamma_{ab} = \gamma_0 f^2$ with $\gamma_0 = 7.2$ dB μs^{-1} GHz^{-2} for silica. The reflection loss rate is given by $\gamma_r = -(v_a/2W) \ln R$, where R is the reflectance of the acoustic wave. For this device, $R = 0.8$. Find the value of γ_a.

d. If the transducer conversion efficiency is $\eta_t = 70\%$, what is the peak modulation loss for an electrical modulation power of $P_e = 1$ W?

8.4.6 Compare traveling-wave and standing-wave acousto-optic modulators. Discuss their differences in device structures, operation principles, performance characteristics, and applications.

8.5.1 In this problem, we consider the phase mismatch $\Delta\mathbf{k} = \Delta k\hat{z}$ in the case of small-angle Bragg diffraction with fixed values of k_i and k_d. With perfect phase matching, the angles of incidence and diffraction are θ_i and θ_d, respectively, and the acoustic wavevector is $\mathbf{K} = K\hat{x}$. It is assumed that $\theta_i \neq 0$ and $\theta_d \neq 0$ in this problem.

a. For a fixed value of K, a deviation of the angle of incidence from the value θ_i required by the phase-matching condition results in a phase mismatch. If the angular deviation of the incident wave is $\delta\theta_i \ll \theta_i$, corresponding to an angle of incidence of $\theta_i' = \theta_i + \delta\theta_i$, show that the angle of diffraction is changed to $\theta_d' = \theta_d + \delta\theta_d$ and

$$\delta\theta_d = \delta\theta_i. \tag{8.171}$$

b. Show that in the situation described in (a), the phase mismatch in the z direction is

$$\Delta k = \mp K\delta\theta_i, \tag{8.172}$$

where the minus sign is for up-shifted diffraction, and the plus sign is for down-shifted diffraction.

c. If the incident angle is fixed at θ_i but the value of K is changed to $K' = K + \delta K$, the angle of diffraction also deviates from θ_d. Show that, to first order,

$$\delta\theta_d = \pm\frac{\delta K}{k_i \cos\theta_i} \approx \pm\frac{\delta K}{k_i}, \tag{8.173}$$

where the plus sign is for up-shifted diffraction, and the minus sign is for down-shifted diffraction.

d. Show that in the situation described in (c), the phase mismatch in the z direction, for either up- or down-shifted diffraction, is, to first order,

$$\Delta k = -\frac{k_d^2 + K^2 - k_i^2}{2k_i K \cos\theta_i}\delta K \approx -\frac{k_d^2 + K^2 - k_i^2}{2k_i K}\delta K. \tag{8.174}$$

e. The phase mismatch caused by a deviation in the incident angle can be compensated by a change in the value of K, and vice versa. Show that in order to maintain phase matching, the amount of δK needed to compensate for the phase mismatch caused by an incident-angle deviation of $\delta\theta_i$ is

$$\delta K = \mp \frac{2k_i K^2 \cos\theta_i}{k_d^2 + K^2 - k_i^2}\delta\theta_i \approx \mp \frac{2k_i K^2}{k_d^2 + K^2 - k_i^2}\delta\theta_i, \tag{8.175}$$

where the minus sign is for up-shifted diffraction, and the plus sign is for down-shifted diffraction.

f. Show that when phase matching is maintained by compensation between $\delta\theta_i$ and δK discussed in (e), the angle of diffraction is changed by an amount of

$$\delta\theta_d = -\frac{k_i^2 + K^2 - k_d^2}{k_d^2 + K^2 - k_i^2}\delta\theta_i \tag{8.176}$$

for either up- or down-shifted diffraction.

8.5.2 Answer the questions in Problem 8.5.1 for acousto-optic Bragg diffraction in an isotropic medium.

8.5.3 In this problem, we consider the optimum bandwidth Δf and transducer length L of a nonbirefringent deflector. Their optimum values depend on the criterion chosen for the acceptable efficiency reduction at the band edges due to phase mismatch. Consider the low-efficiency limit so that the efficiency as a function of phase mismatch can be approximated by (8.170) in Problem 8.3.16.

a. Show that when the criterion is chosen as 3-dB efficiency reduction at the band edges, as is taken to obtain (8.138), the optimum values of Δf and L are those given in (8.141).

b. Find the optimum values of Δf and L for the criteria chosen based on maximum efficiency reductions of 1, 2, 4, and 5 dB, respectively.

8.5.4 Tangential phase matching for Bragg diffraction in the case when $k_i > k_d$ is considered. In the case of perfect phase matching, the acoustic wavevector is $\mathbf{K} = K\hat{x}$ and $\theta_d = 0$. A deviation from the phase-matching condition results in a phase mismatch of $\Delta\mathbf{k} = \Delta k\hat{z}$.

a. If the value of K is fixed, show that a deviation of the angle of incidence from θ_i to $\theta_i' = \theta_i + \delta\theta_i$ results in a change of the angle of diffraction from $\theta_d = 0$ to $\theta_d = \delta\theta_d = \delta\theta_i$ and a phase mismatch $\Delta k = \mp K\delta\theta_i$. These results are the same as those obtained in (8.171) and (8.172) for the general phase-matching condition.

b. The incident angle is fixed at θ_i, but the value of K is changed to $K' = K + \delta K$. Show that in this case the change in the angle of diffraction is $\delta\theta_d = \pm\delta K/k_d$, which is the same as that in (8.173) for the general

phase-matching condition. However, show also that the phase mismatch is

$$\Delta k = [k_{\mathrm{d}}^2 - (\delta K)^2]^{1/2} - k_{\mathrm{d}} \approx -\frac{(\delta K)^2}{2k_{\mathrm{d}}}, \tag{8.177}$$

which is different from that in (8.174) for the general phase-matching condition.

8.5.5 A large bandwidth can be obtained for an acousto-optic deflector using the optimum birefringent phase-matching scheme shown in Fig. 8.14. The value of K for the acoustic wave ranges from $K_0 - \Delta K/2$ to $K_0 + \Delta K/2$, corresponding to the range of frequencies from $f_0 - \Delta f/2$ to $f_0 + \Delta f/2$. In this scheme, perfect phase matching does not occur at the tangential point, or at the center frequency f_0. Instead, perfect phase matching occurs at two different frequencies, which are between the center frequency and the two edges of the bandwidth, respectively. The maximum absolute value of the phase mismatch over the entire bandwidth is minimized if the phase mismatch at center frequency is equal in magnitude but opposite in sign to the phase mismatch at the two edges of the bandwidth. For simplicity, we consider only the case of up-shifted diffraction.

a. Show that in order to minimize the overall phase mismatch according to the discussion above, the incident angle should be chosen to be

$$\theta_{\mathrm{i}} = -\sin^{-1}\frac{K_0}{k_{\mathrm{i}}} = -\sin^{-1}\frac{\lambda f_0}{n_{\mathrm{i}} v_{\mathrm{a}}}, \tag{8.178}$$

so that $\theta_{\mathrm{d}} = 0$ at the center frequency.

b. Show that the phase mismatch at the center frequency is

$$\Delta k = \frac{k_{\mathrm{d}}}{2} - \left[\frac{k_{\mathrm{d}}^2}{4} - \frac{(\Delta K^2)}{16}\right]^{1/2} \approx \frac{(\Delta K)^2}{16k_{\mathrm{d}}} \tag{8.179}$$

if it is made equal in magnitude and opposite in sign to that at the edges of the bandwidth.

c. Show that to satisfy the requirement in (b), the K value corresponding to the center frequency is given by

$$K_0^2 = k_{\mathrm{i}}^2 - (k_{\mathrm{d}} - \Delta k)^2 \approx k_{\mathrm{i}}^2 - k_{\mathrm{d}}^2 + \frac{(\Delta K)^2}{8}. \tag{8.180}$$

d. Show that perfect phase matching occurs at the following two frequencies:

$$f_{\mathrm{PM}} \approx f_0 \pm \frac{\Delta f}{2\sqrt{2}}. \tag{8.181}$$

e. Show that perfect phase matching occurs at the following two angles of diffraction:

$$\theta_{\mathrm{d}}^{\mathrm{PM}} \approx \pm \sin^{-1}\frac{\Delta K}{2\sqrt{2}k_{\mathrm{d}}} = \pm\sin^{-1}\frac{\lambda \Delta f}{2\sqrt{2}n_{\mathrm{d}} v_{\mathrm{a}}}. \tag{8.182}$$

8.5.6 Show that the bandwidth of a birefringent deflector can be chosen to be the maximum value $\Delta f = 2f_0/3$ allowed by (8.134) while at the same time the transducer length can be as large as that given in (8.144) for the tangential phase-matching scheme. Show also that it can be further doubled to that given in (8.147) for the optimum birefringent phase-matching scheme.

8.5.7 The phased-array transducer shown in Fig. 8.15 steers the direction of the **K** vector of the combined acoustic beam through the effect of constructive interference. In this problem, a nonbirefringent acousto-optic deflector using such a phased-array transducer to increase its bandwidth is considered. For simplicity, consider only up-shifted diffraction.

a. Show that the direction of the acoustic beam is steered by the acoustic frequency according to (8.148). Show also that the tip of the **K** vector tracks a line normal to the transducer as the acoustic frequency is varied.

b. Show that if the phase mismatch at the center frequency is made equal in magnitude but opposite in sign to the phase mismatch at the two edges of the bandwidth, the maximum absolute value of the phase mismatch over the entire bandwidth is minimized and has the following value:

$$\Delta k \approx \frac{(\Delta K)^2}{16k}. \tag{8.183}$$

c. Show that in order to accomplish the requirement in (b), the center-to-center distance between adjacent transducer elements has to be

$$d_t \approx \frac{n v_a^2}{\lambda[f_0^2 - (\Delta f)^2/8]}, \tag{8.184}$$

and the incident angle of the optical beam has to be given by

$$\sin\theta_i \approx -\frac{\lambda f_0}{n v_a}\left(1 - \frac{\lambda^2 f_0^2}{8n^2 v_a^2}\right). \tag{8.185}$$

d. Show that perfect phase matching occurs at the two acoustic frequencies of $f_{PM} = f_0 \pm \Delta f/2\sqrt{2}$ for the two corresponding angles of diffraction at $\theta_d^{PM} \approx \pm\sin^{-1}(\Delta K/2\sqrt{2}k)$.

8.5.8 A phased-array transducer is used to optimize the efficiency for the nonbirefringent $LiNbO_3$ deflector described in Example 8.7.

a. What is the optimum center acoustic frequency if it is required that $\Delta f = 1$ GHz? Find the optimum length L of the phased-array transducer and the optimum center-to-center distance d_t between neighboring transducer elements. How many transducer elements are required? What is the minimum transducer height H?

b. What is the required incident angle if up-shifted diffraction is considered? How does it compare with that for the nonbirefringent deflector described in Example 8.7 using a monolithic transducer?

c. Find the peak diffraction efficiency of this device at an acoustic power level of 1 W. At what frequencies does this peak efficiency occur? What are the corresponding diffraction angles?
d. What is the deflection angular range $\Delta\theta_d$?

8.5.9 A nonbirefringent acousto-optic deflector made of fused silica is designed for 632.8 nm optical wavelength. The longitudinal mode of an acoustic wave at a center frequency of $f_0 = 500$ MHz with $v_a = 5.97$ km s^{-1} is used. The time aperture τ_a is limited by the loss of the acoustic wave due to absorption. The absorption loss rate is quadratically proportional to the acoustic frequency and is given by $\gamma_{ab} = \gamma_0 f^2$ with $\gamma_0 = 7.2$ dB μs^{-1} GHz^{-2} for silica. It is desired that the maximum loss of the acoustic power at the far end of the device across the optical beam is less than 1 dB over the entire acoustic frequency range. The optical wave is an optimally shaped elliptical Gaussian beam that is polarized in a direction perpendicular to the **K** vector of the acoustic wave so that $M_2 = 1.5 \times 10^{-15}$ m^2 W^{-1}. The refractive index is $n = 1.457$ at $\lambda = 632.8$ nm.
a. Find the maximum values of the bandwidth Δf and the time aperture τ_a for this device. What is the largest number N of resolvable spots for this device?
b. What are the dimensions of the elliptical beam spot?
c. What are the optimum values of the transducer dimensions L and H?
d. If up-shifted diffraction is considered, what is the required incident angle of the optical beam? What is the deflection angular range $\Delta\theta_d$?
e. What is the peak diffraction efficiency for an acoustic power of 1 W? At what acoustic frequency and diffraction angle does it occur?

8.5.10 A phased-array transducer is used to optimize the efficiency for the nonbirefringent fused silica deflector described in Problem 8.5.9.
a. Find the maximum values of the bandwidth Δf and the time aperture τ_a for this device. What is the largest number N of resolvable spots for this device?
b. What are the dimensions of the elliptical beam spot?
c. Find the optimum length L of the phased-array transducer and the optimum center-to-center distance d_t between the neighboring transducer elements. How many transducer elements are required? What is the minimum transducer height H?
d. If up-shifted diffraction is considered, what is the required incident angle of the optical beam? What is the deflection angular range $\Delta\theta_d$?
e. What is the peak diffraction efficiency for an acoustic power of 1 W? At what acoustic frequencies and diffraction angles does it occur?

8.6.1 Show that in collinear, codirectional Bragg diffraction, phase mismatch as a function of the wavelength and angular deviations of the incident optical beam is that given in (8.150) when both the frequency and the propagation direction of the acoustic wave are fixed.

8.6.2 The $LiNbO_3$ acousto-optic tunable filter described in Example 8.9 is used for 650 nm optical wavelength. At this wavelength, $n_o = 2.284$ and $n_e = 2.198$. The interaction length is fixed at $l = 1.52$ cm as that in Example 8.9. Find the required acoustic frequency for selecting this optical wavelength. What are the resolving power and the spectral linewidth? What are the angular apertures? Compare the results with those found in Example 8.9 for 1.3 μm wavelength.

8.6.3 It is possible to make a collinear $LiNbO_3$ acousto-optic tunable filter with a longitudinal acoustic mode that propagates in the [100] x direction. The acoustic velocity of this mode is $v_a = 6.57$ km s^{-1}. The birefringent collinear interaction with $\mathbf{k}_i \parallel \mathbf{k}_d \parallel \mathbf{K} \parallel \hat{x}$ then couples the ordinary wave polarized with $\hat{e}_o = \hat{y}$ with the extraordinary wave polarized with $\hat{e}_e = \hat{z}$ through the acousto-optic tensor elements $\Delta\epsilon_{yz} = \Delta\epsilon_{zy}$ (see Problem 8.2.5(b)). This filter is used to select an ordinary incident wave spectrally at $\lambda = 1.3$ μm with the same interaction length of $l = 1.52$ cm found in Example 8.9. Find the required acoustic frequency for selecting this optical wavelength. What are the resolving power and the spectral linewidth? What are the angular apertures? Compare the results with those found in Example 8.9 that uses a different acoustic mode.

8.6.4 $LiNbO_3$ has a transparency range between 400 nm and 4 μm. Use the ordinary and extraordinary indices of $LiNbO_3$ as a function of optical wavelength found in (1.190) and (1.191), respectively, to plot the tuning curves of the two collinear $LiNbO_3$ tunable filters showing the required acoustic frequency as a function of optical wavelength in the transparency range of $LiNbO_3$. Also plot the chromatic resolving power as a function of optical wavelength for an interaction length of $l = 1$ cm for both devices.

8.7.1 An optical waveguide is fabricated on a substrate of an isotropic medium. A SAW for acousto-optic interaction with the guided optical waves propagates in the z direction on the surface normal to the y direction as shown in Fig. 8.17.

a. Find the acousto-optic permittivity tensor $\Delta\boldsymbol{\epsilon}(\mathbf{r}, t)$ as a function of the nonvanishing elasto-optic coefficients given in Table 8.1 and the spatially and temporally dependent strain tensor elements associated with the SAW.

b. Under what configurations can acousto-optic interaction of the guided optical waves with this SAW lead to coupling between TE and TM modes of the waveguide?

c. With this SAW, is it possible to make a collinear guided-wave acousto-optic tunable filter or mode converter that involves mode conversion between TE and TM modes? Explain.

8.7.2 What modifications have to be made to the answers to the questions in Problem 8.7.1 if the substrate is instead a cubic crystal with the y and z directions aligned with two of its axes?

8.7.3 Answer the questions in Problem 8.7.1 for a $LiNbO_3$ waveguide with a z-propagating SAW in a y-cut $LiNbO_3$ crystal.

8.7.4 Is it possible to make a collinear guided-wave acousto-optic TE–TM mode converter in a z-cut, y-propagating $LiNbO_3$ waveguide? Is it possible in a y-cut, x-propagating $LiNbO_3$ waveguide?

8.7.5 The fractional bandwidth of a guided-wave acousto-optic deflector using a regular IDT is usually limited to a maximum value of 0.4. Assume that the bandwidth of the driving source is made large enough by optimizing the parameters of the electronic circuit so that the IDT bandwidth is limited primarily by the frequency dependence of the radiation efficiency of the IDT. This IDT is used to generate a z-propagating SAW in a y-cut $LiNbO_3$ waveguide. The velocity of this SAW is $v_a = 3.488$ km s^{-1}. The value of $M_2\Gamma_{id}^2$, including the contributions of both photoelastic and electro-optic effects, for this device operating at $\lambda = 1.3$ μm and a center acoustic frequency of 1 GHz is estimated to be about 2×10^{-14} m^2 W^{-1}. The optical waveguide is weakly guiding so that the refractive indices of the guided optical waves can be approximated by those of the bulk material with $n_o = 2.222$ and $n_e = 2.145$.

a. Find the number of electrodes that corresponds to a 3-dB fractional bandwidth of 0.4 for the IDT. Find the peak efficiency.

b. What are the synchronous frequency of the IDT and the center-to-center distance between two adjacent electrodes in the IDT?

c. To maximize the interaction length, the phase-matching bandwidth of this deflector is chosen to be the same as the IDT bandwidth. If this deflector is used as a nonbirefringent deflector for the TE_0 mode of the waveguide, what are the upper and lower limits of the length L allowed for the electrodes?

d. What is the maximum peak deflection efficiency for $P_e = 10$ mW when the value of L is properly chosen? What is the value of L that gives this maximum efficiency?

e. With the optimum value for L chosen in (d), what is the electrical power P_e needed to drive the IDT for the deflector to have a peak efficiency of 50%?

8.7.6 Compare the characteristics and advantages of different types of IDT that are designed and used for guided-wave acousto-optic deflectors.

SELECT BIBLIOGRAPHY

Banerjee, P. P. and Poon, T. C., *Principles of Applied Optics*. Homewood, IL: Aksen Associates, 1991.

Goutzoulis, A. P. and Pape, D. R., eds., *Design and Fabrication of Acousto-Optic Devices*. New York: Marcel Dekker, 1994.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Iizuka, K., *Elements of Photonics in Free Space and Special Media*, Vol. I. New York: Wiley, 2002.

Kingslake, R. and Thompson, B. J., eds., *Applied Optics and Optical Engineering*, Vol. VI. New York: Academic Press, 1980.

Korpel, A., *Acousto-Optics*, 2nd edn. New York: Marcel Dekker, 1997.

Magdich, L. N. and Molchanov, V. Ya., *Acoustooptic Devices and Their Applications*. New York: Gordon and Breach Science Publishers, 1989.

Nishihara, H., Haruna, M. and Suhara, T., *Optical Integrated Circuits*. New York: McGraw-Hill, 1989.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Sapriel, J., *Acousto-Optics*. New York: Wiley, 1979.

Syms, R. and Cozens, J., *Optical Guided Waves and Devices*. London: McGraw-Hill, 1992.

Tsai, C. S., ed., *Guided-Wave Acousto-Optics: Interactions, Devices, and Applications*. New York: Springer-Verlag, 1990.

Xu, J. and Stroud, R., *Acousto-Optic Devices: Principles, Design, and Applications*. New York: Wiley, 1992.

Yariv, A. and Yeh, P., *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: Wiley, 1984.

ADVANCED READING LIST

Campbell, C. K., "SAW devices satisfy varied wireless system needs," *Applied Microwave and Wireless* **14**(3): 24–36, Mar. 2002.

Collins, J. H., "A short history of microwave acoustics," *IEEE Transactions on Microwave Theory and Techniques* **MTT-32**(9): 1127–1140, Sep. 1984.

Gaylord, T. K. and Moharam, M. G., "Analysis and applications of optical diffraction by gratings," *Proceedings of the IEEE* **73**(5): 894–937, May 1985.

Korpel, A., "Acousto-optics: a review of fundamentals," *Proceedings of the IEEE* **69**(1): 48–53, Jan. 1981.

Morgan, D. P., "A history of surface acoustic wave devices," *International Journal of High Speed Electronics and Systems* **10**(3): 553–602, Sep. 2000.

Sliwinski, A., "Acousto-optics and its perspectives in research and applications," *Ultrasonics* **28**(4): 195–213, July 1990.

9 Nonlinear optical devices

The functioning of electro-optic, magneto-optic, and acousto-optic devices discussed in earlier chapters is based on the fact that the optical properties of a material depend on the strength of an electric, magnetic, or acoustic field that is present in an optical medium. At a sufficiently high optical intensity, the optical properties of a material also become a function of the optical field. Such nonlinear response to the strength of the optical field results in various nonlinear optical effects. Nonlinear optics is an established broad field with applications covering a very wide range. The most important nonlinear optical devices are optical frequency converters. The frequency-converting function of such devices is uniquely nonlinear and is difficult, if not impossible, to accomplish by other means in the absence of optical nonlinearity. Other unique nonlinear optical devices include all-optical switches and modulators. Many interesting nonlinear optical phenomena, such as optical solitons, stimulated Raman scattering, and optical phase conjugation, also find useful applications.

9.1 Optical nonlinearity

The origin of optical nonlinearity is the nonlinear response of electrons in a material to an optical field as the strength of the field is increased. Macroscopically, the nonlinear optical response of a material is described by a polarization that is a nonlinear function of the optical field. In general, such nonlinear dependence on the optical field can take a variety of forms. In particular, it can be very complicated when the optical field becomes extremely strong.

In the situations of most nonlinear optical devices of interest, with the exception of saturable absorbers, the perturbation method can be applied to expand the total optical polarization in terms of a series of linear and nonlinear polarizations:

$$\boldsymbol{P}(\mathbf{r}, t) = \boldsymbol{P}^{(1)}(\mathbf{r}, t) + \boldsymbol{P}^{(2)}(\mathbf{r}, t) + \boldsymbol{P}^{(3)}(\mathbf{r}, t) + \cdots, \tag{9.1}$$

where $\boldsymbol{P}^{(1)}$ is the linear polarization and $\boldsymbol{P}^{(2)}$ and $\boldsymbol{P}^{(3)}$ are the second- and third-order nonlinear polarizations, respectively. Except in some special cases, nonlinear polarizations of fourth order and beyond are usually not important and thus can be ignored.

Linear polarization $\boldsymbol{P}^{(1)}$ is a linear function of the optical field, whereas nonlinear polarizations $\boldsymbol{P}^{(2)}$ and $\boldsymbol{P}^{(3)}$ are, respectively, quadratic and cubic functions of the optical field:

$$\boldsymbol{P}^{(1)}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}' \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}^{(1)}(\mathbf{r}-\mathbf{r}', t-t') \cdot \boldsymbol{E}(\mathbf{r}', t'), \tag{9.2}$$

$$\begin{aligned}&\boldsymbol{P}^{(2)}(\mathbf{r}, t)\\ &= \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}_1 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}_2 \int_{-\infty}^{t} \mathrm{d}t_1 \int_{-\infty}^{t} \mathrm{d}t_2 \boldsymbol{\chi}^{(2)}(\mathbf{r}-\mathbf{r}_1, t-t_1; \mathbf{r}-\mathbf{r}_2, t-t_2) : \boldsymbol{E}(\mathbf{r}_1, t_1)\boldsymbol{E}(\mathbf{r}_2, t_2),\end{aligned} \tag{9.3}$$

$$\begin{aligned}&\boldsymbol{P}^{(3)}(\mathbf{r}, t)\\ &= \epsilon_0 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}_1 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}_2 \int_{-\infty}^{\infty} \mathrm{d}\mathbf{r}_3 \int_{-\infty}^{t} \mathrm{d}t_1 \int_{-\infty}^{t} \mathrm{d}t_2 \int_{-\infty}^{t} \mathrm{d}t_3 \boldsymbol{\chi}^{(3)}(\mathbf{r}-\mathbf{r}_1, t-t_1; \mathbf{r}-\mathbf{r}_2, t-t_2; \mathbf{r}-\mathbf{r}_3, t-t_3)\\ &\qquad \vdots \boldsymbol{E}(\mathbf{r}_1, t_1)\boldsymbol{E}(\mathbf{r}_2, t_2)\boldsymbol{E}(\mathbf{r}_3, t_3),\end{aligned} \tag{9.4}$$

where $\boldsymbol{\chi}^{(1)}$ is the linear susceptibility and $\boldsymbol{\chi}^{(2)}$ and $\boldsymbol{\chi}^{(3)}$ are the second- and third-order *nonlinear susceptibilities*, respectively.[1] The linear susceptibility is that of linear optics discussed in Chapter 1. In general, $\boldsymbol{\chi}^{(1)}$ is a second-rank tensor, $\boldsymbol{\chi}^{(1)} = \left[\chi_{ij}^{(1)}\right]$, as expressed in (1.105), whereas $\boldsymbol{\chi}^{(2)}$ and $\boldsymbol{\chi}^{(3)}$ are, respectively, third- and fourth-rank tensors:

$$\boldsymbol{\chi}^{(2)} = \left[\chi_{ijk}^{(2)}\right] \tag{9.5}$$

and

$$\boldsymbol{\chi}^{(3)} = \left[\chi_{ijkl}^{(3)}\right]. \tag{9.6}$$

The nonlinear susceptibilities, $\boldsymbol{\chi}^{(2)}$ and $\boldsymbol{\chi}^{(3)}$, characterize the nonlinear optical properties of a material. Thus, the relations in (9.3) and (9.4) define the nonlinear polarizations that describe the nonlinear responses of a material to an optical field. Because of the generally anisotropic nature of nonlinear susceptibility tensors, the nonlinear polarizations $\boldsymbol{P}^{(2)}$ and $\boldsymbol{P}^{(3)}$ are expressed in the form of high-order products between the nonlinear susceptibilities and the optical field.

As discussed in Section 1.1, the response of a material to an optical field can be nonlocal in space and noninstantaneous in time. This statement is true for both linear and nonlinear responses. In consideration of this fact, the linear polarization $\boldsymbol{P}^{(1)}$ defined in (9.2) and the nonlinear polarizations $\boldsymbol{P}^{(2)}$ and $\boldsymbol{P}^{(3)}$ defined in (9.3) and (9.4) are generally expressed in the form of convolution integrals over both space and time. For

[1] In defining polarizations and susceptibilities in the form of (9.2)–(9.4), we consider only the electric-dipole contribution to the material response under the *electric-dipole approximation* by neglecting the contributions from magnetic dipoles, electric quadrupoles, and other multipoles to the material response. When the electric-dipole contribution vanishes, other contributions can be important.

material responses that are local in space but not necessarily instantaneous in time, the susceptibilities can be expressed as

$$\boldsymbol{\chi}^{(n)}(\mathbf{r}-\mathbf{r}_1, t-t_1; \mathbf{r}-\mathbf{r}_2, t-t_2; \ldots) = \boldsymbol{\chi}^{(n)}(t-t_1, t-t_2, \ldots)\delta(\mathbf{r}-\mathbf{r}_1)\delta(\mathbf{r}-\mathbf{r}_2)\ldots \tag{9.7}$$

Then, the linear and nonlinear polarizations can be expressed as

$$\boldsymbol{P}^{(1)}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{t} \mathrm{d}t' \boldsymbol{\chi}^{(1)}(t-t') \cdot \boldsymbol{E}(\mathbf{r}, t'), \tag{9.8}$$

$$\boldsymbol{P}^{(2)}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{t} \mathrm{d}t_1 \int_{-\infty}^{t} \mathrm{d}t_2 \boldsymbol{\chi}^{(2)}(t-t_1, t-t_2) : \boldsymbol{E}(\mathbf{r}, t_1)\boldsymbol{E}(\mathbf{r}, t_2), \tag{9.9}$$

$$\boldsymbol{P}^{(3)}(\mathbf{r}, t) = \epsilon_0 \int_{-\infty}^{t} \mathrm{d}t_1 \int_{-\infty}^{t} \mathrm{d}t_2 \int_{-\infty}^{t} \mathrm{d}t_3 \boldsymbol{\chi}^{(3)}(t-t_1, t-t_2, t-t_3) \vdots \boldsymbol{E}(\mathbf{r}, t_1)\boldsymbol{E}(\mathbf{r}, t_2)\boldsymbol{E}(\mathbf{r}, t_3). \tag{9.10}$$

In the momentum space and frequency domain, spatially local but temporally noninstantaneous responses imply that the linear and nonlinear susceptibilities are functions of optical frequencies but are independent of optical wavevectors:

$$\boldsymbol{\chi}^{(n)}(\omega_1, \omega_2, \ldots, \omega_n) = \int_0^{\infty} \mathrm{d}t_1 \int_0^{\infty} \mathrm{d}t_2 \ldots \int_0^{\infty} \mathrm{d}t_n \boldsymbol{\chi}^{(n)}(t_1, t_2, \ldots, t_n) \mathrm{e}^{\mathrm{i}\omega_1 t_1 + \mathrm{i}\omega_2 t_2 + \cdots + \mathrm{i}\omega_n t_n}. \tag{9.11}$$

This situation applies to the interactions discussed in this chapter. Therefore, the discussions in this chapter are restricted to spatially local interactions where the linear and nonlinear polarizations can be expressed in the form of (9.8)–(9.10) and the linear and nonlinear susceptibilities in the momentum space and frequency domain are functions of optical frequencies only.

The polarizations defined in (9.8)–(9.10) above are expressed in terms of real field quantities, just as any basic definitions of electromagnetic field quantities are. However, as we have seen throughout the preceding chapters, it is generally convenient to deal with optical fields in terms of complex field quantities because optical fields are harmonic fields. As seen in Section 1.2, conversion to expressions in terms of complex field quantities is quite straightforward. Maxwell's equations and the wave equation all retain their general form after the conversion. The complex field $\mathbf{E}(\mathbf{r}, t)$ is defined in (1.39) through its relation to the real field $\boldsymbol{E}(\mathbf{r}, t)$ as

$$\boldsymbol{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \text{c.c.} \tag{9.12}$$

In line with this definition, a complex nonlinear polarization, $\mathbf{P}^{(n)}(\mathbf{r}, t)$, in the real space and time domain can be defined through its relation to the real nonlinear

polarization, $\boldsymbol{P}^{(n)}(\mathbf{r}, t)$, as

$$\boldsymbol{P}^{(n)}(\mathbf{r}, t) = \mathbf{P}^{(n)}(\mathbf{r}, t) + \mathbf{P}^{(n)*}(\mathbf{r}, t) = \mathbf{P}^{(n)}(\mathbf{r}, t) + \text{c.c.} \tag{9.13}$$

Note that in our convention, $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{P}^{(n)}(\mathbf{r}, t)$ contain components that vary with time as $\exp(-\mathrm{i}\omega t)$ with positive values of ω, while $\mathbf{E}^*(\mathbf{r}, t)$ and $\mathbf{P}^{(n)*}(\mathbf{r}, t)$ contain those varying with time as $\exp(\mathrm{i}\omega t)$ with positive values of ω or, equivalently, $\exp(-\mathrm{i}\omega t)$ with ω assuming negative values. By substituting (9.12) and (9.13) in (9.8)–(9.10), expressions of complex polarizations in terms of complex fields can be obtained. The relation between the complex linear polarization $\mathbf{P}^{(1)}(\mathbf{r}, t)$ and the complex field $\mathbf{E}(\mathbf{r}, t)$ has the same form as (9.8), as shown in (1.45). However, the complex nonlinear polarizations $\mathbf{P}^{(2)}(\mathbf{r}, t)$ and $\mathbf{P}^{(3)}(\mathbf{r}, t)$ contain products of $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{E}^*(\mathbf{r}, t)$ in addition to those of $\mathbf{E}(\mathbf{r}, t)$ alone. Consequently, they have more complicated expressions than those of the real polarizations in (9.9) and (9.10).

The optical field involved in a nonlinear interaction usually contains multiple, distinct frequency components. Such a field can be expanded in terms of its frequency components as is done in (4.5):

$$\mathbf{E}(\mathbf{r}, t) = \sum_q \mathbf{E}_q(\mathbf{r}) \exp(-\mathrm{i}\omega_q t) = \sum_q \boldsymbol{\mathcal{E}}_q(\mathbf{r}) \exp(\mathrm{i}\mathbf{k}_q \cdot \mathbf{r} - \mathrm{i}\omega_q t), \tag{9.14}$$

where $\boldsymbol{\mathcal{E}}_q(\mathbf{r})$ is the slowly varying amplitude and $\mathbf{k}_q$ is the wavevector of the frequency component ω_q. The nonlinear polarizations also contain multiple frequency components and can be expanded as

$$\mathbf{P}^{(n)}(\mathbf{r}, t) = \sum_q \mathbf{P}_q^{(n)}(\mathbf{r}) \exp(-\mathrm{i}\omega_q t). \tag{9.15}$$

Note that we do not attempt to express $\mathbf{P}_q^{(n)}(\mathbf{r})$ further in terms of a slowly varying polarization amplitude multiplied by a fast varying spatial phase term as is done for $\mathbf{E}_q(\mathbf{r})$. The reason is that, as we shall see later, *the wavevector that characterizes the fast-varying spatial phase of a nonlinear polarization* $\mathbf{P}_q^{(n)}(\mathbf{r})$ *is not simply determined by the frequency* ω_q *but is determined by the fields involved in a particular nonlinear interaction of interest.*

In the discussions of nonlinear polarizations, we also use the notations $\mathbf{E}(\omega_q)$ and $\mathbf{P}^{(n)}(\omega_q)$ defined respectively as

$$\mathbf{E}(\omega_q) = \mathbf{E}_q(\mathbf{r}) \quad \text{and} \quad \mathbf{P}^{(n)}(\omega_q) = \mathbf{P}_q^{(n)}(\mathbf{r}). \tag{9.16}$$

Field and polarization components with negative frequencies are interpreted as

$$\mathbf{E}(-\omega_q) = \mathbf{E}^*(\omega_q) \quad \text{and} \quad \mathbf{P}^{(n)}(-\omega_q) = \mathbf{P}^{(n)*}(\omega_q). \tag{9.17}$$

The following notation for nonlinear susceptibilities is also used:

$$\boldsymbol{\chi}^{(n)}(\omega_q = \omega_1 + \omega_2 + \cdots + \omega_n) = \boldsymbol{\chi}^{(n)}(\omega_1, \omega_2, \ldots, \omega_n) \tag{9.18}$$

for $\omega_1 + \omega_2 + \cdots + \omega_n = \omega_q$, where $\boldsymbol{\chi}^{(n)}(\omega_1, \omega_2, \ldots, \omega_n)$ are the frequency-domain susceptibilities defined in (9.11).

Using the definitions of the complex fields and polarizations in (9.12) and (9.13) as well as their expansions in (9.14) and (9.15), we can obtain, by taking the Fourier transform on (9.9) and (9.10), the following relations:

$$\mathbf{P}^{(2)}(\omega_q) = \epsilon_0 \sum_{m,n} \boldsymbol{\chi}^{(2)}(\omega_q = \omega_m + \omega_n) : \mathbf{E}(\omega_m)\mathbf{E}(\omega_n) \tag{9.19}$$

and

$$\mathbf{P}^{(3)}(\omega_q) = \epsilon_0 \sum_{m,n,p} \boldsymbol{\chi}^{(3)}(\omega_q = \omega_m + \omega_n + \omega_p) \vdots \mathbf{E}(\omega_m)\mathbf{E}(\omega_n)\mathbf{E}(\omega_p). \tag{9.20}$$

The summation is performed over all *positive and negative* values of frequencies that, for a given ω_q, satisfy the constraint of $\omega_m + \omega_n = \omega_q$ in the case of (9.19) and the constraint of $\omega_m + \omega_n + \omega_p = \omega_q$ in the case of (9.20). More explicitly, by performing the summation over *positive frequencies only* and by expanding the product, we have

$$\begin{aligned} P_i^{(2)}(\omega_q) = \epsilon_0 \sum_{j,k} \sum_{\omega_m,\omega_n>0} \Big[& \chi_{ijk}^{(2)}(\omega_q = \omega_m + \omega_n) E_j(\omega_m) E_k(\omega_n) \\ & + \chi_{ijk}^{(2)}(\omega_q = \omega_m - \omega_n) E_j(\omega_m) E_k^*(\omega_n) \\ & + \chi_{ijk}^{(2)}(\omega_q = -\omega_m + \omega_n) E_j^*(\omega_m) E_k(\omega_n) \Big] \end{aligned} \tag{9.21}$$

and

$$\begin{aligned} P_i^{(3)}(\omega_q) = \epsilon_0 \sum_{j,k,l} \sum_{\omega_m,\omega_n,\omega_p>0} \Big[& \chi_{ijkl}^{(3)}(\omega_q = \omega_m + \omega_n + \omega_p) E_j(\omega_m) E_k(\omega_n) E_l(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = \omega_m + \omega_n - \omega_p) E_j(\omega_m) E_k(\omega_n) E_l^*(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = \omega_m - \omega_n + \omega_p) E_j(\omega_m) E_k^*(\omega_n) E_l(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = -\omega_m + \omega_n + \omega_p) E_j^*(\omega_m) E_k(\omega_n) E_l(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = \omega_m - \omega_n - \omega_p) E_j(\omega_m) E_k^*(\omega_n) E_l^*(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = -\omega_m + \omega_n - \omega_p) E_j^*(\omega_m) E_k(\omega_n) E_l^*(\omega_p) \\ & + \chi_{ijkl}^{(3)}(\omega_q = -\omega_m - \omega_n + \omega_p) E_j^*(\omega_m) E_k^*(\omega_n) E_l(\omega_p) \Big]. \end{aligned} \tag{9.22}$$

Usually only a limited number of frequencies participate in a given nonlinear optical interaction. Consequently, only one or a few terms among those listed in (9.21) or (9.22) contribute to a nonlinear polarization of interest.

EXAMPLE 9.1 Three optical fields at the wavelengths of $\lambda_1 = 750$ nm, $\lambda_2 = 600$ nm, and $\lambda_3 = 500$ nm, corresponding to the frequencies of $\omega_1 = 2\pi c/\lambda_1$, $\omega_2 = 2\pi c/\lambda_2$, and $\omega_3 = 2\pi c/\lambda_3$, respectively, are involved in second-order nonlinear optical interactions. Find the nonlinear polarization $\mathbf{P}^{(2)}$ at the frequency of $\omega_4 = 2\pi c/\lambda_4$, where $\lambda_4 = 300$ nm. If $\mathbf{E}(\omega_1) = E_1\hat{x}$, $\mathbf{E}(\omega_2) = E_2(\hat{y} + \hat{z})/\sqrt{2}$, and $\mathbf{E}(\omega_3) = E_3\hat{z}$, what is the x component of $\mathbf{P}^{(2)}(\omega_4)$?

Solution Because $\lambda_1^{-1} + \lambda_3^{-1} = 2\lambda_2^{-1} = \lambda_4^{-1}$, we find that $\omega_4 = \omega_1 + \omega_3 = \omega_2 + \omega_2$. Therefore, using (9.19), we find the following second-order nonlinear polarization at the frequency ω_4:

$$\mathbf{P}^{(2)}(\omega_4) = \epsilon_0 \left[\boldsymbol{\chi}^{(2)}(\omega_4 = \omega_1 + \omega_3) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_3) + \boldsymbol{\chi}^{(2)}(\omega_4 = \omega_3 + \omega_1) : \mathbf{E}(\omega_3)\mathbf{E}(\omega_1) + \boldsymbol{\chi}^{(2)}(\omega_4 = \omega_2 + \omega_2) : \mathbf{E}(\omega_2)\mathbf{E}(\omega_2)\right].$$

Note that there are two terms from the mixing of ω_1 and ω_3 because of permutation, but there is only one term from ω_2 mixing with itself. Using the given fields at the three frequencies, we can express the x component of $\mathbf{P}^{(2)}(\omega_4)$ as

$$\begin{aligned} P_x^{(2)}(\omega_4) &= \epsilon_0 \Big[\chi_{xxz}^{(2)}(\omega_4 = \omega_1 + \omega_3)E_1E_3 + \chi_{xzx}^{(2)}(\omega_4 = \omega_3 + \omega_1)E_3E_1 \\ &\quad + \chi_{xyz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2}{\sqrt{2}}\frac{E_2}{\sqrt{2}} + \chi_{xzy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2}{\sqrt{2}}\frac{E_2}{\sqrt{2}} \\ &\quad + \chi_{xyy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2}{\sqrt{2}}\frac{E_2}{\sqrt{2}} + \chi_{xzz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2}{\sqrt{2}}\frac{E_2}{\sqrt{2}}\Big] \\ &= \epsilon_0 \Big[\chi_{xxz}^{(2)}(\omega_4 = \omega_1 + \omega_3)E_1E_3 + \chi_{xzx}^{(2)}(\omega_4 = \omega_3 + \omega_1)E_3E_1 \\ &\quad + \chi_{xyz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} + \chi_{xzy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} \\ &\quad + \chi_{xyy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} + \chi_{xzz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2}\Big]. \end{aligned}$$

The other two components, $P_y^{(2)}(\omega_4)$ and $P_z^{(2)}(\omega_4)$, of $\mathbf{P}^{(2)}(\omega_4)$ can be explicitly spelled out by following a similar procedure.

9.2 Nonlinear optical susceptibilities

The nonlinear optical properties of a material are characterized by its nonlinear optical susceptibilities. In this section, the general properties of nonlinear optical susceptibilities are discussed.

It can be seen from (9.3) and (9.4) that the space- and time-dependent nonlinear susceptibilities $\boldsymbol{\chi}^{(n)}(\mathbf{r} - \mathbf{r}_1, t - t_1; \mathbf{r} - \mathbf{r}_2, t - t_2; \ldots; \mathbf{r} - \mathbf{r}_n, t - t_n)$ are real tensors because both $\boldsymbol{P}^{(n)}(\mathbf{r}, t)$ and $\boldsymbol{E}(\mathbf{r}, t)$ are real vectors. Though $\boldsymbol{\chi}^{(n)}(\mathbf{r}_1, t_1; \mathbf{r}_2, t_2; \ldots; \mathbf{r}_n, t_n)$ is always a real function of space and time, its Fourier transform is generally complex.

Therefore, the frequency-dependent nonlinear susceptibilities $\boldsymbol{\chi}^{(n)}(\omega_q = \omega_1 + \omega_2 + \cdots + \omega_n)$ defined in the frequency domain are generally complex. This characteristic is common to linear and nonlinear susceptibilities, as can be seen by reviewing the characteristics of the linear susceptibility discussed in Section 1.3. Also common to linear and nonlinear susceptibilities is the fact that *the imaginary part of a frequency-dependent susceptibility signifies the presence of loss or gain in a medium, meaning that there is a net exchange of energy between the optical field and the medium through the interaction described by this susceptibility.* The real part of a frequency-dependent susceptibility, irrespective of whether it is linear or nonlinear, does not cause a net energy exchange between the optical field and the medium.

The linear and nonlinear optical properties of a given material are not independent of each other. Indeed, there are close relations, at both microscopic and macroscopic levels, between the linear and nonlinear optical susceptibilities of the same material. Such connections and their implications can be seen in most of the questions asked in Problems 9.2.5–9.2.11. The reason for such connections is simply that both linear and nonlinear optical properties of a material have their roots in the same microscopic material properties, including the atomic compositions, the energy levels, the resonance frequencies, and the relaxation rates, as determine the optical responses of the material.

Reality condition

The reality condition discussed in Section 1.3 and expressed explicitly in (1.56) for linear susceptibility can be generalized for nonlinear susceptibilities. As mentioned above, nonlinear susceptibilities in the real space and time domain are real functions of space and time. This reality condition leads to the following relation for nonlinear susceptibilities in the momentum space and frequency domain:

$$\boldsymbol{\chi}^{(n)*}(\mathbf{k}_1, \omega_1; \mathbf{k}_2, \omega_2; \ldots; \mathbf{k}_n, \omega_n) = \boldsymbol{\chi}^{(n)}(-\mathbf{k}_1, -\omega_1; -\mathbf{k}_2, -\omega_2; \ldots; -\mathbf{k}_n, -\omega_n). \tag{9.23}$$

In the case of spatially local interaction when the relation in (9.7) is valid, we can use (9.18) to write the reality condition for nonlinear susceptibilities in the following form:

$$\boldsymbol{\chi}^{(n)*}(\omega_q = \omega_1 + \omega_2 + \cdots + \omega_n) = \boldsymbol{\chi}^{(n)}(-\omega_q = -\omega_1 - \omega_2 - \cdots - \omega_n). \tag{9.24}$$

Elements of susceptibility tensors

To gain a general perspective of the susceptibility tensor elements, we first review the properties of the linear susceptibility tensor $\boldsymbol{\chi}^{(1)}$. Because $\boldsymbol{\chi}^{(1)} = \left[\chi_{ij}^{(1)}\right]$ is a second-rank tensor, it consists of nine tensor elements, as shown explicitly in (1.105). Because the linear susceptibility is a function of a single frequency, only one frequency, ω, needs to be specified. When both $\boldsymbol{\chi}^{(1)}(\omega)$ and $\boldsymbol{\chi}^{(1)}(-\omega)$ are considered, the number of elements

doubles. The reality condition, by stating that the elements of $\boldsymbol{\chi}^{(1)}(-\omega)$ are completely determined by those of $\boldsymbol{\chi}^{(1)}(\omega)$, reduces the maximum number of independent elements back to nine. As discussed in Section 1.6, the linear susceptibility tensor $\boldsymbol{\chi}^{(1)}$ of a material can always be diagonalized, thus further reducing the nine tensor elements to only three diagonal elements that represent the eigenvalues of the tensor. Depending on the spatial symmetry of a material, the number of independent linear susceptibility elements needed for characterizing the linear optical properties of the material can be further reduced from three to two or one, as summarized in Table 1.2 in terms of the relations among the three principal refractive indices.

Similar concepts apply in consideration of the properties of nonlinear susceptibilities. However, the complexity increases dramatically due to the fact that the nonlinear susceptibilities are high-rank tensors and are functions of multiple frequencies. Being a third-rank tensor, $\boldsymbol{\chi}^{(2)} = \left[\chi^{(2)}_{ijk}\right]$ has 27 tensor elements. The fourth-rank tensor $\boldsymbol{\chi}^{(3)} = \left[\chi^{(3)}_{ijkl}\right]$ has 81 tensor elements. In the most general situation, three different frequencies are involved in a second-order nonlinear process characterized by $\boldsymbol{\chi}^{(2)}$. The three frequencies, say ω_1, ω_2, and ω_3, are not independent of one another but are subject to the condition: $\omega_3 = \omega_1 + \omega_2$, assuming that $\omega_3 > \omega_1, \omega_2$. For each tensor element $\chi^{(2)}_{ijk}$, there are 3! different permutations of the three frequencies, resulting in the following six different frequency dependencies:

$$\begin{aligned} &\chi^{(2)}_{ijk}(\omega_3 = \omega_1 + \omega_2), \quad \chi^{(2)}_{ijk}(\omega_2 = \omega_3 - \omega_1), \quad \chi^{(2)}_{ijk}(\omega_1 = -\omega_2 + \omega_3), \\ &\chi^{(2)}_{ijk}(\omega_3 = \omega_2 + \omega_1), \quad \chi^{(2)}_{ijk}(\omega_2 = -\omega_1 + \omega_3), \quad \chi^{(2)}_{ijk}(\omega_1 = \omega_3 - \omega_2). \end{aligned} \tag{9.25}$$

The sign of each frequency in every element in (9.25) can be changed simultaneously to have elements such as $\chi^{(2)}_{ijk}(-\omega_3 = -\omega_1 - \omega_2)$, and so on. Fortunately, because of the reality condition expressed in (9.24), this sign change does not result in additional susceptibility elements needed for describing a nonlinear process. Therefore, the total number of frequency-dependent $\boldsymbol{\chi}^{(2)}$ tensor elements needed to describe a second-order nonlinear interaction among three different optical frequencies completely is $27 \times 3! = 162$. For a third-order nonlinear process characterized by $\boldsymbol{\chi}^{(3)}$, there can in general be four different frequencies involved in the interaction. Therefore, the total number of frequency-dependent $\boldsymbol{\chi}^{(3)}$ tensor elements is $81 \times 4! = 1944$.

In most situations of practical interest, the number of independent elements of a nonlinear susceptibility tensor that has to be considered in a particular nonlinear interaction can be greatly reduced by applying the symmetry considerations discussed in the following.

Permutation symmetry

There is an *intrinsic permutation symmetry* that is purely a matter of convention of the notation used for frequency-dependent nonlinear susceptibilities. As an example,

we consider, in the case of $\omega_3 = \omega_1 + \omega_2$, a nonlinear polarization $P_x^{(2)}(\omega_3)$ generated by two orthogonally polarized optical field components $E_y(\omega_1)$ and $E_z(\omega_2)$ through a second-order nonlinear process. According to (9.21), we have

$$P_x^{(2)}(\omega_3) = \epsilon_0 \left[\chi_{xyz}^{(2)}(\omega_3 = \omega_1 + \omega_2)E_y(\omega_1)E_z(\omega_2) + \chi_{xzy}^{(2)}(\omega_3 = \omega_2 + \omega_1)E_z(\omega_2)E_y(\omega_1)\right]. \tag{9.26}$$

Both terms on the right-hand side of (9.26) are needed because of the convention used in (9.21) for expanding the product of (9.19). However, they are equal in magnitude because they represent the same physical process of nonlinear mixing of $E_y(\omega_1)$ and $E_z(\omega_2)$ to generate $P_x^{(2)}(\omega_3)$. Therefore, $\chi_{xyz}^{(2)}(\omega_3 = \omega_1 + \omega_2) = \chi_{xzy}^{(2)}(\omega_3 = \omega_2 + \omega_1)$. Generalization of this result leads to the following intrinsic permutation symmetry:

$$\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2) = \chi_{ikj}^{(2)}(\omega_3 = \omega_2 + \omega_1). \tag{9.27}$$

This intrinsic permutation symmetry permits free permutation of *only the frequencies on the right-hand side* of the equals sign in the argument of a nonlinear susceptibility if the corresponding Cartesian coordinate indices are also permuted simultaneously. It applies to the elements of $\boldsymbol{\chi}^{(3)}$ as well. It reduces the number of independent $\boldsymbol{\chi}^{(2)}$ elements from 162 to 81 and that of independent $\boldsymbol{\chi}^{(3)}$ elements from 1944 to 324 without imposing any qualifying physical conditions.

EXAMPLE 9.2 Simplify the expressions for $\mathbf{P}^{(2)}(\omega_4)$ and $P_x^{(2)}(\omega_4)$ in Example 9.1 by using the intrinsic permutation symmetry of $\boldsymbol{\chi}^{(2)}$. Write out the expressions for $P_y^{(2)}(\omega_4)$ and $P_z^{(2)}(\omega_4)$.

Solution The intrinsic permutation symmetry requires that $\boldsymbol{\chi}^{(2)}(\omega_4 = \omega_1 + \omega_3) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_3) = \boldsymbol{\chi}^{(2)}(\omega_4 = \omega_3 + \omega_1) : \mathbf{E}(\omega_3)\mathbf{E}(\omega_1)$. Therefore, the first two terms in $\mathbf{P}^{(2)}(\omega_4)$ can be combined to have the following expression:

$$\mathbf{P}^{(2)}(\omega_4) = \epsilon_0 \left[2\boldsymbol{\chi}^{(2)}(\omega_4 = \omega_1 + \omega_3) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_3) + \boldsymbol{\chi}^{(2)}(\omega_4 = \omega_2 + \omega_2) : \mathbf{E}(\omega_2)\mathbf{E}(\omega_2)\right].$$

By applying the intrinsic permutation symmetry explicitly to the elements of $\boldsymbol{\chi}^{(2)}$, we can use the relations $\chi_{xxz}^{(2)}(\omega_4 = \omega_1 + \omega_3) = \chi_{xzx}^{(2)}(\omega_4 = \omega_3 + \omega_1)$ and $\chi_{xyz}^{(2)}(\omega_4 = \omega_2 + \omega_2) = \chi_{xzy}^{(2)}(\omega_4 = \omega_2 + \omega_2)$ to express the x component of $\mathbf{P}^{(2)}(\omega_4)$ as follows:

$$P_x^{(2)}(\omega_4) = \epsilon_0 \left[2\chi_{xxz}^{(2)}(\omega_4 = \omega_1 + \omega_3)E_1E_3 + \chi_{xyz}^{(2)}(\omega_4 = \omega_2 + \omega_2)E_2^2 + \chi_{xyy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} + \chi_{xzz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2}\right].$$

The y and z components of $\mathbf{P}^{(2)}(\omega_4)$ are, respectively,

$$P_y^{(2)}(\omega_4) = \epsilon_0\Bigg[2\chi_{yxz}^{(2)}(\omega_4 = \omega_1 + \omega_3)E_1E_3 + \chi_{yyz}^{(2)}(\omega_4 = \omega_2 + \omega_2)E_2^2 + \chi_{yyy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} + \chi_{yzz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2}\Bigg]$$

and

$$P_z^{(2)}(\omega_4) = \epsilon_0\Bigg[2\chi_{zxz}^{(2)}(\omega_4 = \omega_1 + \omega_3)E_1E_3 + \chi_{zyz}^{(2)}(\omega_4 = \omega_2 + \omega_2)E_2^2 + \chi_{zyy}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2} + \chi_{zzz}^{(2)}(\omega_4 = \omega_2 + \omega_2)\frac{E_2^2}{2}\Bigg].$$

A *full permutation symmetry* exists when all of the frequencies contained in a susceptibility are far away from any resonance frequencies of a material so that the material causes no loss or gain to the optical field at those frequencies. Therefore, the full permutation symmetry is valid when the imaginary part of a susceptibility is negligibly small. It breaks down in a nonparametric process, where the imaginary part of the susceptibility is significant. *The full permutation symmetry allows all of the frequencies in a nonlinear susceptibility to be freely permuted if the Cartesian coordinate indices are also permuted accordingly.* It permits the interchange of the frequency on the left-hand side of the equals sign in the argument of a nonlinear susceptibility with any one on the right-hand side, which is not permitted by the intrinsic permutation symmetry. However, the sign of a frequency has to be changed at the time when it is moved across the equals sign in a permutation. For example, $\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2) = \chi_{jik}^{(2)}(-\omega_1 = -\omega_3 + \omega_2)$, and so on. By applying the reality condition given in (9.24) and the fact that the susceptibility is necessarily real when the full permutation symmetry is valid, we then have

$$\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2) = \chi_{jik}^{(2)}(\omega_1 = \omega_3 - \omega_2) = \chi_{kij}^{(2)}(\omega_2 = \omega_3 - \omega_1). \tag{9.28}$$

Similar relations can be written for the $\boldsymbol{\chi}^{(3)}$ elements. This full permutation symmetry further reduces the maximum number of independent $\boldsymbol{\chi}^{(2)}$ elements from 81 to 27 and that of independent $\boldsymbol{\chi}^{(3)}$ elements from 324 to 81.

If, in addition to being lossless so that the full permutation symmetry is valid, a medium is also nondispersive in the entire spectral range that covers all of the frequencies contained in a nonlinear susceptibility, the frequencies in the susceptibility can be freely permuted independently of the Cartesian coordinate indices. Similarly, the Cartesian coordinate indices can also be permuted independently of the frequencies. This permutation symmetry is known as *Kleiman's symmetry condition*. Under this condition, we have

$$\begin{aligned}\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2) &= \chi_{ijk}^{(2)}(\omega_1 = \omega_3 - \omega_2) = \chi_{ijk}^{(2)}(\omega_2 = \omega_3 - \omega_1) \\ &= \chi_{jik}^{(2)}(\omega_3 = \omega_1 + \omega_2) = \chi_{kij}^{(2)}(\omega_3 = \omega_1 + \omega_2),\end{aligned} \tag{9.29}$$

and so on. Kleiman's symmetry condition, when applicable, further reduces the number of independent $\chi^{(2)}$ elements from 27 to a maximum of 10 and that of independent $\chi^{(3)}$ elements from 81 to a maximum of 15.

Spatial symmetry

As we have seen in Section 1.6, the form of the linear susceptibility tensor is determined by the symmetry property of a material. The forms of the nonlinear susceptibility tensors of a material also reflect the spatial symmetry property of the material structure. As a result, some elements in a nonlinear susceptibility tensor may be zero and others may be related in one way or another, greatly reducing the total number of independent tensor elements. However, as mentioned in Section 6.2, *the linear susceptibility tensor has its form determined only by the crystal system of a material, whereas the form of a nonlinear susceptibility tensor further depends on the point group of the material.* Within the 7 crystal systems, there are 32 point groups. Among the 32 point groups, 21 are noncentrosymmetric and 11 are centrosymmetric. The 21 noncentrosymmetric point groups are those listed in Table 9.1. The 11 centrosymmetric point groups are triclinic $\overline{1}$, monoclinic $3/m$, orthorhombic mmm, tetragonal $4/m$ and $4/mmm$, trigonal $\overline{3}$ and $\overline{3}m$, hexagonal $6/m$ and $6/mmm$, and cubic $m3$ and $m3m$.

Many materials, including gases, liquids, amorphous solids, and many crystals that belong to the 11 centrosymmetric point groups, possess space-inversion symmetry. In the electric-dipole approximation, nonlinear optical effects of all even orders, but not those of the odd orders, vanish identically in a centrosymmetric material.[2] Therefore, $\chi^{(2)}$ *contributed by electric-dipole interaction is identically zero in a centrosymmetric material, whereas a nonzero* $\chi^{(3)}$ *exists in any material.* This fact can be easily verified by considering the effect of space inversion on the nonlinear polarizations $\boldsymbol{P}^{(2)}$ and $\boldsymbol{P}^{(3)}$ given in (9.3) and (9.4), respectively. The space-inversion transformation can be performed on a centrosymmetric material without changing the properties of the material. Being polar vectors, $\boldsymbol{P}^{(2)}$, $\boldsymbol{P}^{(3)}$, and $\boldsymbol{E}$ all change sign under such a transformation. From (9.3), we then find that $\boldsymbol{P}^{(2)} = -\boldsymbol{P}^{(2)}$. Therefore, $\boldsymbol{P}^{(2)}$ cannot exist and $\chi^{(2)}$ has to vanish identically in a centrosymmetric material. No such conclusion is drawn for $\boldsymbol{P}^{(3)}$ and $\chi^{(3)}$ as we examine (9.4) following the same procedure (see Problem 9.2.3).

The discussion above about the vanishing electric-dipole $\chi^{(2)}$ for a centrosymmetric material is valid only for the bulk nonlinear optical property of the material but does not apply to the surface or interface of the material. Centrosymmetry does not exist on the surface of any material or at an interface between two different materials even when

[2] Nonlinear optical effects of even orders that are contributed by magnetic-dipole and electric-quadrupole interactions can still exist in a centrosymmetric material. Nonlinear optical effects of even orders contributed by electric-dipole interaction can also exist at the surfaces or interfaces of centrosymmetric materials where the centrosymmetry is broken.

Table 9.1 *Nonvanishing elements of the second-order nonlinear susceptibility tensor for noncentrosymmetric point groups*

System	Point group	Nonvanishing tensor elements
Triclinic	1	All elements are independent and nonvanishing
Monoclinic	2	$xyz, xzy, xxy, xyx, yxx, yyy, yzz, yzx, yxz,$ zyz, zzy, zxy, zyx (two-fold axis parallel to $\hat{y}$)
	m	$xxx, xyy, xzz, xzx, xxz, yyz, yzy, yxy, yyx,$ zxx, zyy, zzz, zzx, zxz (mirror plane perpendicular to $\hat{y}$)
Orthorhombic	222	$xyz, xzy, yzx, yxz, zxy, zyx$
	$mm2$	$xzx, xxz, yyz, yzy, zxx, zyy, zzz$
Tetragonal	4	$xyz = -yxz, xzy = -yzx, xzx = yzy, xxz = yyz,$ $zxx = zyy, zzz, zxy = -zyx$
	$\overline{4}$	$xyz = yxz, xzy = yzx, xzx = -yzy, xxz = -yyz,$ $zxx = -zyy, zxy = zyx$
	422	$xyz = -yxz, xzy = -yzx, zxy = -zyx$
	$4mm$	$xzx = yzy, xxz = yyz, zxx = zyy, zzz$
	$\overline{4}2m$	$xyz = yxz, xzy = yzx, zxy = zyx$
Trigonal	3	$xxx = -xyy = -yxy = -yyx, xyz = -yxz, xzy = -yzx,$ $xzx = yzy, xxz = yyz, yyy = -yxx = -xxy = -xyx,$ $zxx = zyy, zzz, zxy = -zyx$
	32	$xxx = -xyy = -yxy = -yyx, xyz = -yxz, xzy = -yzx,$ $zxy = -zyx$
	$3m$	$xzx = yzy, xxz = yyz, yyy = -yxx = -xxy = -xyx,$ $zxx = zyy, zzz$ (mirror plane perpendicular to $\hat{x}$)
Hexagonal	6	$xyz = -yxz, xzy = -yzx, xzx = yzy, xxz = yyz,$ $zxx = zyy, zzz, zxy = -zyx$
	$\overline{6}$	$xxx = -xyy = -yxy = -yyx,$ $yyy = -yxx = -xxy = -xyx$
	622	$xyz = -yxz, xzy = -yzx, zxy = -zyx$
	$6mm$	$xzx = yzy, xxz = yyz, zxx = zyy, zzz$
	$\overline{6}m2$	$yyy = -yxx = -xxy = -xyx$
Cubic	432	$xyz = yzx = zxy = -xzy = -yxz = -zyx$
	23	$xyz = yzx = zxy, xzy = yxz = zyx$
	$\overline{4}3m$	$xyz = yzx = zxy = xzy = yxz = zyx$

the materials themselves are centrosymmetric. Therefore, $\chi^{(2)}$ contributed by electric-dipole interaction exists at any material surface or interface. As a result, second-order nonlinear processes that normally do not occur in the bulk of a certain material, such as silicon, which is centrosymmetric, can take place on its surface or interface. The surface $\chi^{(2)}$ also depends on the structure of the material surface.

We have seen in Section 6.1 that the Pockels effect exists only in noncentrosymmetric materials while the electro-optic Kerr effect exists in all materials. Indeed, the Pockels

effect can be considered a special second-order nonlinear optical effect, and the electro-optic Kerr effect is a special third-order nonlinear optical effect. According to (9.21), a nonlinear polarization $P_i^{(2)}(\omega)$ induced by the interaction between a static electric field $E_{0k} = E_k(0)$ polarized along the k direction and an optical field $E_j(\omega)$ polarized along the j direction can be expressed as

$$\begin{aligned} P_i^{(2)}(\omega) &= \epsilon_0 \left[\chi_{ijk}^{(2)}(\omega = \omega + 0) E_j(\omega) E_k(0) + \chi_{ikj}^{(2)}(\omega = 0 + \omega) E_k(0) E_j(\omega) \right] \\ &= 2\epsilon_0 \chi_{ijk}^{(2)}(\omega = \omega + 0) E_j(\omega) E_{0k}. \end{aligned} \tag{9.30}$$

Using (6.18) and identifying $\Delta\epsilon_{ij}(\omega)$ with $P_i^{(2)}(\omega)/E_j(\omega) = 2\epsilon_0 \chi_{ijk}^{(2)}(\omega = \omega + 0) E_{0k}$, we find that the Pockels coefficients are related to the $\boldsymbol{\chi}^{(2)}$ elements as (see Problem 9.2.4)

$$r_{ijk} = -\frac{2}{n_i^2 n_j^2} \chi_{ijk}^{(2)}(\omega = \omega + 0) = -\frac{2}{n_i^2 n_j^2} \chi_{kij}^{(2)}(0 = \omega - \omega), \tag{9.31}$$

where the full permutation symmetry is used in moving the zero frequency to the left-hand side of the equals sign in the argument of $\boldsymbol{\chi}^{(2)}$. Similarly, for the electro-optic Kerr coefficients, we have (see Problem 9.2.4)

$$s_{ijkl} = -\frac{3}{n_i^2 n_j^2} \chi_{ijkl}^{(3)}(\omega = \omega + 0 + 0). \tag{9.32}$$

It can be seen from the above discussions that though not all noncentrosymmetric crystals are useful, any material that can support a second-order nonlinear process through electric-dipole interaction is necessarily a noncentrosymmetric crystal. The nonvanishing $\boldsymbol{\chi}^{(2)}$ tensor elements and the relations among them for each of the 21 noncentrosymmetric point groups are listed in Table 9.1.

EXAMPLE 9.3 The $\hat{x}$, $\hat{y}$, and $\hat{z}$ directional unit vectors used to define the electric field polarizations in Examples 9.1 and 9.2 are aligned with the principal x, y, and z axes of a crystal. (a) Use the result obtained in Example 9.2 to find the nonvanishing terms in the three components of $\mathbf{P}^{(2)}(\omega_4)$ if the nonlinear interaction takes place in a $\overline{4}3m$ crystal, such as GaAs. (b) Find the nonvanishing terms in the case of a crystal of $mm2$ point group, such as KTP.

Solution (a) From Table 9.1, the only nonvanishing $\boldsymbol{\chi}^{(2)}$ elements for the $\overline{4}3m$ point group are $\chi_{xyz}^{(2)} = \chi_{yzx}^{(2)} = \chi_{zxy}^{(2)} = \chi_{xzy}^{(2)} = \chi_{yxz}^{(2)} = \chi_{zyx}^{(2)}$. From the expressions for the components of $\mathbf{P}^{(2)}(\omega_4)$ obtained in Example 9.2, we have, for a $\overline{4}3m$ crystal,

$$\begin{aligned} P_x^{(2)}(\omega_4) &= \epsilon_0 \chi_{xyz}^{(2)}(\omega_4 = \omega_2 + \omega_2) E_2^2, \\ P_y^{(2)}(\omega_4) &= 2\epsilon_0 \chi_{yxz}^{(2)}(\omega_4 = \omega_1 + \omega_3) E_1 E_3, \\ P_z^{(2)}(\omega_4) &= 0. \end{aligned}$$

(b) For the $mm2$ point group, the only nonvanishing $\boldsymbol{\chi}^{(2)}$ elements are $\chi^{(2)}_{xzx}$, $\chi^{(2)}_{xxz}$, $\chi^{(2)}_{yyz}$, $\chi^{(2)}_{yzy}$, $\chi^{(2)}_{zxx}$, $\chi^{(2)}_{zyy}$, and $\chi^{(2)}_{zzz}$, according to Table 9.1. Then, the expressions for the components of $\mathbf{P}^{(2)}(\omega_4)$ obtained in Example 9.2 reduce to

$$P_x^{(2)}(\omega_4) = 2\epsilon_0 \chi^{(2)}_{xxz}(\omega_4 = \omega_1 + \omega_3) E_1 E_3,$$
$$P_y^{(2)}(\omega_4) = \epsilon_0 \chi^{(2)}_{yyz}(\omega_4 = \omega_2 + \omega_2) E_2^2,$$
$$P_z^{(2)}(\omega_4) = \epsilon_0 \left[\chi^{(2)}_{zyy}(\omega_4 = \omega_2 + \omega_2) \frac{E_2^2}{2} + \chi^{(2)}_{zzz}(\omega_4 = \omega_2 + \omega_2) \frac{E_2^2}{2} \right].$$

Because $\boldsymbol{\chi}^{(3)}$ exists in all materials, the materials used for the devices that are based on third-order nonlinear optical processes are usually isotropic noncrystalline materials, such as glasses, or cubic crystals, such as the III–V semiconductors. Only occasionally are noncubic crystals used for such devices. Third-order nonlinear processes are particularly important for isotropic materials because $\boldsymbol{\chi}^{(2)}$ vanishes identically so that $\boldsymbol{\chi}^{(3)}$ becomes the leading nonlinear susceptibility of such materials. Table 9.2 lists the nonvanishing $\boldsymbol{\chi}^{(3)}$ tensor elements and the relations among them for the cubic crystal system and for isotropic materials. It can be seen that for all of the point groups in the cubic system and for isotropic materials, there are only 21 nonvanishing $\boldsymbol{\chi}^{(3)}$ tensor elements. For the 23 and $m3$ point groups, there are 7 independent $\boldsymbol{\chi}^{(3)}$ elements. For the 432, $\bar{4}3m$, and $m3m$ point groups, there are only 4 independent elements of the types $\chi^{(3)}_{1111}$, $\chi^{(3)}_{1122}$, $\chi^{(3)}_{1212}$, and $\chi^{(3)}_{1221}$. If Kleiman's symmetry condition is valid, the number of independent elements reduces to 2 of the types $\chi^{(3)}_{1111}$ and $\chi^{(3)}_{1122} = \chi^{(3)}_{1212} = \chi^{(3)}_{1221}$ for all point groups in the cubic system. For an isotropic material, there are only

Table 9.2 *Nonvanishing elements of the third-order nonlinear susceptibility tensor for cubic and isotropic materials*

System	Point group	Nonvanishing tensor elements
Cubic	$23, m3$	$xxxx = yyyy = zzzz,$
		$xxyy = yyzz = zzxx, yyxx = zzyy = xxzz,$
		$xyxy = yzyz = zxzx, yxyx = zyzy = xzxz,$
		$xyyx = yzzy = zxxz, yxxy = zyyz = xzzx$
	$432, \bar{4}3m, m3m$	$xxxx = yyyy = zzzz,$
		$xxyy = yyxx = yyzz = zzyy = zzxx = xxzz,$
		$xyxy = yxyx = yzyz = zyzy = zxzx = xzxz,$
		$xyyx = yxxy = yzzy = zyyz = zxxz = xzzx$
Isotropic		$xxxx = yyyy = zzzz = xxyy + xyxy + xyyx,$
		$xxyy = yyxx = yyzz = zzyy = zzxx = xxzz,$
		$xyxy = yxyx = yzyz = zyzy = zxzx = xzxz,$
		$xyyx = yxxy = yzzy = zyyz = zxxz = xzzx$

3 independent $\boldsymbol{\chi}^{(3)}$ elements among the 4 types of nonvanishing elements because $\chi^{(3)}_{1111} = \chi^{(3)}_{1122} + \chi^{(3)}_{1212} + \chi^{(3)}_{1221}$. If Kleiman's symmetry condition is valid in an isotropic material, we have

$$\chi^{(3)}_{1122} = \chi^{(3)}_{1212} = \chi^{(3)}_{1221} = \frac{1}{3}\chi^{(3)}_{1111}, \tag{9.33}$$

reducing the number of independent $\boldsymbol{\chi}^{(3)}$ elements to only 1.

Nonlinear optical d coefficients

In the literature, experimentally measured values of second-order nonlinear susceptibilities of a material are commonly quoted in terms of nonlinear coefficients d_{ijk}, or $d_{i\alpha}$ under index contraction.[3] The relation between the d coefficients and the $\boldsymbol{\chi}^{(2)}$ elements is simply

$$d_{ijk} = \frac{1}{2}\chi^{(2)}_{ijk} \tag{9.34}$$

if neither index j nor index k is associated with a DC field.

Index contraction

In certain situations, the rule of index contraction expressed in (1.115) can be applied to $\chi^{(2)}_{ijk}$ and d_{ijk} on the last two indices j and k by replacing jk with α. Then the 27 elements of $\chi^{(2)}_{ijk}$, or d_{ijk}, are reduced to 18 elements of $\chi^{(2)}_{i\alpha}$, or $d_{i\alpha}$, for $i = 1, 2, 3$ and $\alpha = 1, 2, \ldots, 6$. Clearly, the condition for index contraction to be applicable is when there is no physical significance in interchanging the last two indices j and k independently of the frequencies in $\chi^{(2)}_{ijk}$.

For $\boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2)$ in general, index contraction applies only when Kleiman's symmetry condition is valid so that $\chi^{(2)}_{ijk}(\omega_3 = \omega_1 + \omega_2) = \chi^{(2)}_{ikj}(\omega_3 = \omega_1 + \omega_2)$. However, index contraction applies without the requirement of Kleiman's symmetry condition in the special cases of $\boldsymbol{\chi}^{(2)}(2\omega = \omega + \omega)$ and $\boldsymbol{\chi}^{(2)}(0 = \omega - \omega)$.

For $\boldsymbol{\chi}^{(2)}(2\omega = \omega + \omega)$, which characterizes the process of *second-harmonic generation*, index contraction always applies because $\chi^{(2)}_{ijk}(2\omega = \omega + \omega) = \chi^{(2)}_{ikj}(2\omega = \omega + \omega)$ by the definition of the intrinsic permutation symmetry.

For $\boldsymbol{\chi}^{(2)}(0 = \omega - \omega)$, which characterizes the process of *optical rectification* for the generation of a DC electric field by an optical field, index contraction applies only

[3] The nonlinear optical d coefficients are not to be confused with the piezoelectric d coefficients though both are second-rank tensors and they have the same matrix form. The piezoelectric d coefficients define the piezoelectric polarization induced by a strain tensor.

when the medium is lossless at the frequency ω so that $\chi^{(2)}_{ijk}(0 = \omega - \omega) = \chi^{(2)}_{ijk}(0 = -\omega + \omega) = \chi^{(2)}_{ikj}(0 = \omega - \omega)$ due to the reality condition and the intrinsic permutation symmetry. From (9.31), we find that

$$r_{\alpha k} = -\frac{2}{n_i^2 n_j^2}\chi^{(2)}_{k\alpha} = -\frac{4}{n_i^2 n_j^2}d_{k\alpha}, \tag{9.35}$$

where $\chi^{(2)}_{k\alpha} = \chi^{(2)}_{kij}(0 = \omega - \omega)$ and the index k is associated with the DC electric field. Note that Kleiman's symmetry condition is never valid for $\boldsymbol{\chi}^{(2)}(0 = \omega - \omega)$ because no material can be completely nondispersive in the entire spectral range from DC to the optical frequencies.

With index contraction, the second-order nonlinear susceptibilities $\chi^{(2)}_{i\alpha}$ and, correspondingly, the nonlinear coefficients $d_{i\alpha}$ can be expressed in the form of a 3×6 matrix. From the relation in (9.35), it is clear that the matrix form of $\chi^{(2)}_{i\alpha}$ and $d_{i\alpha}$ for each of the noncentrosymmetric point groups is exactly the *transpose* of the matrix of the Pockels coefficients listed in Table 6.1. If Kleiman's symmetry condition is valid, the matrix form of $\chi^{(2)}_{i\alpha}$ and $d_{i\alpha}$ is further simplified to result in a maximum of only 10 independent parameters. For example, $d_{14} = d_{25} = d_{36}$ under Kleiman's symmetry condition. From Table 6.1, we find that the only nonlinear coefficients for the 422 and 622 point groups are $d_{25} = -d_{14}$, which have to vanish identically under Kleiman's symmetry condition though the Pockels coefficients $r_{52} = -r_{41}$ do not have to vanish. We also find that under Kleiman's symmetry condition the 3 independent nonlinear coefficients d_{14}, d_{25}, and d_{36} for the 222 point group reduce to 1 identical parameter, and the 2 independent parameters $d_{14} = d_{25}$ and d_{36} for the $\bar{4}2m$ point group also reduce to a single parameter. The properties of some important nonlinear crystals are listed in Table 9.3.

Using (9.21) and (9.34), the second-order nonlinear polarization can be expressed in terms of the $d_{i\alpha}$ matrix. In the general case of $\omega_1 + \omega_2 = \omega_3$ with $\omega_1 \neq \omega_2$, we have

$$\begin{bmatrix} P^{(2)}_x(\omega_3) \\ P^{(2)}_y(\omega_3) \\ P^{(2)}_z(\omega_3) \end{bmatrix} = 4\epsilon_0 \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{bmatrix}$$

$$\times \begin{bmatrix} E_x(\omega_1)E_x(\omega_2) \\ E_y(\omega_1)E_y(\omega_2) \\ E_z(\omega_1)E_z(\omega_2) \\ E_y(\omega_1)E_z(\omega_2) + E_z(\omega_1)E_y(\omega_2) \\ E_z(\omega_1)E_x(\omega_2) + E_x(\omega_1)E_z(\omega_2) \\ E_x(\omega_1)E_y(\omega_2) + E_y(\omega_1)E_x(\omega_2) \end{bmatrix}. \tag{9.36}$$

Table 9.3 *Properties of representative nonlinear crystals*[a]

Crystal	$LiNbO_3$ (LN)	β-BaB_2O_4 (BBO)	LiB_3O_5 (LBO)	$KTiOPO_4$ (KTP)	$KTiOAsO_4$ (KTA)	$AgGaS_2$	$AgGaSe_2$	$ZnGeP_2$	$LiIO_3$
Point group	$3m$	$3m$	$mm2$	$mm2$	$mm2$	$\bar{4}2m$	$\bar{4}2m$	$\bar{4}2m$	6
$n_x(n_o)$: see Note *b*									
A	4.913	2.735 9	2.454 3	3.006 48	3.142 40	5.728	6.854 40	4.473 30	3.414
B	0.118 8	0.018 78	0.011 413	0.038 81	0.045 89	0.241 07	0.408 15	5.265 76	0.046 29
C	0.045 97	0.018 22	0.009 498 1	0.043 52	0.044 48	0.087 03	0.192 64	0.133 81	0.037 11
D	0.027 8	0.013 54	0.013 900	0.013 20	0.010 60	0.002 10	0.001 283 8	1.490 85	0.007 603 2
E								662.55	
n_y: see Note *b*									
A			2.538 2	3.030 42	3.167 90				
B	Same as	Same as	0.012 83	0.041 76	0.044 05	Same as	Same as	Same as	Same as
C	n_x	n_x	0.011 387	0.047 53	0.056 55	n_x	n_x	n_x	n_x
D			0.017 034	0.013 27	0.015 00				
E									
$n_z(n_e)$: see Note *b*									
A	4.579 8	2.375 3	2.585 4	3.313 72	3.448 30	5.497	6.691 30	4.633 18	2.922 82
B	0.099 4	0.012 24	0.013 065	0.056 80	0.063 53	0.202 59	0.395 83	5.342 15	0.032 87
C	0.042 35	0.016 67	0.011 617	0.056 79	0.057 70	0.130 70	0.283 65	0.142 55	0.033 335
D	0.022 4	0.015 16	0.018 146	0.016 79	0.017 40	0.002 33	0.001 343 0	0.145 795	0.004 262 3
E								662.55	
d_{31}(pm V^{-1})	−4.4	0.04	−1.09	3.7	2.8	0	0	0	−4.64
d_{32}	$= d_{31}$	$= d_{31}$	1.17	2.2	4.2	0	0	0	$= d_{31}$
d_{33}	−25.2	—	0.065	14.6	16.2	0	0	0	−4.84
d_{24}	$= d_{31}$	$= d_{31}$	−1.1	1.9	4.24	0	0	0	—
d_{15}	$= d_{24}$	$= d_{24}$	1.00	3.7	2.24	0	0	0	$= d_{24}$
d_{36}	0	0	0	0	0	28.7	49.3	75	0
d_{14}	0	0	0	0	0	$= d_{36}$	$= d_{36}$	—	0.31
d_{25}	0	0	0	0	0	$= d_{14}$	$= d_{14}$	—	$= -d_{14}$
$d_{22} = -d_{21} = -d_{16}$	2.4	2.22	0	0	0	0	0	0	0
Transparency (nm)	400–5000	190–3000	160–2600	350–4500	400–5000	500–13000	780–18000	700–12000	300–5500
Damage threshold (GW cm^{-2})	1	13	25	1	1.2	0.02	0.02	>1	0.5
dn_x/dT ($\times 10^{-5}$ K^{-1})	2.0	−1.7	−0.19	1.1		5	4	14.3	−8.9
dn_y/dT ($\times 10^{-5}$ K^{-1})	2.0	−1.7	−1.3	1.3		5	4	14.3	−8.9
dn_z/dT ($\times 10^{-5}$ K^{-1})	7.6	−0.93	−0.83	1.6		5	4	15.0	−7.8

[a] The data are collected from various sources in the literature. Many of the crystal properties are constantly being revised. The refractive indices of some crystals vary with crystal growth and preparation procedures. The nonlinear susceptibilities can vary with optical wavelength and temperature, too.

[b] The parameters listed here for the refractive indices are those at 300 K. For all crystals except $ZnGeP_2$, the refractive indices are calculated using the following Sellmeier equation:

$$n^2 = A + \frac{B}{\lambda^2 - C} - D\lambda^2.$$

For $ZnGeP_2$, the following Sellmeier equation has to be used:

$$n^2 = A + \frac{B}{1 - C/\lambda^2} + \frac{D}{1 - E/\lambda^2}.$$

In both formulas, λ is in micrometers.

In the case of second-harmonic generation, we have

$$\begin{bmatrix} P_x^{(2)}(2\omega) \\ P_y^{(2)}(2\omega) \\ P_z^{(2)}(2\omega) \end{bmatrix} = 2\epsilon_0 \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{bmatrix} \begin{bmatrix} E_x^2(\omega) \\ E_y^2(\omega) \\ E_z^2(\omega) \\ 2E_y(\omega)E_z(\omega) \\ 2E_z(\omega)E_x(\omega) \\ 2E_x(\omega)E_y(\omega) \end{bmatrix}. \tag{9.37}$$

Note that each d coefficient in (9.36) has the same value as the corresponding one in (9.37) if we ignore dispersion due to frequency differences between (9.36) and (9.37). It is true that in the case when $\omega_1 \neq \omega_2$, $\mathbf{P}^{(2)}(\omega_3) = 2\mathbf{P}^{(2)}(2\omega)$ if $\mathbf{E}(\omega_1) = \mathbf{E}(\omega_2) = \mathbf{E}(\omega)$ as is seen by comparing (9.36) and (9.37).

Unit conversion

The SI system, which is essentially the MKSA system, is used consistently in this book. Nevertheless, the Gaussian system is also used quite often in the literature. In the Gaussian system, cgs units are used, but the electric field quantities and the susceptibilities are normally given the units of esu, meaning electrostatic units, without explicitly spelling out their true dimensions. In the SI system, the units are explicit. Unit conversion for susceptibilities between the SI and Gaussian systems follows the following relations:

$$\begin{aligned} &\text{SI} && \text{Gaussian} \\ &\chi^{(1)}(\text{dimensionless}) = 4\pi\,\chi^{(1)}(\text{dimensionless}), && (9.38) \\ &\chi^{(2)}(\text{m V}^{-1}) = \frac{4\pi}{3\times 10^4}\chi^{(2)}(\text{esu}), && (9.39) \\ &\chi^{(3)}(\text{m}^2\,\text{V}^{-2}) = \frac{4\pi}{9\times 10^8}\chi^{(3)}(\text{esu}). && (9.40) \end{aligned}$$

Unit conversion for the d coefficient is the same as that for $\chi^{(2)}$ given in (9.39) because the relation in (9.34) is independent of the unit system used.

9.3 Nonlinear optical interactions

As discussed in the preceding section, optical susceptibilities in the frequency domain are, in general, complex quantities. Following the same convention, used in Sections 1.3 and 1.10 for complex linear susceptibility in the frequency domain, complex nonlinear susceptibilities in the frequency domain can be expressed as $\boldsymbol{\chi}^{(2)} = \boldsymbol{\chi}^{(2)\prime} + \mathrm{i}\boldsymbol{\chi}^{(2)\prime\prime}$ and $\boldsymbol{\chi}^{(3)} = \boldsymbol{\chi}^{(3)\prime} + \mathrm{i}\boldsymbol{\chi}^{(3)\prime\prime}$ to define their real and imaginary parts clearly. Similarly to the case of linear susceptibility discussed in Section 1.10, the imaginary part of a nonlinear

susceptibility is always associated with the intrinsic resonances of a material. Such resonances signify the transitions between different energy levels of the material. Near the resonance frequencies that are relevant to a particular susceptibility, the imaginary part of the susceptibility peaks while the real part displays a highly dispersive behavior (see Problems 9.2.5 and 9.2.8). We shall not get into discussions on the quantum-mechanical characteristics of nonlinear susceptibilities as a function of energy levels of a material, or on the intrinsic relationship between the real and imaginary parts of a nonlinear susceptibility. Such subjects are of special interest in studies of material properties as well as in the application of nonlinear optics as a spectroscopic tool. For our purpose in this chapter, it suffices to understand that the imaginary part of a nonlinear susceptibility is strongly coupled to the transition resonances of a material in a way similar to the dependence of the imaginary part of the linear susceptibility on the material resonances. In addition, the real and imaginary parts of a nonlinear susceptibility are also related in a way similar to the relationship, expressed in the Kramers–Kronig relation, between those of the linear susceptibility.

An optical interaction that is characterized by a real frequency-dependent susceptibility, or the real part of a complex susceptibility, is generally classified as *parametric*, whereas one that is associated with the imaginary part of a complex frequency-dependent susceptibility is *nonparametric*. In a nonparametric process, the state of the material changes, and the total optical energy also changes accordingly, because the process is connected to resonant transitions in the material. In a purely parametric process, however, both the state of the material in the medium and the total optical energy remain unchanged because the process does not cause any net exchange of energy between the optical field and the medium.

In a parametric *linear* process, the energy of any given optical frequency component is conserved because a linear process does not couple fields of different frequencies, as can be seen from the fact that linear susceptibility $\chi^{(1)}(\omega)$ is a function of a single frequency. In a parametric *nonlinear* process, energy exchange among different frequency components caused by nonlinear coupling among them usually occurs though the sum of energies from all of the interacting frequency components is conserved. For example, in a parametric process characterized by a real $\chi^{(2)}(\omega_3 = \omega_1 + \omega_2)$, optical energy can be transferred from the frequency components at ω_1 and ω_2 to the component at ω_3, or vice versa. Therefore, the energy in each individual frequency component may change, but the total optical energy contained in all three frequency components is conserved because there is no net exchange of energy between the optical field and the medium in a parametric process.

There are two features that are unique to nonlinear optical processes: one is optical frequency conversion, and the other is field-dependent modification of a certain material property. All nonlinear optical processes exhibit at least one, though not necessarily both, of these two features. All functional nonlinear optical devices take advantage of one or both of these two unique features.

As we have seen again and again in earlier chapters, a very important condition for efficient coupling among optical waves or among optical modes is phase matching, regardless of which physical mechanism is responsible for the coupling. Phase matching is also most important for efficient nonlinear optical interactions. The phase-matching condition for a second-order nonlinear interaction that is characterized by the relation $\omega_3 = \omega_1 + \omega_2$ among the interacting frequencies is

$$\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2. \tag{9.41}$$

The phase-matching condition for a third-order interaction that is characterized by the relation $\omega_4 = \omega_1 + \omega_2 + \omega_3$ among the interacting frequencies is

$$\mathbf{k}_4 = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3. \tag{9.42}$$

When a particular frequency changes sign, the corresponding wavevector in the phase-matching condition also changes sign. The phase-matching condition is not automatically satisfied in a parametric process that leads to the conversion of optical frequencies. However, it is automatically satisfied in a parametric process that does not convert energy from one optical frequency to another. It is also automatically satisfied in any nonparametric process where an exchange of energy between the optical field and the medium occurs.

Second-order nonlinear optical processes

Table 9.4 lists the second-order nonlinear processes. As indicated in this table, the contributing susceptibility for each of these processes is the real part, $\chi^{(2)'}$, of a frequency-dependent second-order susceptibility. Therefore, all of them are parametric in nature, and the frequencies involved in a second-order process of interest are generally far away from any resonance frequencies of the nonlinear medium.

Although each process listed in Table 9.4 demonstrates a unique phenomenon and has its own specific applications, all of them are basically parametric frequency conversion

Table 9.4 *Second-order nonlinear optical processes*

Process	Susceptibility	Phase matching
Second-harmonic generation (SHG)	$\chi^{(2)'}(2\omega = \omega + \omega)$	Required
Sum-frequency generation (SFG)	$\chi^{(2)'}(\omega_3 = \omega_1 + \omega_2)$	Required
Difference-frequency generation (DFG)	$\chi^{(2)'}(\omega_2 = \omega_3 - \omega_1)$	Required
Optical parametric amplification (OPA)	$\chi^{(2)'}(\omega_2 = \omega_3 - \omega_1)$	Required
Optical parametric generation (OPG)	$\chi^{(2)'}(\omega_3 = \omega_1 + \omega_2)$	Required
Optical rectification	$\chi^{(2)'}(0 = \omega - \omega)$	Automatic
Pockels effect	$\chi^{(2)'}(\omega = \omega + 0)$	Automatic

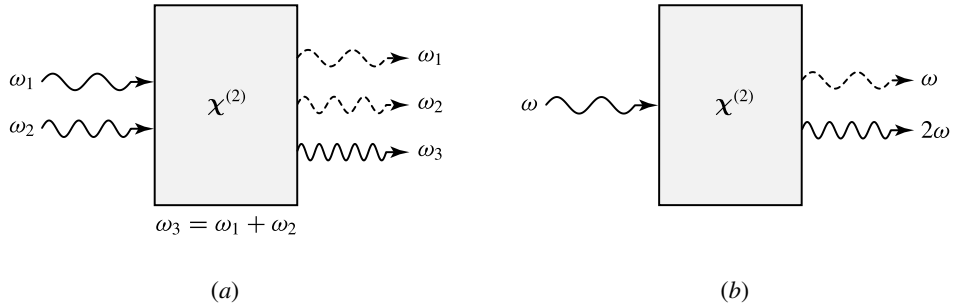


Figure 9.1 (*a*) Sum-frequency generation and (*b*) second-harmonic generation. Second-harmonic generation is the degenerate case of sum-frequency generation.

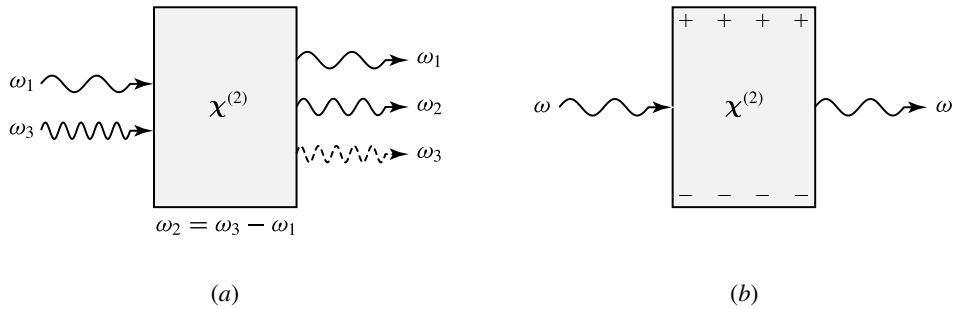


Figure 9.2 (*a*) Difference-frequency generation and (*b*) optical rectification. Optical rectification is the degenerate case of difference-frequency generation.

processes generally characterized by a real susceptibility of the form $\chi^{(2)}(\omega_3 = \omega_1 + \omega_2)$. The differences among different processes come either from different experimental conditions or from different subjective purposes of application.

The full permutation symmetry expressed in (9.28) applies to the real susceptibilities characterizing parametric second-order processes. Therefore, the processes of sum-frequency generation, difference-frequency generation, and optical parametric generation listed in Table 9.4 have the same nonlinear susceptibility. In sum-frequency generation, shown in Fig. 9.1(*a*), an optical wave at a high frequency, ω_3, is generated through nonlinear interaction of optical waves at two lower frequencies, ω_1 and ω_2, with the nonlinear medium. Second-harmonic generation is the degenerate case of sum-frequency generation for $\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$, as shown in Fig. 9.1(*b*). In difference-frequency generation, shown in Fig. 9.2(*a*), two optical waves at frequencies ω_3 and ω_1 interact with the nonlinear medium to generate an optical wave at the difference frequency $\omega_2 = \omega_3 - \omega_1$. Optical rectification, shown in Fig. 9.2(*b*), is the degenerate case of difference-frequency generation for $\omega_3 = \omega_1 = \omega$ and $\omega_2 = 0$. In nondegenerate sum- and difference-frequency generation, two optical waves at different frequencies have to be supplied at the input. In second-harmonic generation and optical

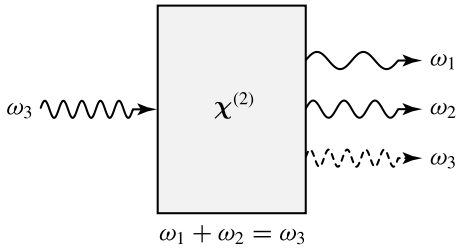


Figure 9.3 Optical parametric generation.

rectification, only one optical wave is needed at the input. Second-harmonic generation performs the function of *frequency doubling* for the input optical wave, whereas optical rectification generates a DC electric field in a nonlinear medium through the interaction of an optical wave with the medium.

In sum-frequency generation, energy conservation implied by the relation $\omega_3 = \omega_1 + \omega_2$ requires that one photon at ω_1 and another at ω_2 be annihilated simultaneously when one photon at ω_3 is generated. The intensities of both input waves at ω_1 and ω_2 decrease as the intensity of the output wave at ω_3 grows. In difference-frequency generation characterized by the relation $\omega_2 = \omega_3 - \omega_1$, however, one photon at the input frequency ω_1 is *generated* simultaneously with one photon at the difference frequency ω_2 when one photon at ω_3 is annihilated. Therefore, as the intensity of the output wave at ω_2 grows, the intensity of the low-frequency input wave at ω_1 also increases though that of the high-frequency input wave at ω_3 diminishes.

The reverse process of sum-frequency generation is optical parametric generation, in which an optical wave at a high frequency ω_3 interacts with the nonlinear medium and generates two optical waves at the lower frequencies ω_1 and ω_2, as shown in Fig. 9.3. In this process, one photon at ω_3 is annihilated when two photons at ω_1 and ω_2, respectively, are generated. This process can occur spontaneously in the form of *parametric fluorescence*. In practical applications, it normally takes place with a feedback or with another input wave at either ω_1 or ω_2. When a feedback is provided, usually by placing the nonlinear crystal in a resonant optical cavity, *optical parametric oscillation* is possible with only one input wave at ω_3. If an input wave at one of the parametric frequencies, say ω_1, is also provided, the process is basically the same as that of difference-frequency generation except that the purpose is now the amplification of the signal at the input frequency ω_1 rather than the generation of the wave at the difference frequency ω_2. For this reason, this process is called *optical parametric amplification*.

The special case of the Pockels effect can also be considered as parametric mixing of an optical field at ω with a DC or low-frequency electric field though it is generally described in terms of a modification on the permittivity tensor of a crystal by the DC or low-frequency electric field.

EXAMPLE 9.4 Starting with a Nd:YAG laser that emits at a single wavelength of 1.064 μm, what different optical wavelengths can possibly be generated through second-order nonlinear optical processes in a single step? What can be generated in two cascaded steps?

Solution Starting from $\lambda_\omega = 1.064$ μm, we can generate its second harmonic at $\lambda_{2\omega} = \lambda_\omega/2 = 532$ nm by frequency doubling in one step through $\chi^{(2)}(2\omega = \omega + \omega)$ in a nonlinear crystal. We can also use optical parametric generation to generate pairs of tunable wavelengths at $\lambda_1, \lambda_2 > 1.064$ μm for $1/\lambda_1 + 1/\lambda_2 = 1/\lambda_\omega$ in one step through $\chi^{(2)}(\omega = \omega_1 + \omega_2)$ in a nonlinear crystal. Therefore, it is possible to obtain a visible wavelength at 532 nm and a range of tunable infrared wavelengths longer than 1.064 μm.

In two cascaded steps, many more wavelengths can be generated. The second harmonic can be further frequency doubled to the fourth harmonic at $\lambda_{4\omega} = \lambda_\omega/4 =$ 266 nm. We can also mix the fundamental and the second harmonic in a nonlinear crystal to generate the third harmonic at $\lambda_{3\omega} = \lambda_\omega/3 = 354.7$ nm by sum-frequency generation through $\chi^{(2)}(3\omega = \omega + 2\omega)$ in a nonlinear crystal. The second harmonic at 532 nm generated in the first step can be used to generate a range of tunable wavelengths longer than 532 nm through optical parametric generation. This range of tunable wavelengths longer than 532 nm can also be covered by frequency doubling the tunable infrared wavelengths generated through optical parametric generation in the first step, as well as by mixing the tunable infrared wavelengths with the fundamental at 1.064 μm through sum-frequency generation.

We can see further along this line that if we take merely three steps of second-order nonlinear optical mixing processes, either all in cascade or mixed in parallel/cascade, it is possible to cover a wide spectral range from the deep ultraviolet to the far infrared by starting with a single laser wavelength. The requirements for these processes to take place efficiently, as well as their limitations, are discussed in Sections 9.5 and 9.6.

Third-order nonlinear optical processes

The third-order nonlinear processes of common interest are listed in Table 9.5. Among these third-order processes, some, such as third-harmonic generation, parametric frequency conversion, and the optical Kerr effect, are parametric, whereas others, such as absorption saturation and stimulated Raman scattering, are nonparametric. As can be seen in Table 9.5, the contributing susceptibility of a parametric third-order process is the real part, $\chi^{(3)\prime}$, and that of a nonparametric process is the imaginary part, $\chi^{(3)\prime\prime}$, of a frequency-dependent third-order susceptibility.

As many as four different optical frequencies can participate in a third-order parametric frequency conversion process. As shown in Fig. 9.4, there are a few different variations of this process. In a scenario similar to sum-frequency generation, three

Table 9.5 *Third-order nonlinear optical processes*

Process	Susceptibility	Phase matching
Third-harmonic generation (THG)	$\chi^{(3)\prime}(3\omega = \omega + \omega + \omega)$	Required
Parametric frequency conversion	$\chi^{(3)\prime}(\omega_4 = \omega_1 + \omega_2 + \omega_3)$ $\chi^{(3)\prime}(\omega_4 = \omega_1 + \omega_2 - \omega_3)$ $\chi^{(3)\prime}(\omega_3 = \omega_4 - \omega_1 - \omega_2)$	Required
Optical Kerr effect	$\chi^{(3)\prime}(\omega = \omega + \omega - \omega)$ $\chi^{(3)\prime}(\omega = \omega + \omega' - \omega')$	Automatic
Self-phase modulation (SPM)	$\chi^{(3)\prime}(\omega = \omega + \omega - \omega)$	Automatic
Cross-phase modulation (XPM)	$\chi^{(3)\prime}(\omega = \omega + \omega' - \omega')$	Automatic
Electro-optic Kerr effect	$\chi^{(3)\prime}(\omega = \omega + 0 + 0)$	Automatic
Absorption saturation	$\chi^{(3)\prime\prime}(\omega = \omega + \omega - \omega)$	Automatic
Gain saturation	$\chi^{(3)\prime\prime}(\omega = \omega + \omega - \omega)$	Automatic
Two-photon absorption (TPA)	$\chi^{(3)\prime\prime}(\omega_1 = \omega_1 + \omega_2 - \omega_2)$	Automatic
Stimulated Raman scattering (SRS)	$\chi^{(3)\prime\prime}(\omega_S = \omega_S + \omega_p - \omega_p)$	Automatic
Stimulated Brillouin scattering (SBS)	$\chi^{(3)\prime\prime}(\omega_S = \omega_S + \omega_p - \omega_p)$	Automatic

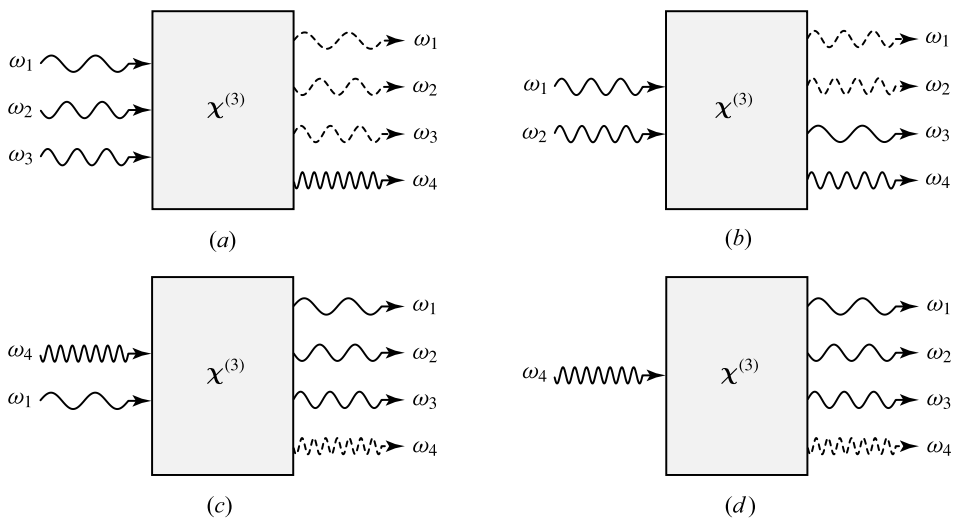


Figure 9.4 Third-order parametric frequency conversion processes. (*a*) $\omega_1 + \omega_2 + \omega_3 \rightarrow \omega_4$. (*b*) $\omega_1 + \omega_2 \rightarrow \omega_3 + \omega_4$. (*c*) $\omega_4 - \omega_1 \rightarrow \omega_2 + \omega_3$. (*d*) $\omega_4 \rightarrow \omega_1 + \omega_2 + \omega_3$.

photons at frequencies ω_1, ω_2, and ω_3 combine to generate a photon at a higher frequency $\omega_4 = \omega_1 + \omega_2 + \omega_3$, as shown in Fig. 9.4(*a*). In another scenario, shown in Fig. 9.4(*b*), two photons at ω_1 and ω_2 combine to generate two other photons at ω_3 and ω_4. In yet another scenario, which is similar to difference-frequency generation, shown in Fig. 9.4(*c*), one photon at ω_4 breaks into three photons at ω_1, ω_2, and ω_3 through the interaction of the wave at ω_4 with another wave at a lower frequency, say ω_1. In a process similar to parametric generation, it is also possible for a photon at ω_4 to break into three photons at ω_1, ω_2, and ω_3 without an input at any of the lower frequencies,

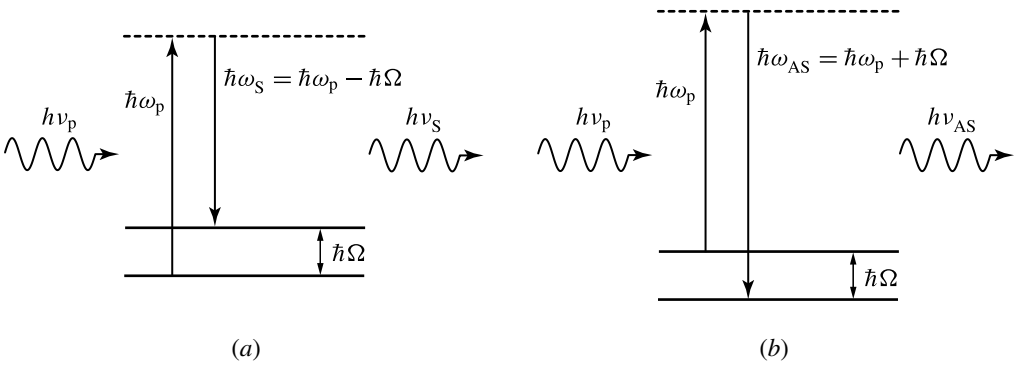


Figure 9.5 (*a*) Stokes and (*b*) anti-Stokes transitions for stimulated Raman scattering.

as shown in Fig. 9.4(*d*). Third-harmonic generation is simply the degenerate case of the first scenario for $\omega_1 = \omega_2 = \omega_3 = \omega$ and $\omega_4 = 3\omega$. Other partially degenerate cases exist also. For example, for any of the three scenarios shown in Fig. 9.4, it is possible that $\omega_1 = \omega_2 = \omega$ but $\omega_3 \neq \omega$ and $\omega_3 \neq \omega_4$.

The parametric frequency conversion processes are not the only third-order nonlinear processes that result in optical frequency conversion. Among the nonparametric processes, stimulated Raman scattering and stimulated Brillouin scattering also lead to optical frequency conversion. However, being nonparametric, these processes are connected to the intrinsic resonances of the medium and are dependent on the initial state of the material. If the material is originally in the ground state of the relevant transition, annihilation of a photon at the pump frequency, ω_p, of the incident optical wave creates an excitation in the material and a photon at a down-shifted *Stokes frequency* $\omega_S = \omega_p - \Omega$, as illustrated in Fig. 9.5(*a*). If the material is originally in an excited state, it is possible to create a photon at an up-shifted *anti-Stokes frequency* $\omega_{AS} = \omega_p + \Omega$ while the material simultaneously makes a transition from the excited state to the ground state, as illustrated in Fig. 9.5(*b*). The Stokes susceptibility is $\chi^{(3)''}(\omega_S = \omega_S + \omega_p - \omega_p)$, whereas the anti-Stokes susceptibility is $\chi^{(3)''}(\omega_{AS} = \omega_{AS} + \omega_p - \omega_p)$. The amount of frequency shift, $\Omega = \omega_p - \omega_S = \omega_{AS} - \omega_p$, is determined by the excitation responsible for a given process and is a characteristic of the material. The fundamental difference between a Raman process and a Brillouin process is the mode of excitation in the material that participates in the interaction. In stimulated Raman scattering, the interaction is associated with the excitation at the molecular or atomic level, such as the optical phonons of a medium or the vibrational modes of molecules. In stimulated Brillouin scattering, the interaction is associated with the long-range excitation characterized by the acoustic phonons, or the acoustic wave, of a medium.

Other third-order processes listed in Table 9.5 do not cause optical frequency conversion but have the characteristic of inducing field-dependent changes in the optical properties of a material. These processes are characterized by either the real or imaginary part of a susceptibility of the form $\chi^{(3)}(\omega = \omega + \omega' - \omega')$. When $\omega' = \omega$, there can be only one beam in the interaction, as illustrated in Fig. 9.6(*a*), but there can

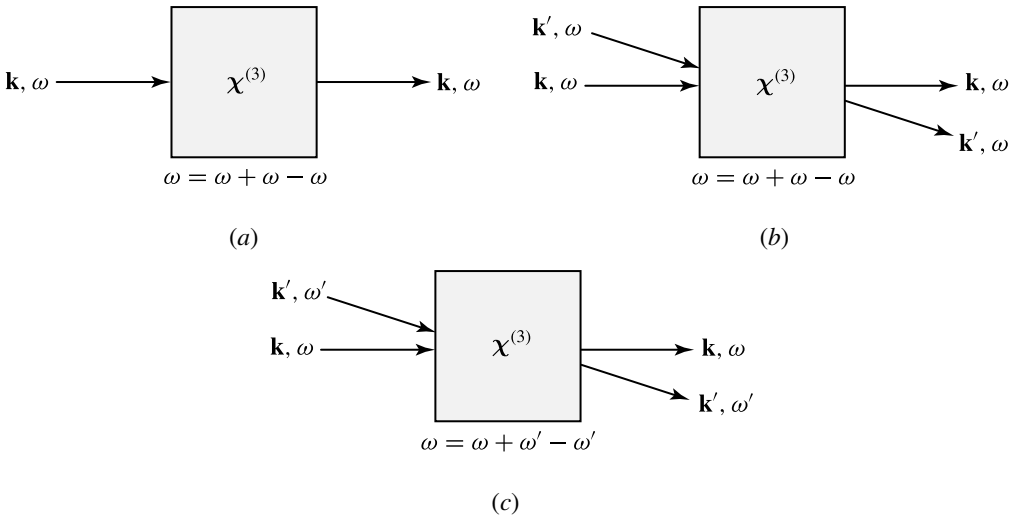


Figure 9.6 Third-order processes for field-induced susceptibility changes: (*a*) one-beam interaction, (*b*) interaction of two beams of the same frequency, and (*c*) interaction of two beams of different frequencies.

also be two physically distinguishable beams of the same frequency, as illustrated in Fig. 9.6(*b*). When $\omega' \neq \omega$, there are always two optical beams in the interaction, as illustrated in Fig. 9.6(*c*). In the case of one-beam interaction, we find that

$$P_i^{(3)}(\omega) = 3\epsilon_0 \sum_{j,k,l} \chi_{ijkl}^{(3)}(\omega = \omega + \omega - \omega) E_j(\omega) E_k(\omega) E_l^*(\omega) \tag{9.43}$$

using (9.22) and the intrinsic permutation symmetry. In the case of two-beam interaction, for either $\omega' = \omega$ or $\omega' \neq \omega$, we have

$$\begin{aligned} P_i^{(3)}(\omega) = {} & 3\epsilon_0 \sum_{j,k,l} \chi_{ijkl}^{(3)}(\omega = \omega + \omega - \omega) E_j(\omega) E_k(\omega) E_l^*(\omega) \\ & + 6\epsilon_0 \sum_{j,k,l} \chi_{ijkl}^{(3)}(\omega = \omega + \omega' - \omega') E_j(\omega) E_k(\omega') E_l^*(\omega') \end{aligned} \tag{9.44}$$

and a similar expression for $P_i^{(3)}(\omega')$ in terms of $\chi_{ijkl}^{(3)}(\omega' = \omega' + \omega' - \omega')$ and $\chi_{ijkl}^{(3)}(\omega' = \omega' + \omega - \omega)$. By identifying the total polarization at the frequency ω as $P_i(\omega) = P_i^{(1)}(\omega) + P_i^{(3)}(\omega)$, we find that the total *field-dependent permittivity tensor* can be expressed as

$$\epsilon_{ij}(\omega, \mathbf{E}) = \epsilon_{ij}(\omega) + \Delta\epsilon_{ij}(\omega, \mathbf{E}), \tag{9.45}$$

where $\epsilon_{ij}(\omega) = \epsilon_0 \left[1 + \chi_{ij}^{(1)}(\omega)\right]$ represents the field-independent linear permittivity tensor of the medium and $\Delta\epsilon_{ij}(\omega, \mathbf{E})$ accounts for the field-dependent change induced

by nonlinear optical interaction. For one-beam interaction,

$$\Delta\epsilon_{ij}(\omega, \mathbf{E}) = 3\epsilon_0 \sum_{k,l} \chi^{(3)}_{ijkl}(\omega = \omega + \omega - \omega) E_k(\omega) E^*_l(\omega). \tag{9.46}$$

For two-beam interaction,

$$\begin{aligned}\Delta\epsilon_{ij}(\omega, \mathbf{E}) = {} & 3\epsilon_0 \sum_{k,l} \chi^{(3)}_{ijkl}(\omega = \omega + \omega - \omega) E_k(\omega) E^*_l(\omega) \\ & + 6\epsilon_0 \sum_{k,l} \chi^{(3)}_{ijkl}(\omega = \omega + \omega' - \omega') E_k(\omega') E^*_l(\omega').\end{aligned} \tag{9.47}$$

The field-dependent permittivity of the form described here is the basis of many nonlinear optical phenomena that have important practical applications.

The nonlinear process discussed here generally leads to an *optical-field-induced birefringence* because $\Delta\epsilon_{ij}$ is a tensor. The simplest case involves a single linearly polarized optical wave in an isotropic medium with the optical field polarized in any fixed direction, or in a cubic crystal with the optical field polarized along one of the crystal axes. Then $\mathbf{P}^{(3)}$ is parallel to $\mathbf{E}$ of the optical field, and the only susceptibility element that contributes in this type of interaction is $\chi^{(3)}_{1111}(\omega = \omega + \omega - \omega)$. Therefore, the permittivity seen by the optical field is

$$\begin{aligned}\epsilon(\omega, \mathbf{E}) &= \epsilon(\omega) + 3\epsilon_0 \chi^{(3)}_{1111} |E(\omega)|^2 \\ &= \epsilon(\omega) + \frac{3\chi^{(3)}_{1111}}{2cn_0} I(\omega),\end{aligned} \tag{9.48}$$

where n_0 is the linear refractive index of the medium and $I(\omega)$ is the intensity of the optical beam. We see from this relation that the real part of $\chi^{(3)}_{1111}(\omega = \omega + \omega - \omega)$ leads to the following *intensity-dependent index of refraction*:[4]

$$n = n_0 + n_2 I(\omega), \tag{9.49}$$

where

$$n_2 = \frac{3\chi^{(3)'}_{1111}}{4c\epsilon_0 n_0^2} \ (\mathrm{m}^2\ \mathrm{W}^{-1}). \tag{9.50}$$

This intensity-dependent index of refraction represents the simplest case of the *optical Kerr effect*. Depending on the material properties and the experimental conditions, it leads to the phenomena of *self-phase modulation*, *self focusing*, and *self defocusing*.

The value of n_2 for a given material varies with optical wavelength, impurities, and temperature. In particular, it can be significantly enhanced by transition resonances

[4] In the literature, we sometimes see a different expression of $n = n_0 + 2n_2|E|^2$. The value of n_2 defined by this expression is accordingly different from that given in (9.50).

in a manner like the resonant enhancement of the linear refractive index. Its value also depends on its response speed because n_2 of a given material can be contributed by many different physical mechanisms that have different relaxation times. For example, the value of n_2 at room temperature of a semiconductor, such as GaAs or AlGaAs, can range from the order of 1×10^{-17} m^2 W^{-1} for wavelengths far away from the absorption bandgap to the order of 2×10^{-10} m^2 W^{-1} for wavelengths near the bandgap with exciton enhancement, and then further to the order of 1×10^{-8} m^2 W^{-1} with exciton enhancement near the band edge in GaAs/AlGaAs quantum wells.

EXAMPLE 9.5 Silica glass has an electronically contributed nonlinear susceptibility of $\chi^{(3)}_{1111}(\omega = \omega + \omega - \omega) = 1.8 \times 10^{-22}$ m^2 V^{-2} that causes an intensity-dependent index change in optical fibers. Its linear refractive index is $n_0 \approx 1.45$ in the visible and near infrared spectral regions of interest for most applications of optical fibers. Find the value of n_2 for silica fibers. For an ultrashort optical pulse that has a pulsewidth on the order of picoseconds or femtoseconds, the peak power can easily be a few kilowatts. Take a femtosecond pulse of a 10 kW peak power that propagates in a fiber of a 10 μm core diameter. What is the optical-field-induced index change seen by this pulse?

Solution Using (9.50), we find that

$$n_2 = \frac{3 \times 1.8 \times 10^{-22}}{4 \times 3 \times 10^8 \times 8.85 \times 10^{-12} \times 1.45^2} \text{ m}^2 \text{ W}^{-1} = 2.4 \times 10^{-20} \text{ m}^2 \text{ W}^{-1}.$$

Therefore, $n_2 = 2.4 \times 10^{-20}$ m^2 W^{-1} for silica fibers. The optical-field-induced index change seen by the pulse is

$$\Delta n = n_2 I = 2.4 \times 10^{-20} \times \frac{10 \times 10^3}{\pi \times (10 \times 10^{-6}/2)^2} = 3 \times 10^{-6}.$$

We see that the optical-field-induced index change is very small even for a pulse of 10 kW peak power confined in a very small fiber core to reach a very high intensity. Nevertheless, this small index change can have very significant effects on the characteristics of the optical pulse through the process of self-phase modulation, and on other pulses through cross-phase modulation. It is the root of such fascinating phenomena as *optical solitons* and *pulse spectral broadening*. The value of n_2 for optical fibers varies with the dopants in the fiber and with the optical wavelength. In the literature, the value of $n_2 = 3.2 \times 10^{-20}$ m^2 W^{-1} is often quoted for optical fibers. In Er-doped fibers, the value of n_2 at a resonance wavelength of 980 nm can be increased to the order of 1 to 3×10^{-15} m^2 W^{-1}, depending on the doping concentration, due to resonance enhancement.

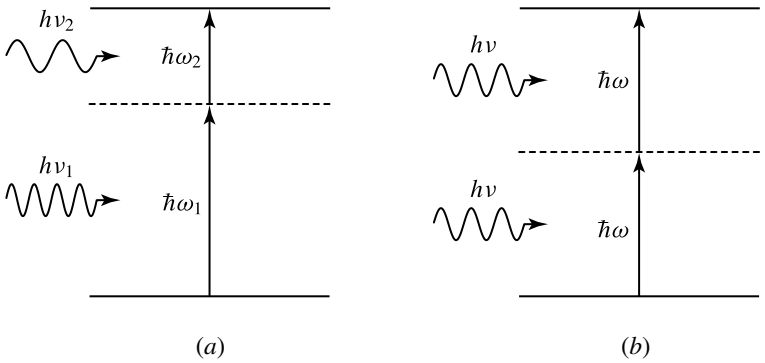


Figure 9.7 Resonant transitions for two-photon absorption at (*a*) $\omega_1 + \omega_2$ and (*b*) 2ω.

From (9.48), we also see that the imaginary part of the total susceptibility is

$$\chi'' = \chi^{(1)''} + \frac{3\chi^{(3)''}_{1111}}{2cn_0} I(\omega). \tag{9.51}$$

Therefore, the imaginary part of $\chi^{(3)}_{1111}(\omega = \omega + \omega - \omega)$ leads to an intensity-dependent change in the loss or gain of a medium. As a general rule, the sign of $\chi^{(3)''}_{1111}(\omega = \omega + \omega - \omega)$ that is contributed by a single-photon transition of a resonance frequency at or near ω is always the opposite of that of $\chi^{(1)''}(\omega)$. When $\chi^{(1)''} > 0$, the medium has a linear loss but $\chi^{(3)''}$ is negative and contributes to an intensity-dependent reduction of the loss, resulting in *absorption saturation*. When $\chi^{(1)''} < 0$, the medium has a gain. Then, $\chi^{(3)''}$ is positive and causes intensity-dependent *gain saturation*.

The characteristics of a nonparametric third-order process are determined by the resonant transition that is responsible for the process. In Table 9.5, we see that the susceptibility for *two-photon absorption* has the same form as that for stimulated Raman scattering. However, the resonance frequency of the transition for stimulated Raman scattering is the difference $\omega_p - \omega_S$, whereas that for two-photon absorption is the sum $\omega_1 + \omega_2$, as shown in Fig. 9.7(*a*). In the special case when $\omega_1 = \omega_2 = \omega$, both the susceptibility for two-photon absorption and that for absorption saturation or gain saturation have the form $\chi^{(3)''}(\omega = \omega + \omega - \omega)$. The difference between them is that the nonlinear susceptibility responsible for absorption saturation or gain saturation is resonant at ω but that for two-photon absorption is resonant at 2ω, as shown in Fig. 9.7(*b*).

A nonlinear process that characterizes the interaction of four optical waves is generally referred to as *four-wave mixing*. In nondegenerate four-wave mixing, all four optical waves have different frequencies. The process becomes partially degenerate if there are only two or three distinct frequencies. In *degenerate four-wave mixing*, all of the participating waves have the same frequency. With the exception of the electro-optic Kerr effect, all of the third-order nonlinear processes can be described as four-wave mixing processes.

9.4 Coupled-wave analysis

The coupled-wave theory developed in Section 4.1 is used for analysis of nonlinear optical interactions in a homogeneous medium. In nonlinear optical waveguides, coupled-mode theory can be used if there is no mixing between different optical frequencies, but a combination of coupled-wave and coupled-mode formalisms has to be used if there is frequency mixing in the interaction. In applying coupled-wave or coupled-mode theory to the analysis of a nonlinear interaction, the perturbing polarization, generally expressed as $\Delta\mathbf{P}$ in Chapter 4, is identified as the characteristic nonlinear polarization, $\mathbf{P}^{(n)}$, of the specific interaction.

A nonlinear optical interaction often takes place in an anisotropic crystal, as can be expected from the fact that $\chi^{(2)}$ vanishes identically in the bulk of an isotropic medium under the electric-dipole approximation. Even when a nonlinear interaction takes place in an isotropic medium, a longitudinal field component can sometimes be generated because of a field-dependent birefringence induced by a third-order nonlinear process such as the optical Kerr effect discussed above. For these reasons, the correct coupled-wave equation to be used, under the slowly varying amplitude approximation, for the analysis of nonlinear optical interactions is the one in (4.18):

$$(\mathbf{k}_q \cdot \boldsymbol{\nabla})\boldsymbol{\mathcal{E}}_{q,\mathrm{T}} \approx \frac{\mathrm{i}\omega_q^2\mu_0}{2}\mathbf{P}_{q,\mathrm{T}}^{(n)}\mathrm{e}^{-\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}}, \tag{9.52}$$

where $\mathbf{P}_q^{(n)}$ is identified with $\mathbf{P}_q^{(2)}$, or $\mathbf{P}^{(2)}(\omega_q)$ of (9.19), for a second-order process and with $\mathbf{P}_q^{(3)}$, or $\mathbf{P}^{(3)}(\omega_q)$ of (9.20), for a third-order process and, according to (9.14) and (9.16), the field amplitude $\boldsymbol{\mathcal{E}}_q$ is defined by the following relation:

$$\mathbf{E}(\omega_q) = \mathbf{E}_q(\mathbf{r}) = \boldsymbol{\mathcal{E}}_q(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}} = \hat{e}_q\mathcal{E}_q(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}}. \tag{9.53}$$

In most cases of interest, the amplitudes of all of the interacting waves vary along the same direction, which is designated the z direction. Then, the coupled-wave equation can be written as

$$\frac{\mathrm{d}\boldsymbol{\mathcal{E}}_{q,\mathrm{T}}(z)}{\mathrm{d}z} \approx \frac{\mathrm{i}\omega_q^2\mu_0}{2k_{q,z}}\mathbf{P}_{q,\mathrm{T}}^{(n)}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}}. \tag{9.54}$$

Note that the propagation direction, which is the direction normal to the wavefront and is defined by the wavevector $\mathbf{k}$, of each wave is not necessarily the same as the direction along which the field amplitude varies. In general, the nonlinear polarization $\mathbf{P}_q^{(n)}$ may also have variations along other directions.

Except in some unusual cases, the longitudinal field components of the interacting optical waves are small and unimportant though they may exist. Then, $\mathbf{P}_{q,\mathrm{T}}^{(n)}$ in (9.52) and (9.54) can be replaced by $\mathbf{P}_q^{(n)}$, further simplifying the coupled-wave equation. When

this simplification is done, we can multiply both sides of (9.54) by the unit vector $\hat{e}_q^*$ to write the coupled-wave equation as

$$\frac{\mathrm{d}\mathcal{E}_q}{\mathrm{d}z} = \frac{\mathrm{i}\omega_q^2\mu_0}{2k_{q,z}}\hat{e}_q^* \cdot \mathbf{P}_q^{(n)}\mathrm{e}^{-\mathrm{i}\mathbf{k}_q\cdot\mathbf{r}}. \tag{9.55}$$

In the analysis of a nonlinear optical interaction, a coupled-wave equation is written for each of the interacting waves. The nonlinear polarization on the right-hand side of each equation couples the equations for different waves, resulting in an array of coupled nonlinear equations.

Parametric interactions

To illustrate several important concepts using a concrete example, we consider the coupled equations that describe a parametric second-order interaction of three different frequencies ω_1, ω_2, and ω_3 with the relation $\omega_3 = \omega_1 + \omega_2$. We also take the approximations that allow us to use (9.55). Using (9.19) and the intrinsic permutation symmetry, we find that

$$\hat{e}_3^* \cdot \mathbf{P}_3^{(2)} = 2\epsilon_0\hat{e}_3^* \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_1\hat{e}_2\mathcal{E}_1\mathcal{E}_2\mathrm{e}^{\mathrm{i}(\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r}}, \tag{9.56}$$

$$\hat{e}_1^* \cdot \mathbf{P}_1^{(2)} = 2\epsilon_0\hat{e}_1^* \cdot \boldsymbol{\chi}^{(2)}(\omega_1 = \omega_3 - \omega_2) : \hat{e}_3\hat{e}_2^*\mathcal{E}_3\mathcal{E}_2^*\mathrm{e}^{\mathrm{i}(\mathbf{k}_3-\mathbf{k}_2)\cdot\mathbf{r}}, \tag{9.57}$$

$$\hat{e}_2^* \cdot \mathbf{P}_2^{(2)} = 2\epsilon_0\hat{e}_2^* \cdot \boldsymbol{\chi}^{(2)}(\omega_2 = \omega_3 - \omega_1) : \hat{e}_3\hat{e}_1^*\mathcal{E}_3\mathcal{E}_1^*\mathrm{e}^{\mathrm{i}(\mathbf{k}_3-\mathbf{k}_1)\cdot\mathbf{r}}. \tag{9.58}$$

The full permutation symmetry is valid for the real $\boldsymbol{\chi}^{(2)}$ that characterizes the parametric process. Therefore, we can define an *effective nonlinear susceptibility* as

$$\begin{aligned}\chi_{\mathrm{eff}} &= \hat{e}_3^* \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_1\hat{e}_2 \\ &= \hat{e}_1 \cdot \boldsymbol{\chi}^{(2)}(\omega_1 = \omega_3 - \omega_2) : \hat{e}_3^*\hat{e}_2 \\ &= \hat{e}_2 \cdot \boldsymbol{\chi}^{(2)}(\omega_2 = \omega_3 - \omega_1) : \hat{e}_3^*\hat{e}_1.\end{aligned} \tag{9.59}$$

Following the relation given in (9.34), the effective d coefficient for this interaction is simply $d_{\mathrm{eff}} = \chi_{\mathrm{eff}}/2$. We have the following coupled equations for a parametric second-order interaction:

$$\frac{\mathrm{d}\mathcal{E}_3}{\mathrm{d}z} = \frac{\mathrm{i}\omega_3^2}{c^2k_{3,z}}\chi_{\mathrm{eff}}\mathcal{E}_1\mathcal{E}_2\mathrm{e}^{\mathrm{i}\Delta kz}, \tag{9.60}$$

$$\frac{\mathrm{d}\mathcal{E}_1}{\mathrm{d}z} = \frac{\mathrm{i}\omega_1^2}{c^2k_{1,z}}\chi_{\mathrm{eff}}^*\mathcal{E}_3\mathcal{E}_2^*\mathrm{e}^{-\mathrm{i}\Delta kz}, \tag{9.61}$$

$$\frac{\mathrm{d}\mathcal{E}_2}{\mathrm{d}z} = \frac{\mathrm{i}\omega_2^2}{c^2k_{2,z}}\chi_{\mathrm{eff}}^*\mathcal{E}_3\mathcal{E}_1^*\mathrm{e}^{-\mathrm{i}\Delta kz}, \tag{9.62}$$

where $\Delta\mathbf{k} = \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3 = \Delta k\hat{z}$ is the phase mismatch. For linearly polarized waves, we have $\hat{e}^* = \hat{e}$, and $\chi_{\mathrm{eff}}^* = \chi_{\mathrm{eff}}$ is a real quantity. Otherwise, χ_{eff} can be complex.

EXAMPLE 9.6 Take $\hat{x}$, $\hat{y}$, and $\hat{z}$ to be the principal axes of $LiNbO_3$. Find the effective nonlinear susceptibilities for second-harmonic generation in $LiNbO_3$ with (a) a linearly x-polarized fundamental wave propagating in the z direction and (b) a linearly z-polarized fundamental wave propagating in the x direction. The propagation direction of the second harmonic is in practice determined by many factors. Here we make it the same as that of the fundamental wave.

Solution For second-harmonic generation, we have the degenerate case of $\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$. From Table 9.3, we find that the only nonvanishing second-order nonlinear susceptibility elements of $LiNbO_3$ are $d_{31} = d_{32} = d_{24} = d_{15} = -4.4$ pm V^{-1}, $d_{22} = -d_{21} = -d_{16} = 2.4$ pm V^{-1}, and $d_{33} = -25.2$ pm V^{-1}. We also know that $\chi^{(2)}_{i\alpha} = 2d_{i\alpha}$, according to (9.34).

In case (a), we have $\mathbf{k}_{2\omega} \parallel \mathbf{k}_\omega \parallel \hat{z}$ and $\hat{e}_\omega = \hat{x}$. We then find from the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$ for $LiNbO_3$ that $\mathbf{P}^{(2)}(2\omega)$ has only two components in the y and z directions contributed by $\chi^{(2)}_{21}$ and $\chi^{(2)}_{31}$, respectively. However, because $\mathbf{k}_{2\omega}$ is forced to be in the z direction, $P^{(2)}_z$ is a longitudinal component that cannot contribute to the propagation of the second-harmonic wave. In this situation, $\hat{e}_{2\omega} = \hat{y}$ because only the transverse component $P^{(2)}_y$ is useful for generating the second-harmonic wave. Therefore,

$$\chi_{\text{eff}} = \hat{e}^*_{2\omega} \cdot \boldsymbol{\chi}^{(2)} : \hat{e}_\omega \hat{e}_\omega = \hat{y} \cdot \boldsymbol{\chi}^{(2)} : \hat{x}\hat{x} = \chi^{(2)}_{21} = 2d_{21} = -4.8 \text{ pm V}^{-1}.$$

In case (b), we have $\mathbf{k}_{2\omega} \parallel \mathbf{k}_\omega \parallel \hat{x}$ and $\hat{e}_\omega = \hat{z}$. We find from the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$ for $LiNbO_3$ that $\mathbf{P}^{(2)}(2\omega)$ has only one component in the z direction contributed by $\chi^{(2)}_{33}$. Therefore, $\hat{e}_{2\omega} = \hat{z}$ and

$$\chi_{\text{eff}} = \hat{e}^*_{2\omega} \cdot \boldsymbol{\chi}^{(2)} : \hat{e}_\omega \hat{e}_\omega = \hat{z} \cdot \boldsymbol{\chi}^{(2)} : \hat{z}\hat{z} = \chi^{(2)}_{33} = 2d_{33} = -50.4 \text{ pm V}^{-1}.$$

We see from this example that the value of χ_{eff} can vary significantly depending on the polarization directions of the interacting waves, which in turn are constrained by the wave propagation directions.

Note that though $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ individually may not be parallel to $\hat{z}$, the phase mismatch $\Delta\mathbf{k}$ has to be parallel to $\hat{z}$ if the field amplitudes are to vary only along the z direction. This fact is required mathematically in (9.60)–(9.62) because $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ are all functions of z only. Physically, the boundary conditions, which are dictated by Maxwell's equations, at the surface of a nonlinear crystal where the input waves enter the crystal require that the tangential component, but not the normal component, of $\mathbf{k}_1 + \mathbf{k}_2$ be equal to that of $\mathbf{k}_3$ for an interaction defined by the relation $\omega_3 = \omega_1 + \omega_2$. Therefore, any phase mismatch $\Delta\mathbf{k}$ occurs only in the direction normal to the input surface of the nonlinear crystal, as illustrated in Fig. 9.8(a). This condition can always be satisfied because only one or two of the interacting waves are provided at the input and

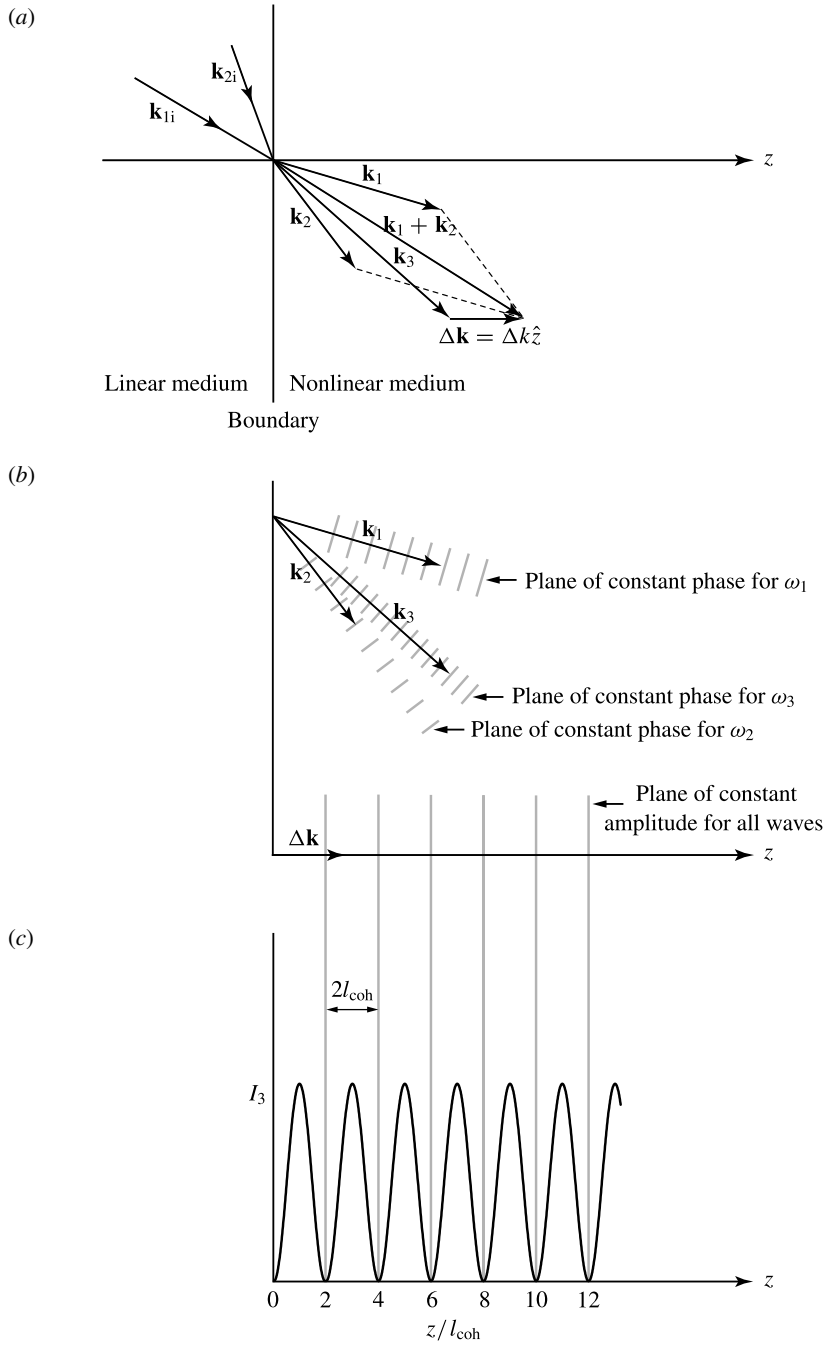


Figure 9.8 (*a*) In a parametric interaction, the boundary conditions at the input surface of a nonlinear crystal require that the phase mismatch $\Delta\mathbf{k}$, as well as the z direction, along which the field amplitudes vary, be normal to the input surface. (*b*) The wavefront, defined by the plane of constant phase, of each wave is normal to its wavevector, but the plane of constant field amplitude is parallel to the input surface and is normal to $\hat{z}$. (*c*) Periodic variation of the intensity of a nonlinearly generated wave in the presence of a phase mismatch.

only their wavevectors are initially given. For example, in sum-frequency generation, $\mathbf{k}_1$ and $\mathbf{k}_2$ are determined by the propagation directions of the input waves at ω_1 and ω_2, respectively. The propagation direction, $\mathbf{k}_3$, of the generated sum-frequency wave is then determined by two conditions: (1) its magnitude, $k_3 = \omega_3 n_3/c$, is determined by the dispersion and birefringence of the nonlinear crystal; (2) its projection on the crystal surface has to be equal to the projection of $\mathbf{k}_1 + \mathbf{k}_2$ on the crystal surface, as Fig. 9.8(a) shows. Because $\Delta\mathbf{k} = \Delta k\hat{z}$, the z direction, along which the field amplitudes vary, is normal to the input surface of the nonlinear crystal. Figure 9.8(b) shows the fact that though the Poynting vector of each wave may not line up with $\hat{z}$, its magnitude varies only along the z direction.

The intensity of a wave at a frequency ω_q, projected on the plane of constant amplitude that is normal to the z direction, is given by

$$I_q = |\overline{\boldsymbol{S}}_q \cdot \hat{z}| \approx \frac{2k_{q,z}}{\omega_q \mu_0}|\mathcal{E}_q|^2 = 2c\epsilon_0 n_{q,z}|\mathcal{E}_q|^2, \tag{9.63}$$

where $\boldsymbol{S}_q$ is the Poynting vector, $n_{q,z} = ck_{q,z}/\omega_q = n_q \cos\theta_q$, and θ_q is the angle between $\mathbf{k}_q$ and $\hat{z}$. In a birefringent crystal, a possible walk-off between the vectors $\boldsymbol{S}_q$ and $\mathbf{k}_q$ is neglected by taking the approximation in (9.63). We find that (9.60)–(9.62) lead to

$$\frac{\mathrm{d}I_3}{\mathrm{d}z} = -\frac{2\omega_3|\chi_{\mathrm{eff}}|}{(2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z})^{1/2}} I_1^{1/2} I_2^{1/2} I_3^{1/2} \sin\varphi, \tag{9.64}$$

$$\frac{\mathrm{d}I_1}{\mathrm{d}z} = \frac{2\omega_1|\chi_{\mathrm{eff}}|}{(2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z})^{1/2}} I_1^{1/2} I_2^{1/2} I_3^{1/2} \sin\varphi, \tag{9.65}$$

$$\frac{\mathrm{d}I_2}{\mathrm{d}z} = \frac{2\omega_2|\chi_{\mathrm{eff}}|}{(2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z})^{1/2}} I_1^{1/2} I_2^{1/2} I_3^{1/2} \sin\varphi, \tag{9.66}$$

where

$$\varphi = \varphi_\chi + \varphi_1 + \varphi_2 - \varphi_3 + \Delta kz, \tag{9.67}$$

φ_χ is the phase of χ_{eff} defined as $\chi_{\mathrm{eff}} = |\chi_{\mathrm{eff}}|\mathrm{e}^{\mathrm{i}\varphi_\chi}$, and φ_1, φ_2, and φ_3 are the phases of $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, respectively, defined as $\mathcal{E}_q = |\mathcal{E}_q|\mathrm{e}^{\mathrm{i}\varphi_q}$.

The total intensity in the three interacting waves is $I = I_1 + I_2 + I_3$. Using the relation $\omega_3 = \omega_1 + \omega_2$, we find from (9.64)–(9.66) that

$$\frac{\mathrm{d}I}{\mathrm{d}z} = \frac{\mathrm{d}(I_1 + I_2 + I_3)}{\mathrm{d}z} = 0. \tag{9.68}$$

Consequently, the total optical energy is conserved in a parametric process, as is expected. In addition, we also find that

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_1}{\omega_1}\right) = \frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_2}{\omega_2}\right) = -\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_3}{\omega_3}\right). \tag{9.69}$$

Therefore, every time a photon at ω_3 is annihilated, one photon at ω_1 and another at ω_2 are generated simultaneously, and vice versa. The relations in (9.68) and (9.69) are known as the *Manley–Rowe relations*.

The coupled equations and the Manley–Rowe relations formulated above apply to all parametric second-order interactions that involve three different optical frequencies. We see from (9.64)–(9.66) that optical energy is converted from ω_3 to ω_1 and ω_2 if $\sin\varphi > 0$, whereas it is converted from ω_1 and ω_2 to ω_3 if $\sin\varphi < 0$. If φ_χ, φ_1, φ_2, and φ_3 are fixed or vary slowly with z, as is normally the case, then $\sin\varphi$ changes sign periodically with z because of the phase mismatch Δk. This periodic change of sign in $\sin\varphi$ results in periodic reversal of a parametric process. Therefore, in the presence of a phase mismatch, the maximum interaction length a given frequency-conversion process can take without a reversal of the process is the *coherence length*:

$$l_{\text{coh}} = \frac{\pi}{|\Delta k|}. \tag{9.70}$$

From the above discussions and an examination of (9.64)–(9.66), we can see that the intensity of a wave generated by a parametric nonlinear process in the presence of a finite phase mismatch varies periodically along the direction normal to the input surface of the nonlinear crystal with a half period of l_{coh}, as illustrated in Fig. 9.8(*c*). The intensities of other waves in the interaction also vary with the same period along the z direction. Therefore, an interaction length larger than l_{coh} is not useful and can even be detrimental. Clearly, phase matching is very important for an efficient parametric interaction.

EXAMPLE 9.7 With a fundamental wave at $\lambda = 1.10$ μm, find the coherence length for each of the two cases of second-harmonic generation in $LiNbO_3$ discussed in Example 9.6. At room temperature, $LiNbO_3$ has $n_o = 2.2319$ and $n_e = 2.1536$ at 1.10 μm wavelength and $n_o = 2.3168$ and $n_e = 2.2260$ at 550 nm wavelength.

Solution Because $\mathbf{k}_{2\omega} \parallel \mathbf{k}_\omega$, we have $\Delta k = 2k_\omega - k_{2\omega} = 4\pi(n_\omega - n_{2\omega})/\lambda$. In case (a), we have $n_\omega = n_o(\omega) = 2.2319$ and $n_{2\omega} = n_o(2\omega) = 2.3168$ because $\hat{e}_\omega = \hat{x}$ and $\hat{e}_{2\omega} = \hat{y}$. Then,

$$l_{\text{coh}} = \frac{\pi}{|\Delta k|} = \frac{\lambda}{4 \times |n_\omega^{\text{o}} - n_{2\omega}^{\text{o}}|} = \frac{1.1}{4 \times 0.0849}\ \mu\text{m} = 3.24\ \mu\text{m}.$$

In case (b), we have $n_\omega = n_e(\omega) = 2.1536$ and $n_{2\omega} = n_e(2\omega) = 2.2260$ because $\hat{e}_\omega = \hat{z}$ and $\hat{e}_{2\omega} = \hat{z}$. Then,

$$l_{\text{coh}} = \frac{\pi}{|\Delta k|} = \frac{\lambda}{4 \times |n_\omega^{\text{e}} - n_{2\omega}^{\text{e}}|} = \frac{1.1}{4 \times 0.0724}\ \mu\text{m} = 3.80\ \mu\text{m}.$$

We see from this example that the coherence length is very small for both cases. Clearly, the interaction would not be efficient. The reason for this undesirable situation

is that we have arbitrarily chosen in Example 9.6 some convenient propagation and polarization directions for the optical waves involved in the second-harmonic generation process without any consideration of the requirement for phase matching. These two examples together illustrate that it is possible to obtain a decent value of $|\chi_{\text{eff}}|$, while the interaction is still very inefficient because of phase mismatch. Methods for phase matching are discussed in Section 9.5. Phase-matched second-harmonic generation processes in $LiNbO_3$ with properly chosen propagation and polarization directions of the optical waves are demonstrated in Examples 9.8 and 9.9.

For a parametric interaction among linearly polarized waves in a homogeneous bulk crystal, $\varphi_\chi = 0$ or π, depending on the sign of χ_{eff}. Then, $\varphi = \varphi_1 + \varphi_2 - \varphi_3$ or $\varphi = \pi + \varphi_1 + \varphi_2 - \varphi_3$ in the case of perfect phase matching. In this situation, it is possible for a frequency-conversion process to continue over the entire length of a crystal. Which parametric process occurs is determined completely by the value of φ. For sum-frequency generation, we need $\varphi = -\pi/2$ so that optical energy is converted most efficiently in the direction $\omega_1 + \omega_2 \rightarrow \omega_3$. For difference-frequency generation, optical parametric amplification, or optical parametric generation, $\varphi = \pi/2$ is needed to have the highest efficiency for the conversion of optical energy in the direction $\omega_3 \rightarrow \omega_1 + \omega_2$.

In real experimental settings, a desired process is controlled by the input conditions. Normally only one or two waves in a parametric three-wave interaction are supplied at the input; therefore, only one or two phases are set, and at least one phase is arbitrary. Consider the situation where χ_{eff} is real and positive so that $\varphi_\chi = 0$. If the input waves are at ω_1 and ω_2 and the phase-matching condition $\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2$ is satisfied, sum-frequency generation occurs with the generation of a wave at ω_3 that automatically picks a phase of $\varphi_3 = \varphi_1 + \varphi_2 + \pi/2$. If the same phase-matching condition is satisfied but the input waves are at ω_3 and ω_2, a wave at ω_1 is generated with a phase of $\varphi_1 = \varphi_3 - \varphi_2 + \pi/2$, resulting in difference-frequency generation, or optical parametric amplification in the case when the amplification of the signal at ω_2 is the objective. If only a wave at ω_3 is supplied at the input, optical parametric generation occurs with $\varphi_1 + \varphi_2 = \varphi_3 + \pi/2$. In this situation, the values of ω_1 and ω_2 are determined by the phase-matching condition subject to the condition that $\omega_3 = \omega_1 + \omega_2$.

An interesting question is whether it is possible for other parametric processes, such as sum-frequency generation for $\omega_1 + \omega_3$ and difference-frequency generation for $\omega_1 - \omega_2$, and so on, to occur once all three waves at ω_1, ω_2, and ω_3 exist in a crystal, say, through a sum-frequency generation process of $\omega_1 + \omega_2 \rightarrow \omega_3$. From the above discussions, it is clear that any parametric process can occur if it (1) has a nonvanishing χ_{eff}, (2) is phase matched, and (3) has the correct initial value of the phase φ. It is thus possible to have simultaneous multiple parametric processes if all of them satisfy the required conditions. In normal situations, however, it is highly unusual for two or more different processes to occur in a single experimental arrangement

because of the difficulty of satisfing their respective phase-matching conditions all at once.

Nonparametric interactions

When writing the coupled-wave equations for any nonlinear process, it is important to clearly understand the properties of the nonlinear susceptibility that characterizes the process under consideration first. For a parametric process, the full permutation symmetry is valid for the susceptibility; this fact is used in defining the effective susceptibility given in (9.59). In general, the susceptibility for a nonparametric process does not have the full permutation symmetry because its imaginary part is significant for the process. The susceptibilities for different nonparametric processes generally have different permutation properties because they are related to different resonant transitions in the material. For example, the susceptibility, $\boldsymbol{\chi}^{(3)}(\omega_S = \omega_S + \omega_p - \omega_p)$, for the Stokes process of stimulated Raman scattering and the susceptibility, $\boldsymbol{\chi}^{(3)}(\omega_1 = \omega_1 + \omega_2 - \omega_2)$, for two-photon absorption look the same, but they have different microscopic forms and thus very different properties because the former is resonant at $\omega_p - \omega_S$ while the latter is resonant at $\omega_1 + \omega_2$. If ω_p, ω_S, and $\omega_p + \omega_S$ are all far from any resonant transition frequencies while $\omega_p - \omega_S$ is in resonance with a transition in the material, the following property applies to the Raman susceptibility:

$$\begin{aligned}\chi^{(3)}_{ijkl}(\omega_S = \omega_S + \omega_p - \omega_p) &= \chi^{(3)^*}_{klij}(\omega_p = \omega_p + \omega_S - \omega_S)\\ &= \chi^{(3)}_{jilk}(\omega_S = \omega_S + \omega_p - \omega_p)\\ &= \chi^{(3)^*}_{lkji}(\omega_p = \omega_p + \omega_S - \omega_S).\end{aligned} \tag{9.71}$$

In contrast, the susceptibility for two-photon absorption has the following property:

$$\begin{aligned}\chi^{(3)}_{ijkl}(\omega_1 = \omega_1 + \omega_2 - \omega_2) &= \chi^{(3)}_{klij}(\omega_2 = \omega_2 + \omega_1 - \omega_1)\\ &= \chi^{(3)}_{jilk}(\omega_1 = \omega_1 + \omega_2 - \omega_2)\\ &= \chi^{(3)}_{lkji}(\omega_2 = \omega_2 + \omega_1 - \omega_1)\end{aligned} \tag{9.72}$$

if ω_1, ω_2, and $|\omega_1 - \omega_2|$ are all far from any resonant transition frequencies while $\omega_1 + \omega_2$ is in resonance with a transition. The difference between the relations in (9.71) and (9.72) is significant because the imaginary parts of these susceptibilities are responsible for the nonlinear processes under consideration.

The coupled-wave equations for the process of stimulated Raman scattering are considered. Using (9.20) and the intrinsic permutation symmetry, we can write

$$\hat{e}^*_S \cdot \mathbf{P}^{(3)}_S = 6\epsilon_0 \hat{e}^*_S \cdot \boldsymbol{\chi}^{(3)}(\omega_S = \omega_S + \omega_p - \omega_p) \vdots \hat{e}_S \hat{e}_p \hat{e}^*_p \mathcal{E}_S |\mathcal{E}_p|^2 e^{i\mathbf{k}_S \cdot \mathbf{r}}, \tag{9.73}$$

$$\hat{e}^*_p \cdot \mathbf{P}^{(3)}_p = 6\epsilon_0 \hat{e}^*_p \cdot \boldsymbol{\chi}^{(3)}(\omega_p = \omega_p + \omega_S - \omega_S) \vdots \hat{e}_p \hat{e}_S \hat{e}^*_S \mathcal{E}_p |\mathcal{E}_S|^2 e^{i\mathbf{k}_p \cdot \mathbf{r}}. \tag{9.74}$$

Applying the relation in (9.71), we can define an effective Raman susceptibility:

$$\chi_{\mathrm{R}} = \hat{e}_{\mathrm{S}}^{*} \cdot \boldsymbol{\chi}^{(3)}(\omega_{\mathrm{S}} = \omega_{\mathrm{S}} + \omega_{\mathrm{p}} - \omega_{\mathrm{p}}) \vdots \hat{e}_{\mathrm{S}} \hat{e}_{\mathrm{p}} \hat{e}_{\mathrm{p}}^{*}$$
$$= \hat{e}_{\mathrm{p}} \cdot \boldsymbol{\chi}^{(3)*}(\omega_{\mathrm{p}} = \omega_{\mathrm{p}} + \omega_{\mathrm{S}} - \omega_{\mathrm{S}}) \vdots \hat{e}_{\mathrm{p}}^{*} \hat{e}_{\mathrm{S}}^{*} \hat{e}_{\mathrm{S}}. \tag{9.75}$$

The coupled-wave equations for stimulated Raman scattering are

$$\frac{\mathrm{d}\mathcal{E}_{\mathrm{S}}}{\mathrm{d}z} = \frac{\mathrm{i}3\omega_{\mathrm{S}}^{2}}{c^{2}k_{\mathrm{S},z}} \chi_{\mathrm{R}} \mathcal{E}_{\mathrm{S}} |\mathcal{E}_{\mathrm{p}}|^{2}, \tag{9.76}$$

$$\frac{\mathrm{d}\mathcal{E}_{\mathrm{p}}}{\mathrm{d}z} = \frac{\mathrm{i}3\omega_{\mathrm{p}}^{2}}{c^{2}k_{\mathrm{p},z}} \chi_{\mathrm{R}}^{*} \mathcal{E}_{\mathrm{p}} |\mathcal{E}_{\mathrm{S}}|^{2}. \tag{9.77}$$

By comparing these two equations to the three equations given in (9.60)–(9.62) for the parametric second-order interaction, we see clearly that the nonparametric process of stimulated Raman scattering is automatically phase matched, as is discussed in the preceding section.

The relation in (9.63) can be used to transform (9.76) and (9.77) into

$$\frac{\mathrm{d}I_{\mathrm{S}}}{\mathrm{d}z} = -\frac{3\omega_{\mathrm{S}}\mu_{0}}{n_{\mathrm{S},z}n_{\mathrm{p},z}} \chi_{\mathrm{R}}'' I_{\mathrm{S}} I_{\mathrm{p}}, \tag{9.78}$$

$$\frac{\mathrm{d}I_{\mathrm{p}}}{\mathrm{d}z} = \frac{3\omega_{\mathrm{p}}\mu_{0}}{n_{\mathrm{S},z}n_{\mathrm{p},z}} \chi_{\mathrm{R}}'' I_{\mathrm{S}} I_{\mathrm{p}}. \tag{9.79}$$

We find that the total light intensity, $I = I_{\mathrm{S}} + I_{\mathrm{p}}$, varies as

$$\frac{\mathrm{d}I}{\mathrm{d}z} = \frac{\mathrm{d}(I_{\mathrm{S}} + I_{\mathrm{p}})}{\mathrm{d}z} = \frac{3\mu_{0}}{n_{\mathrm{S},z}n_{\mathrm{p},z}} (\omega_{\mathrm{p}} - \omega_{\mathrm{S}}) \chi_{\mathrm{R}}'' I_{\mathrm{S}} I_{\mathrm{p}}. \tag{9.80}$$

Therefore, optical energy is not conserved in the nonparametric Raman process because there is energy exchange with the material due to the resonant transition at the frequency $\Omega = \omega_{\mathrm{p}} - \omega_{\mathrm{S}}$. Nevertheless, one Stokes photon is created for every pump photon that is annihilated. Therefore, in the absence of other loss mechanisms, we still have the following Manley–Rowe relation:

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_{\mathrm{S}}}{\omega_{\mathrm{S}}}\right) = -\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_{\mathrm{p}}}{\omega_{\mathrm{p}}}\right). \tag{9.81}$$

We see from (9.78) and (9.79) that the direction of energy flow in the Raman process is determined by the sign of χ_{R}'', which depends on the state of the material. If the material is in the ground state of the Raman transition, the imaginary part of $\boldsymbol{\chi}^{(3)}(\omega_{\mathrm{S}} = \omega_{\mathrm{S}} + \omega_{\mathrm{p}} - \omega_{\mathrm{p}})$ is negative, resulting in $\chi_{\mathrm{R}}'' < 0$ according to (9.75). In this situation, energy is converted from the pump wave to the Stokes wave. We also see from (9.80) that in a Stokes process, there is a net loss in the total optical intensity. The energy corresponding to this loss is absorbed by the material in making the Stokes Raman transition from the ground state to the excited state. In case the excited state of the Raman transition is more

populated than the ground state, the imaginary part of $\chi^{(3)}(\omega_S = \omega_S + \omega_p - \omega_p)$ for $\Omega = \omega_p - \omega_S$ becomes positive. Then $\chi_R'' > 0$. In this situation, the anti-Stokes process occurs with the wave at ω_S acting as the pump wave and the wave at ω_p acting as the anti-Stokes wave. In the anti-Stokes process, the total optical intensity has a net gain corresponding to the energy released by the material in making the anti-Stokes Raman transition from the excited state to the ground state.

From the above discussions, we see that the characteristics of a nonparametric process are completely determined by the state of the material. Phase matching for the interacting optical waves is automatically satisfied in a nonparametric process because any difference in the momenta of the interacting photons can be absorbed by the material if there is energy exchange between the optical field and the medium. For the same reason, the phase relationship among the interacting waves, which determines the direction of frequency conversion in a parametric process, plays no role in a nonparametric process.

9.5 Phase matching

We have seen in the preceding section the importance of phase matching for parametric nonlinear processes. If a parametric interaction is phase matched, optical power can be converted efficiently from one frequency to another. Otherwise, the process is periodically reversed, and the optical power shuttles back and forth among the interacting waves, as Fig. 9.8(*c*) shows. No matter how long the crystal is, the best efficiency we can expect from a parametric interaction that is not perfectly phase matched is that contributed by the interaction over a coherence length. Therefore, phase matching is one of the most important technical issues that have to be addressed in designing any efficient nonlinear optical device based on a parametric frequency conversion process. In this section, we discuss phase matching for second-order nonlinear optical processes.

The phase-matching condition of a nonlinear optical process is a relation among the wavevectors of the interacting waves. When more than two different wavevectors are involved, phase matching can be either *collinear* or *noncollinear*, as illustrated in Figs. 9.9(*a*) and (*b*) for a second-order process. All of the wavevectors are parallel to one another in collinear phase matching, but they are not in noncollinear phase matching. In the case of second-harmonic generation with only one input fundamental wave, phase matching is always collinear because there are only two wavevectors, $\mathbf{k}_\omega$ and $\mathbf{k}_{2\omega}$, involved in the process, as shown in Fig. 9.9(*c*). However, noncollinear phase matching for second-harmonic generation is also possible if there are two spatially separated fundamental waves at the input, as shown in Fig. 9.9(*d*). In the latter situation, the wavevectors of the two distinct fundamental waves can have different magnitudes if the two waves are not polarized in the same direction inside a birefringent crystal.

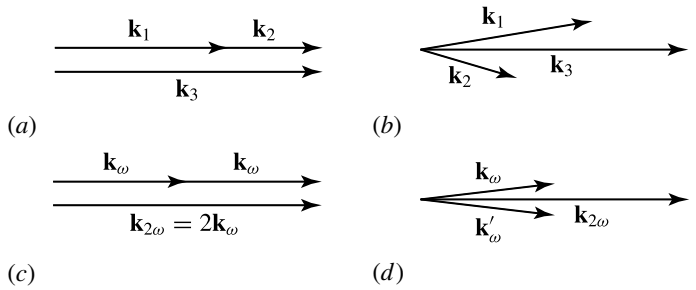


Figure 9.9 (*a*) Collinear and (*b*) noncollinear phase matching for a second-order process with the phase-matching condition: $\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2$. (*c*) Collinear and (*d*) noncollinear phase matching for second-harmonic generation.

With noncollinear phase matching, the interaction length is limited by the finite distance over which the beams overlap in space. Therefore, the collinear phase-matching arrangement is employed in most nonlinear optical devices though noncollinear phase matching is also useful in some special applications. In the following, we consider only collinear phase matching for second-order nonlinear processes, but the general concepts can be easily extended to noncollinear phase matching. Phase matching for the third-order processes that are not automatically phase matched is difficult and rarely worth the effort because such processes are relatively inefficient.

With collinear beams, the phase-matching condition for a general second-order process reduces to the following simple scalar relation:

$$k_3 = k_1 + k_2, \qquad \text{or} \qquad n_3\omega_3 = n_1\omega_1 + n_2\omega_2. \tag{9.82}$$

Efficient parametric interactions are normally carried out in a spectral region away from transition resonances of a medium to avoid attenuation of the optical beams due to resonant absorption of the medium. As discussed in Section 1.10 and shown in Fig. 1.22, a material has normal dispersion in a spectral region away from resonances, meaning that $n(\omega_3) > n(\omega_1), n(\omega_2)$. Clearly, collinear phase matching is not possible in an isotropic material or a cubic crystal within a spectral region of normal dispersion, nor is it possible in a birefringent crystal if all of the interacting waves have the same polarization. An isotropic material is of no practical use for second-order nonlinear interactions because $\chi^{(2)} = 0$ in the electric-dipole approximation. A noncentrosymmetric cubic crystal, such as GaAs, has a decent $\chi^{(2)}$. However, such a crystal in bulk, homogeneous form is also not useful for second-order nonlinear optical interactions because of its inability to support collinear phase matching in the normal dispersion region where the crystal is transparent to the interacting optical waves. It is useful for second-order processes that are automatically phase matched, namely, the processes of optical rectification and Pockels effect.

Collinear phase matching can be achieved through the use of (1) anomalous dispersion near the resonance frequency of a material, (2) birefringence in a nonlinear

crystal, (3) periodic spatial modulation in the nonlinear coefficient of a medium, or (4) modal dispersion of an optical waveguide. Among these possibilities, the use of anomalous dispersion is not very practical for device applications because of strong material absorption near a resonance frequency. The modal dispersion of a waveguide is usually not strong; thus it is also of limited usefulness.

Birefringent phase matching

The most commonly used method of obtaining collinear phase matching for a second-order nonlinear optical process employs the birefringence of a uniaxial or biaxial crystal. In the following, phase matching in uniaxial crystals is addressed specifically because it is less complicated than that in biaxial crystals. The same principles apply to phase matching in biaxial crystals.

As discussed in Section 1.6 and illustrated in Fig. 1.10, there are two normal mode polarizations, $\hat{e}_o$ and $\hat{e}_e$, associated with each direction $\hat{k}$ of wave propagation in a uniaxial crystal. The ordinary wave with polarization $\hat{e}_o$ has an ordinary refractive index n_o independent of the direction of $\hat{k}$, whereas the extraordinary wave with polarization $\hat{e}_e$ has an extraordinary refractive index $n_e(\theta)$ that is given in (1.125) and is a function of the angle θ between the $\hat{k}$ vector and the optical axis. To satisfy the phase-matching condition in (9.82) in a spectral region of normal dispersion, the wave at the highest frequency, ω_3, has to be associated with the smaller of the two indices. Consequently, in a positive uniaxial crystal the wave at ω_3, or that at 2ω in the case of second-harmonic generation, has to be an ordinary wave, whereas in a negative uniaxial crystal it has to be an extraordinary wave.

There are two different types of birefringent phase-matching methods. In *type I phase matching* the two low-frequency waves have the same polarization, whereas in *type II phase matching* they have orthogonal polarizations. Note that in collinear phase matching, the **k** vectors of the interacting waves are all parallel to one another. Therefore, their normal modes also have the same $\hat{e}_o$ and $\hat{e}_e$ vectors. Table 9.6 summarizes the characteristics of type I and type II phase-matching methods for uniaxial crystals. In Table 9.6, we have arbitrarily assigned for type II phase matching the wave at ω_1 to be the ordinary wave and that at ω_2 to be the extraordinary wave. When $\omega_1 \neq \omega_2$, there are two different possibilities of type II phase matching, with the ordinary wave at ω_1

Table 9.6 *Two types of birefringent phase matching for uniaxial crystals*

	Positive uniaxial ($n_e > n_o$)	Negative uniaxial ($n_e < n_o$)
Type I	$n_3^o\omega_3 = n_1^e(\theta_{PM})\omega_1 + n_2^e(\theta_{PM})\omega_2$ $\chi_{eff} = \hat{e}_o \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_e\hat{e}_e$	$n_3^e(\theta_{PM})\omega_3 = n_1^o\omega_1 + n_2^o\omega_2$ $\chi_{eff} = \hat{e}_e \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_o\hat{e}_o$
Type II	$n_3^o\omega_3 = n_1^o\omega_1 + n_2^e(\theta_{PM})\omega_2$ $\chi_{eff} = \hat{e}_o \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_o\hat{e}_e$	$n_3^e(\theta_{PM})\omega_3 = n_1^o\omega_1 + n_2^e(\theta_{PM})\omega_2$ $\chi_{eff} = \hat{e}_e \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{e}_o\hat{e}_e$

being the one at either the lower or the higher frequency. The angle θ_{PM} between vector $\hat{k}$ and the optical axis that allows a particular phase-matching condition to be satisfied is known as the *phase-matching angle*.

In collinearly phase-matched second-harmonic generation, there is only one fundamental wave. In type I phase matching, the fundamental wave is completely polarized along one of the normal mode polarizations; thus the phase-matching condition is simply

$$n_{2\omega} = n_{\omega}. \tag{9.83}$$

In type II phase matching, the fundamental wave consists of components in both normal mode polarizations. Therefore, the phase-matching condition becomes

$$n^{\text{o}}_{2\omega} = \frac{1}{2}\left[n^{\text{o}}_{\omega} + n^{\text{e}}_{\omega}(\theta_{\text{PM}})\right] \tag{9.84}$$

for a positive uniaxial crystal, or

$$n^{\text{e}}_{2\omega}(\theta_{\text{PM}}) = \frac{1}{2}\left[n^{\text{o}}_{\omega} + n^{\text{e}}_{\omega}(\theta_{\text{PM}})\right] \tag{9.85}$$

for a negative uniaxial crystal. These different phase-matching methods for second-harmonic generation are illustrated in Fig. 9.10.

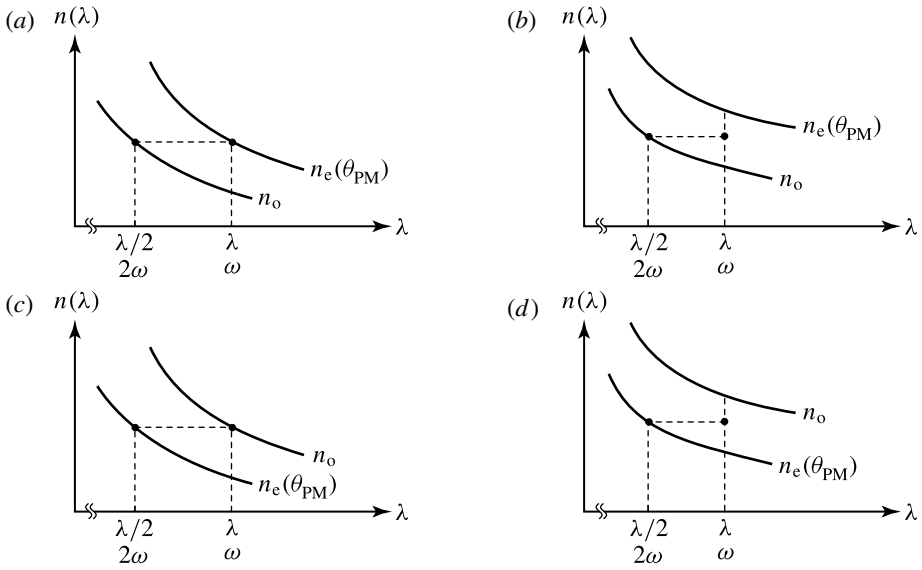


Figure 9.10 Different phase-matching methods in the region of normal dispersion for second-harmonic generation: (*a*) type I and (*b*) type II phase matching in a positive uniaxial crystal, and (*c*) type I and (*d*) type II phase matching in a negative uniaxial crystal.

We see from (1.122) that vector $\hat{k}$ is a function of both θ and ϕ. In a phase-matched interaction, the value of θ is the phase-matching angle θ_{PM}, which is obtained by solving the condition for a specific phase-matching method. Phase matching in a uniaxial crystal is independent of angle ϕ. Therefore, the value of θ_{PM} is determined without knowledge of ϕ. The value of χ_{eff} is usually a function of both θ and ϕ, however. For example, KDP is a negative uniaxial crystal of $\bar{4}2m$ symmetry whose only nonvanishing second-order nonlinear susceptibility elements are $\chi_{14}^{(2)} = \chi_{25}^{(2)}$ and $\chi_{36}^{(2)}$ under the conditions for index contraction. Using (1.123) and (1.124) for $\hat{e}_{\mathrm{o}}$ and $\hat{e}_{\mathrm{e}}$, respectively, we find that $\chi_{\mathrm{eff}} = -\chi_{36}^{(2)} \sin\theta \sin 2\phi$, or $d_{\mathrm{eff}} = -d_{36} \sin\theta \sin 2\phi$, for type I phase matching and $\chi_{\mathrm{eff}} = \left(\chi_{14}^{(2)} + \chi_{36}^{(2)}\right) \sin\theta \cos\theta \cos 2\phi$, or $d_{\mathrm{eff}} = (d_{14} + d_{36}) \sin\theta \cos\theta \cos 2\phi$, for type II phase matching. Therefore, to maximize the value of $|\chi_{\mathrm{eff}}|$ so that a second-order interaction in KDP is most efficient, ϕ has to be chosen to have one of the values among $\pi/4$, $-\pi/4$, $3\pi/4$, and $-3\pi/4$ in the case of type I phase matching and one of the values among 0, $\pi/2$, $-\pi/2$, and π in the case of type II phase matching. In some crystals, notably uniaxial crystals of symmetry classes 4, 6, 422, 622, $4mm$, and $6mm$, χ_{eff} is independent of angle ϕ but is a function of θ only. Then the value of ϕ can be chosen arbitrarily, though that of θ is still determined by the phase-matching condition (see Problem 9.5.1).

For a specific nonlinear interaction in a given crystal, type I and type II phase-matching methods generally have different phase-matching angles and different effective nonlinear susceptibilities. In certain cases, only one type of phase matching is possible. Sometimes, both types are not possible in a particular crystal within a certain spectral range. In case it is possible to have both type I and type II phase matching, the choice between the two depends on many practical considerations, including efficiency, angular tolerance, temperature sensitivity, and beam walk-off. Usually the one with the larger value of $|\chi_{\mathrm{eff}}|$ is chosen if it has no significant disadvantages from other considerations. Sometimes, χ_{eff} vanishes when phase matching is achieved. Clearly, such phase matching is of no practical usefulness. A simple example is type II phase matching in KDP with $\theta_{\mathrm{PM}} = \pi/2$.

In summary, *the condition for phase matching and the value of* χ_{eff} *have to be considered at the same time when designing a practical device.* Phase matching by itself does not guarantee a desirable value of χ_{eff} and, in some special cases, can even lead to a vanishing χ_{eff}. For a collinearly phase-matched interaction in a uniaxial crystal, the value of θ is determined by phase matching to be $\theta = \theta_{\mathrm{PM}}$, while that of ϕ is determined by maximizing the value of $|\chi_{\mathrm{eff}}|$. In a biaxial crystal, the angle θ is not independent of the angle ϕ, but they are determined by similar considerations. When both θ and ϕ are determined, the propagation direction $\hat{k}$, which is common to all of the interacting waves in a collinear interaction, is fixed. In practice, a nonlinear crystal intended for a particular application is normally cut with the knowledge of the correct values of θ and ϕ for the application in a way that vector $\hat{k}$ is normal to the input surface

of the crystal and the $\hat{e}_o$ and $\hat{e}_e$ polarizations are along certain convenient directions in the experimental setup.

EXAMPLE 9.8 Both type I and type II configurations of collinear birefringent phase matching are considered for second-harmonic generation in $LiNbO_3$ with a fundamental wave at 1.10 μm. Find the polarization directions of the interacting waves, the phase-matching angle, and the effective nonlinear susceptibility for each type.

Solution $LiNbO_3$ is a negative uniaxial crystal of $3m$ symmetry. The polarizations of the ordinary and extraordinary waves in a uniaxial crystal are $\hat{e}_o = \hat{x}\sin\phi - \hat{y}\cos\phi$ and $\hat{e}_e = -\hat{x}\cos\theta\cos\phi - \hat{y}\cos\theta\sin\phi + \hat{z}\sin\theta$, given in (1.123) and (1.124), respectively. According to Example 9.6, the nonvanishing nonlinear susceptibility tensor elements of $LiNbO_3$ are $d_{31} = d_{32} = d_{24} = d_{15} = -4.4$ pm V^{-1}, $d_{22} = -d_{21} = -d_{16} = 2.4$ pm V^{-1}, and $d_{33} = -25.2$ pm V^{-1}. According to Example 9.7, the refractive indices of $LiNbO_3$ are $n_o = 2.2319$ and $n_e = 2.1536$ at the fundamental wavelength of 1.10 μm and $n_o = 2.3168$ and $n_e = 2.2260$ at the second-harmonic wavelength of 550 nm. The extraordinary index at an angle θ is given by (1.125) as $n_e^{-2}(\theta) = n_o^{-2}\cos^2\theta + n_e^{-2}\sin^2\theta$.

For type I phase matching, we find from Table 9.6 that the fundamental is an ordinary wave with $\hat{e}_\omega = \hat{e}_o$ and the second harmonic has to be an extraordinary wave with $\hat{e}_{2\omega} = \hat{e}_e$. With the given nonvanishing nonlinear susceptibility elements, we find that the effective nonlinear susceptibility is

$$\chi^{I}_{eff} = \hat{e}_e \cdot \boldsymbol{\chi}^{(2)} : \hat{e}_o\hat{e}_o = \chi^{(2)}_{31}\sin\theta - \chi^{(2)}_{22}\cos\theta\sin 3\phi, \tag{9.86}$$

or, equivalently, $d^{I}_{eff} = d_{31}\sin\theta - d_{22}\cos\theta\sin 3\phi$. The phase-matching angle θ^{I}_{PM} can be found by using the relation in (9.83) for $n^{e}_{2\omega}(\theta^{I}_{PM}) = n^{o}_{\omega}$. Using the formula for $n_e(\theta)$, we find that

$$\theta^{I}_{PM} = \sin^{-1}\left[\frac{(n^{o}_{\omega})^{-2} - (n^{o}_{2\omega})^{-2}}{(n^{e}_{2\omega})^{-2} - (n^{o}_{2\omega})^{-2}}\right]^{1/2} = \sin^{-1}\left(\frac{2.2319^{-2} - 2.3168^{-2}}{2.2260^{-2} - 2.3168^{-2}}\right)^{1/2} = 74.8^\circ. \tag{9.87}$$

The angle ϕ is chosen so that $|d^{I}_{eff}|$ is maximized because ϕ is irrelevant to phase matching in a uniaxial crystal. Because $d_{31} < 0$, $d_{22} > 0$, and $0^\circ \le \theta \le 90^\circ$, we can maximize $|d^{I}_{eff}|$ by simply making $\sin 3\phi = 1$ in (9.86). Therefore, ϕ is chosen to be -90°, 30°, or 150°. We then find that $d^{I}_{eff} = -4.88$ pm V^{-1} and $\chi^{I}_{eff} = -9.76$ pm V^{-1} for $\theta = 74.8^\circ$ and $\phi = -90^\circ$, 30°, or 150°.

For type II phase matching, we find from Table 9.6 that the fundamental is required to have both ordinary and extraordinary components but the second harmonic is an extraordinary wave with $\hat{e}_{2\omega} = \hat{e}_e$. With the given nonvanishing nonlinear susceptibility

elements, we find that the effective nonlinear susceptibility is

$$\chi_{\text{eff}}^{\text{II}} = \hat{e}_{\text{e}} \cdot \boldsymbol{\chi}^{(2)} : \hat{e}_{\text{o}}\hat{e}_{\text{e}} = \chi_{22}^{(2)} \cos^2\theta \cos 3\phi, \tag{9.88}$$

or, equivalently, $d_{\text{eff}}^{\text{II}} = d_{22} \cos^2\theta \cos 3\phi$. The phase-matching angle $\theta_{\text{PM}}^{\text{II}}$ can be found by using the relation in (9.85). Plugging the formula for $n_{\text{e}}(\theta)$ into (9.85) results in a complicated algebraic relation, which can be either solved graphically or numerically to find that there is no solution for $\theta_{\text{PM}}^{\text{II}}$ in the range from 0° to 90°. Therefore, type II phase matching is not possible for second-harmonic generation in $LiNbO_3$ at $\lambda = 1.10$ μm. The angle ϕ can still be chosen so that $|d_{\text{eff}}^{\text{I}}|$ is maximized because ϕ is irrelevant to phase matching in a uniaxial crystal. In type II phase matching, $|d_{\text{eff}}^{\text{II}}|$ can be maximized by making $\cos 3\phi = \pm 1$ so that $|\cos 3\phi| = 1$. For this purpose, ϕ can be chosen as one of the following values: 0°, ±60°, ±120°, or 180°. Because phase matching is not possible, maximizing $|d_{\text{eff}}^{\text{II}}|$ in this situation serves no practical purpose.

We can compare $\chi_{\text{eff}}^{\text{I}}$ and $\chi_{\text{eff}}^{\text{II}}$ obtained in (9.86) and (9.88), respectively, to see that for $LiNbO_3$ type I interaction is more efficient than type II interaction.

Angle tuning

The phase-matching angle for a specific interaction in a given nonlinear crystal is a function of the frequencies, or the wavelengths, of the interacting waves. When the frequencies of the interacting waves are varied, the angle θ has to be varied accordingly for the interaction to remain phase matched. In practice, this angle tuning is normally carried out by rotating the crystal while maintaining the beam propagation direction though it can also be achieved by varying the beam propagation direction while fixing the orientation of the crystal. One situation where this tuning is necessary is in an application with a wavelength-tunable input wave, such as in the generation of a wavelength-tunable difference- or sum-frequency wave or in the frequency doubling of the output from a wavelength-tunable laser. In optical parametric generation where only the pump-wave frequency at ω_3 is fixed, the parametrically generated frequencies ω_1 and ω_2 are varied when the parameters for phase matching are varied. Therefore, angle tuning of a nonlinear crystal is a convenient means for tuning the parametric wavelengths. Figure 9.11 shows as an example the *angle-tuning curves* of the parametric wavelengths for type I and type II collinear phase matching in $LiNbO_3$ with a fixed pump wavelength at 527 nm. $LiNbO_3$ is a negative uniaxial crystal of $3m$ symmetry, in which type I interaction is more efficient than type II interaction. The effective nonlinear susceptibilities for type I and type II phase matching in $LiNbO_3$ are found in Example 9.8 and are given in (9.86) and (9.88), respectively.

When the frequencies of the interacting waves are fixed, any deviation of the wave propagation direction away from the phase-matched direction results in a phase mismatch. The amount of this phase mismatch can be calculated by expanding

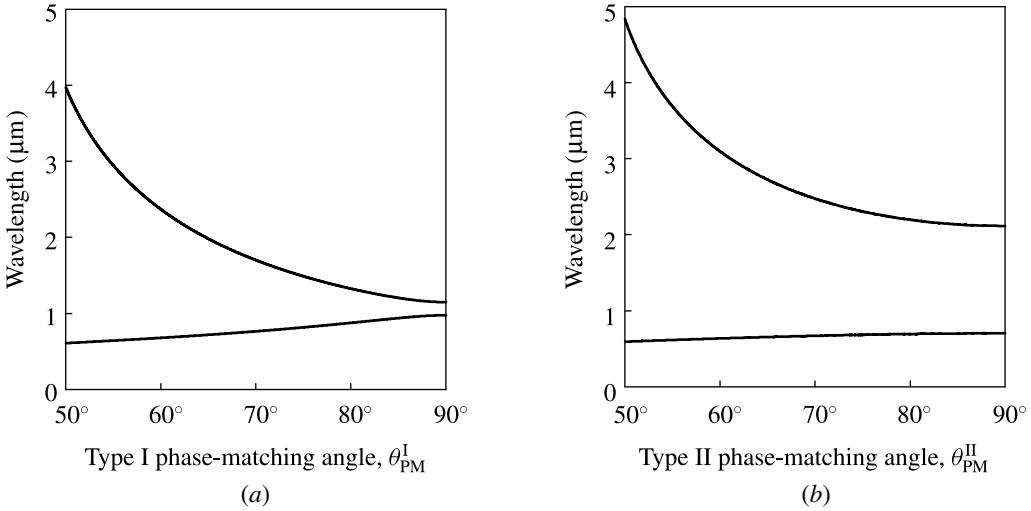


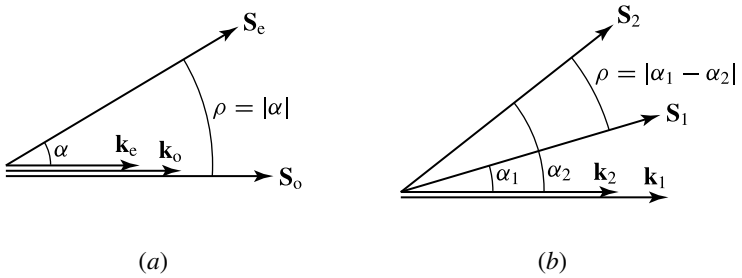
Figure 9.11 Angle-tuning curves showing parametric wavelengths as a function of the phase-matching angle for (*a*) type I and (*b*) type II collinear phase matching in $LiNbO_3$ with a fixed pump wavelength at 527 nm. The data listed in Table 9.3 are used to generate these curves.

$\Delta k = k_1 + k_2 - k_3$ around the phase-matching angle:

$$\begin{aligned}\Delta k &= (k_1 + k_2 - k_3)_{\theta_{PM}} + \Delta\theta \left[\frac{d}{d\theta}(k_1 + k_2 - k_3)\right]_{\theta_{PM}} \\ &\quad + \frac{(\Delta\theta)^2}{2}\left[\frac{d^2}{d\theta^2}(k_1 + k_2 - k_3)\right]_{\theta_{PM}} + \cdots \\ &= \frac{\Delta\theta}{c}\left(\omega_1\frac{dn_1}{d\theta} + \omega_2\frac{dn_2}{d\theta} - \omega_3\frac{dn_3}{d\theta}\right)_{\theta_{PM}} \\ &\quad + \frac{(\Delta\theta)^2}{2c}\left(\omega_1\frac{d^2n_1}{d\theta^2} + \omega_2\frac{d^2n_2}{d\theta^2} - \omega_3\frac{d^2n_3}{d\theta^2}\right)_{\theta_{PM}} + \cdots \end{aligned} \tag{9.89}$$

The acceptable angular tolerance in a nonlinear interaction is set by the amount of the acceptable phase mismatch. A common rule for setting this tolerance is $\Delta kl < \pi$. In most applications of nonlinear optical devices, the interacting beams are focused to increase the efficiency. Because focusing increases the divergence, thus the angular spread, of a beam, an interaction that has a small angular tolerance requires the interacting beams to be well collimated and critically aligned.

Because $n_e(0^\circ) = n_o$, a phase-matching angle in a uniaxial crystal cannot have the value of 0°. Therefore, it can be shown by using (1.125) that $(dn_e(\theta)/d\theta)_{\theta_{PM}} \neq 0$ except when $\theta_{PM} = 90^\circ$. If $\theta_{PM} \neq 90^\circ$, the first-order term in (9.89) exists; thus $\Delta k \propto \Delta\theta$ approximately. For phase matching with $\theta_{PM} = 90^\circ$, known as 90° *phase matching*, the first-order term in (9.89) vanishes; thus $\Delta k \propto (\Delta\theta)^2$. Because $\Delta\theta \ll 1$, 90° phase matching has a smaller phase mismatch for a given angular deviation or, equivalently,



(*a*) (*b*)

Figure 9.12 Walk-off between (*a*) an ordinary beam and an extraordinary beam and (*b*) two extraordinary beams when the beams propagate collinearly.

a larger angular tolerance for a given acceptable phase mismatch than phase matching with $\theta_{PM} \neq 90^\circ$. In 90° phase matching, an extraordinary wave is polarized along the optical axis, and $n_e(90^\circ) = n_e$.

As can be seen from Table 9.6, for any method of birefringent phase matching, there is always at least one extraordinary wave involved in the interaction. For an extraordinary wave that is not polarized along a principal axis of a crystal, there is a walk-off angle α given by (1.131) between its direction of propagation, defined by $\hat{k}$, and its direction of power flow, defined by its Poynting vector $\mathbf{S}$. In a collinear interaction, all of the interacting waves have the same direction of propagation, but not necessarily the same direction of power flow. When two interacting beams have different directions of power flow, there is a walk-off angle ρ between these two beams, which is defined as the angle between their Poynting vectors. Note the fine difference between the walk-off angle α and the walk-off angle ρ. As is shown in Fig. 9.12(*a*), the walk-off angle between an ordinary beam and an extraordinary beam is simply $\rho = |\alpha|$, which is determined only by the walk-off angle α of the extraordinary beam. However, as illustrated in Fig. 9.12(*b*), the walk-off angle between two collinear extraordinary beams is the difference of the walk-off angles of these two beams: $\rho = |\alpha_1 - \alpha_2|$, which exists between two extraordinary beams of different frequencies because of dispersion.

Because optical beams have finite transverse dimensions in reality, the existence of a walk-off angle ρ between two interacting beams limits the effective interaction length, as Fig. 9.12 shows. For Gaussian beams with a radius w_0 at the beam waist, the effective interaction length between two beams subject to the limitation of beam walk-off is the *aperture distance*:

$$l_a = \frac{\pi^{1/2} w_0}{\rho}. \tag{9.90}$$

Clearly, the aperture distance decreases as the beams are increasingly focused.

When $\theta = 90^\circ$, an extraordinary wave is polarized along the extraordinary principal axis and thus has no walk-off, as can be verified with (1.131). Consequently, there is

no walk-off between any two interacting beams in the case of 90° phase matching. For this reason and for the reason discussed above that it has a larger angular tolerance than phase matching with $\theta_{PM} \neq 90°$ has, 90° phase matching is also called *noncritical phase matching*.

EXAMPLE 9.9 A Gaussian beam of fundamental wave at 1.10 μm is used for second-harmonic generation with type I collinear angle phase matching in $LiNbO_3$ discussed in Example 9.8. Find the walk-off angle ρ between the fundamental and second-harmonic beams. If the fundamental beam is focused to a waist size of $w_0 = 50$ μm, what is the aperture distance limited by the walk-off effect?

Solution For type I phase matching, only the second-harmonic beam, which is an extraordinary wave, has walk-off with an angle α between $\mathbf{S}_{2\omega}$ and $\mathbf{k}_{2\omega}$. The fundamental beam is an ordinary wave with $\mathbf{S}_\omega \parallel \mathbf{k}_\omega$. Therefore, the walk-off between the two Poynting vectors $\mathbf{S}_\omega$ and $\mathbf{S}_{2\omega}$, which is what matters in this interaction, is $\rho = |\alpha|$ for collinear phase matching with $\mathbf{k}_\omega \parallel \mathbf{k}_{2\omega}$. By using (1.131) for α and taking the refractive indices to be $n_o = 2.3168$ and $n_e = 2.2260$ at the second-harmonic wavelength of 550 nm, we find, with $\theta = \theta_{PM}^I = 74.8°$, the following walk-off angle:

$$\rho = |\alpha| = \left| \tan^{-1}\left(\frac{n_o^2}{n_e^2} \tan\theta_{PM}^I \right) - \theta_{PM}^I \right|$$
$$= \left| \tan^{-1}\left(\frac{2.3168^2}{2.2260^2} \tan 74.8° \right) - 74.8° \right| = 1.12° = 19.5 \text{ mrad}.$$

For $w_0 = 50$ μm, the aperture distance is

$$l_a = \frac{\pi^{1/2} w_0}{\rho} = \frac{\pi^{1/2} \times 50 \times 10^{-6}}{19.5 \times 10^{-3}} \text{ m} = 4.54 \text{ mm}.$$

The waist size of the second-harmonic beam is normally different from that of the fundamental beam. In the presence of walk-off, a second-harmonic beam generated by a circular Gaussian fundamental beam can have an elliptical spot shape. Such complications are ignored here.

Temperature tuning

It is clear from the above discussions that 90° phase matching is most desirable for both type I and type II phase-matching methods. In general, the ordinary and the extraordinary indices of a uniaxial crystal have different temperature dependencies, and the three principal indices of a biaxial crystal also change with temperature at different rates. In certain cases, it is possible to fix the angle θ at 90° while varying the temperature to achieve phase matching. The temperature, T_{PM}, at which 90° phase matching is achieved in a crystal is called the *phase-matching temperature*. Whether 90°

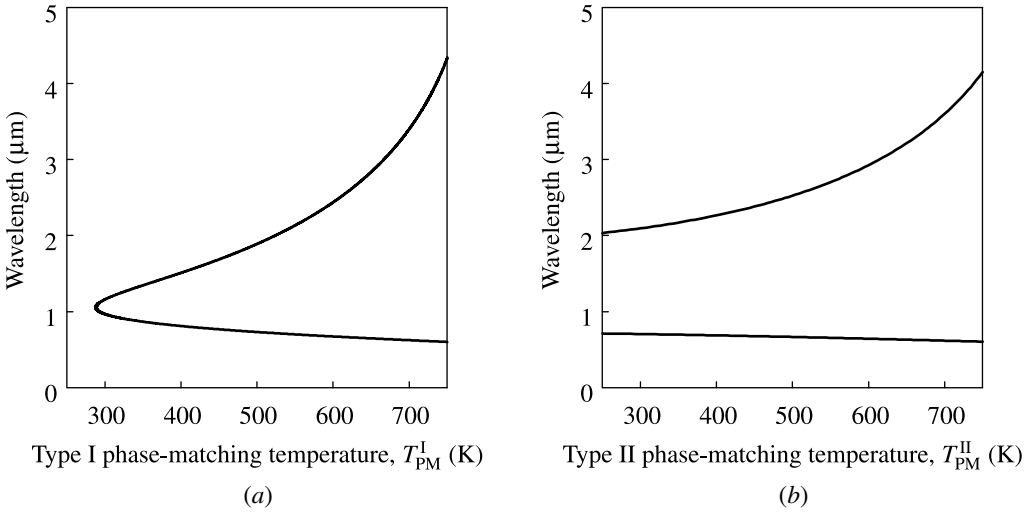


Figure 9.13 Temperature-tuning curves showing parametric wavelengths as a function of the phase-matching temperature for (*a*) type I and (*b*) type II collinear phase matching in $LiNbO_3$ with a fixed pump wavelength at 527 nm. The temperature and wavelength dependencies of refractive indices given in (9.92) and (9.93) are used to generate these curves.

phase matching by tuning the temperature is possible or not depends on the temperature dependence of the birefringence of a crystal, as well as on the wavelengths of the interacting waves in a given nonlinear process.

As discussed above, it is important to examine the value of χ_{eff} at $\theta = 90^\circ$ for a given phase-matching method also. It turns out that $\chi_{\text{eff}} = 0$ for 90° phase matching in any uniaxial crystal if there are two extraordinary waves and one ordinary wave in the interaction, as well as in crystals of certain symmetry classes when there are two ordinary waves and one extraordinary wave in the interaction. Specifically, among all uniaxial crystals, 90° phase matching with $\chi_{\text{eff}} \neq 0$ is possible only for type I phase matching in a negative uniaxial crystal and for type II phase matching in a positive uniaxial crystal, but only if the crystal belongs to one of the symmetry classes 3, 4, 6, $4mm$, $6mm$, $3m$, $\overline{4}$, and $\overline{4}2m$ in either case (see Problem 9.5.2). For biaxial crystals, the situation is more complicated.

Because the phase-matching temperature for a specific nonlinear interaction in a given crystal is a function of the frequencies of the interacting waves, it has to be tuned when the wavelengths of the waves are varied. Alternatively, in optical parametric generation with a fixed pump wavelength, the parametrically generated wavelengths can be tuned by tuning the temperature while keeping both the propagation direction of the beams and the orientation of the crystal fixed. Figure 9.13 shows as an example the *temperature-tuning curves* of the parametric wavelengths for type I and type II collinear phase matching in $LiNbO_3$ with a fixed pump wavelength at 527 nm. As shown in Fig. 9.13(*b*), temperature tuning with type II collinear phase matching for

$LiNbO_3$ is possible, but it is not practically useful because χ_{eff} vanishes for type II 90° phase matching in $LiNbO_3$. Therefore, temperature tuning for $LiNbO_3$ is useful only with type I phase matching.

In temperature phase matching, deviation from the phase-matching temperature results in a phase mismatch given by

$$\begin{aligned}\Delta k &= (k_1 + k_2 - k_3)_{T_{\text{PM}}} + \Delta T \left[\frac{\mathrm{d}}{\mathrm{d}T}(k_1 + k_2 - k_3)\right]_{T_{\text{PM}}} + \cdots \\ &= \frac{\Delta T}{c}\left(\omega_1 \frac{\mathrm{d}n_1}{\mathrm{d}T} + \omega_2 \frac{\mathrm{d}n_2}{\mathrm{d}T} - \omega_3 \frac{\mathrm{d}n_3}{\mathrm{d}T}\right)_{T_{\text{PM}}} + \cdots \end{aligned} \tag{9.91}$$

For a practical device that is temperature tuned, the first-order term in (9.91) normally does not vanish. Therefore, $\Delta k \propto \Delta T$ to first order.

In comparison with angle tuning, temperature tuning has all of the advantages of 90° phase matching discussed above. In addition, because the crystal orientation and the beam propagation direction are both fixed in a temperature-tuned device, temperature tuning also eliminates all of the troubles that come with angle tuning in mechanically changing the crystal orientation while trying to keep the optical beams aligned.

EXAMPLE 9.10 In this example, we consider 90° phase matching in both type I and type II configurations for second-harmonic generation in $LiNbO_3$ with a fundamental wave at 1.10 μm. The ordinary and extraordinary refractive indices of $LiNbO_3$ as a function of wavelength and temperature are given by

$$n_{\text{o}}^2 = 4.9130 + \frac{0.1173 + 1.65 \times 10^{-8} T^2}{\lambda^2 - (0.212 + 2.7 \times 10^{-8} T^2)^2} - 0.0278\lambda^2, \tag{9.92}$$

$$\begin{aligned} n_{\text{e}}^2 = {} & 4.5567 + 2.605 \times 10^{-7} T^2 \\ & + \frac{0.0970 + 2.70 \times 10^{-8} T^2}{\lambda^2 - (0.201 + 5.4 \times 10^{-8} T^2)^2} - 0.0224\lambda^2, \end{aligned} \tag{9.93}$$

where λ is the optical wavelength in micrometers and T is the temperature in kelvins. Use the data in Example 9.8 to find the polarization directions of the interacting waves, the phase-matching temperature, and the effective nonlinear susceptibility for each type.

Solution For type I phase matching, the fundamental is an ordinary wave with $\hat{e}_\omega = \hat{e}_{\text{o}}$, and the second harmonic has to be an extraordinary wave with $\hat{e}_{2\omega} = \hat{e}_{\text{e}}$. The temperature, T_{PM}^{I}, for 90° type I phase matching is found by solving $n_{2\omega}^{\text{e}}(T_{\text{PM}}^{\text{I}}) = n_{\omega}^{\text{o}}(T_{\text{PM}}^{\text{I}})$. Using the relations given in (9.92) and (9.93), we find the phase-matching temperature to be $T_{\text{PM}}^{\text{I}} = 396.7$ K, or 123.7 °C. Because $\theta_{\text{PM}} = 90°$ for 90° phase matching, we find from Example 9.8 that $d_{\text{eff}}^{\text{I}} = d_{31} = -4.4$ pm V^{-1} and $\chi_{\text{eff}}^{\text{I}} = \chi_{31}^{(2)} = -8.8$ pm V^{-1}.

Because $\chi^{\mathrm{I}}_{\mathrm{eff}}$ does not depend on angle ϕ in this situation, the value of ϕ can be chosen arbitrarily.

For type II phase matching, the fundamental is required to have both ordinary and extraordinary components, and the second harmonic is an extraordinary wave with $\hat{e}_{2\omega} = \hat{e}_{\mathrm{e}}$. The temperature, $T^{\mathrm{II}}_{\mathrm{PM}}$, for 90° type II phase matching is found by solving $n^{\mathrm{e}}_{2\omega}(T^{\mathrm{II}}_{\mathrm{PM}}) = (n^{\mathrm{o}}_{\omega}(T^{\mathrm{II}}_{\mathrm{PM}}) + n^{\mathrm{e}}_{\omega}(T^{\mathrm{II}}_{\mathrm{PM}}))/2$. Using the relations given in (9.92) and (9.93), we find that there is no solution for $T^{\mathrm{II}}_{\mathrm{PM}}$. Therefore, 90° type II phase matching is not possible for second-harmonic generation in $LiNbO_3$ at a fundamental wavelength of 1.10 μm. Combining this finding and that in Example 9.8, we find that type II phase matching is simply not possible for second-harmonic generation in $LiNbO_3$ at 1.10 μm fundamental wavelength.

As demonstrated above in Fig. 9.12, birefringent type II phase matching in $LiNbO_3$ is possible for parametric generation at certain wavelengths. However, 90° type II phase matching in $LiNbO_3$, and in any other crystal of $3m$ symmetry alike, is useless anyway because $\chi^{\mathrm{II}}_{\mathrm{eff}} = 0$ for $\theta = 90°$, as discussed in the text above and can be seen from (9.88).

For 90° collinear phase matching, there is no walk-off between $\mathbf{S}_{\omega}$ and $\mathbf{S}_{2\omega}$ because $\mathbf{S}_{\omega} \parallel \mathbf{k}_{\omega} \parallel \mathbf{k}_{2\omega} \parallel \mathbf{S}_{2\omega}$. In this situation, the interaction is not limited by an aperture length because $l_{\mathrm{a}} = \infty$ effectively.

Quasi-phase matching

A very different phase-matching technique involves the introduction of a periodic modulation in a nonlinear medium. This approach is known as *quasi-phase matching* because phase mismatch is not eliminated within each modulation period but is compensated periodically. In principle, the periodic modulation can be on either the linear or the nonlinear susceptibility of the medium. In practice, however, modulating the linear susceptibility is less efficient than modulating the nonlinear susceptibility.

The principle of quasi-phase matching can be understood intuitively by following the discussions in Section 9.4 on the physical significance of the phase φ given in (9.67). The existence of a phase mismatch Δk leads to a change of the phase φ by an amount of π over a coherence length, resulting in a change of sign in $\sin\varphi$ and a reversal of the direction of energy flow in a parametric process. If the nonlinear susceptibility is periodically modulated such that a phase $\varphi_{\chi} = \pi$ is introduced over each coherence length, the total phase φ is reset to its initial value so that the reversal of the process is prevented. Then energy can continue to flow in the desired direction. The simplest and most effective approach to implementing such a periodic modulation is to change the sign of $\chi^{(2)}$ periodically, as illustrated in Fig. 9.14(a). In ferroelectric nonlinear crystals, such as $LiNbO_3$, $LiTaO_3$, and KTP, the periodic sign change in $\chi^{(2)}$ can be achieved by periodic poling with an external electric field for periodic ferroelectric domain reversal. Periodically poled $LiNbO_3$ (PPLN) and periodically poled KTP (PPKTP) are of great interest.

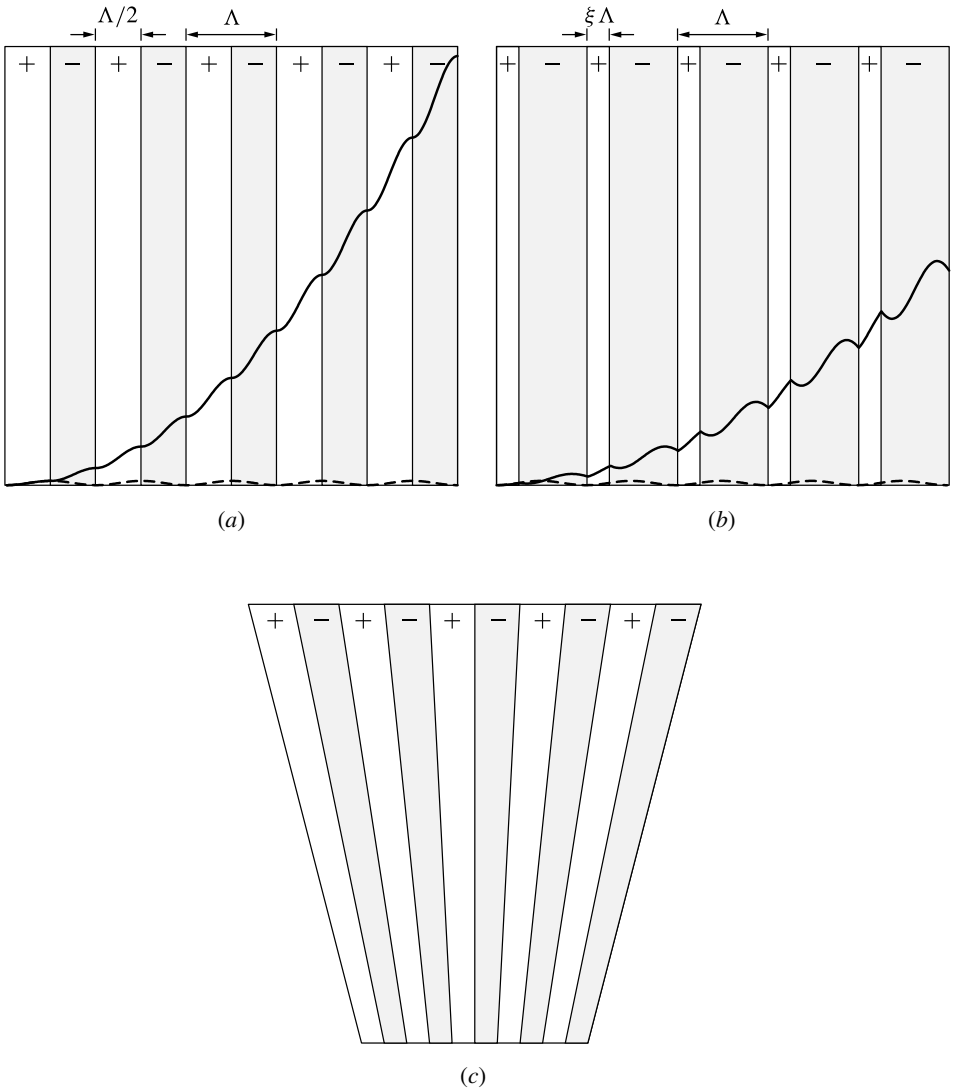


Figure 9.14 Structures with periodic sign reversal of the nonlinear susceptibility for quasi-phase matching: (*a*) first-order structure with a 50% duty factor, (*b*) general structure with a duty factor of ξ, and (*c*) fanned structure for wavelength tuning. Also shown in (*a*) and (*b*) is the growth of the second-harmonic intensity with quasi-phase matching, solid curves, as a function of propagation distance when a first-order structure with $\Lambda = 2l_{\text{coh}}$ is used for second-harmonic generation. Shown for comparison, dashed curves, is the second-harmonic intensity without quasi-phase matching. The plus and minus signs refer to the sign of $\chi^{(2)}$ in each region.

Any periodic modulation can be viewed as a grating. Therefore, the effect of a periodically modulated nonlinear susceptibility can be formally analyzed with a procedure similar to that used in the analysis of grating couplers in Section 5.1. In the presence of a periodic spatial modulation, the effective susceptibility defined in (9.59) becomes

a periodic function of z. It can be expressed in terms of a Fourier series expansion as

$$\chi_{\text{eff}}(z) = \sum_q \chi_{\text{eff}}(q) \mathrm{e}^{\mathrm{i}qKz}, \tag{9.94}$$

where $K = 2\pi/\Lambda$, Λ is the modulation period, and

$$\chi_{\text{eff}}(q) = \frac{1}{\Lambda} \int_0^{\Lambda} \chi_{\text{eff}}(z) \mathrm{e}^{-\mathrm{i}qKz} \mathrm{d}z. \tag{9.95}$$

By substituting $\chi_{\text{eff}}(z)$ of (9.94) for χ_{eff} in (9.60), we have

$$\begin{aligned} \frac{\mathrm{d}\mathcal{E}_3}{\mathrm{d}z} &= \frac{\mathrm{i}\omega_3^2}{c^2 k_{3,z}} \mathcal{E}_1 \mathcal{E}_2 \sum_q \chi_{\text{eff}}(q) \mathrm{e}^{\mathrm{i}(\Delta k + qK)z} \\ &\approx \frac{\mathrm{i}\omega_3^2}{c^2 k_{3,z}} \chi_{\text{Q}} \mathcal{E}_1 \mathcal{E}_2 \mathrm{e}^{\mathrm{i}\Delta k_{\text{Q}} z}, \end{aligned} \tag{9.96}$$

where

$$\chi_{\text{Q}} = \chi_{\text{eff}}(q) \tag{9.97}$$

and

$$\Delta k_{\text{Q}} = \Delta k + qK \tag{9.98}$$

for an integer q that minimizes the value of $|\Delta k + qK|$. The other two coupled parametric equations in (9.61) and (9.62) can also be transformed in a similar manner. Therefore, all of the results obtained for parametric interactions discussed in the preceding section are still valid in the case of quasi-phase matching after making the substitution of χ_{Q} and Δk_{Q} for χ_{eff} and Δk, respectively.

Perfect quasi-phase matching is achieved when $\Delta k_{\text{Q}} = 0$. This happens when the modulation period is chosen to be

$$\Lambda = -q \frac{2\pi}{\Delta k} = |q| \cdot 2l_{\text{coh}}. \tag{9.99}$$

Therefore, *first-order quasi-phase matching* occurs when $\Lambda = 2l_{\text{coh}}$ for $q = 1$ or -1. Quasi-phase matching at a high order is also possible.

When designing a periodic structure for quasi-phase matching, it is important to maximize the value of χ_{Q} to obtain the best efficiency for an interaction. In principle, any periodic structure is potentially useful. The simplest structure is one in which the sign of the nonlinear susceptibility is periodically reversed at abrupt boundaries. It is efficient and easy to fabricate. For such a structure with a duty factor ξ, as shown in

Fig. 9.14(*b*), we find that

$$\chi_{\mathrm{Q}} = \frac{1}{\Lambda}\left[\int_0^{\xi\Lambda} \chi_{\mathrm{eff}} \mathrm{e}^{-\mathrm{i}qKz}\mathrm{d}z - \int_{\xi\Lambda}^{\Lambda} \chi_{\mathrm{eff}} \mathrm{e}^{-\mathrm{i}qKz}\mathrm{d}z\right]$$
$$= 2\chi_{\mathrm{eff}} \frac{\sin \xi q\pi}{q\pi} \mathrm{e}^{-\mathrm{i}\xi q\pi}, \tag{9.100}$$

which has a form similar to the coupling coefficient of a square grating found in (5.18). Note that the optimum value for the duty factor ξ depends on the order q of a structure. Clearly, a first-order structure with a 50% duty factor ($\xi = 1/2$), which is shown in Fig. 9.14(*a*), has the largest effective nonlinear susceptibility:

$$|\chi_{\mathrm{Q}}| = \frac{2}{\pi}|\chi_{\mathrm{eff}}|. \tag{9.101}$$

For a given interaction, $|\chi_{\mathrm{Q}}|$ is always small than $|\chi_{\mathrm{eff}}|$.

It seems that an interaction with birefringent phase matching is always more efficient than that with quasi-phase matching. This is not true, however. In an interaction with birefringent phase matching, χ_{eff} is subject to the constraints imposed by the phase-matching configuration on the propagation direction and the polarizations of the interacting waves. Therefore, other than choosing a proper angle ϕ for the wave propagation direction and considering the difference between type I and type II phase-matching methods, there is little freedom in maximizing the effective nonlinear susceptibility when using birefringent phase matching. Quasi-phase matching is not subject to such constraints because it depends on an externally imposed structure, rather than intrinsic material properties, for phase matching. Therefore, there is much freedom in seeking a high susceptibility in an interaction with quasi-phase matching. For example, for $LiNbO_3$, $|\chi_{22}^{(2)}| < |\chi_{31}^{(2)}| \approx |\chi_{33}^{(2)}|/6$. From Examples 9.8 and 9.10, we find that $|\chi_{\mathrm{eff}}^{\mathrm{I}}| < (|\chi_{31}^{(2)}|^2 + |\chi_{22}^{(2)}|^2)^{1/2}$ and $|\chi_{\mathrm{eff}}^{\mathrm{II}}| < |\chi_{31}^{(2)}|$ for type I and type II birefringent phase matching in $LiNbO_3$, respectively. With birefringent phase matching, it is not possible to exploit the largest nonlinear susceptibility element $\chi_{33}^{(2)}$ in $LiNbO_3$ because $\chi_{33}^{(2)}$ can be used only when all of the interacting waves are polarized along the extraordinary principal axis. In contrast, with quasi-phase matching, all of the interacting waves can be polarized along the extraordinary axis so that $\chi_{\mathrm{eff}} = \chi_{33}^{(2)}$. If a first-order periodic modulation with a 50% duty factor is used for quasi-phase matching to compensate for the phase mismatch among the extraordinary waves, we have $|\chi_{\mathrm{Q}}| = 2|\chi_{\mathrm{eff}}|/\pi = 2|\chi_{33}^{(2)}|/\pi$, which is about four times the value of $|\chi_{\mathrm{eff}}|$ for the most efficient interaction with type I or type II birefringent phase matching.

From the above discussions, it is clear that one important advantage of quasi-phase matching is that it makes possible efficient nonlinear interactions for which birefringent phase matching is not possible. Nonlinear interactions in nonbirefringent nonlinear materials, such as III–V semiconductors, can also be phase matched with quasi-phase matching. The polarization directions of the interacting waves are not restricted in

quasi-phase matching as they are in birefringent phase matching. This flexibility allows a collinear interaction within the transparency range of a nonlinear material to be noncritically phase matched with no beam walk-off at any temperature. High efficiency is possible by arranging the polarization directions of the waves for an interaction to use the largest susceptibility element of a nonlinear crystal. For wavelength tuning, the modulation period Λ has to be varied. With a fanned structure such as that shown in Fig. 9.14(c), continuous wavelength tuning can be accomplished by translating the crystal transversely through the beam path.

EXAMPLE 9.11 A PPLN crystal is used for second-harmonic generation of a fundamental beam at 1.10 μm wavelength. Find the required grating period for quasi-phase matching and the largest effective nonlinear susceptibility available for this interaction.

Solution The largest nonlinear susceptibility element of $LiNbO_3$ is $d_{33} = -25.2$ pm V^{-1}, thus $\chi^{(2)}_{33} = 2d_{33} = -50.4$ pm V^{-1}. From the above discussions in the text, we know that both fundamental and second-harmonic waves have to be extraordinary waves polarized in the z direction in order to obtain the largest value of $|\chi_Q|$ for a PPLN crystal. Therefore, we have to take $n^e_\omega = 2.1536$ for the fundamental wave at $\lambda = 1.10$ μm and $n^e_{2\omega} = 2.2260$ for the second harmonic at $\lambda/2 = 550$ nm to calculate the phase mismatch Δk and the coherence length l_{coh} as in case (b) of Example 9.7:

$$l_{coh} = \frac{\pi}{|\Delta k|} = \frac{\lambda}{4|n^e_\omega - n^e_{2\omega}|} = \frac{1.10 \times 10^{-6}}{4 \times |2.1536 - 2.2260|}\ \text{m} = 3.80\ \mu\text{m}.$$

From the discussions following (9.100), we know that the largest value for $|\chi_Q|$ is that given in (9.101) obtained with a first-order structure with a 50% duty factor. Therefore, the required grating period is

$$\Lambda = 2l_{coh} = 7.60\ \mu\text{m}$$

for $|q| = 1$, and the effective nonlinear susceptibility is

$$|\chi_Q| = \frac{2}{\pi}|\chi^{(2)}_{33}| = 32.08\ \text{pm V}^{-1}$$

or, equivalently, $|d_Q| = 16.04$ pm V^{-1}.

In this scheme of quasi-phase matching, $\mathbf{S}_\omega \parallel \mathbf{k}_\omega \parallel \mathbf{k}_{2\omega} \parallel \mathbf{S}_{2\omega}$ because both fundamental and second-harmonic fields are polarized along the principal z axis. Therefore, there is no walk-off between $\mathbf{S}_\omega$ and $\mathbf{S}_{2\omega}$. This interaction is not limited by an aperture length, which is effectively $l_a = \infty$.

9.6 Optical frequency converters

A very important class of nonlinear optical devices is the optical frequency converters. Nonlinear optical frequency conversion is the only means for *direct* conversion of optical energy from one frequency to another. Indeed, the discipline of nonlinear optics was born out of the first observation of second-harmonic generation in 1961.

There are basically two types of nonlinear optical frequency converters. The majority are based on parametric processes, particularly the parametric second-order processes, that require phase matching. Sum-frequency generators, difference-frequency generators, harmonic generators, and parametric amplifiers and oscillators belong to this type. Devices that use the nonparametric third-order processes of stimulated Raman or Brillouin scattering to shift the optical frequency are the other type. In this section, we consider only those based on parametric processes. Devices based on stimulated Raman or Brillouin scattering are discussed in Section 9.9.

Sum-frequency generators

The basic function of a sum-frequency generator is the generation of an optical wave at a high frequency, ω_3, by mixing two optical waves at low frequencies, ω_1 and ω_2, as schematically shown in Fig. 9.1(a). The general application of a sum-frequency generator is straightforward. It is most often used to obtain, through mixing available optical waves at long wavelengths, a coherent optical beam at a desired short wavelength that is not readily available from other sources. If one of the two input waves is tunable in wavelength, a wavelength-tunable sum-frequency output wave is obtained. For example, a wavelength-tunable optical beam in the ultraviolet spectral region can be obtained with a sum-frequency generator that mixes the output of a tunable laser in the visible spectral region with that of another laser at a fixed wavelength also in the visible spectral region.

The process of sum-frequency generation is generally described by the coupled equations in (9.60)–(9.62) with the condition that $\mathcal{E}_1(0) \neq 0$ and $\mathcal{E}_2(0) \neq 0$ but $\mathcal{E}_3(0) = 0$ at the input surface, $z = 0$, of a nonlinear crystal. The general solutions to these coupled equations can be found in terms of the Jacobi elliptic functions. However, simpler, and often more useful, solutions can be found for specific experimental conditions of interest.

The simplest situation is when the efficiency of a sum-frequency generator is low so that the intensities of both input waves at ω_1 and ω_2 are not depleted significantly throughout the interaction. We can then assume $\mathcal{E}_1$ and $\mathcal{E}_2$ to be independent of z, ignore (9.61) and (9.62) in the coupled equations, and integrate (9.60) directly to find $\mathcal{E}_3(z)$. Using the relation in (9.63) for light intensity, we find that, in the *low-efficiency limit*, the intensity of the wave at the sum frequency as a function of the interaction length l

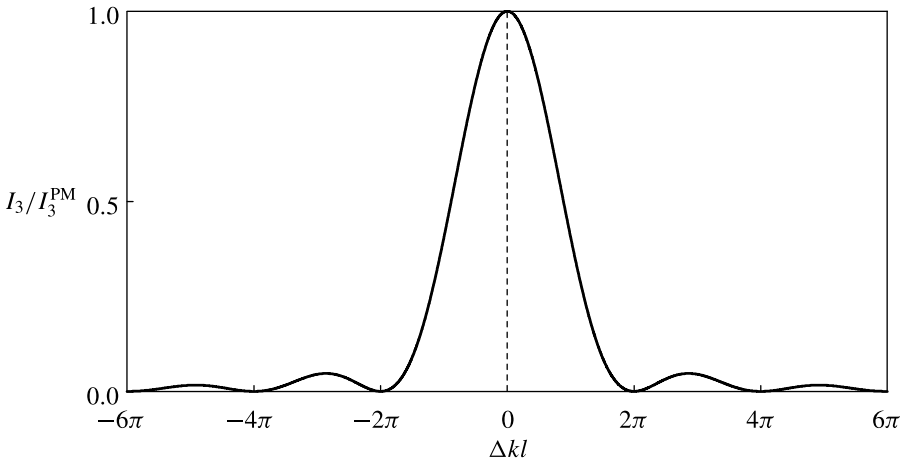


Figure 9.15 Effect of phase mismatch on the efficiency of sum-frequency generation in the low-efficiency limit.

can be expressed as

$$I_3(l) = \frac{\omega_3^2|\chi_{\text{eff}}|^2}{2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z}} I_1 I_2 l^2 \frac{\sin^2(\Delta kl/2)}{(\Delta kl/2)^2}$$
$$= \frac{8\pi^2|d_{\text{eff}}|^2}{c\epsilon_0 n_{1,z}n_{2,z}n_{3,z}\lambda_3^2} I_1 I_2 l^2 \frac{\sin^2(\Delta kl/2)}{(\Delta kl/2)^2}, \tag{9.102}$$

where $d_{\text{eff}} = \chi_{\text{eff}}/2$, and $\lambda_3 = 2\pi c/\omega_3$ is the wavelength of the sum-frequency wave in free space. In the case of quasi-phase matching, χ_{eff} and d_{eff} in (9.102) are replaced by χ_{Q} and d_{Q}, respectively.

The effect of phase mismatch is characterized by a function of the form

$$\frac{I_3}{I_3^{\text{PM}}} = \frac{\sin^2(\Delta kl/2)}{(\Delta kl/2)^2}, \tag{9.103}$$

which is plotted in Fig. 9.15. When $\Delta k \neq 0$, it does not pay to have a crystal longer than the coherence length of the interaction, as discussed in the preceding section and as illustrated in Fig. 9.8(*c*) (see Problem 9.6.1). When perfect phase matching is accomplished, the intensity of the sum-frequency wave grows quadratically with interaction length as $I_3 = I_3^{\text{PM}} \propto l^2/\lambda_3^2$.

We see from (9.102) that $I_3 \propto |d_{\text{eff}}|^2 I_1 I_2$ in the low-efficiency limit. Therefore, if the purpose of an application is to produce a significant intensity for the sum-frequency wave, the two input waves need to have high and comparable intensities.

In the above, we have assumed that the interacting waves are perfect plane waves. In reality, each optical beam has a finite cross section and a nonuniform intensity distribution. This and other spatial effects have to be carefully considered in a detailed

analysis of a sum-frequency generation process, as well as in that of any other nonlinear process to be discussed later. Without carrying out such an analysis, we point out the important, yet easily understood, fact that interaction between two or more optical beams takes place only in the area where those beams overlap spatially and, if the beams are optical pulses, also temporally. Therefore, in terms of optical power, I_1, I_2, and I_3 in (9.102) have to be replaced by $P_1/\mathcal{A}_1$, $P_2/\mathcal{A}_2$, and $P_3/\mathcal{A}_3$, respectively, where P_q is the total power of the wave at the frequency ω_q and $\mathcal{A}_q$ is its effective cross-sectional area. In the low-efficiency limit, we then have $P_3 \propto |d_{\text{eff}}|^2 P_1 P_2 \mathcal{A}_3/\mathcal{A}_1\mathcal{A}_2$. It is important to realize that $\mathcal{A}_3 \leq \min(\mathcal{A}_1, \mathcal{A}_2)$ because the sum-frequency wave is generated only in the area where the two input waves overlap.

We thus arrive at the following conclusions: (1) to maximize the efficiency of sum-frequency generation with two input waves at given power levels, it is important to collimate these two beams to the same cross-sectional area and to have them overlap uniformly so that $\mathcal{A}_3$ is maximized; (2) it is possible to increase the conversion efficiency by focusing the input waves to reduce $\mathcal{A}_1$ and $\mathcal{A}_2$ simultaneously so long as the effective interaction length is not reduced due to the increased divergence of the focused beams; (3) it does not pay to just focus one input beam or to focus the two input beams unevenly because doing so results in a corresponding reduction in $\mathcal{A}_3$.

Difference-frequency generators

By mixing two optical waves, taken to be at ω_3 and ω_1, respectively, a difference-frequency generator produces a third optical wave at the difference frequency $\omega_2 = \omega_3 - \omega_1$, as schematically shown in Fig. 9.2(*a*). Difference-frequency generators are the simplest devices for the generation of coherent infrared radiation, particularly the radiation in the mid to far infrared region where efficient laser materials are rare. For this purpose, both input waves can be in the visible region, or one in the visible and another in the near infrared region, where many efficient lasers sources are available. Wavelength-tunable infrared radiation can be obtained if one of the input waves is from a wavelength-tunable source.

The equations for the description of the difference-frequency generation process are also those given in (9.60)–(9.62), but the boundary conditions are $\mathcal{E}_3(0) \neq 0, \mathcal{E}_1(0) \neq 0$, and $\mathcal{E}_2(0) = 0$ at the input surface of a nonlinear crystal. Similarly to the case of sum-frequency generation, general solutions of the coupled equations with the boundary conditions for difference-frequency generation can be found in terms of elliptic functions. However, also similarly to the case of sum-frequency generation, simple solutions under special situations are often more useful.

In the low-efficiency limit, depletion of the intensities of the two input waves is negligible. By taking the two input fields, $\mathcal{E}_3$ and $\mathcal{E}_1$, to be independent of z, (9.61) can be integrated directly for field $\mathcal{E}_2(z)$ at the difference frequency ω_2. The following

solution for the intensity of the difference-frequency wave is found:

$$I_2(l) = \frac{\omega_2^2 |\chi_{\text{eff}}|^2}{2c^3 \epsilon_0 n_{1,z} n_{2,z} n_{3,z}} I_3 I_1 l^2 \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2}$$
$$= \frac{8\pi^2 |d_{\text{eff}}|^2}{c\epsilon_0 n_{1,z} n_{2,z} n_{3,z} \lambda_2^2} I_3 I_1 l^2 \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2}, \tag{9.104}$$

where $\lambda_2 = 2\pi c/\omega_2$ is the wavelength of the difference-frequency wave in free space. In the case of quasi-phase matching, χ_{eff} and d_{eff} in (9.104) are replaced by χ_{Q} and d_{Q}, respectively.

The relation in (9.104) has the same form as that in (9.102). The effect of phase mismatch is also that shown in Fig. 9.15. With perfect phase matching, $I_2 = I_2^{\text{PM}} \propto l^2/\lambda_2^2$. To produce a significant intensity for the difference-frequency wave, it is also desirable to have two strong input waves with comparable intensities because $I_2 \propto |d_{\text{eff}}|^2 I_3 I_1$. In terms of optical power, we have $P_2 \propto |d_{\text{eff}}|^2 P_3 P_1 \mathcal{A}_2/\mathcal{A}_3 \mathcal{A}_1$, where $\mathcal{A}_2 \leq \min(\mathcal{A}_3, \mathcal{A}_1)$. Therefore, in the low-efficiency limit, the wave produced by a difference-frequency generator has the same general characteristics as discussed for the wave produced by a sum-frequency generator.

One word of caution in the application of a difference-frequency generator goes to the generation of far infrared radiation. When the wavelength of the difference-frequency wave in the far infrared region becomes comparable to, or even larger than, one of the cross-sectional beam diameters of the input waves, the diffraction effect of the long-wavelength difference-frequency wave becomes significant. As a result, the relation in (9.104) is no longer valid. Instead, spatially nonuniform distribution of the difference-frequency wave caused by this diffraction has to be considered though the total power integrated over the entire cross section of the difference-frequency wave is not changed by the diffraction effect.

Second-harmonic generators

By far the most widely used nonlinear optical devices are the second-harmonic generators. An optical harmonic generator produces an optical wave at a frequency that is an integral multiple of the frequency of the input wave. A second-harmonic generator produces a wave at double the frequency of the input wave; thus it is also called an *optical frequency doubler*. In the application of a second-harmonic generator, only two optical waves are involved in the interaction: one input wave at the *fundamental frequency* of ω and a nonlinearly generated wave at the *second-harmonic frequency* of 2ω, as schematically illustrated in Fig. 9.1(*b*).

Following a procedure similar to that leading to the coupled equations of (9.60)–(9.62), we find the following two coupled equations for second-harmonic

generation:

$$\frac{\mathrm{d}\mathcal{E}_{2\omega}}{\mathrm{d}z} = \frac{\mathrm{i}(2\omega)^2}{2c^2 k_{2\omega,z}} \chi_{\mathrm{eff}} \mathcal{E}_{\omega}^2 \mathrm{e}^{\mathrm{i}\Delta k z} = \frac{\mathrm{i}\omega}{c n_{2\omega,z}} \chi_{\mathrm{eff}} \mathcal{E}_{\omega}^2 \mathrm{e}^{\mathrm{i}\Delta k z}, \tag{9.105}$$

$$\frac{\mathrm{d}\mathcal{E}_{\omega}}{\mathrm{d}z} = \frac{\mathrm{i}\omega^2}{c^2 k_{\omega,z}} \chi_{\mathrm{eff}}^* \mathcal{E}_{2\omega} \mathcal{E}_{\omega}^* \mathrm{e}^{-\mathrm{i}\Delta k z} = \frac{\mathrm{i}\omega}{c n_{\omega,z}} \chi_{\mathrm{eff}}^* \mathcal{E}_{2\omega} \mathcal{E}_{\omega}^* \mathrm{e}^{-\mathrm{i}\Delta k z}, \tag{9.106}$$

where $\chi_{\mathrm{eff}} = \hat{e}_{2\omega}^* \cdot \boldsymbol{\chi}^{(2)}(2\omega = \omega + \omega) : \hat{e}_{\omega}\hat{e}_{\omega} = \hat{e}_{\omega} \cdot \boldsymbol{\chi}^{(2)}(\omega = 2\omega - \omega) : \hat{e}_{2\omega}^* \hat{e}_{\omega}$ and $\Delta\mathbf{k} = 2\mathbf{k}_{\omega} - \mathbf{k}_{2\omega} = \Delta k \hat{z}$. Using the relation in (9.63), we find that (9.105) and (9.106) lead to the following relation for the intensities of the fundamental and the second-harmonic waves:

$$\frac{\mathrm{d}I_{2\omega}}{\mathrm{d}z} = -\frac{\mathrm{d}I_{\omega}}{\mathrm{d}z} = -\frac{2\omega|\chi_{\mathrm{eff}}|}{(2c^3\epsilon_0 n_{\omega,z}^2 n_{2\omega,z})^{1/2}} I_{\omega} I_{2\omega}^{1/2} \sin\varphi, \tag{9.107}$$

where $\varphi = \varphi_{\chi} + 2\varphi_{\omega} - \varphi_{2\omega} + \Delta k z$. Therefore, we find the following Manley–Rowe relations for second-harmonic generation that involves only two optical beams:

$$\frac{\mathrm{d}I}{\mathrm{d}z} = \frac{\mathrm{d}(I_{\omega} + I_{2\omega})}{\mathrm{d}z} = 0 \tag{9.108}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_{\omega}}{\omega}\right) = -2\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_{2\omega}}{2\omega}\right). \tag{9.109}$$

As expected, two photons at the fundamental frequency are annihilated to create each photon at the second-harmonic frequency.

In the low-efficiency limit, depletion of the intensity of the fundamental beam can be neglected. Then, (9.105) can be integrated directly for $\mathcal{E}_{2\omega}(z)$ by taking $\mathcal{E}_{\omega}$ to be independent of z. The result, expressed in terms of second-harmonic intensity as a function of interaction length, is

$$\begin{aligned} I_{2\omega}(l) &= \frac{\omega^2|\chi_{\mathrm{eff}}|^2}{2c^3\epsilon_0 n_{\omega,z}^2 n_{2\omega,z}} I_{\omega}^2 l^2 \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2} \\ &= \frac{8\pi^2|d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^2 n_{2\omega,z}\lambda^2} I_{\omega}^2 l^2 \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2}, \end{aligned} \tag{9.110}$$

where $\lambda = 2\pi/\omega$ is the wavelength of the fundamental wave in free space. In the low-efficiency limit, $I_{2\omega} \propto |d_{\mathrm{eff}}|^2 I_{\omega}^2$, or $P_{2\omega} \propto |d_{\mathrm{eff}}|^2 P_{\omega}^2 \mathcal{A}_{2\omega}/\mathcal{A}_{\omega}^2$, where $\mathcal{A}_{\omega}$ and $\mathcal{A}_{2\omega}$ are the effective cross-sectional areas of the fundamental and second-harmonic beams, respectively, and $\mathcal{A}_{2\omega} \leq \mathcal{A}_{\omega}$ due to the nonlinear nature of the second-harmonic generation process.

Perfect phase matching is required if a high efficiency for second-harmonic generation is desired. In addition, according to the discussions in Section 9.4, it is also necessary to have $\varphi = -\pi/2$. This condition is automatically satisfied if perfect phase

matching is accomplished and if the input consists of only the fundamental wave because, without any coherent second-harmonic field at the input, only the second-harmonic field that has the most favorable phase is generated and subsequently amplified. The Manley–Rowe relation in (9.108) states that the total intensity of the fundamental and second-harmonic waves remains constant throughout the interaction: $I = I_\omega + I_{2\omega} = I_\omega(0)$ for $I_{2\omega}(0) = 0$. Under these conditions, (9.107) leads to

$$\frac{\mathrm{d}I_{2\omega}}{\mathrm{d}z} = \frac{2\omega|\chi_{\mathrm{eff}}|}{(2c^3\epsilon_0 n_{\omega,z}^3)^{1/2}}[I_\omega(0) - I_{2\omega}]I_{2\omega}^{1/2}. \tag{9.111}$$

Note that with perfect phase matching, $n_{\omega,z} = n_{2\omega,z}$. By making the change of variable $u^2 = I_{2\omega}/I_\omega(0)$ and by using the fact that $u = \tanh \kappa z$ is the solution of the equation $\mathrm{d}u/\mathrm{d}z = \kappa(1 - u^2)$, we can solve (9.111) to obtain the following general results for second-harmonic generation with perfect phase matching:

$$I_{2\omega}(l) = I_\omega(0)\tanh^2 \kappa l, \tag{9.112}$$

$$I_\omega(l) = I_\omega(0)\,\mathrm{sech}^2 \kappa l, \tag{9.113}$$

where

$$\kappa = \left[\frac{\omega^2|\chi_{\mathrm{eff}}|^2}{2c^3\epsilon_0 n_{\omega,z}^3} I_\omega(0)\right]^{1/2} = \left[\frac{8\pi^2|d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^3\lambda^2} I_\omega(0)\right]^{1/2}. \tag{9.114}$$

These results are plotted in Fig. 9.16. With perfect phase matching, it is theoretically possible to convert all of the fundamental power to the second harmonic if the interaction length is sufficiently long. In the case of quasi-phase matching, χ_{eff} and d_{eff} in (9.114) are replaced by χ_{Q} and d_{Q}, respectively.

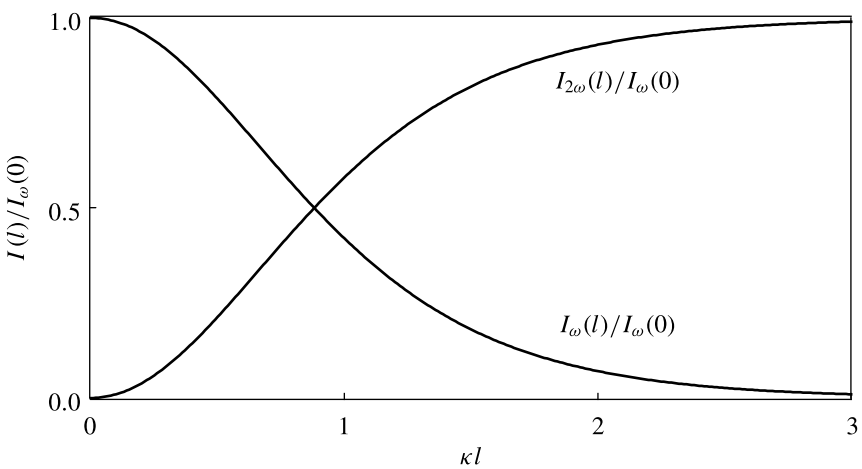


Figure 9.16 Intensities of the fundamental and second-harmonic waves, normalized to the total intensity, as a function of interaction length in a second-harmonic generator with perfect phase matching.

The conversion efficiency of a second-harmonic generator is commonly defined as

$$\eta_{\mathrm{SH}} = \frac{P_{2\omega}(l)}{P_{\omega}(0)}. \tag{9.115}$$

In the low-efficiency limit with perfect phase matching,

$$\eta_{\mathrm{SH}} = \frac{8\pi^2 |d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^3 \lambda^2} \frac{\mathcal{A}_{2\omega}}{\mathcal{A}_{\omega}^2} P_{\omega}(0) l^2. \tag{9.116}$$

Because η_{SH} in the low-efficiency limit is linearly proportional to the fundamental power, it is convenient to define a *normalized second-harmonic conversion efficiency* as

$$\hat{\eta}_{\mathrm{SH}} = \frac{\eta_{\mathrm{SH}}}{P_{\omega}(0)} = \frac{8\pi^2 |d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^3 \lambda^2} \frac{\mathcal{A}_{2\omega}}{\mathcal{A}_{\omega}^2} l^2. \tag{9.117}$$

There is a relation between $\mathcal{A}_{2\omega}$ and $\mathcal{A}_{\omega}$ that depends on the cross-sectional profile of the fundamental beam. For example, $\mathcal{A}_{2\omega} = \mathcal{A}_{\omega}/2$ if the beam has a Gaussian profile, but $\mathcal{A}_{2\omega} = \mathcal{A}_{\omega}$ if the beam has a uniform profile. We see from (9.117) that the conversion efficiency can be raised by focusing the fundamental beam to reduce its cross-sectional area, provided that the beam remains well collimated. Focusing the beam too tightly increases the beam divergence, thus reducing its intensity outside the Rayleigh range from the beam waist. In addition, the conversion efficiency can be reduced by any walk-off between the interacting beams.

The second-harmonic generation efficiency of a focused Gaussian beam is a function of three characteristic lengths: the crystal length l, the confocal parameter $b = 2\pi n w_0^2/\lambda$, and the aperture length $l_{\mathrm{a}} = \pi^{1/2} w_0/\rho$ defined in (9.90). For $b \gg l$ and $l_{\mathrm{a}} \gg l$, the dependence of $\hat{\eta}_{\mathrm{SH}}$ on l^2 seen in (9.117) is valid, and $\hat{\eta}_{\mathrm{SH}}$ can be expressed in the following form:

$$\hat{\eta}_{\mathrm{SH}} = \frac{\eta_{\mathrm{SH}}}{P_{\omega}(0)} = \frac{16\pi^2 |d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^2 \lambda^3} \frac{l^2}{b}. \tag{9.118}$$

For $b < l < 10b$ or $l_{\mathrm{a}} < l$, (9.117) and (9.118) are not valid, but the conversion efficiency can be approximated by

$$\hat{\eta}_{\mathrm{SH}} = \frac{\eta_{\mathrm{SH}}}{P_{\omega}(0)} = \frac{16\pi^2 |d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^2 \lambda^3} \frac{1.068 l}{1 + lb/l_{\mathrm{a}}^2}. \tag{9.119}$$

We see that if $lb \gg l_{\mathrm{a}}^2$, the conversion efficiency is independent of crystal length as $\eta_{\mathrm{SH}} \propto l_{\mathrm{a}}^2/b$ in this situation. The best efficiency that can be obtained with an optimally focused Gaussian beam is

$$\hat{\eta}_{\mathrm{SH}} = \frac{\eta_{\mathrm{SH}}}{P_{\omega}(0)} = \frac{16\pi^2 |d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{\omega,z}^2 \lambda^3} (1.068 l), \tag{9.120}$$

which occurs under the conditions of no walk-off so that $l_{\mathrm{a}} = \infty$ and $l = 2.84b$. We see that, with the fundamental beam optimally focused for the best efficiency, the conversion

efficiency increases only linearly with crystal length but the focused beam waist spot area has to vary linearly with crystal length to maintain this optimum condition. Note that in the case of quasi-phase matching, d_{eff} that appears in the expressions of η_{SH} and $\hat{\eta}_{\text{SH}}$ given in (9.116)–(9.120) has to be replaced by d_{Q}.

We also see that the conversion efficiency increases linearly with an increase in the input power of the fundamental beam. This statement is true as long as we stay in the low-efficiency limit so that depletion of the fundamental beam is negligible. In the high-efficiency regime, conversion efficiency increases sublinearly with the input power of the fundamental beam. According to (9.112), it is theoretically possible to have 100% conversion efficiency for second-harmonic generation if the input power is sufficiently high and the interaction length is sufficiently large. However, the conversion efficiency of a practical device is usually limited by the damage threshold of a nonlinear crystal, as well as by many complicated spatial and temporal effects.

For many practical applications, it is often necessary to generate the third harmonic or the fourth harmonic of a fundamental wave. As discussed in Section 9.3, the third harmonic can be generated with a parametric third-order nonlinear process characterized by $\chi^{(3)}(3\omega = \omega + \omega + \omega)$, However, a third-harmonic generator using a third-order nonlinear process is of little practical usefulness for two reasons: (1) the value of $\chi^{(3)}$, though always nonvanishing, is orders of magnitude smaller than the value of $\chi^{(2)}$ of any commonly used nonlinear crystals; (2) phase matching is very difficult for such a process. In practice, efficient third-harmonic generation is normally carried out by following second-harmonic generation with sum-frequency generation for $\omega + 2\omega \rightarrow 3\omega$, as shown in Fig. 9.17($a$). Similarly, fourth-harmonic generation is accomplished by cascading two second-harmonic generators by first doubling ω and then doubling 2ω to obtain 4ω, as shown in Fig. 9.17(b). These possibilities are already demonstrated in Example 9.4.

EXAMPLE 9.12 In this example, we consider the second-harmonic conversion efficiency with a focused Gaussian beam at $\lambda = 1.10$ μm in $LiNbO_3$ under different phase-matching conditions discussed in Examples 9.8–9.11. Perfect phase matching is assumed for each case, with $\Delta k_{\text{Q}} = 0$ in the case of quasi-phase matching. The fundamental beam is focused to have its beam waist located at the center of a crystal of $l = 1$ cm length. (a) With angle phase matching as described in Example 9.8, what is the normalized efficiency $\hat{\eta}_{\text{SH}}$ if the beam is focused to have a beam waist radius of $w_0 = 50$ μm? (b) With 90° phase matching by temperature tuning as described in Example 9.10, what is $\hat{\eta}_{\text{SH}}$ for $w_0 = 50$ μm? (c) With 90° phase matching, the conversion efficiency can be increased by optimum focusing. What is the optimum beam waist radius for this purpose? What is the best conversion efficiency? (d) With quasi-phase matching in a PPLN crystal as described in Example 9.11, what is the best attainable conversion efficiency?

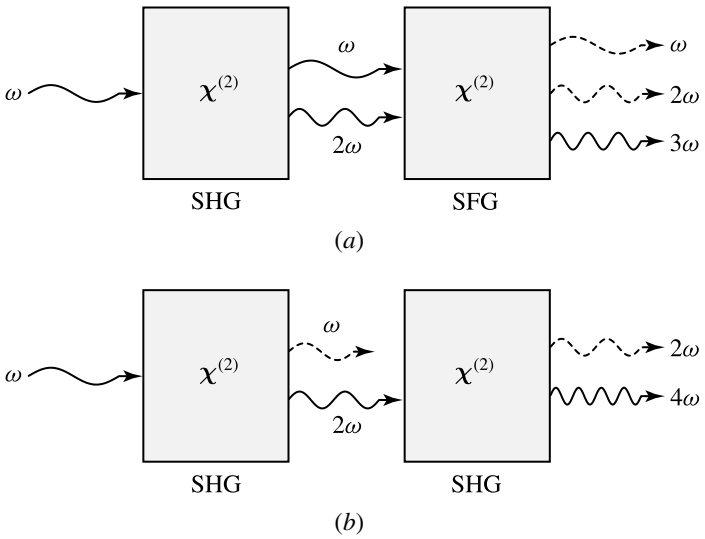


Figure 9.17 (*a*) A third-harmonic generator consisting of a second-harmonic generator (SHG) and a sum-frequency generator (SFG) in cascade. (*b*) A fourth-harmonic generator consisting of two second-harmonic generators in cascade.

Solution (a) With type I angle phase matching, we have $|d_{\text{eff}}| = 4.88$ pm V^{-1} and $n_{\omega,z} = n_\omega^{\text{o}} = 2.2319$ from Example 9.8. With $w_0 = 50$ μm and $\lambda = 1.10$ μm, we find that $l_{\text{a}} = 4.54$ mm from Example 9.9 and that $b = 3.19$ cm. Because $l_{\text{a}} < l$, (9.119) has to be used to estimate the efficiency. Because $lb \gg l_{\text{a}}^2$, we find that

$$\begin{aligned}\hat{\eta}_{\text{SH}} &\approx \frac{16\pi^2|d_{\text{eff}}|^2(1.068l_{\text{a}}^2)}{c\epsilon_0 n_{\omega,z}^2\lambda^3 b} \\ &= \frac{16\pi^2 \times (4.88 \times 10^{-12})^2 \times 1.068 \times (4.54 \times 10^{-3})^2}{3 \times 10^8 \times 8.85 \times 10^{-12} \times 2.2319^2 \times (1.10 \times 10^{-6})^3 \times 3.19 \times 10^{-2}} \text{ W}^{-1} \\ &= 0.015\% \text{ W}^{-1}.\end{aligned}$$

(b) With 90° type I phase matching, we have $|d_{\text{eff}}| = 4.4$ pm V^{-1}, $n_{\omega,z} = n_\omega^{\text{o}} = 2.2321$ at $T = 396.7$ K, and $l_{\text{a}} = \infty$ from Example 9.10. With $w_0 = 50$ μm, $\lambda = 1.10$ μm, and $n = 2.2321$, we still have $b = 3.19$ cm. In this case, (9.118) is valid because $b > 3l$ and $l_{\text{a}} \gg l$. Therefore, we find that

$$\begin{aligned}\hat{\eta}_{\text{SH}} &= \frac{16\pi^2|d_{\text{eff}}|^2 l^2}{c\epsilon_0 n_{\omega,z}^2\lambda^3 b} \\ &= \frac{16\pi^2 \times (4.4 \times 10^{-12})^2 \times (1 \times 10^{-2})^2}{3 \times 10^8 \times 8.85 \times 10^{-12} \times 2.2321^2 \times (1.10 \times 10^{-6})^3 \times 3.19 \times 10^{-2}} \text{ W}^{-1} \\ &= 0.054\% \text{ W}^{-1}.\end{aligned}$$

We see that the conversion efficiency is 3.6 times that found in (a) by using 90° phase matching to eliminate the walk-off effect.

(c) In the absence of walk-off for 90° phase matching, the best efficiency can be obtained by making $b = l/2.84 = 3.52$ mm, which can be accomplished by focusing the fundamental beam to the following beam waist radius:

$$w_0 = \left(\frac{\lambda b}{2\pi n}\right)^{1/2} = \left(\frac{1.10 \times 10^{-6} \times 3.52 \times 10^{-3}}{2\pi \times 2.2321}\right)^{1/2} \text{ m} = 16.6\ \mu\text{m}.$$

The efficiency is found by using (9.120) to be

$$\begin{aligned}\hat{\eta}_{\text{SH}} &= \frac{16\pi^2 |d_{\text{eff}}|^2}{c\epsilon_0 n_{\omega,z}^2 \lambda^3}(1.068l) \\ &= \frac{16\pi^2 \times (4.4 \times 10^{-12})^2 \times 1.068 \times 1 \times 10^{-2}}{3 \times 10^8 \times 8.85 \times 10^{-12} \times 2.2321^2 \times (1.10 \times 10^{-6})^3}\ \text{W}^{-1} \\ &= 0.186\%\ \text{W}^{-1}.\end{aligned}$$

We see that this efficiency is more than three times that found in (b) by optimally focusing the beam in the absence of walk-off.

(d) Because there is no walk-off in the case of quasi-phase matching in a PPLN crystal described in Example 9.11, we can still take $b = l/2.84 = 3.52$ mm. Because $n_{\omega,z} = n_\omega^{\text{e}} = 2.1536$ and $n_{2\omega}^{\text{e}} = 2.2260$ in this situation, we find that $w_0 = 16.9$ μm, which is slightly larger than that found in (c). The best efficiency is still found by using (9.120) but with $|d_{\text{eff}}|$ replaced by $|d_{\text{Q}}| = 16.04$ pm V^{-1} found in Example 9.11. Therefore,

$$\begin{aligned}\hat{\eta}_{\text{SH}} &= \frac{16\pi^2 |d_{\text{Q}}|^2}{c\epsilon_0 n_{\omega,z} n_{2\omega,z} \lambda^3}(1.068l) \\ &= \frac{16\pi^2 \times (16.04 \times 10^{-12})^2 \times 1.068 \times 1 \times 10^{-2}}{3 \times 10^8 \times 8.85 \times 10^{-12} \times 2.1536 \times 2.2260 \times (1.10 \times 10^{-6})^3}\ \text{W}^{-1} \\ &= 2.56\%\ \text{W}^{-1}.\end{aligned}$$

This conversion efficiency is about 14 times that found in (c) because quasi-phase matching using a PPLN crystal allows us to take advantage of the largest nonlinear susceptibility element d_{33} of $LiNbO_3$.

This example illustrates how the efficiency of a second-harmonic generator can be substantially increased by a combination of optimization procedures. Further increase of efficiency is possible using a waveguide structure, as illustrated later in Example 9.23, or by using short optical pulses to increase the peak intensity at a given average power level. The same techniques can be applied generally to other nonlinear frequency converters discussed in this section for increasing their conversion efficiencies.

Optical parametric frequency converters

The function of an optical parametric frequency converter is the conversion of a signal-carrying optical wave from one carrier frequency to another through *parametric*

up-conversion or *parametric down-conversion*. Parametric up-conversion is a special case of sum-frequency generation with the objective of converting a signal-carrying optical wave at a low frequency, typically in the mid or far infrared region, where sensitive detectors are not available, to an optical wave carrying the same signal at a frequency in the visible region, where efficient detection can be easily made. Parametric down-conversion is a special case of difference-frequency generation in which a signal-carrying optical wave at a high frequency, often in the ultraviolet region, is converted to one at a low frequency in the visible or the infrared region. The signal-carrying input wave, which is called the *signal*, is generally very weak in comparison to the other input wave, which is called the *pump*. In the following analysis, the strong pump wave is taken to be at ω_2. The signal is taken to be at ω_1 for up-conversion and is taken to be at ω_3 for down-conversion. The relation $\omega_3 = \omega_1 + \omega_2$ applies to both cases.

Because the pump is much stronger than the signal, the intensity of the pump can be considered to be constant though that of the signal is not. As a result, we have the following coupled equations for parametric conversion processes:

$$\frac{\mathrm{d}\mathcal{E}_3}{\mathrm{d}z} = \mathrm{i}\left(\frac{\omega_3^2}{c^2 k_{3,z}}\chi_{\mathrm{eff}}\mathcal{E}_2\right)\mathcal{E}_1 \mathrm{e}^{\mathrm{i}\Delta k z} = \mathrm{i}\kappa_{31}\mathcal{E}_1 \mathrm{e}^{\mathrm{i}\Delta k z}, \tag{9.121}$$

$$\frac{\mathrm{d}\mathcal{E}_1}{\mathrm{d}z} = \mathrm{i}\left(\frac{\omega_1^2}{c^2 k_{1,z}}\chi_{\mathrm{eff}}^{*}\mathcal{E}_2^{*}\right)\mathcal{E}_3 \mathrm{e}^{-\mathrm{i}\Delta k z} = \mathrm{i}\kappa_{13}\mathcal{E}_3 \mathrm{e}^{-\mathrm{i}\Delta k z}. \tag{9.122}$$

These two equations have the form of the coupled equations of (4.57) and (4.58), which are solved in Section 4.3. Because the signal is normally weak in the application of a parametric converter, a high conversion efficiency is most desirable. Therefore, the device is normally used under the condition of perfect phase matching (see Problem 9.6.8).

For up-conversion, the boundary conditions are $\mathcal{E}_1(0) \neq 0$ and $\mathcal{E}_3(0) = 0$. The solutions under the condition of perfect phase matching are

$$\mathcal{E}_1(l) = \mathcal{E}_1(0)\cos\kappa l, \tag{9.123}$$

$$\mathcal{E}_3(l) = \frac{\mathrm{i}\kappa_{31}}{\kappa}\mathcal{E}_1(0)\sin\kappa l, \tag{9.124}$$

where

$$\kappa = (\kappa_{31}\kappa_{13})^{1/2} = \left(\frac{\omega_1\omega_3|\chi_{\mathrm{eff}}|^2}{2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z}}I_2\right)^{1/2} = \left(\frac{8\pi^2|d_{\mathrm{eff}}|^2}{c\epsilon_0 n_{1,z}n_{2,z}n_{3,z}\lambda_1\lambda_3}I_2\right)^{1/2}. \tag{9.125}$$

In the case of quasi-phase matching, χ_{eff} and d_{eff} in (9.125) are replaced by χ_{Q} and d_{Q}, respectively.

The schematic diagram of an optical parametric up-converter is shown in Fig. 9.18(*a*). For a parametric up-converter with perfect phase matching, the intensities of the three

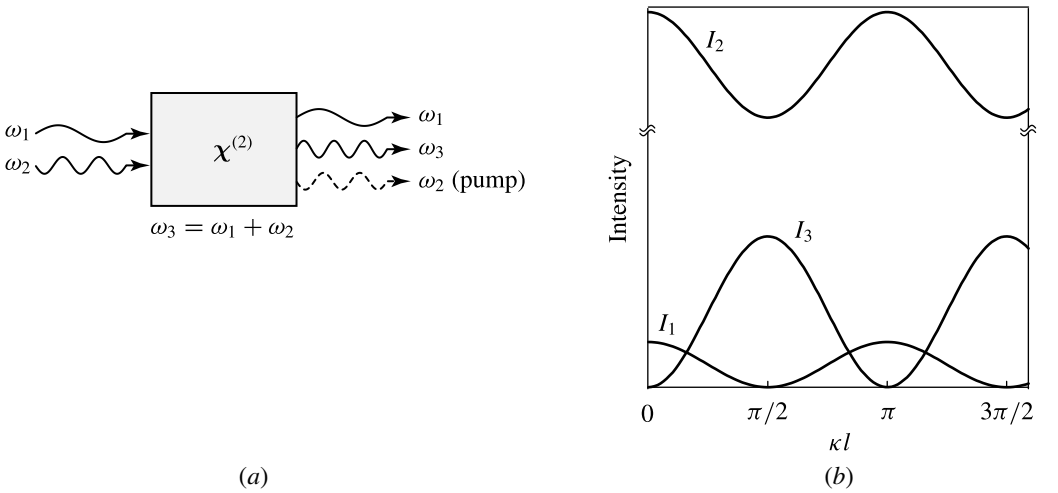


Figure 9.18 (*a*) Schematics of an optical parametric up-converter. (*b*) Intensity variations of the interacting optical waves as a function of interaction length. The pump wave is at ω_2. The signal wave is at ω_1 for up-conversion.

interacting beams vary with interaction length as

$$I_3(l) = \frac{\omega_3}{\omega_1} I_1(0) \sin^2 \kappa l, \tag{9.126}$$

$$I_1(l) = I_1(0) \cos^2 \kappa l, \tag{9.127}$$

$$I_2(l) = I_2(0) - \frac{\omega_2}{\omega_1} I_1(0) \sin^2 \kappa l \approx I_2(0). \tag{9.128}$$

Figure 9.18(*b*) illustrates these intensity variations. Complete up-conversion of the signal occurs at an interaction length of $l_c^{\text{PM}} = \pi/2\kappa$, as expected of phase-matched codirectional coupling. The value of this length can be varied by varying the pump intensity because the value of κ depends on that of I_2. Note that when the signal intensity is completely depleted by up-conversion, the intensity of the sum-frequency wave reaches a maximum value of $I_3^{\max} = I_1(0)\omega_3/\omega_1 > I_1(0)$ because the total number of sum-frequency photons that are created is equal to the total number of signal photons that are annihilated.

Parametric down-conversion is simply the reverse process of up-conversion, and vice versa. The same parametric converter can function as either an up-converter or a down-converter. The only difference is the initial conditions at the input. If the initial conditions are $\mathcal{E}_1(0) = 0$ and $\mathcal{E}_3(0) \neq 0$, the device functions as a down-converter. In Fig. 9.18, we see clearly that when the intensity of the wave at ω_1 is completely depleted, for example, at a distance of $l = l_c^{\text{PM}}$, further interaction in the parameter converter leads to down-conversion from the wave at ω_3 back to the wave at ω_1.

Optical parametric amplifiers

The physical process involved in an optical parametric amplifier, commonly called an OPA, is basically the same as that in a difference-frequency generator. The only difference is in the usage of the device. In either case, there are two input waves at ω_1 and ω_3. While the usage of a difference-frequency generator is for generation of a wave at the difference frequency $\omega_2 = \omega_3 - \omega_1$, that of an OPA is for amplification of the input wave at ω_1. The wave at the difference frequency ω_2 is still generated in an OPA though it is not the purpose of this application. Therefore, the high-frequency input wave at ω_3 is called the *pump* wave, the low-frequency input wave at ω_1 is called the *signal* wave, and the side product at ω_2 is called the *idler* wave, as shown in Fig. 9.19(*a*).

Normally the pump wave of an OPA is much stronger than the signal wave and can be considered constant throughout the interaction. Therefore, only (9.61) and (9.62) have to be considered, and the initial conditions are $\mathcal{E}_1(0) \neq 0$ and $\mathcal{E}_2(0) = 0$. We have the following coupled equations:

$$\frac{\mathrm{d}\mathcal{E}_1}{\mathrm{d}z} = \mathrm{i}\left(\frac{\omega_1^2}{c^2 k_{1,z}}\chi_{\mathrm{eff}}^*\mathcal{E}_3\right)\mathcal{E}_2^*\mathrm{e}^{-\mathrm{i}\Delta k z} = \mathrm{i}\kappa_{12}\mathcal{E}_2^*\mathrm{e}^{-\mathrm{i}\Delta k z}, \tag{9.129}$$

$$\frac{\mathrm{d}\mathcal{E}_2^*}{\mathrm{d}z} = \mathrm{i}\left(-\frac{\omega_2^2}{c^2 k_{2,z}}\chi_{\mathrm{eff}}\mathcal{E}_3^*\right)\mathcal{E}_1\mathrm{e}^{\mathrm{i}\Delta k z} = \mathrm{i}\kappa_{21}\mathcal{E}_1\mathrm{e}^{\mathrm{i}\Delta k z}, \tag{9.130}$$

where (9.130) is obtained by taking the complex conjugate of (9.62). Again, these two coupled equations have the form of the coupled equations of (4.57) and (4.58), and the solutions in Sections 4.3 can be applied directly. For efficient parametric amplification, phase matching is required. By identifying $\beta_c = (\kappa_{12}\kappa_{21})^{1/2} = \mathrm{i}\kappa$ in the case of perfect

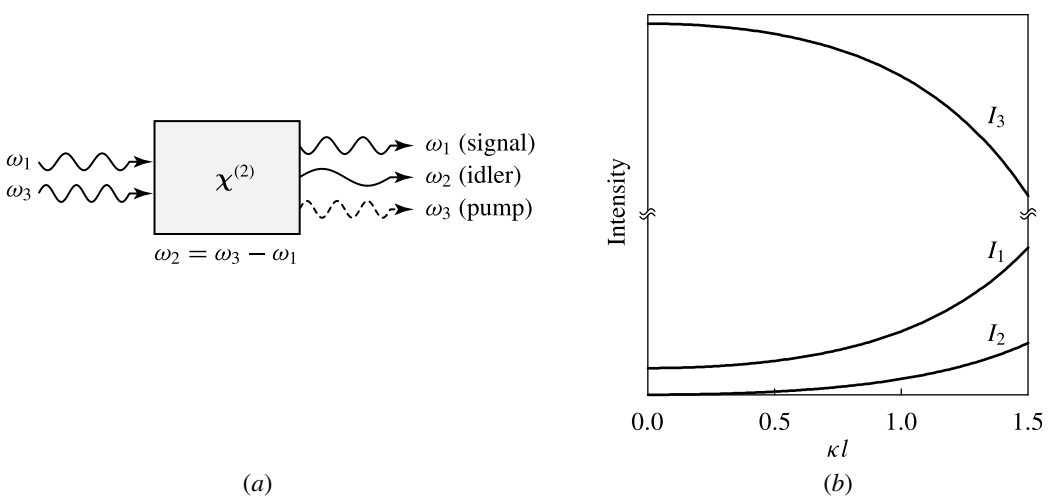


Figure 9.19 (*a*) Schematics of an OPA. (*b*) Intensity variations of the pump, signal, and idler waves of an OPA with a strong pump as a function of interaction length in the case of perfect phase matching.

phase matching, we have the following solutions:

$$\mathcal{E}_1(z) = \mathcal{E}_1(0)\cos\beta_c z = \mathcal{E}_1(0)\cosh\kappa z, \tag{9.131}$$

$$\mathcal{E}_2^*(z) = \frac{i\kappa_{21}}{\beta_c}\mathcal{E}_1(0)\sin\beta_c z = \frac{i\kappa_{21}}{\kappa}\mathcal{E}_1(0)\sinh\kappa z, \tag{9.132}$$

where

$$\kappa = \left(\frac{\omega_1\omega_2|\chi_{\text{eff}}|^2}{2c^3\epsilon_0 n_{1,z}n_{2,z}n_{3,z}}I_3\right)^{1/2} = \left(\frac{8\pi^2|d_{\text{eff}}|^2}{c\epsilon_0 n_{1,z}n_{2,z}n_{3,z}\lambda_1\lambda_2}I_3\right)^{1/2}. \tag{9.133}$$

In the case of quasi-phase matching, χ_{eff} and d_{eff} in (9.133) are replaced by χ_Q and d_Q, respectively.

With perfect phase matching, the intensities of the signal, idler, and pump waves vary with interaction length as

$$I_1(l) = I_1(0)\cosh^2\kappa l, \tag{9.134}$$

$$I_2(l) = \frac{\omega_2}{\omega_1}I_1(0)\sinh^2\kappa l, \tag{9.135}$$

$$I_3(l) = I_3(0) - \frac{\omega_3}{\omega_1}I_1(0)\sinh^2\kappa l \approx I_3(0), \tag{9.136}$$

which are plotted in Fig. 9.19(*b*). We see that while the intensity of the signal wave grows as a result of parametric amplification, the intensity of the idler wave also increases because an idler photon is generated simultaneously with each additional signal photon generated in the parametric process.

With perfect phase matching, the *amplification factor*, or the *intensity gain*, of the signal wave for a single pass through an OPA is (see Problem 9.6.10)

$$G = \frac{I_1(l)}{I_1(0)} = \cosh^2\kappa l \approx \begin{cases} 1+\kappa^2 l^2, & \text{in the low-gain limit,} \\ \dfrac{e^{2\kappa l}}{4}, & \text{in the high-gain limit.} \end{cases} \tag{9.137}$$

Note that a large gain factor does not necessarily imply a high conversion efficiency from the pump to the signal and idler because the input signal can be extremely weak. Therefore, it is possible that the pump is not much depleted when the signal is amplified by a large gain factor but the conversion efficiency is low. When the input signal is strong, however, it is also possible that pump depletion is significant but the gain factor is small.

EXAMPLE 9.13 An OPA for a signal wavelength at $\lambda_1 = 1.55$ μm consists of a PPLN crystal that has a length of $l = 1$ cm. It is pumped with a Gaussian beam at $\lambda_3 = 527$ nm, which is focused to a beam waist radius of $w_0 = 50$ μm. (a) What is the idler wavelength? (b) What is the required first-order grating period for quasi-phase matching? (c) What is the amplification factor for the signal if the power of the pump beam is $P = 1$ W? (d) What is the required pump power for an amplification factor of $G = 10^3$? Consider only the situation where the pump is not much depleted even when $G = 10^3$.

Solution (a) The wavelengths of the interacting beams in a parametric amplifier have the relation $\lambda_3^{-1} = \lambda_1^{-1} + \lambda_2^{-1}$ given in (9.138) below. Therefore, the idler wavelength is

$$\lambda_2 = \left(\frac{1}{\lambda_3} - \frac{1}{\lambda_1}\right)^{-1} = \left(\frac{1}{527 \times 10^{-9}} - \frac{1}{1.55 \times 10^{-6}}\right)^{-1} \text{ m} = 798 \text{ nm}.$$

(b) For the most efficient interaction in a PPLN crystal, all of the interacting waves have to be extraordinary waves polarized in the z direction. Using the data given in Table 9.3 for the Sellmeier equation of $LiNbO_3$, we find that $n_3^e = 2.2351$ at $\lambda_3 = 527$ nm, $n_1^e = 2.1373$ at $\lambda_1 = 1.55$ μm, and $n_2^e = 2.1755$ at $\lambda_3 = 798$ nm. The phase mismatch is $\Delta k = k_1 + k_2 - k_3 = 2\pi(n_1/\lambda_1 + n_2/\lambda_2 - n_3/\lambda_3)$ for collinear interaction. Therefore, according to (9.99), the required first-order grating period is

$$\begin{aligned}\Lambda &= \frac{2\pi}{|\Delta k|} = \left|\frac{n_1}{\lambda_1} + \frac{n_2}{\lambda_2} - \frac{n_3}{\lambda_3}\right|^{-1} \\ &= \left|\frac{2.1373}{1.55 \times 10^{-6}} + \frac{2.1755}{798 \times 10^{-9}} - \frac{2.2351}{527 \times 10^{-9}}\right|^{-1} \text{ m} \\ &= 7.35 \text{ μm}.\end{aligned}$$

(c) For a Gaussian pump beam that is focused to a waist size of $w_0 = 50$ μm, we find that its confocal parameter is $b = 2\pi n_3^e w_0^2/\lambda_3 = 6.66$ cm. Because $b \gg l = 1$ cm, we can ignore the complicated effect of focusing and take $I_3 = P_3/\mathcal{A}_3 = 2P/\pi w_0^2$ over the entire length of the PPLN crystal. For this interaction, we have $|d_Q| = |2d_{33}/\pi| = 16.04$ pm V^{-1} from Example 9.11. Then, using (9.133) with d_{eff} replaced by d_Q, we can express $\kappa^2 l^2$ as a function of the pump power:

$$\kappa^2 l^2 = \frac{16\pi |d_Q|^2 l^2}{c\epsilon_0 n_1^e n_2^e n_3^e \lambda_1 \lambda_2 w_0^2} P_3 = 0.015 P_3 \text{ W}^{-1}.$$

For $P_3 = 1$ W, the single-pass amplification factor is $G \approx 1 + \kappa^2 l^2 = 1.015$ according to (9.137). The signal intensity grows only 1.5% in a single pass through the parametric amplifier.

(d) For an amplification factor of $G = 10^3$, we find by using the high-gain limit of (9.137) that $\kappa l = 4.147$ is required. From the dependence of $\kappa^2 l^2$ on P_3 found in (c), we find that the required pump power for $G = 10^3$ is

$$P_3 = \frac{\kappa^2 l^2}{0.015} \text{ W} = \frac{4.147^2}{0.015} \text{ W} = 1.15 \text{ kW}.$$

This pump power looks unrealistically high. It is indeed unrealistic if we consider only the possibility of CW pump beams. It is not if we consider pulse pumping. For example, by using a Q-switched laser pulse of duration $\Delta t_{ps} = 100$ ns, such a pump power requires a very common pump pulse energy of $U_{ps} = P_{pk}\Delta t_{ps} = 115$ μJ. As another example, if mode-locked pulses of pulsewidth $\Delta t_{ps} = 10$ ps at a repetition rate

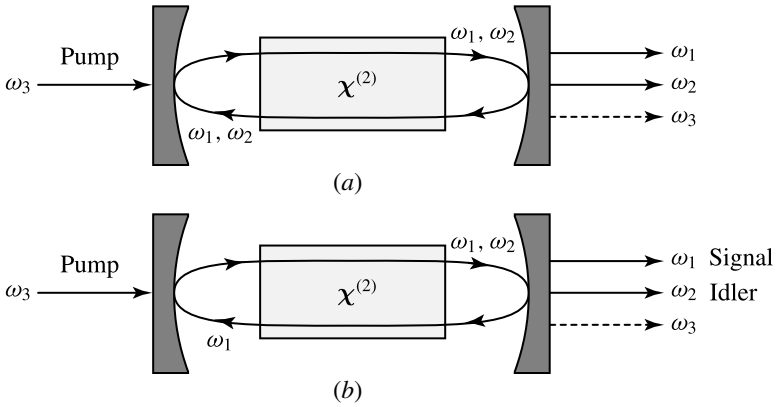


Figure 9.20 Schematic diagrams of (*a*) a doubly resonant OPO, in which both ω_1 and ω_2 are resonated, and (*b*) a singly resonant OPO, in which only ω_1 is resonated.

of $f_{ps} = 100$ MHz are used to pump the amplifier, the average power of the pulsed pump beam is again at a realistic level of $\overline{P} = P_{pk}\Delta t_{ps} f_{ps} = 1.15$ W.

Optical parametric oscillators

The parametric gain can be utilized to construct an optical parametric oscillator, commonly called an OPO, by placing a parametric amplifier in a resonant optical cavity that provides feedback to the parametric amplifier. There are basically two different types of OPOs. In a *doubly resonant* OPO, both waves at ω_1 and ω_2 are resonated because the mirrors of the optical cavity are highly reflective at both frequencies, as shown in Fig. 9.20(*a*). In a *singly resonant* OPO, the mirrors of the optical cavity are highly reflective at only one frequency, either ω_1 or ω_2, and only one wave is resonated, as shown in Fig. 9.20(*b*). The cavity mirrors are transparent to the pump wave and, in the singly resonant case, also to the nonresonant parametric wave.

The input to an OPO consists of only the pump wave at ω_3 to pump the nonlinear crystal for a parametric gain. When the parametric gain is high enough so that the round-trip loss in the optical resonator is compensated by the parametric gain, the oscillator reaches its *threshold* and parametric oscillation occurs. Because the parametric gain is a function of the pump intensity, the threshold parametric gain for an OPO translates into a threshold pump intensity required of the pump beam. Resonant oscillation builds up from the spontaneous emission noise of parametric fluorescence. No signal input is needed.

Because both low-frequency parametric waves at ω_1 and ω_2 are generated in the oscillator without a signal input, either of them can be called the signal or the idler. The designation of one particular wave to be called the signal is purely a matter of one's subjective interest. However, the choice of the resonating frequency in a singly resonant OPO is usually not arbitrary but is based on many practical considerations,

such as the availability of high-quality cavity mirrors at either of the two parametric frequencies, the spectral characteristics of the transmittance of the nonlinear crystal, and other wavelength-dependent characteristics of the optical cavity.

The frequencies and, correspondingly, the wavelengths of a parametric oscillator are subject to the following conditions:

$$\omega_3 = \omega_1 + \omega_2 \quad \text{and} \quad \frac{1}{\lambda_3} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}, \tag{9.138}$$

which are required by conservation of energy because one photon at ω_3 splits into a pair of photons at ω_1 and ω_2. The exact frequencies to be generated by the oscillator are further dictated by the following two conditions: (1) the phase-matching condition

$$\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2, \tag{9.139}$$

which is determined by the properties and the physical arrangement of the nonlinear crystal; and (2) the resonance condition of the optical cavity, which depends on the physical parameters of the cavity and determines the resonance optical frequencies. The peak parametric gain appears at frequencies that satisfy the phase-matching condition exactly. The oscillation frequencies are those, subject to the condition in (9.138), that satisfy the resonance condition of the optical resonator with the least amount of phase mismatch. Therefore, the signal and idler frequencies of an OPO can be simultaneously tuned, though in opposite directions due to the constraint of (9.138), by varying the phase-matching condition in the crystal while the pump frequency is fixed. This wavelength tunability is one of the most important characteristics of OPOs. Another important characteristic is that the parametric gain is not tied to any resonant transitions in the gain medium because the gain medium is a parametric nonlinear crystal. These two key characteristics make the OPOs unique devices for the generation of wavelength-tunable coherent optical waves in any spectral ranges where efficient laser materials do not exist, provided that an efficient nonlinear crystal and a commonly available laser source at a higher frequency to serve as the pump can be found.

A doubly resonant OPO generally has a lower oscillation threshold than a singly resonant one of comparable physical parameters. However, a doubly resonant OPO is difficult to operate because of its intrinsic instability. To resonate both signal and idler waves, both frequencies ω_1 and ω_2 have to satisfy the resonance condition of the optical cavity. With the constraint of (9.138), this requirement cannot be met with an arbitrary cavity length but only with some specific values of the cavity length. This situation limits the tunability of the parametric oscillator. In addition, any variations in the cavity length due to mechanical or thermal fluctuations can lead to instability in the oscillation frequencies and the amplitudes of the optical fields. These problems do not exist in a singly resonant optical parametric resonator. Therefore, most OPOs designed for practical applications are of the singly resonant type.

EXAMPLE 9.14 The PPLN parametric amplifier described in Example 9.13 is placed in a properly designed optical cavity to make a singly resonant OPO. When the OPO is sufficiently pumped above threshold with a pump beam at 527 nm of a pump power of $P_3 = 2$ W, it is found that 5% of the pump power is converted to the combined output power of the signal and idler. What are the output powers of the signal and idler beams, respectively?

Solution The total output power from this OPO is $P_{\text{out}} = 0.05P_3 = 100$ mW. In a parametric conversion process, an idler photon is simultaneously generated each time a signal photon is generated while a pump photon is annihilated because of the relation $\omega_3 = \omega_1 + \omega_2$ required by (9.138). As a consequence, the total number of signal photons has to be equal to that of idler photons because there are no input signal or idler photons to an OPO. If the signal and idler photons are subject to the same fractional loss, the power ratio between the signal and the idler is

$$\frac{P_1^{\text{out}}}{P_2^{\text{out}}} = \frac{\omega_1}{\omega_2} = \frac{\lambda_2}{\lambda_1}, \tag{9.140}$$

which leads to the following power split:

$$P_1^{\text{out}} = \frac{\lambda_2}{\lambda_1 + \lambda_2} P_{\text{out}}, \qquad P_2^{\text{out}} = \frac{\lambda_1}{\lambda_1 + \lambda_2} P_{\text{out}}. \tag{9.141}$$

With $P_{\text{out}} = 100$ mW, we find that $P_1^{\text{out}} = 34$ mW for the signal at $\lambda_1 = 1.55$ μm and $P_2^{\text{out}} = 66$ mW for the idler at $\lambda_2 = 798$ nm.

For the split of output power expressed in (9.141), it is assumed that the signal and the idler suffer the same fractional loss in the OPO. In practice, this assumption may not be true, particularly when the wavelengths of the signal and the idler are far apart from each other. When the signal and the idler experience significantly disparate losses, the output power split can be very different from that described by (9.141). Even in this situation, it is still true that equal numbers of signal and idler photons are generated from converting the same number of pump photons when they interact in the nonlinear crystal.

Many lasers are available in the visible and near infrared wavelength regions for pumping second-harmonic generators to generate short-wavelength optical waves well into the deep ultraviolet region and for pumping OPOs to generate wavelength-tunable optical waves in a broad infrared region. With the advances in laser sources and crystal technology, the wavelengths that can be reached by nonlinear frequency conversion are basically only limited by the transmission windows of available nonlinear crystals. Figure 9.21 shows the transmission windows of various nonlinear optical crystals that can be chosen for frequency converters and the wavelengths of several lasers that can be used as pump sources. From this figure, we see that second-order nonlinear frequency conversion can cover a spectral range from about 200 nm in the deep ultraviolet to

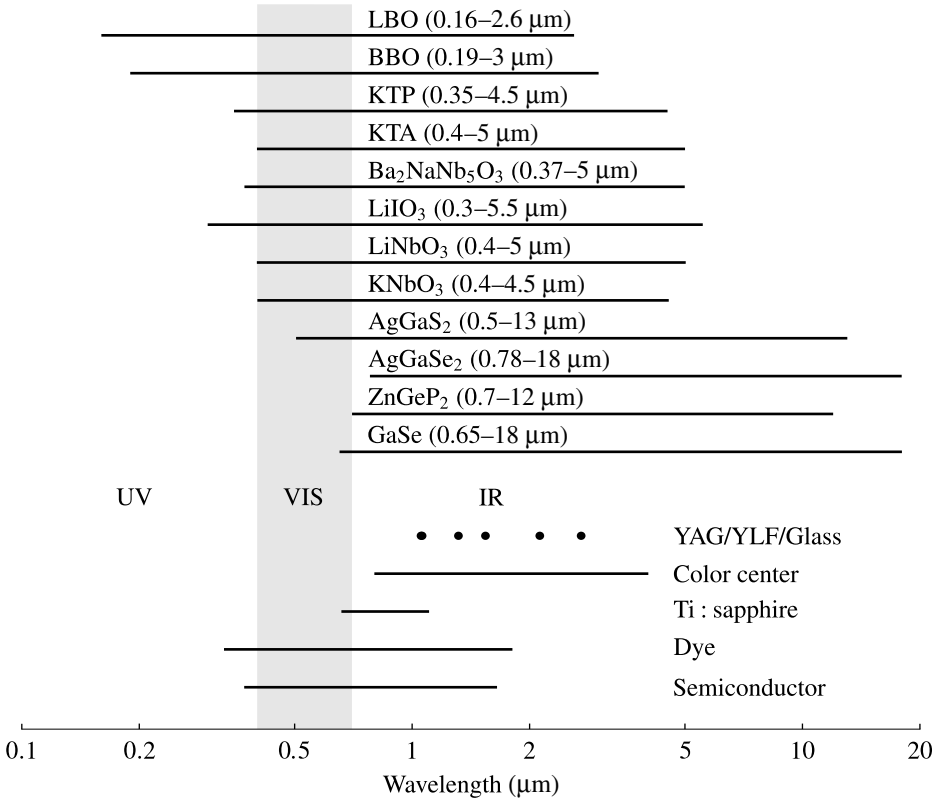


Figure 9.21 Transmission windows of various nonlinear optical crystals for frequency converters and wavelengths of several lasers that can be used as pump sources. (Based on data from assorted sources.)

about 18 μm in the mid infrared. Depending on the pump lasers used, the coherent optical waves generated by these frequency converters in this wide spectral range cover the entire range of temporal characteristics, from CW beams through nanosecond Q-switched pulses to picosecond and femtosecond mode-locked pulses. Through these nonlinear optical devices, optical sources over a wide range of spectral and temporal characteristics are made available and flexible for many applications.

9.7 Nonlinear optical modulators and switches

In a nonlinear optical modulator, the modulation of an optical wave is accomplished through a nonlinear optical process. A nonlinear optical modulator can be based on either *self modulation* or *cross modulation*. In the case of self modulation, only one optical beam is present, and the modulation on the beam is a function of the characteristics of the beam itself. In the case of cross modulation, two or more optical beams are

present, and the beam of interest is modulated by one or more other beams that carry the modulation signals. In either case, no electric, magnetic, or acoustic field is needed. Therefore, nonlinear optical modulators and switches are also known as *all-optical modulators* and *all-optical switches*, respectively.

There are two fundamentally different types of nonlinear optical modulators and switches. One is the *dispersive*, or *refractive*, type, which is based on the optical Kerr effect due to optical-field-induced changes in the real part of the permittivity of a material. Another is the *absorptive* type, which relies on an intensity-dependent absorption coefficient caused by the nonlinear characteristics of the imaginary part of the permittivity of a material.

Kerr lenses

We first consider the simplest case of the optical Kerr effect discussed in Section 9.3 in which $\mathbf{P}^{(3)}$ is parallel to a linearly polarized optical field $\mathbf{E}$ so that the net effect is an intensity-dependent index of refraction given in (9.49). For a plane optical wave, this optical Kerr effect merely causes a uniform intensity-dependent phase shift across the wavefront. Thus the beam remains a plane wave without any change in its spatial intensity distribution. If an optical beam has a nonuniform intensity distribution, the intensity-dependent index of refraction leads to a nonuniform phase shift across the wavefront as the beam propagates through the nonlinear medium. This beam will then be focused or defocused as a result of distortion in its phase front.

For simplicity, we consider the propagation of a circular beam, which has a transverse spatial intensity distribution $I(r)$. After such a beam propagates through a *thin* nonlinear medium of a thickness l, the total intensity-dependent phase shift can be approximated by

$$\varphi(r) = \frac{\omega}{c}[n_0 + n_2 I(r)]l. \tag{9.142}$$

The intensity-dependent Kerr phase change given by

$$\varphi_{\mathrm{K}}(r) = \frac{\omega}{c} n_2 l I(r) \tag{9.143}$$

is known as *self-phase modulation* because it is imposed by an optical beam on itself through the optical Kerr effect.

Recall that the effect of a thin spherical lens of a focal length f is to cause a spatially varying phase shift of

$$\varphi(r) = -k\frac{r^2}{2f} = -\frac{\omega}{c}\frac{r^2}{2f} \tag{9.144}$$

in an optical wave passing through the lens, where r is the transverse radial distance from the center of the lens. Therefore, if the intensity-dependent phase shift given in (9.142) has a quadratic dependence on the transverse radial coordinate, the optical Kerr

effect in the thin nonlinear medium would be equivalent to the effect of a thin lens. A thin nonlinear medium with such a function is called a *Kerr lens*. In reality, no optical beam has an ideal quadratic spatial intensity distribution. However, if the intensity distribution of a circular beam is approximately quadratic in r near the beam center, the effective focal length of the Kerr lens can be given by

$$\frac{1}{f_{\mathrm{K}}} = -a\frac{c}{\omega}\frac{\mathrm{d}^2\varphi}{\mathrm{d}r^2}\bigg|_{r=0} = -an_2 l\frac{\mathrm{d}^2 I(r)}{\mathrm{d}r^2}\bigg|_{r=0}, \tag{9.145}$$

where a is a correction factor to account for the difference between the true beam profile and the ideal quadratic profile. Using this relation, we find that the effective focal length of the Kerr lens for a circular Gaussian beam with an intensity distribution of $I(r) = I_0 \exp(-2r^2/w^2)$ is

$$f_{\mathrm{K}} = \frac{w^2}{4an_2 l I_0} = \frac{\pi w^4}{8an_2 l P}, \tag{9.146}$$

where w is the beam radius at the location of the Kerr medium, I_0 is the intensity at the beam center, and P is the power of the beam. For a circular Gaussian beam, $a = 1.723$, and the thin-lens condition for (9.146) to be valid is $l < z_{\mathrm{R}} = \pi n w_0^2/\lambda$. Note that n_2 can be either positive or negative because $\chi^{(3)'}$ can be either positive or negative. Therefore, a Kerr lens can either focus or defocus a beam, depending on the sign of its effective focal length.

Most applications of Kerr lenses are based on the fact that the effective focal length f_{K} of a thin Kerr lens is inversely proportional to the peak intensity I_0 of an optical beam. As a result of this characteristic, the divergence of the beam after passing through a Kerr lens is a function of the intensity of the beam. In addition, the beam divergence also depends on the sign of n_2 and the location of the Kerr lens with respect to the beam waist, as illustrated in Fig. 9.22.

A Kerr lens is often used as an *optical power limiter* for the protection of a sensitive optical detector. In this application, the action of the Kerr lens is to increase the beam divergence as the input intensity of a beam is increased, thereby increasing the spread and reducing the intensity of the beam at the surface of the detector. As demonstrated in Figs. 9.22(*a*), (*b*), (*e*), and (*f*), with proper arrangement, either a Kerr lens of a positive effective focal length, $f_{\mathrm{K}} > 0$, or one with a negative effective focal length, $f_{\mathrm{K}} < 0$, can be used for this purpose. When a Kerr lens in such an arrangement is used as an optical power limiter, only a fraction of the diverging optical beam within a finite central cross-sectional area that is defined either by the area of a small detector or by a hole in a beam block is allowed to reach the detector. Because the divergence of the beam increases with its intensity, the optical power passing through the finite area to be received by the detector will saturate at a certain level as the input power of the beam continues to increase. Without the Kerr lens, the beam divergence does not change with its intensity. Then, the optical power received by the detector increases linearly with

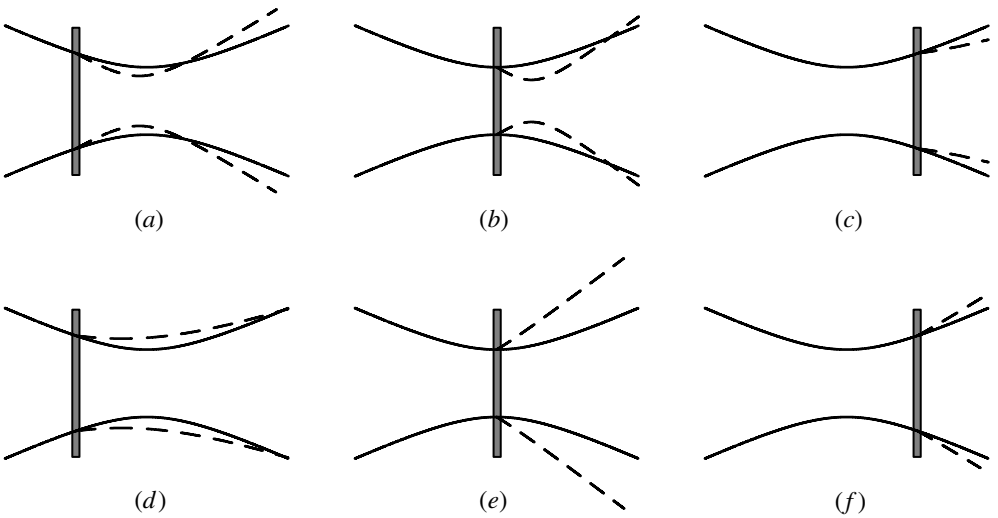


Figure 9.22 Nonlinear refraction caused by a Kerr lens as a function of beam intensity and the location of the Kerr lens with respect to the beam waist for (*a*), (*b*), and (*c*) $f_K > 0$, and for (*d*), (*e*), and (*f*) $f_K < 0$. The solid curves are the propagation lines of a Gaussian beam at a low intensity for which the effect of the Kerr lens is negligible, and the dashed curves are those at a high intensity for which the effect of the Kerr lens is significant.

the input power of the beam without a limit until the detector is damaged even if the detector has a very small area to intercept only a tiny fraction of the beam.

A Kerr lens can also be used as a passive optical switch or an *optical thresholding device*. For this purpose, an arrangement, such as that shown in Fig. 9.22(*c*) or (*d*), that leads to a reduction in beam divergence with an increase in input beam intensity is used. Similarly to the setup of a power limiter, only a portion of the beam within a finite central area of the beam cross section is allowed to pass. However, instead of a saturation, the optical power passing through this area increases nonlinearly with the input power of the beam. This behavior can be used to provide a nonlinear feedback to an optical system or to switch on an optical device at a certain threshold. It has been used as the *passive mode locker* in a technique known as *Kerr-lens mode locking* for the generation of ultrashort laser pulses.

EXAMPLE 9.15 A Ti : sapphire laser generates a train of laser pulses of wavelength $\lambda = 780$ nm and pulsewidth $\Delta t_{ps} = 100$ fs at a repetition rate of $f_{ps} = 100$ MHz. A beam of such pulses at an average power of $\overline{P} = 50$ mW is focused tightly on a thin silica plate of thickness $l = 1$ mm. The nonlinear response time of silica is much faster than 100 fs so that the optical Kerr effect can be considered instantaneous in response to the temporal variation of each pulse. Silica has a linear refractive index of $n_0 = 1.4537$ at

$\lambda = 780$ nm and, according to Example 9.5, a nonlinear refractive index of $n_2 = 2.4 \times 10^{-20}$ m^2 W^{-1}. If the laser beam is focused with its waist on the silica plate as tightly as allowed by the thin-lens condition, what is the effective focal length of the Kerr lens caused by self-phase modulation at the peaks of the optical pulses?

Solution The peak power of the pulses is

$$P_{\text{pk}} = \frac{\overline{P}}{f_{\text{ps}} \Delta t_{\text{ps}}} = \frac{50 \times 10^{-3}}{100 \times 10^{6} \times 100 \times 10^{-15}} \text{ W} = 5 \text{ kW}.$$

The thin-lens condition, $l < z_{\text{R}} = \pi n w_0^2 / \lambda$, requires that

$$w_0 > \left(\frac{l\lambda}{\pi n} \right)^{1/2} = \left(\frac{1 \times 10^{-3} \times 780 \times 10^{-9}}{\pi \times 1.4537} \right)^{1/2} \text{ m} = 13 \text{ μm}.$$

By focusing the beam to the limit of $w_0 = 13$ μm allowed by the thin-lens condition and by placing the beam waist on the silica plate, we have the following Kerr focal length at the peak of each pulse:

$$f_{\text{K}} = \frac{\pi w_0^4}{8 a n_2 l P_{\text{pk}}} = \frac{\pi \times (13 \times 10^{-6})^4}{8 \times 1.723 \times 2.4 \times 10^{-20} \times 1 \times 10^{-3} \times 5 \times 10^{3}} \text{ m} = 5.42 \text{ cm}.$$

Note that this is the Kerr focal length only at the temporal peak of each pulse. Because f_{K} is inversely proportional to the optical power and because the nonlinear refractive response of silica is much faster than the 100 fs duration of each pulse, we can easily see that the value of f_{K} varies in time through the duration of a pulse. As a consequence of this temporally varying f_{K}, the divergence of the pulse after the silica plate is a function of time over the pulse duration. Kerr-lens mode locking of lasers takes advantage of this interesting phenomenon.

Polarization and amplitude modulators

The optical-field-induced birefringence of the optical Kerr effect can be used for polarization modulation of an optical wave. Such polarization modulation can be either self induced in a one-beam interaction or cross induced in a two-beam interaction. For simplicity, we consider the interactions in an isotropic medium. The same principle applies to nonlinear optical polarization modulators using anisotropic crystals.

We have already seen from the discussions on Kerr lenses that in a one-beam interaction in an isotropic medium, the induced $\mathbf{P}^{(3)}$ and the optical field $\mathbf{E}$ have the same polarization state if the optical field is linearly polarized. This is also true for a circularly polarized optical field. Therefore, the optical Kerr effect does not change the polarization state of a linearly or circularly polarized optical wave that propagates alone in an isotropic medium. The situation is different for an elliptically polarized optical wave in a one-beam interaction, as well as for a linearly or circularly polarized

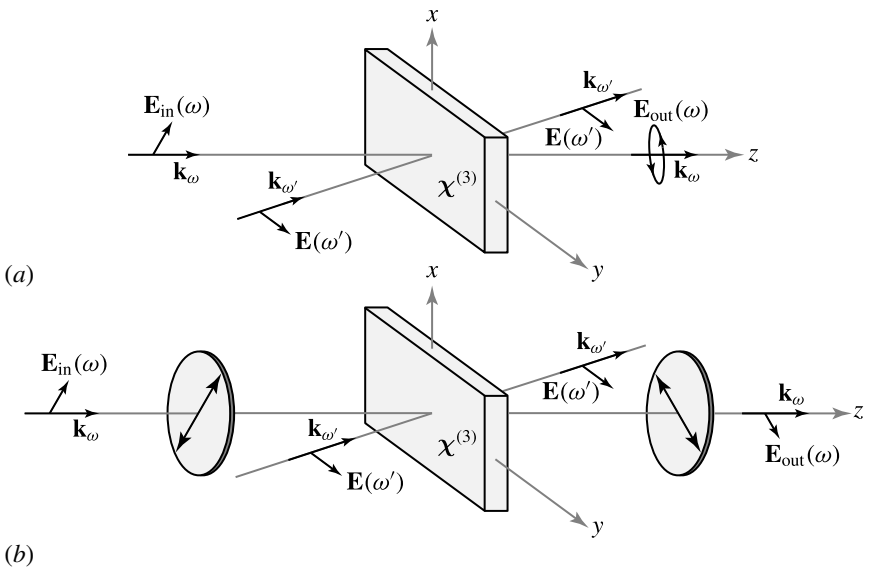


Figure 9.23 (a) Nonlinear optical polarization modulator and (b) nonlinear optical amplitude modulator.

optical wave in a two-beam interaction. In a one-beam interaction with an elliptically polarized optical wave, the polarization state of the induced $\mathbf{P}^{(3)}$ is different from that of the optical field $\mathbf{E}$, causing the polarization of the optical field to change. The result is a phenomenon known as *ellipse rotation* because the axes of the ellipse defined by the tip of the elliptically polarized optical field continue to rotate in space as the wave propagates through the nonlinear medium (see Problem 9.7.1).

In the interaction of two linearly polarized optical waves, polarization modulation on one wave by the other through the optical Kerr effect is possible if the polarizations of the two waves are neither parallel nor orthogonal to each other. The optical beam being modulated is called the signal or the probe, and that creating the modulation is called the pump. In an isotropic medium, the coordinate axes can be chosen arbitrarily. With a signal beam at a frequency ω and a pump beam at a frequency ω', we choose the xy plane to be that defined by the two linearly polarized field vectors $\mathbf{E}(\omega)$ and $\mathbf{E}(\omega')$ and the y axis to be in the direction of $\mathbf{E}(\omega')$, as shown in Fig. 9.23(a). While the signal beam propagates in the z direction, the pump beam propagates in a direction within the zx plane that may or may not be collinear with the propagation direction of the signal beam, as also shown in Fig. 9.23(a).

The optical-field-induced birefringence seen by the signal beam is described by $\Delta\epsilon_{ij}(\omega, \mathbf{E})$ given in (9.47). In a practical application, the intensity of the signal beam is much lower than that of the pump beam: $I(\omega) \ll I(\omega')$. Therefore, the first term on the right-hand side of (9.47), which accounts for the self modulation of the signal beam, can be neglected in comparison to the second term, which accounts for the cross

modulation on the signal by the pump. With $\mathbf{E}(\omega') \parallel \hat{y}$, we then have

$$\Delta\epsilon_{xx} \approx 6\epsilon_0\chi^{(3)}_{1122}|E(\omega')|^2 = \frac{3\chi^{(3)}_{1122}}{cn_0}I(\omega'), \tag{9.147}$$

$$\Delta\epsilon_{yy} \approx 6\epsilon_0\chi^{(3)}_{1111}|E(\omega')|^2 = \frac{3\chi^{(3)}_{1111}}{cn_0}I(\omega'). \tag{9.148}$$

This optical-field-induced birefringence leads to the following intensity-dependent indices of refraction:

$$n_x = n_0 + \frac{3\chi^{(3)}_{1122}}{2c\epsilon_0 n_0^2}I(\omega'), \tag{9.149}$$

$$n_y = n_0 + \frac{3\chi^{(3)}_{1111}}{2c\epsilon_0 n_0^2}I(\omega'). \tag{9.150}$$

If the signal beam has a field of $\mathbf{E}(\omega) = (\hat{x}\mathcal{E}_x + \hat{y}\mathcal{E}_y)\mathrm{e}^{-\mathrm{i}\omega t}$ at the input surface of the nonlinear medium that has a thickness of l, its field at the output is

$$\mathbf{E}(\omega) = (\hat{x}\mathcal{E}_x + \hat{y}\mathcal{E}_y\mathrm{e}^{\mathrm{i}\Delta\varphi})\mathrm{e}^{\mathrm{i}k^x l - \mathrm{i}\omega t}, \tag{9.151}$$

where $k^x = n_x\omega/c$ and

$$\Delta\varphi = \frac{3\pi(\chi^{(3)}_{1111} - \chi^{(3)}_{1122})l}{c\epsilon_0 n_0^2\lambda}I(\omega') \tag{9.152}$$

is the phase retardation between the x and y components of the signal field. Because this phase retardation is linearly proportional to the pump intensity, the polarization state of the signal beam at the output can be modulated by varying the pump intensity if $\mathbf{E}(\omega)$ is neither parallel nor perpendicular to $\mathbf{E}(\omega')$ so that both $\mathcal{E}_x$ and $\mathcal{E}_y$ have nonvanishing values.

In comparison to the electro-optic polarization modulators discussed in Section 6.3, the only difference is that the nonlinear optical polarization modulators discussed here are controlled by a pump optical beam rather than by a voltage. Other than this difference, these two types of polarization modulators have the same function and serve the same purpose.

As seen in Section 6.3, an amplitude modulator can be easily constructed by placing a polarization modulator between two polarizers. This approach is also applicable to the construction of a nonlinear optical amplitude modulator using a nonlinear optical polarization modulator, as illustrated in Fig. 9.23(*b*). A nonlinear optical amplitude modulator and an electro-optic amplitude modulator have the same transmission characteristics, which are discussed in Section 6.3, if they are set up in the same manner. When an ultrashort optical pulse is used as the pump beam, a nonlinear optical amplitude modulator can function as a fast *optical gate*, or a fast all-optical switch, for switching the signal beam within a very short time.

Saturable absorbers

A saturable absorber has an absorption coefficient that decreases with increasing light intensity, such as that characterized by (9.51) with $\chi^{(1)''} > 0$ and $\chi^{(3)''} < 0$. Note, however, that the relation in (9.51) is rooted in the power series expansion of (9.1). Because absorption saturation necessarily occurs at a resonant transition between two energy levels, the perturbation approach taken for power series expansion is not valid at sufficiently high intensities. Instead, a full analysis of the resonant absorption has to be carried out. Such an analysis results in an intensity-dependent absorption coefficient characterized by the relation[5]

$$\alpha = \frac{\alpha_0}{1 + I/I_{\text{sat}}}, \tag{9.153}$$

where α_0 is the unsaturated absorption coefficient and I_{sat} is known as the *saturation intensity*. The saturation intensity is a characteristic of the resonant transition that is responsible for the absorption under consideration. For $I < I_{\text{sat}}$, the relation in (9.153) can be expanded:

$$\alpha = \alpha_0 \left[1 - \frac{I}{I_{\text{sat}}} + \left(\frac{I}{I_{\text{sat}}} \right)^2 - \left(\frac{I}{I_{\text{sat}}} \right)^3 + \cdots \right]. \tag{9.154}$$

Only when $I \ll I_{\text{sat}}$ can α be accurately approximated by the first two terms of this expansion, resulting in a linear dependence on I like the relation in (9.51). In general, the relation in (9.153) has to be used because the light intensity encountered in a practical device that uses a saturable absorber can easily be comparable to or higher than I_{sat}.

The propagation of an optical wave through a saturable absorber that has an absorption coefficient given in (9.153) is described by

$$\frac{\mathrm{d}I}{\mathrm{d}z} = -\frac{\alpha_0}{1 + I/I_{\text{sat}}} I. \tag{9.155}$$

This equation can be integrated to obtain the following relation:

$$I(z)\mathrm{e}^{I(z)/I_{\text{sat}}} = I(0)\mathrm{e}^{I(0)/I_{\text{sat}}}\mathrm{e}^{-\alpha_0 z}, \tag{9.156}$$

where $I(0)$ is the input light intensity at $z = 0$. The transmittance of an optical wave through a saturable absorber of a thickness l is $T = I_{\text{out}}/I_{\text{in}} = I(l)/I(0)$, which can be calculated by numerically solving (9.156). It is plotted in Fig. 9.24 as a function of the input light intensity, normalized to the saturation intensity, for a few different values of $\alpha_0 l$ represented in terms of $T_0 = \mathrm{e}^{-\alpha_0 l}$. As Fig. 9.24 shows, the optical transmittance through a saturable absorber increases nonlinearly as the input intensity is increased and approaches unity at high input intensities. In a particular application of a saturable absorber, the value of $\alpha_0 l$ has to be properly chosen for a desired difference between

[5] The absorption coefficient described by (9.153) is that for a homogeneously broadened medium. For an inhomogeneously broadened medium, the relation is $\alpha = \alpha_0/(1 + I/I_{\text{sat}})^{1/2}$.

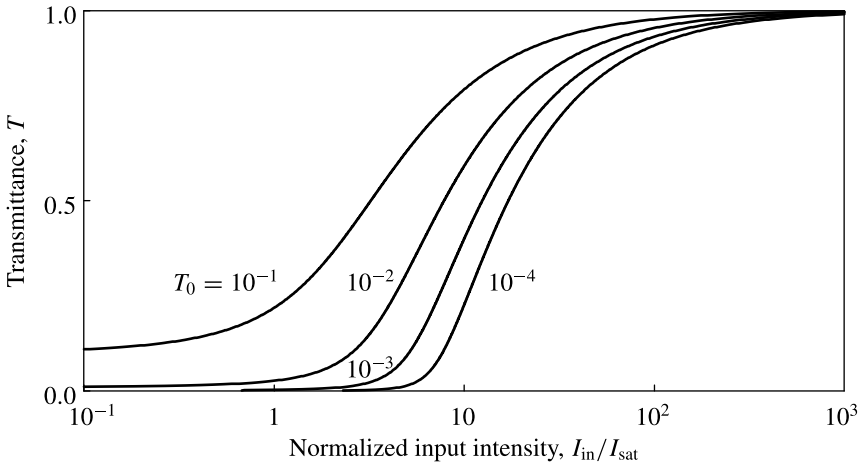


Figure 9.24 Transmittance of an optical wave through a saturable absorber that has a thickness l and an unsaturated absorption coefficient α_0 as a function of the input light intensity normalized to the saturation intensity. The curves are plotted for different values of $\alpha_0 l$ in terms of $T_0 = e^{-\alpha_0 l}$.

the maximum transmittance at high intensities and the minimum transmittance at low intensities.

A saturable absorber can be used as a *spatial light filter*, which blocks low-intensity stray light or background optical noise but transmits a high-intensity signal beam. It can be used as an *optical discriminator*, which transmits optical pulses of intensities above a certain threshold and suppresses those below. A saturable absorber is also commonly used as a *passive Q switch* in a *Q*-switched laser or as a *passive mode locker* in a mode-locked laser for the generation of very short laser pulses. The saturable absorber in this kind of application functions as a passive optical switch in the time domain. It is switched open by the rising intensity of a laser pulse and closes through its own relaxation after the passing of the pulse. Therefore, the relaxation time of a saturable absorber is also an important factor to be considered in its application as a *Q* switch or a mode locker.

9.8 Bistable optical devices

A bistable device has two stable output states under one input condition. Because of this *binary* feature, bistable devices can be used for many digital operations, such as switches, memories, registers, and flip-flops. Bistable electronic circuits and devices have become indispensable components in a wide range of applications that require the storage of binary information. Bistable optical devices can be important for their applications as optical logic, memories, and analog-to-digital converters in optical signal processing systems. In addition, they can also be used as optical pulse discriminators and optical power limiters.

The output parameter of a bistable device is a multivalued function of its input parameter. Any system with such a multivalued characteristic is by definition a nonlinear system. Therefore, optical nonlinearity is absolutely required for a bistable optical device. Optical nonlinearity alone is not sufficient for bistability, however. As seen from earlier sections, the propagation characteristics of an optical beam through a nonlinear medium vary nonlinearly but also *monotonically* with the beam intensity. Bistability is not possible with monotonic nonlinearity alone because a monotonic characteristic does not lead to a multivalued dependence of the output on an input parameter. The required nonmonotonic characteristics for optical bistability can be made possible only with proper feedback.

The necessary conditions for optical bistability are *optical nonlinearity* and *positive feedback*. Depending on whether the optical nonlinearity responsible for the bistable function comes from the real or the imaginary part of a nonlinear susceptibility, a bistable optical device can be classified as either *dispersive* or *absorptive*. In some devices, this distinction is not clear, however, because both refractive and absorptive nonlinear mechanisms may be present. Depending on the type of feedback, a bistable optical device can also be classified as either *intrinsic* or *hybrid*. In an intrinsic bistable device, both the interaction and the feedback are *all optical*. In a hybrid bistable device, electrical feedback is used to modify the optical interaction, thereby creating an artificial optical nonlinearity.

Figure 9.25(*a*) shows a generic characteristic for *intensity bistability* of a bistable optical device. For each input intensity within the range between $I_{\text{in}}^{\text{down}}$ and $I_{\text{in}}^{\text{up}}$, there are three values for the output intensity. Only the two values that lie on the upper and the lower branches of the curve are stable output values. The one that lies on the middle

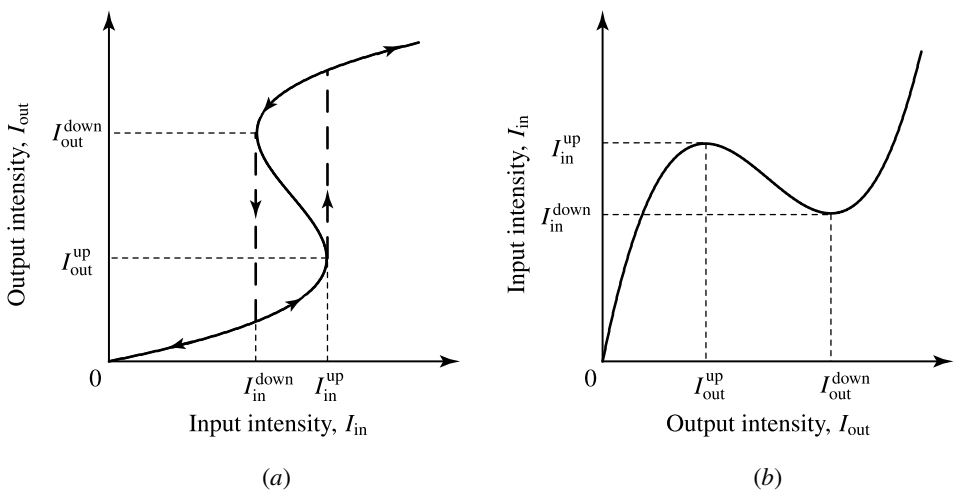


Figure 9.25 Generic characteristic for intensity bistability (*a*) plotted with output intensity I_{out} as a function of input intensity I_{in} and (*b*) plotted with I_{in} as a function of I_{out}.

branch is unstable because the middle branch has a negative slope: $dI_{out}/dI_{in} < 0$. When the input intensity is gradually increased from zero, the output intensity traces the lower branch of the curve until the input intensity reaches the *up-transition point* at I_{in}^{up}, where the output makes a sudden jump to the upper branch. Once the system is in a state that lies on the upper branch, it can be brought back to the lower branch only when the input intensity is lowered to the *down-transition point* at I_{in}^{down}. If the input intensity is set at a value within the bistable region, the output can be in either stable state depending on the history of the system. With a proper external excitation, it can be switched from one of the stable states to the other. Otherwise, it stays in one state indefinitely.

We see from Fig. 9.25(*a*) that the slope of the characteristic curve for bistability changes sign at both up- and down-transition points. This fact can be exploited to find the condition for bistability and the transition points. Though I_{out} is a multivalued function of I_{in}, I_{in} is a single-valued function of I_{out}. Therefore, it is convenient to express I_{in} as a function of I_{out}, as shown in Fig. 9.25(*b*). From the curve shown in Fig. 9.25(*b*), we find that the condition for the existence of a bistable region is the existence of a region of negative slope, $dI_{in}/dI_{out} < 0$, between regions of positive slope. Because both I_{in} and I_{out} are real and positive quantities, this condition can be satisfied only when the relation

$$\frac{dI_{in}}{dI_{out}} = 0 \tag{9.157}$$

has *two nondegenerate real and positive solutions*. These two solutions correspond to the two transition points (I_{in}^{up}, I_{out}^{up}) and (I_{in}^{down}, I_{out}^{down}), as can be seen by an examination of Figs. 9.25(*a*) and (*b*).

In principle, it is possible to construct bistable optical devices using a variety of different nonlinear effects discussed in Section 9.3 if the device parameters are properly chosen. In practice, however, the nonlinear optical media that are most commonly used for bistable devices are either nonabsorptive Kerr media, for the dispersive type, or saturable absorbers, for the absorptive type. A simple bistable optical device of the intrinsic type can be constructed by placing a nonlinear optical medium inside a *Fabry–Perot* cavity, as shown in Fig. 9.26(*a*), or inside a ring cavity, as shown in Fig. 9.26(*b*). The mirrors of the cavity provide the needed optical feedback to the nonlinear optical interaction. The only difference between the two configurations in Fig. 9.26 is that the optical wave in a Fabry–Perot cavity travels through the nonlinear medium twice in each round trip and forms a standing wave pattern, but the wave in a ring cavity is a traveling wave that travels through the nonlinear medium only once in each round trip. Otherwise, the basic principle and the characteristics of optical bistability are the same for the two configurations. Other optical feedback configurations for intrinsic bistable optical devices are based on this concept as well.

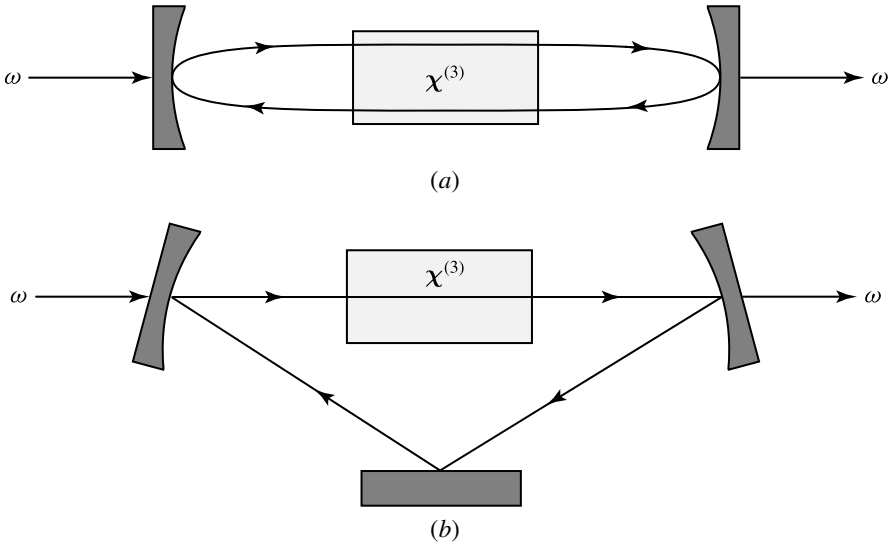


Figure 9.26 Intrinsic bistable optical devices using optical feedback in the configurations of (*a*) a Fabry–Perot cavity and (*b*) a ring cavity.

In the following, we consider bistable devices that have the configuration of the Fabry–Perot cavity shown in Fig. 9.26(*a*). For simplicity, we assume that the mirrors of the Fabry–Perot cavity are identical and lossless. The mirrors have a real reflection coefficient r, which can be either positive or negative, but their transmission coefficient t can be complex because of the finite thickness of the mirrors. The intensity reflectance is $R = r^2$, and the intensity transmittance is $T = |t|^2 = 1 - R$, for $R + T = 1$. We also assume that the nonlinear medium fills up the entire space inside the Fabry–Perot cavity of length l. We ignore transverse spatial variations by considering only plane optical waves. When the steady state is reached, the forward-traveling field, $\mathcal{E}_\mathrm{f}$, and the backward-traveling field, $\mathcal{E}_\mathrm{b}$, inside the cavity satisfy the following relations at the input end, $z = 0$:

$$\mathcal{E}_\mathrm{f}(0) = t\mathcal{E}_\mathrm{in} + r\mathcal{E}_\mathrm{b}(0), \tag{9.158}$$

$$\mathcal{E}_\mathrm{b}(0) = r\mathcal{E}_\mathrm{f}(0)\mathrm{e}^{\mathrm{i}2kl-\alpha l}, \tag{9.159}$$

where k and α are the propagation constant and the absorption coefficient, respectively, in the medium. At the output end, $z = l$, we have

$$\mathcal{E}_\mathrm{out} = t\mathcal{E}_\mathrm{f}(l) = t\mathcal{E}_\mathrm{f}(0)\mathrm{e}^{\mathrm{i}kl-\alpha l/2}. \tag{9.160}$$

Using these relations, we find that

$$\mathcal{E}_\mathrm{out} = \frac{t^2\mathrm{e}^{\mathrm{i}kl-\alpha l/2}}{1 - r^2\mathrm{e}^{\mathrm{i}2kl-\alpha l}}\mathcal{E}_\mathrm{in}, \tag{9.161}$$

which gives the following relation between the input and output intensities:

$$I_{\text{out}} = \frac{(1-R)^2 \mathrm{e}^{-\alpha l}}{(1-R\mathrm{e}^{-\alpha l})^2 + 4R\mathrm{e}^{-\alpha l}\sin^2 kl} I_{\text{in}}. \tag{9.162}$$

Dispersive bistable optical devices

We first consider dispersive bistability in a Fabry–Perot cavity filled with a nonlinear medium that has an intensity-dependent index of refraction due to the optical Kerr effect. For simplicity, we ignore the standing wave pattern in the cavity and take the average intracavity intensity $I_{\text{c}} \approx I_{\text{f}} + I_b \approx 2I_{\text{out}}/(1-R)$. The intensity-dependent index of refraction is

$$n = n_0 + n_2 I_{\text{c}} \approx n_0 + \frac{2n_2 I_{\text{out}}}{1-R}. \tag{9.163}$$

Then, the total phase shift over a round trip in the cavity can be expressed as

$$2kl = \frac{2n_0\omega l}{c} + \frac{4n_2\omega l}{c(1-R)} I_{\text{out}} = 2m\pi + \varphi, \tag{9.164}$$

where m is a properly chosen integer such that

$$\varphi = \varphi_0 + \varphi_2 I_{\text{out}} \tag{9.165}$$

for $|\varphi_0| < \pi$ and

$$\varphi_2 = \frac{4n_2\omega l}{c(1-R)} = \frac{8\pi n_2 l}{\lambda(1-R)}. \tag{9.166}$$

Note that φ_0 is a bias phase that can be chosen at will by slightly varying the cavity length l for a given optical frequency ω or by varying the optical frequency for a fixed cavity length.

For the device under consideration, we can rearrange (9.162) as

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \frac{F^2/F_0^2}{1 + 4(F^2/\pi^2)\sin^2(\varphi/2)}, \tag{9.167}$$

where

$$F = \frac{\pi\sqrt{R\mathrm{e}^{-\alpha l}}}{1-R\mathrm{e}^{-\alpha l}} \tag{9.168}$$

is the *finesse* of a generic lossy Fabry–Perot cavity, and

$$F_0 = \frac{\pi\sqrt{R}}{1-R} \tag{9.169}$$

is the finesse of a lossless Fabry–Perot cavity. The characteristic described by (9.167) has resonance peaks at $\varphi = 0, \pm 2\pi, \pm 4\pi, \ldots$, each of which has the same characteristics as those of the peak shown in Fig. 9.27.

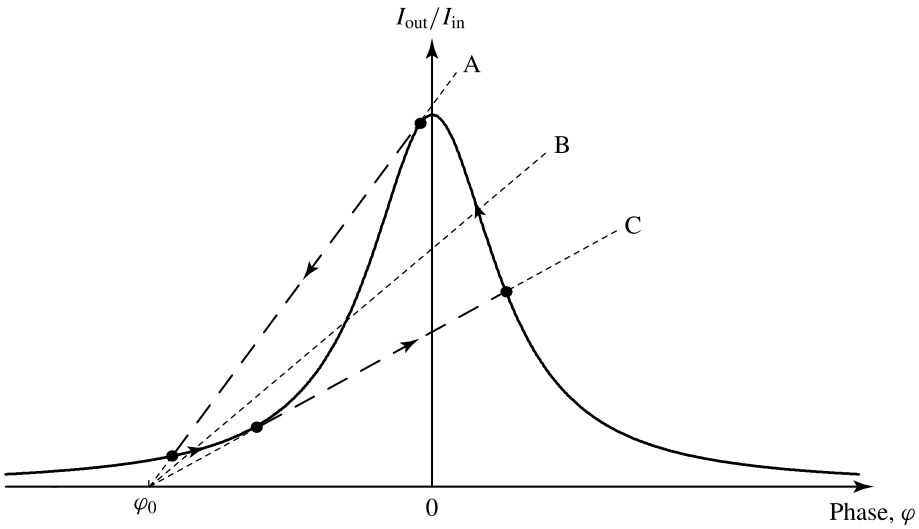


Figure 9.27 Graphic illustration of the bistable characteristic of a dispersive bistable device with a Fabry–Perot cavity. The peaked curve is the transmission characteristic of the Fabry–Perot cavity at a resonance condition. Other resonance peaks are far away and are not shown. Line A represents the input intensity at I_{in}^{down}. Line B represents an input intensity in the bistable region. Line C represents the input intensity at I_{in}^{up}.

For φ being intensity dependent as given in (9.165), the system can exhibit intensity bistability under proper conditions. A graphic solution can be obtained by expressing (9.165) in the form of

$$\frac{I_{out}}{I_{in}} = \frac{\varphi - \varphi_0}{\varphi_2 I_{in}} \tag{9.170}$$

and plotting it as straight lines for various values of I_{in} to find the intersections between these lines and the curve representing (9.167). An example of the graphic solution for $\varphi_0 < 0$ and $n_2 > 0$ is shown in Fig. 9.27. It can be seen from this illustration that up-transition corresponds to line C, which is tangent to the curve at its heel, whereas down-transition is described by line A, which is tangent to the curve near its peak.

Analytical solution is possible if $|\varphi| < 1$ so that $\sin^2(\varphi/2) \approx \varphi^2/4$. Then, by combining (9.165) and (9.167), we have

$$\frac{F^2}{F_0^2} I_{in} = \left[1 + \frac{F^2}{\pi^2}(\varphi_0 + \varphi_2 I_{out})^2\right] I_{out}. \tag{9.171}$$

By demanding that $dI_{in}/dI_{out} = 0$ has two nondegenerate, real and positive solutions for I_{out}, we find that the conditions for bistability under the assumption that $|\varphi| < 1$ are (see Problem 9.8.1)

$$\varphi_0 n_2 < 0 \tag{9.172}$$

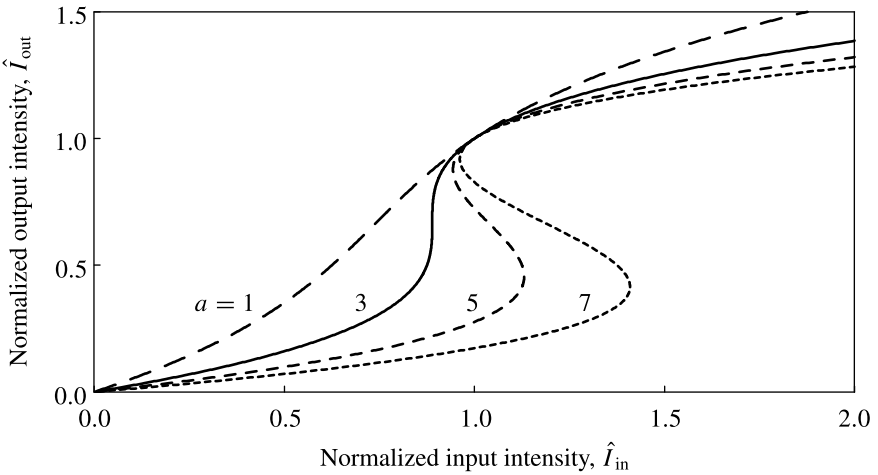


Figure 9.28 Characteristics of a dispersive nonlinear device with an optical Kerr medium in a Fabry–Perot cavity, where $\hat{I}_{in} = -F^2\varphi_2 I_{in}/(F_0^2\varphi_0)$ and $\hat{I}_{out} = -\varphi_2 I_{out}/\varphi_0$. Bistability exists only when $\varphi_2/\varphi_0 < 0$ and $a = F^2\varphi_0^2/\pi^2 > 3$. The solid curve represents the threshold condition for bistability at $a = 3$.

and

$$|\varphi_0| > \frac{\sqrt{3}\pi}{F}. \tag{9.173}$$

Once the conditions for bistability are satisfied, the up- and down-transition points, as well as the region of bistability, are found using the two nondegenerate solutions for I_{out}. Figure 9.28 shows the characteristics of this dispersive device for a few different values of the characteristic parameter $a = F^2\varphi_0^2/\pi^2$ under the condition that (9.172) is satisfied. It can be seen that bistability exists only when $a > 3$ so that (9.173) is satisfied.

There is a threshold input intensity, I_{th}, required for bistability to be possible in a device. This threshold input intensity is (see Problem 9.8.1)

$$I_{th} = \frac{\sqrt{3}\lambda(1-R)}{9F|n_2|l}\frac{F_0^2}{F^2}. \tag{9.174}$$

This *bistability threshold intensity* corresponds to the switching intensity for $a = 3$. If the input intensity is limited to below this threshold, the device can never reach its switching point even when both F and φ_0 are made large enough so that $a > 3$. When the input intensity exceeds this threshold, the device can enter its bistable regime for a properly chosen value of $a > 3$. However, the minimum input intensity required for a given device to operate properly in its bistable states increases as the value of φ_0 and, correspondingly, the value of a for a given finesse F increase (see Problem 9.8.2).

Ideally for a dispersive bistable device, the medium should be completely lossless. In practice, however, there are always some absorption or scattering losses in the medium.

Such losses in a dispersive device are usually very small so that $\alpha l \ll 1$. Though the losses are small, they can reduce the finesse of the cavity, thus significantly increasing the input intensity required for the operation of the device.

EXAMPLE 9.16 A vertical cavity InGaAsP bistable device consists of an active InGaAsP layer of $l = 1$ μm between highly reflective DBR mirrors of $R = 99\%$. It operates at $\lambda = 1.55$ μm, which is close to the bandgap wavelength $\lambda_g = 1.49$ μm of the InGaAsP layer. The nonlinear refractive index at this wavelength is found to be $n_2 = -9 \times 10^{-11}$ m^2 W^{-1}. (a) Find the bistability threshold intensity for this device assuming that the medium is lossless. If an optical beam of a circular Gaussian profile is focused to a spot size of $w_0 = 20$ μm on the device, what is the bistability threshold power? (b) The medium is found to have an absorption coefficient of $\alpha = 1.5 \times 10^4$ m^{-1}. Accounting for this loss, what are the realistic threshold intensity and threshold power for the device?

Solution (a) For $R = 99\%$, we find that the finesse of the cavity without loss is

$$F_0 = \frac{\pi\sqrt{0.99}}{1 - 0.99} = 312.6.$$

Using (9.174) with $F = F_0$, we find the following bistability threshold intensity:

$$I_{\text{th}} = \frac{\sqrt{3} \times 1.55 \times 10^{-6} \times (1 - 0.99)}{9 \times 312.6 \times 9 \times 10^{-11} \times 1 \times 10^{-6}} \text{ W m}^{-2} = 106 \text{ kW m}^{-2}.$$

For a spot size of $w_0 = 20$ μm, the bistability threshold power is

$$P_{\text{th}} = \frac{\pi w_0^2}{2} I_{\text{th}} = \frac{\pi \times (20 \times 10^{-6})^2}{2} \times 106 \times 10^3 \text{ W} = 66.6 \text{ μW}.$$

(b) With an absorption coefficient of $\alpha = 1.5 \times 10^4$ m^{-1}, we find that $e^{-\alpha l} = 0.9851$ for $l = 1$ μm. This amounts to a small single-pass loss of only 1.49%, but the finesse of the cavity is significantly reduced to

$$F = \frac{\pi\sqrt{0.99 \times 0.9851}}{1 - 0.99 \times 0.9851} = 125.3.$$

From (9.174), we see that the threshold intensity for such a lossy cavity is increased by a factor $F_0^3/F^3 = 15.5$ over that obtained above for a lossless cavity. Therefore, the realistic threshold intensity for this device is $I_{\text{th}} = 15.5 \times 106$ kW m^{-2} ≈ 1.64 MW m^{-2}, and the realistic threshold power is $P_{\text{th}} = 15.5 \times 66.6$ μW ≈ 1.03 mW. We see from this example that the loss in the cavity has a very significant effect on increasing the threshold of a dispersive bistable device.

Absorptive bistable optical devices

For a purely absorptive bistable device, we consider a Fabry–Perot cavity filled with a saturable absorber. The absorption coefficient is

$$\alpha = \frac{\alpha_0}{1 + I_c/I_{sat}} = \frac{\alpha_0}{1 + 2I_{out}/I_{sat}(1 - R)}. \tag{9.175}$$

The real part of the index of refraction is assumed to be independent of the light intensity. Therefore, the round-trip phase shift is a constant. We fix it at $2kl = 2m\pi$, which corresponds to a resonance peak of the Fabry–Perot cavity and can be done by tuning the cavity length at a given optical frequency. For a useful device, the total absorption has to be small in order to reduce the loss. Therefore, we consider only the limit of $\alpha l \ll 1$.

Under the conditions described above, (9.162) becomes

$$\frac{I_{out}}{I_{in}} \approx \frac{(1 - R)^2}{(1 - R + R\alpha l)^2} = \frac{1}{[1 + R\alpha l/(1 - R)]^2}. \tag{9.176}$$

The characteristics of this device are obtained by solving (9.175) and (9.176) together. A graphic solution is not necessary because the analytic solution is relatively simple.

Using (9.175), we can express the relation in (9.176) in the following form:

$$I_{in} = \left[1 + \frac{R\alpha_0 l}{1 - R}\frac{1}{1 + 2I_{out}/I_{sat}(1 - R)}\right]^2 I_{out}. \tag{9.177}$$

By demanding that $dI_{in}/dI_{out} = 0$ have two nondegenerate, real and positive solutions, we find the following condition for bistability (see Problem 9.8.5):

$$C_0 = \frac{R\alpha_0 l}{1 - R} > 8. \tag{9.178}$$

The transition points and the bistable region are found from the two nondegenerate solutions. Figure 9.29 shows the characteristics of this absorptive device for a few different values of the characteristic parameter $C_0 = R\alpha_0 l/(1 - R)$. It can be seen that bistability exists only when $C_0 > 8$.

Other bistable devices

The bistable optical devices discussed above are passive, intrinsic devices with intensity bistability. Besides these, there are many other optical devices that also exhibit optical bistability. As mentioned earlier, there are also hybrid bistable devices, which employ electrical feedback to create bistability in their optical characteristics. In addition, optical bistability also occurs in lasers. Intensity bistability is only one form of optical bistability. Indeed, optical bistability can take many different forms, including intensity bistability, phase bistability, frequency bistability, and polarization bistability.

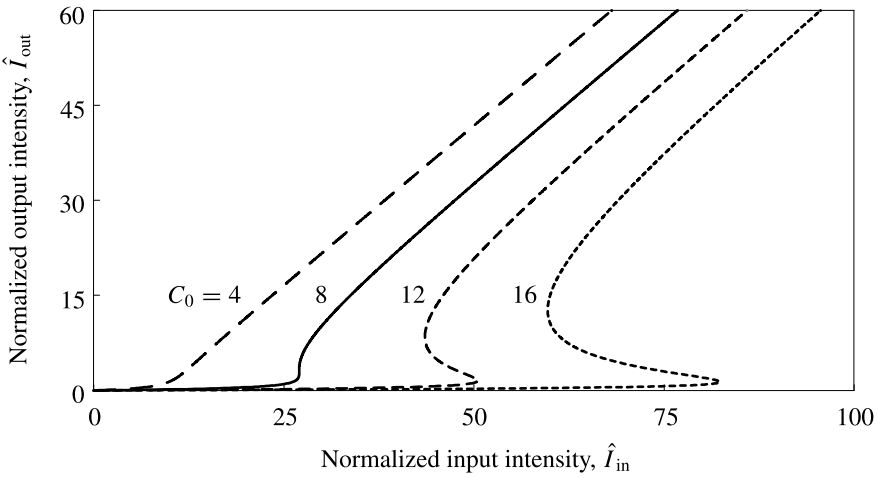


Figure 9.29 Characteristics of an absorptive nonlinear device with a saturable absorber in a Fabry–Perot cavity, where $\hat{I}_{in} = 2I_{in}/((1-R)I_{sat})$ and $\hat{I}_{out} = 2I_{out}/((1-R)I_{sat})$. Bistability exists only when $C_0 > 8$. The solid curve represents the threshold condition for bistability at $C_0 = 8$.

9.9 Raman and Brillouin devices

The nonparametric processes of stimulated Raman scattering and stimulated Brillouin scattering both cause a shift of the optical frequency, leading to a loss for the pump beam and a gain for a Stokes beam if the material is not originally excited or a gain for an anti-Stokes beam if it is excited. On the positive side, such processes can be utilized for optical frequency conversion and optical signal amplification. On the negative side, however, they also place some serious limitations on the performance of certain optical devices and systems.

Both stimulated Raman scattering and stimulated Brillouin scattering can be characterized by the imaginary part of a complex third-order nonlinear susceptibility of the form $\chi^{(3)}(\omega_S = \omega_S + \omega_p - \omega_p)$ for the Stokes interaction and one of the form $\chi^{(3)}(\omega_{AS} = \omega_{AS} + \omega_p - \omega_p)$ for the anti-Stokes interaction. These susceptibilities are in resonance with a frequency $\Omega = \omega_p - \omega_S = \omega_{AS} - \omega_p$ that characterizes a material excitation and, consequently, have the property given in (9.71). In most device applications using a Raman or Brillouin process, a material is initially in its normal state without being excited. Therefore, we shall consider only the Stokes process in the following discussions.

Being nonparametric, the Raman and Brillouin processes are automatically phase matched. The corresponding material excitation in an interaction picks up any phase mismatch between the interacting optical waves. Indeed, a Raman or Brillouin Stokes process can be viewed as a *parametric* interaction among a pump wave, a Stokes wave, and a *material excitation wave*. The material excitation wave is characterized by a

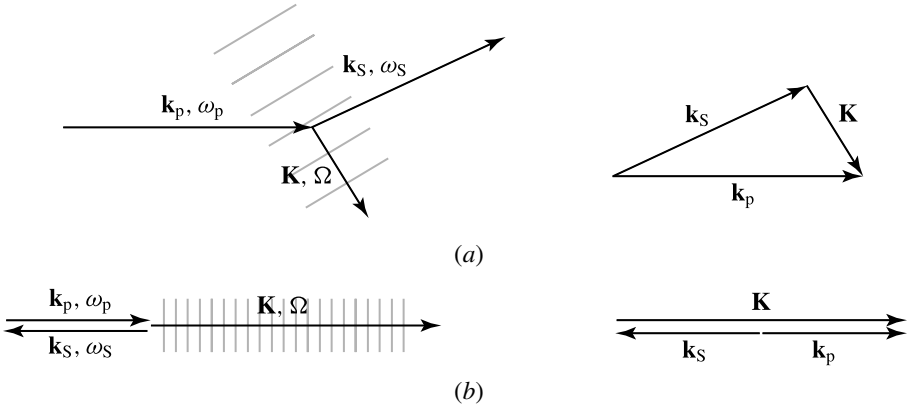


Figure 9.30 Generation of a Stokes optical wave and a material excitation wave by a pump optical wave with phase-matching condition in (*a*) a Raman Stokes process and (*b*) a Brillouin Stokes process. For the Raman process, the Stokes wave can be generated in any direction. For the Brillouin process, the Stokes wave is generated only in the backward direction.

frequency Ω and a wavevector $\mathbf{K}$. From this viewpoint, it is easy to see that a Stokes interaction is governed by the following conditions:

$$\omega_S = \omega_p - \Omega, \tag{9.179}$$

$$\mathbf{k}_S = \mathbf{k}_p - \mathbf{K}. \tag{9.180}$$

Clearly, phase matching among the pump wave, the Stokes wave, and the material excitation wave is needed, but it is automatically achieved when the pump wave generates a material excitation that allows a Raman or Brillouin process to occur. Figure 9.30 illustrates the relations among the three interacting waves in a Stokes process.

As mentioned in Section 9.3, the fundamental difference between the Raman and the Brillouin processes lies in the different mode of material excitation associated with each process. This difference leads to very different considerations for these two processes.

Raman gain

Because an excitation that is responsible for Raman scattering is associated with a transition at the molecular or atomic level, the Raman frequency shift is determined by the resonance frequency, Ω_R, of the Raman transition. This *Raman frequency* is an intrinsic property of a material and is independent of the frequency of the pump optical wave. Such an excitation is also nondispersive. Because a nondispersive excitation can take any wavevector $\mathbf{K}$ independently of its frequency, the phase-matching condition in (9.180) is satisfied for any combination of $\mathbf{k}_p$ and $\mathbf{k}_S$ independently of the condition in (9.179). As a consequence, Raman scattering in all directions has the same frequency shift that is specific to a given material. Spontaneous Raman scattering has a nearly isotropic emission pattern in all directions, whereas stimulated Raman

scattering occurs predominantly in the forward and backward directions due to the fact that a stimulated signal grows in strength as the interaction length increases. The Raman frequency shift, which is usually quoted per centimeter, is typically in the range of 300–3000 cm^{-1} for $f_R/c = \Omega_R/2\pi c$, equivalent to 10–100 THz for f_R, for most materials (1 cm^{-1} is equivalent to 30 GHz).

When a material is initially in its ground state of a Raman transition, the effective Raman susceptibility defined in (9.75) has a negative imaginary part: $\chi_R'' < 0$. This situation leads to a gain for the Raman Stokes signal at the expense of the pump wave. From (9.78), we find that the *Raman gain factor* for the Stokes signal is given by

$$\tilde{g}_R = -\frac{3\omega_S \mu_0}{n_{S,z} n_{p,z}} \chi_R'', \tag{9.181}$$

which has a positive value when $\chi_R'' < 0$. The unit of $\tilde{g}_R$ is meters per watt. The Raman susceptibility is only very weakly dependent on the individual optical frequencies, ω_p and ω_S, but it is a strong function of $\Omega = \omega_p - \omega_S$ with a resonance at $\Omega = \Omega_R$. In the simple case when there is only one Raman resonance frequency in a material, χ_R'' as a function of Ω has a Lorentzian lineshape as that of the linear susceptibility given in (1.176). Therefore, according to (9.181), the corresponding Raman gain factor has the form:

$$\tilde{g}_R = \tilde{g}_{R0} \frac{\gamma_R^2}{(\Omega - \Omega_R)^2 + \gamma_R^2} = \tilde{g}_{R0} \frac{\gamma_R^2}{(\omega_p - \omega_S - \Omega_R)^2 + \gamma_R^2}, \tag{9.182}$$

where $\tilde{g}_{R0}$ is the peak Raman gain factor and γ_R is the relaxation constant for the Raman excitation. This Raman gain factor has a FWHM linewidth given by $\Delta\Omega_R = 2\gamma_R$, or $\Delta f_R = \gamma_R/\pi$. Note that both the Raman frequency f_R and the Raman spectral linewidth Δf_R are independent of the pump and the Stokes optical frequencies, but the peak Raman gain factor varies linearly with the Stokes optical frequency: $\tilde{g}_{R0} \propto \omega_S \propto 1/\lambda_S$.

The response time of a Raman process is measured by $\tau_R = \gamma_R^{-1}$, which is the relaxation lifetime of the Raman excitation such as an optical phonon or a molecular vibration. Typical Raman response times range from a few hundred picoseconds in molecules through a few picoseconds in crystalline solids to tens of femtoseconds in amorphous solids such as glasses. Accordingly, the Raman linewidth Δf_R ranges from a few gigahertz to the order of 10 THz, depending on the properties of the materials. A Raman process can efficiently respond only to an optical signal that has a bandwidth narrower than the Raman linewidth. For an optical pulse, this means that its pulsewidth has to be greater than the Raman response time. Therefore, depending on the specific material used, it is possible for a Raman device to function in steady state or in quasi-steady state with optical signals ranging from CW waves to picosecond or even subpicosecond optical pulses. When an optical signal varies faster than the Raman relaxation time, the interaction is characterized by *transient stimulated Raman*

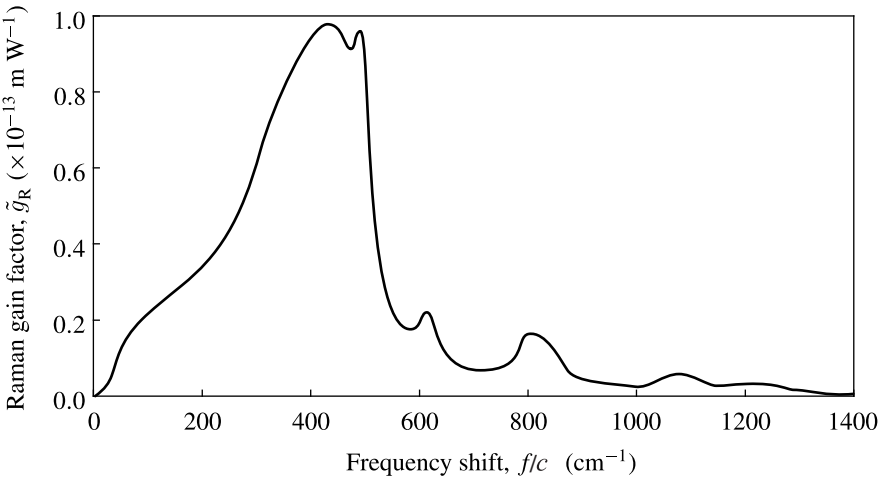


Figure 9.31 Spectrum of the Raman gain factor of fused silica measured at a pump wavelength of 1 μm for pump and Stokes waves that are linearly polarized in the same direction. Note that 1 cm^{-1} ≡ 30 GHz. (Adapted from Stolen, R. H., "Nonlinearity in fiber transmission," *Proceedings of the IEEE* **68**: 1232–1236, Oct. 1980.)

scattering with a reduced Raman gain among other features that are different from those of steady-state stimulated Raman scattering. In this section, we consider only Raman devices operating in the steady-state regime.

Forward and backward Raman interactions have the same Raman gain factor. However, the value of $\tilde{g}_{R0}$ is a function of the polarization states of the pump and the Stokes waves because $\boldsymbol{\chi}^{(3)}$ is a tensor and the effective Raman susceptibility defined in (9.75) depends on $\hat{e}_p$ and $\hat{e}_S$. In an isotropic medium, the maximum value of $\tilde{g}_{R0}$ is found when the pump and the Stokes waves are linearly polarized in the same direction. Therefore, special attention has to be paid to the polarization states of the optical waves throughout the course of interaction in evaluating the efficiency of a Raman process.

As an example, Fig. 9.31 shows the spectral dependence of the Raman gain factor of fused silica glass measured at a pump optical wavelength of 1 μm for pump and Stokes waves that are linearly polarized in the same direction. This spectrum is very broad and does not have the ideal Lorentzian lineshape because there are many closely clustered Raman resonances in such an amorphous solid material. This Raman spectral shape remains more or less the same for other pump wavelengths, but its peak value scales with the pump wavelength as $\tilde{g}_{R0} \approx (1 \times 10^{-13}/\lambda_S)$ m W^{-1}, where λ_S is in micrometers. The Raman gain factor of fused silica is relatively small compared to those of many molecular substances such as benzene and CS_2. Many amorphous glass materials, such as GeO_2, B_2O_3, and P_2O_5, which are commonly used to dope silica fibers also have peak Raman gain factors that are five to ten times that of pure silica with corresponding frequency shifts ranging from 400 to 1400 cm^{-1}. In particular, the peak Raman gain factor of GeO_2 is 9.2 times that of pure silica at a frequency shift of

420 cm^{-1}. Therefore, the peak value, the frequency shift corresponding to the peak, and the spectral shape of the Raman gain factor of a particular fiber all depend on the type and concentration of the dopants in the fiber.

EXAMPLE 9.17 An optical wave at 1.55 μm wavelength propagates in a silica fiber that has a peak Raman gain factor as described in the text above at a Raman frequency shift of 460 cm^{-1}. If the optical intensity is sufficiently high to generate a Raman Stokes signal, what is the wavelength of the Stokes signal? What is the Raman frequency shift in hertz? What is the Raman gain factor for this signal?

Solution Because $\omega_S = \omega_p - \Omega_R$, the Stokes wavelength can be found by using the relation

$$\frac{1}{\lambda_S} = \frac{1}{\lambda_p} - \frac{f_R}{c}.$$

The Raman frequency shift quoted per centimeter is actually f_R/c in the above relation. For the fiber considered here, we have $f_R/c = 460\ \text{cm}^{-1} = 4.6 \times 10^4\ \text{m}^{-1}$. Therefore, the wavelength of the Stokes signal is

$$\lambda_S = \left(\frac{1}{1.55 \times 10^{-6}} - 4.6 \times 10^4\right)^{-1}\ \text{m} = 1.669\ \mu\text{m}.$$

Because 1 $\text{cm}^{-1} \equiv 30$ GHz, the Raman frequency shift is $f_R = 460\ \text{cm}^{-1} = 13.8$ THz. The Raman gain factor at this wavelength is

$$\tilde{g}_{R0} = \frac{1 \times 10^{-13}}{1.669}\ \text{m W}^{-1} = 5.99 \times 10^{-14}\ \text{m W}^{-1}.$$

Brillouin gain

For Brillouin scattering, the relevant excitation is a long-range acoustic wave, which has a linear dispersion relation between the magnitude of its wavevector and its frequency as that given in (8.3): $K = \Omega/v_a$. The conditions in (9.179) and (9.180) for Brillouin Stokes scattering are the same as those for the first-order down-shifted Bragg diffraction discussed in Section 8.3, except that in Brillouin scattering the acoustic wave is generated by the pump optical wave whereas in acousto-optic Bragg diffraction the acoustic wave is externally applied to the medium. Therefore, the amount of frequency shift in Brillouin scattering is a function of the pump optical frequency and the scattering angle, θ, between $\mathbf{k}_S$ and $\mathbf{k}_p$. In general, the Brillouin frequency shift is a few orders of magnitude smaller than the pump and the Stokes optical frequencies. For Brillouin scattering in an isotropic medium, the approximation $k_S \approx k_p$ is valid. Then, by using (9.179) and (9.180) together with the dispersion relation of the acoustic wave, we find

the following angle-dependent frequency shift:

$$\Omega = 2v_a k_p \sin\frac{\theta_{def}}{2} = \frac{2nv_a}{c}\omega_p \sin\frac{\theta_{def}}{2}, \tag{9.183}$$

where n is the index of refraction at the optical frequency ω_p, v_a is the acoustic velocity in the medium, and $\theta_{def} = \theta_d - \theta_i$ is the deflection angle of the acousto-optic Bragg diffraction as defined in Section 8.3. We see that, very differently from Raman scattering, Brillouin scattering does not have a constant frequency shift in all directions. In particular, there is no Brillouin Stokes scattering in the forward direction because Ω given in (9.183) vanishes for $\theta_{def} = 0$. Spontaneous Brillouin scattering appears in all other directions with a frequency shift that varies with the scattering angle. Stimulated Brillouin scattering occurs predominantly in the backward direction with a maximum frequency shift, known as the *Brillouin frequency*, which is determined by the phase-matching condition given in (9.180) to be

$$\Omega_B = \frac{nv_a}{c}(\omega_p + \omega_S) = \frac{2nv_a/c}{1 + nv_a/c}\omega_p \approx \frac{2nv_a}{c}\omega, \tag{9.184}$$

where we have used the fact that $\omega_p \approx \omega_S = \omega \gg \Omega_B$, or

$$f_B = \frac{\Omega_B}{2\pi} = \frac{2nv_a}{\lambda}. \tag{9.185}$$

With a pump beam in the optical spectral region, the Brillouin frequency f_B falls in the hypersonic region, typically in the range of 1–50 GHz for a large variety of materials.

The *Brillouin gain factor* of a material can be expressed in a form similar to that of the Raman gain factor. For backward interaction at a Brillouin frequency Ω_B, we have

$$\tilde{g}_B = \tilde{g}_{B0}\frac{\gamma_B^2}{(\Omega - \Omega_B)^2 + \gamma_B^2} = \tilde{g}_{B0}\frac{\gamma_B^2}{(\omega_p - \omega_S - \Omega_B)^2 + \gamma_B^2}. \tag{9.186}$$

This Brillouin gain factor has a FWHM linewidth of $\Delta\Omega_B = 2\gamma_B$, or $\Delta f_B = \gamma_B/\pi$, which is associated with a response time of γ_B^{-1} for the acoustic excitation that is responsible for the Brillouin process. Because the Brillouin response time in a common material is typically on the order of nanoseconds, the Brillouin linewidth Δf_B is typically in the range of 10 MHz to 1 GHz. Therefore, a Brillouin device does not respond efficiently to very short optical pulses, or to any optical waves that have spectral widths in the gigahertz range or above. For a pump optical wave that has a Lorentzian spectral shape with a FWHM linewidth $\Delta\nu_p$, the peak Brillouin gain factor of a medium scales as

$$\tilde{g}_{B0} = \frac{\Delta f_B}{\Delta f_B + \Delta\nu_p}\tilde{g}_{B0}^{max}, \tag{9.187}$$

where $\tilde{g}_{B0}^{max}$ is the peak Brillouin gain factor for an idealistic CW wave of zero linewidth with $\Delta\nu_p = 0$. Clearly, when $\Delta\nu_p \gg \Delta f_B$, the peak Brillouin gain factor is greatly

reduced. The Brillouin gain factor has other characteristics that are different from those of the Raman gain factor due to the fact that the Brillouin frequency is dictated by the phase-matching condition of (9.180). We have seen in (9.184) that $\Omega_B \propto \omega$. In addition, $\gamma_B \propto \omega^2$, but $\tilde{g}_{B0}$ is independent of optical frequency. For fused silica, $f_B \approx (17.3/\lambda)$ GHz and $\Delta f_B \approx (38.4/\lambda^2)$ MHz, where λ is in micrometers, and $\tilde{g}_{B0}^{max} = 4.5 \times 10^{-11}$ m W^{-1}.

Many gases, such as H_2, N_2, O_2, Ar, and Xe, have useful Raman and Brillouin gains and frequency shifts for practical applications. A gas for such applications is normally contained in a high-pressure cell, often called a *Raman cell* or a *Brillouin cell* depending on its intended application. One significant difference between a gaseous medium and a liquid or solid medium is that $\tilde{g}_{R0}$, Δf_R, $\tilde{g}_{B0}$, and Δf_B of a gaseous medium depend on the density of the molecules in the medium, which can be varied by varying the gas pressure in a cell of fixed length and volume. The value of $\tilde{g}_{R0}$ scales linearly with the density of the gas molecules at low pressures until it saturates at a certain pressure. In comparison, the value of $\tilde{g}_{B0}$ scales quadratically with the density of the gas molecules. Therefore, for a given gaseous medium, $\tilde{g}_{R0}$ can be larger than $\tilde{g}_{B0}$ at low pressures, but $\tilde{g}_{B0}$ eventually becomes larger than $\tilde{g}_{R0}$ at a sufficiently high pressure. In addition, $\tilde{g}_{B0}$ and Δf_B also depend on temperature.

EXAMPLE 9.18 For the optical wave at 1.55 μm wavelength propagating in a silica fiber as described in Example 9.17, what are the Brillouin frequency shift, the Brillouin linewidth, and the peak Brillouin gain factor if the optical wave has a linewidth of 1 MHz? What is the peak Brillouin gain factor if the optical wave has a linewidth of 100 MHz?

Solution According to the characteristics of fused silica described in the text, the Brillouin frequency shift is $f_B = (17.3/1.55)$ GHz $= 11.16$ GHz, and the Brillouin linewidth is $\Delta f_B = (38.4/1.55^2)$ MHz $= 15.98$ MHz. Though $\tilde{g}_{B0}^{max} = 4.5 \times 10^{-11}$ m W^{-1} for the silica fiber is quite independent of the optical wavelength, the peak Brillouin gain factor varies with the linewidth $\Delta\nu_p$ of the optical wave according to (9.187). Therefore, the peak Brillouin gain factor is $\tilde{g}_{B0} = (15.98/16.98) \times 4.5 \times 10^{-11}$ m W^{-1} $\approx 4.23 \times 10^{-11}$ m W^{-1} if the optical wave has a narrow linewidth of $\Delta\nu_p = 1$ MHz. If the linewidth of the optical wave is increased to $\Delta\nu_p = 100$ MHz, the peak Brillouin gain is reduced to $\tilde{g}_{B0} = (15.98/115.98) \times 4.5 \times 10^{-11}$ m W^{-1} $\approx 6.2 \times 10^{-12}$ m W^{-1}. Further increase of the linewidth of the optical wave will further reduce the peak Brillouin gain.

Raman amplifiers

The Raman gain can be utilized to amplify an optical signal at a Stokes frequency ω_S through the process of stimulated Raman scattering by choosing a proper pump wave

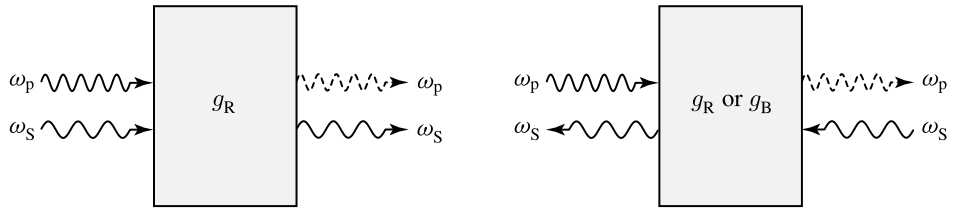


Figure 9.32 (*a*) Codirectional Raman amplifier and (*b*) contradirectional Raman or Brillouin amplifier. A Brillouin amplifier cannot take the codirectional configuration in (*a*). For Raman amplification $\omega_p = \omega_S + \Omega_R$, while for Brillouin amplification $\omega_p = \omega_S + \Omega_B$.

at the frequency $\omega_p = \omega_S + \Omega$ with a frequency shift Ω that is within the Raman gain spectrum, ideally at the gain-peak frequency Ω_R.

Because stimulated Raman scattering has the same gain factor in forward and backward directions, the pump wave and the Stokes signal wave can propagate either codirectionally, as shown in Fig. 9.32(*a*), or contradirectionally, as shown in Fig. 9.32(*b*), in a Raman amplifier. However, though the gain factor is the same for the two configurations, a Raman amplifier with a contradirectional configuration would have little or no efficiency for short optical pulses because of the short interaction length between the pump and the Stokes pulses that propagate in opposite directions. Here we consider only Raman amplification in a codirectional configuration. The general formulation and characteristics for Raman amplification in a contradirectional configuration are similar to those for Brillouin amplification discussed later.

Following (9.78) and (9.79) and allowing for the existence of linear absorption loss in a medium, we have the following coupled equations for Raman amplification in a codirectional configuration:

$$\frac{dI_S}{dz} + \alpha_S I_S = \tilde{g}_R I_S I_p, \tag{9.188}$$

$$\frac{dI_p}{dz} + \alpha_p I_p = -\frac{\omega_p}{\omega_S}\tilde{g}_R I_S I_p, \tag{9.189}$$

with given values of $I_p(0)$ and $I_S(0)$ at the input end, $z = 0$, of the amplifier as the initial conditions. The parameters α_S and α_p are the linear absorption coefficients of the medium at the Stokes and the pump frequencies, respectively.

The coupled equations in (9.188) and (9.189) have an exact analytical solution when $\alpha_p = \alpha_S$ (see Problem 9.9.3). In the amplification of a weak signal when the depletion of the pump intensity due to Raman interaction can be neglected, a simple approximate solution can be obtained by ignoring the term on the right-hand side of (9.189). Then, for a Raman amplifier of length l, we have the signal at the output end of the amplifier

given by

$$I_S(l) = I_S(0) \exp\left(g_R l_{eff} - \alpha_S l\right), \tag{9.190}$$

where g_R is the *Raman gain coefficient* defined as

$$g_R = \tilde{g}_R I_p(0) = -\frac{3\omega_S \mu_0}{n_{S,z} n_{p,z}} \chi_R'' I_p(0), \tag{9.191}$$

and l_{eff} is the effective interaction length of Raman amplification given by

$$l_{eff} = \frac{1 - e^{-\alpha_p l}}{\alpha_p}. \tag{9.192}$$

Note that the Raman gain coefficient increases with pump intensity. Such dependence on an optical intensity is characteristic of an optical gain that is contributed by a nonlinear optical process. The amplification factor, or the *Raman amplifier gain*, for the Stokes signal in the case of negligible pump depletion is then

$$G_R = \frac{I_S(l)}{I_S(0)\exp(-\alpha_S l)} = \exp(g_R l_{eff}) = \exp\left[\tilde{g}_R I_p(0) l_{eff}\right]. \tag{9.193}$$

For a given Raman amplifier with a fixed length, the amplifier gain can be controlled by varying the pump intensity.

EXAMPLE 9.19 An optical fiber that has the Raman gain characteristics described in Example 9.17 is used as a fiber Raman amplifier for codirectional amplification of an optical signal at $\lambda_S = 1.55$ μm. The input power of the signal is −15 dBm, and the desired output power is 0 dBm. The fiber has an absorption coefficient of $\alpha = 0.2$ dB km^{-1} at this signal wavelength and a length of $l = 25$ km. Its core has an effective cross-sectional area of $\mathcal{A}_{eff} = 5 \times 10^{-11}$ m^2. What is the pump wavelength for the largest Raman gain? What is the required pump power if the absorption coefficient at the pump wavelength is the same as that at the signal wavelength?

Solution Because the Raman gain spectrum of an optical fiber is very broad, it is in general only necessary to pick a pump wavelength so that the signal wavelength falls within the Raman gain spectral range of the pump. However, to have the largest Raman gain, we need to choose a pump wavelength properly so that the Raman gain peak appears at the signal wavelength. From Example 9.17, we have $f_R/c = 460$ cm^{-1} = 4.6×10^4 m^{-1} at the peak of the Raman spectrum. Therefore, the pump wavelength for the largest Raman gain is

$$\lambda_p = \left(\frac{1}{1.55 \times 10^{-6}} + 4.6 \times 10^4\right)^{-1} \text{ m} = 1.4468 \text{ μm}.$$

The peak Raman gain factor at $\lambda_S = 1.55$ μm is

$$\tilde{g}_{\mathrm{R}} = \frac{1 \times 10^{-13}}{1.55}\ \mathrm{m\ W^{-1}} = 6.45 \times 10^{-14}\ \mathrm{m\ W^{-1}}.$$

With $\alpha = 0.2\ \mathrm{dB\ km^{-1}} = 0.046\ \mathrm{km^{-1}}$ and $l = 25$ km, we have

$$l_{\mathrm{eff}} = \frac{1 - \mathrm{e}^{-0.046 \times 25}}{0.046}\ \mathrm{km} = 14.86\ \mathrm{km}.$$

Because $P_{\mathrm{S}}^{\mathrm{in}} = I_{\mathrm{S}}(0)\mathcal{A}_{\mathrm{eff}}$ and $P_{\mathrm{S}}^{\mathrm{out}} = I_{\mathrm{S}}(l)\mathcal{A}_{\mathrm{eff}}$, we find from (9.193) that the required Raman amplifier gain in decibels is

$$G_{\mathrm{R}} = P_{\mathrm{S}}^{\mathrm{out}}(\mathrm{dBm}) - P_{\mathrm{S}}^{\mathrm{in}}(\mathrm{dBm}) + \alpha(\mathrm{dB\ km^{-1}})l(\mathrm{km}) = 20\ \mathrm{dB}. \tag{9.194}$$

Therefore, $G_{\mathrm{R}} = 20\ \mathrm{dB} = 100$. Identifying the pump power $P_{\mathrm{p}} = I_{\mathrm{p}}(0)\mathcal{A}_{\mathrm{eff}}$ and using (9.193), we find the following required pump power:

$$P_{\mathrm{p}} = \frac{\mathcal{A}_{\mathrm{eff}}}{l_{\mathrm{eff}}} \frac{\ln G_{\mathrm{R}}}{\tilde{g}_{\mathrm{R}}} = \frac{5 \times 10^{-11} \times \ln 100}{14.86 \times 10^{3} \times 6.45 \times 10^{-14}}\ \mathrm{W} = 240\ \mathrm{mW}. \tag{9.195}$$

Note that in using (9.193) to obtain (9.194) and (9.195), we have implicitly assumed that depletion of the pump power due to its conversion to the signal power is negligible. This assumption is clearly valid here because the pump power obtained under such an assumption is 240 mW while the output signal is only 0 dBm, which is 1 mW. Stimulated Brillouin scattering has to be suppressed in a Raman amplifier because it can deplete the pump power for Raman amplification. Because the Brillouin gain factor decreases with the linewidth of the pump, an optical source of a linewidth that is large enough to suppress stimulated Brillouin scattering is normally used for pumping a Raman amplifier. Multimode semiconductor lasers can serve such a purpose for pumping fiber Raman amplifiers.

Raman generators

When there is a pump optical wave in a medium that has a Raman susceptibility, spontaneous Raman scattering that generates incoherent Stokes and anti-Stokes emission in all directions always occurs. In a Raman amplifier where a coherent input signal is amplified, such incoherent spontaneous emission contributes to the noise in the amplifier. In the absence of an input signal, however, the ubiquitous spontaneous Raman emission can be the seed for the generation of a Stokes or anti-Stokes wave through stimulated amplification under the right conditions. A Raman generator is normally used for the generation of the Stokes wave at the down-shifted Stokes frequency of $\omega_{\mathrm{S}} = \omega_{\mathrm{p}} - \Omega_{\mathrm{R}}$ with the medium initially unexcited. A Raman generator can simply be a Raman amplifier without an input signal but with a pump of sufficient intensity for significant power conversion from the pump frequency to the Stokes frequency in a single pass through the medium.

In a Raman generator, the Stokes wave grows from stimulated amplification of the spontaneous Stokes emission. Because spontaneous Stokes emission occurs along the entire length of the generator, the total Stokes power at the output is the result of the cumulative amplification of all spontaneous Stokes emission over the length of the generator. A detailed analysis that takes into account such cumulative amplification can be carried out. For forward Raman interaction, the net result is equivalent to treating the generator as an amplifier with the injection of an effective Stokes signal $I_S^{\text{eff}}(0)$ at $z = 0$ while ignoring all of the spontaneous Stokes emission in the generator. For backward interaction, it is equivalent to injection of an effective Stokes signal $I_S^{\text{eff}}(l)$ at $z = l$ while ignoring all of the spontaneous Stokes emission. The values of the effective signals depend on the Raman characteristics, particularly $\tilde{g}_{R0}$ and Δf_R, of the medium, as well as on the pump intensity. Besides, due to the difference between the forward and the backward interactions in the geometric relation of the pump and the Stokes waves, the effective signal $I_S^{\text{eff}}(l)$ for backward interaction is significantly smaller than the effective signal $I_S^{\text{eff}}(0)$ for forward interaction at a given pump intensity in a given medium. This difference leads to a higher threshold for backward Raman generation than that for forward Raman generation. As a result, only a Stokes wave in the forward direction is generated in a Raman generator. No backward generation occurs.

Because significant power conversion from the pump wave to the Stokes wave is desired in the application of a Raman generator, pump depletion cannot be neglected. If we assume for simplicity that $\alpha_S = \alpha_p$ and consider the fact that $I_p(0) \gg I_S^{\text{eff}}(0)$ for a Raman generator, the complete solution of the coupled equations in (9.188) and (9.189) leads to the relation (see Problem 9.9.3)

$$\frac{I_S(l)}{I_p(l)} \approx \frac{I_S^{\text{eff}}(0)}{I_p(0)} \exp\left[\tilde{g}_R I_p(0) l_{\text{eff}}\right] = \frac{I_S^{\text{eff}}(0)}{I_p(0)} G_R. \tag{9.196}$$

The threshold of a Raman generator can be defined as the condition for $I_S(l) = I_p(l)$. Then, the following threshold amplification factor is obtained at the *Raman threshold*:

$$G_R^{\text{th}} = \exp\left[\tilde{g}_R I_p^{\text{th}}(0) l_{\text{eff}}\right] = \frac{I_p^{\text{th}}(0)}{I_{S,\text{th}}^{\text{eff}}(0)}. \tag{9.197}$$

The physical meaning of this relation is that the Raman threshold is reached when stimulated amplification of the spontaneous Stokes emission brings the Stokes intensity at the output to the same level as that of the pump intensity. Because the value of $I_S^{\text{eff}}(0)$ depends on the characteristics of the medium, the value of G_R^{th} is also a function of the characteristics of the medium. Therefore, the threshold pump intensity for forward Raman Stokes generation in a medium of length l can be calculated using the following relation:

$$I_p^{\text{th}}(0) \approx \frac{\ln G_R^{\text{th}}}{\tilde{g}_R l_{\text{eff}}}. \tag{9.198}$$

For example, $\ln G_{\mathrm{R}}^{\mathrm{th}} \approx 16$ for forward Raman Stokes generation in single-mode silica fibers. For backward Raman Stokes generation in a single-mode silica fiber, $\ln G_{\mathrm{R}}^{\mathrm{th}} \approx 20$. Backward Raman Stokes generation normally does not occur because it has a much higher threshold than forward generation.

In the case when $\alpha_{\mathrm{S}} = \alpha_{\mathrm{p}} = \alpha$, a simple relation for calculating the conversion efficiency of a Raman generator can be obtained by assuming that $I_{\mathrm{p}}(0) \gg I_{\mathrm{S}}^{\mathrm{eff}}(0) = I_{\mathrm{S,th}}^{\mathrm{eff}}(0)$ for any pump intensity. Then, the Raman conversion efficiency from the pump to the Stokes is found to be (see Problem 9.9.4)

$$\eta_{\mathrm{R}} = \frac{I_{\mathrm{S}}(l)}{I_{\mathrm{p}}(0)} = \frac{\omega_{\mathrm{S}}}{\omega_{\mathrm{p}}} \frac{1}{1 + (\omega_{\mathrm{S}}/\omega_{\mathrm{p}}) r (G_{\mathrm{R}}^{\mathrm{th}})^{1-r}} \mathrm{e}^{-\alpha l}, \tag{9.199}$$

where $r = I_{\mathrm{p}}(0)/I_{\mathrm{p}}^{\mathrm{th}}(0)$ is the pump ratio with respect to the threshold pump intensity. The threshold of a Raman generator is very sharp. Below the threshold, η_{R} quickly approaches zero, but it quickly approaches its maximum value of $(\omega_{\mathrm{S}}/\omega_{\mathrm{p}})\mathrm{e}^{-\alpha l}$ above the threshold. Therefore, (9.199) can be used to find the Raman conversion efficiency quite accurately for any value of r irrespective of the assumption used in obtaining it. By using (9.199) to calculate the Raman Stokes generation in a single-mode silica fiber, it is found that a reduction in the pump intensity by 1 dB below the threshold reduces the output Stokes intensity by more than 10 dB, but an increase in the pump intensity by 1 dB above the threshold causes the conversion from the pump to the Stokes to be more than 98% complete.

If the pump intensity is many times above the threshold, complete conversion of power from the pump to the Stokes occurs within a very short distance from the input end. This first Stokes wave at the frequency $\omega_{\mathrm{S1}} = \omega_{\mathrm{p}} - \Omega_{\mathrm{R}}$ can then serve as a pump to generate the second Stokes wave at the frequency $\omega_{\mathrm{S2}} = \omega_{\mathrm{S1}} - \Omega_{\mathrm{R}} = \omega_{\mathrm{p}} - 2\Omega_{\mathrm{R}}$. This cascading process continues until the waves reach the end of the generator. Therefore, with proper choices of generator length and pump intensity, a high-order Stokes wave can be generated at a frequency that is down-shifted from the pump frequency by an integral multiple of the Raman frequency. However, such complete power conversion from the pump to the first Stokes and from a low-order Stokes to a high-order Stokes is possible only for CW waves or very long optical pulses. For short optical pulses, power conversion from one order to another is normally not complete due to temporal walk-off between the interacting pulses of different wavelengths caused by group-velocity dispersion in the medium, but generation of multiple Stokes orders is still possible with high-intensity pulses.

Sometimes, in addition to the Stokes wave, an anti-Stokes wave at the up-shifted anti-Stokes frequency of $\omega_{\mathrm{AS}} = \omega_{\mathrm{p}} + \Omega_{\mathrm{R}}$ can also be generated through Stokes–anti-Stokes coupling and/or parametric four-wave mixing with the pump if the required phase-matching conditions for such parametric processes are satisfied. In practical applications, however, a Raman generator is normally used as a nonparametric

frequency converter to convert the optical power at a high-frequency pump wave to a low-frequency Stokes wave.

EXAMPLE 9.20 The fiber Raman amplifier described in Example 9.19 can be used as a fiber Raman generator for a Stokes signal at $\lambda_S = 1.55$ μm without an input signal at this wavelength by raising the pump power at $\lambda_p = 1.4468$ μm. Find the threshold pump power for this fiber Raman generator.

Solution Identifying $P_p^{th} = I_p^{th}(0)\mathcal{A}_{eff}$ and using $\ln G_R^{th} \approx 16$, we have, from (9.198),

$$P_p^{th} = \frac{\mathcal{A}_{eff}}{l_{eff}} \frac{\ln G_R^{th}}{\tilde{g}_R} = \frac{16\mathcal{A}_{eff}}{\tilde{g}_R l_{eff}} \tag{9.200}$$

for the threshold pump power of a fiber Raman generator. Using the parameters obtained in Example 9.19, we find that

$$P_p^{th} = \frac{16 \times 5 \times 10^{-11}}{6.45 \times 10^{-14} \times 14.86 \times 10^3} \text{ W} = 835 \text{ mW}.$$

When the pump power is below P_p^{th}, very little power is converted to the Stokes signal in a Raman generator. When the pump power exceeds P_p^{th} at a certain level, it is completely converted to the Stokes. If the pump power continues to increase, the power is converted to a successively higher order of Stokes at the output.

Brillouin amplifiers

The Brillouin gain in a medium can also be utilized to amplify an optical signal at a frequency that is down-shifted from the pump frequency by an amount equal to the Brillouin frequency.

Due to the fundamental differences between the Raman and the Brillouin processes discussed above, the Brillouin amplifiers have several characteristics that are very different from those of the Raman amplifiers. First, only the contradirectional configuration shown in Fig. 9.32(*b*) is acceptable for a Brillouin amplifier because there is no forward Brillouin scattering. Second, the Brillouin linewidth is relatively narrow. Therefore, a Brillouin amplifier is useful only for the amplification of narrow-band signals, whereas a Raman amplifier can be used for broadband signals or short-pulse signals because of the large Raman linewidth. Third, the peak Brillouin gain factor, $\tilde{g}_{B0}$, of a solid or liquid medium, or a high-pressure gaseous medium, is usually much larger than the peak Raman gain factor, $\tilde{g}_{R0}$, of the same medium. Therefore, a Brillouin amplifier usually requires a much lower pump intensity than what a Raman amplifier needs to have the same amplification factor for the signal.

Because of the contradirectional configuration of a Brillouin amplifier, the signal propagates in the $-z$ direction while the pump propagates in the z direction. Therefore,

Brillouin amplification is described by the following coupled equations:

$$-\frac{\mathrm{d}I_\mathrm{S}}{\mathrm{d}z} + \alpha_\mathrm{S} I_\mathrm{S} = \tilde{g}_\mathrm{B} I_\mathrm{S} I_\mathrm{p}, \tag{9.201}$$

$$\frac{\mathrm{d}I_\mathrm{p}}{\mathrm{d}z} + \alpha_\mathrm{p} I_\mathrm{p} = -\frac{\omega_\mathrm{p}}{\omega_\mathrm{S}} \tilde{g}_\mathrm{B} I_\mathrm{S} I_\mathrm{p}, \tag{9.202}$$

with the input pump intensity $I_\mathrm{p}(0)$ at $z = 0$ and the input signal intensity $I_\mathrm{S}(l)$ at $z = l$ given as the boundary conditions.

The exact solution for this backward amplification differs from that for the forward amplification. It can be found when $\alpha_\mathrm{S} = \alpha_\mathrm{p} = 0$ (see Problem 9.9.11). In the application of an optical amplifier for the amplification of a weak signal, however, there is little pump depletion due to nonlinear Brillouin interaction. Then, the right-hand side of (9.202) can be ignored to obtain the following solution for the output intensity of the signal at $z = 0$:

$$I_\mathrm{S}(0) = I_\mathrm{S}(l) \exp\left(g_\mathrm{B} l_\mathrm{eff} - \alpha_\mathrm{S} l\right), \tag{9.203}$$

where l_eff is the effective interaction length of the same form as that defined in (9.192) and g_B is the *Brillouin gain coefficient* defined as

$$g_\mathrm{B} = \tilde{g}_\mathrm{B} I_\mathrm{p}(0). \tag{9.204}$$

Therefore, in the case of negligible pump depletion, the amplification factor of a Brillouin amplifier, or the *Brillouin amplifier gain*, is

$$G_\mathrm{B} = \frac{I_\mathrm{S}(0)}{I_\mathrm{S}(l)\exp(-\alpha_\mathrm{S} l)} = \exp(g_\mathrm{B} l_\mathrm{eff}) = \exp\left[\tilde{g}_\mathrm{B} I_\mathrm{p}(0) l_\mathrm{eff}\right]. \tag{9.205}$$

EXAMPLE 9.21 If the fiber Raman amplifier described in Example 9.19 is turned into a Brillouin amplifier for the same input signal and the same desired output signal, what should the pump wavelength be? If the pump wave has a linewidth of 100 MHz, what is the required pump power?

Solution From Example 9.18, we know that $f_\mathrm{B} = 11.16$ GHz. Therefore, $f_\mathrm{R}/c = 37.2\ \mathrm{m}^{-1}$, and the pump wavelength is

$$\lambda_\mathrm{p} = \left(\frac{1}{1.55 \times 10^{-6}} + 37.2\right)^{-1}\ \mathrm{m} = 1.5499\ \mu\mathrm{m}.$$

The pump wavelength is very close to the signal wavelength because of the small Brillouin frequency shift. We find from Example 9.18 that the peak Brillouin gain for this amplifier is $\tilde{g}_\mathrm{B} = 6.2 \times 10^{-12}\ \mathrm{m\ W^{-1}}$ because the pump has a linewidth of 100 MHz. Because a Brillouin amplifier functions only in the contradirectional configuration, we identify $P_\mathrm{S}^\mathrm{in} = I_\mathrm{S}(l)\mathcal{A}_\mathrm{eff}$ and $P_\mathrm{S}^\mathrm{out} = I_\mathrm{S}(0)\mathcal{A}_\mathrm{eff}$. Then we find from (9.205) that the

required Brillouin amplifier gain in decibels is

$$G_{\mathrm{B}} = P_{\mathrm{S}}^{\mathrm{out}}(\mathrm{dBm}) - P_{\mathrm{S}}^{\mathrm{in}}(\mathrm{dBm}) + \alpha(\mathrm{dB\ km^{-1}})l(\mathrm{km}) = 20\ \mathrm{dB}, \tag{9.206}$$

which is the same as the Raman amplifier gain in Example 9.19. Therefore, $G_{\mathrm{B}} = 100$. From (9.205) we find by identifying the pump power as $P_{\mathrm{p}} = I_{\mathrm{p}}(0)\mathcal{A}_{\mathrm{eff}}$ that

$$P_{\mathrm{p}} = \frac{\mathcal{A}_{\mathrm{eff}}}{l_{\mathrm{eff}}}\frac{\ln G_{\mathrm{B}}}{\tilde{g}_{\mathrm{B}}} = \frac{5 \times 10^{-11} \times \ln 100}{14.86 \times 10^{3} \times 6.2 \times 10^{-12}}\ \mathrm{W} = 2.5\ \mathrm{mW}. \tag{9.207}$$

By comparing (9.195) with (9.207), we find that for the same amplifier gain, $G_{\mathrm{B}} = G_{\mathrm{R}}$, the pump power required for a Brillouin amplifier is scaled from that for a Raman amplifier by a factor of $P_{\mathrm{p}}^{\mathrm{B}}/P_{\mathrm{p}}^{\mathrm{R}} = \tilde{g}_{\mathrm{R}}/\tilde{g}_{\mathrm{B}}$. Because $\tilde{g}_{\mathrm{B}} \gg \tilde{g}_{\mathrm{R}}$ by about two orders of magnitude in this example, the pump power is reduced by as much.

Note that in using (9.205) to obtain (9.206) and (9.207), we have implicitly assumed that depletion of the pump power due to its conversion to the signal power is negligible. This assumption is not really valid here because the pump power obtained under such an assumption is 2.5 mW but the output signal is 1 mW. A more detailed analysis with the effect of pump depletion taken into consideration is required to obtain the accurate result.

Brillouin generators

Similarly to the situation in a Raman generator, the emission from spontaneous Brillouin scattering can also seed the generation of a Brillouin Stokes frequency in the presence of a pump above a *Brillouin threshold* but in the absence of an input Stokes signal. Besides the fundamental differences in terms of the frequency shift and the generation efficiency, an important difference between a Brillouin generator and a Raman generator is that the Brillouin Stokes wave is generated only in the backward direction but the Raman Stokes is generated only in the forward direction.

As discussed above, for backward generation, the net result of the cumulative backward amplification of spontaneous emission over the entire length of interaction is equivalent to the injection of an effective backward-propagating Stokes signal $I_{\mathrm{S}}^{\mathrm{eff}}(l)$ at $z = l$. Considering the physical implication of the threshold amplification factor given in (9.197) for a Raman generator, the Brillouin threshold can be defined as the condition in which stimulated Brillouin amplification of the spontaneous Brillouin Stokes emission brings the Stokes intensity to the level of the pump intensity. Because the effective Stokes signal at $z = l$ is related to the pump intensity at $z = l$, we then have the following threshold amplification factor for a Brillouin generator:

$$G_{\mathrm{B}}^{\mathrm{th}} = \exp\left[\tilde{g}_{\mathrm{B}} I_{\mathrm{p}}^{\mathrm{th}}(0) l_{\mathrm{eff}}\right] = \frac{I_{\mathrm{p}}^{\mathrm{th}}(l)}{I_{\mathrm{S,th}}^{\mathrm{eff}}(l)}, \tag{9.208}$$

where $I_p^{th}(0)$ and $I_p^{th}(l)$ are the input pump intensity at $z = 0$ and the remaining pump intensity at $z = l$, respectively, at the threshold of the Brillouin generator. The value of G_B^{th} is a function of the characteristics of the medium and is generally larger than that of G_R^{th} for the same medium, primarily because of the fact that the Brillouin Stokes is generated in the backward direction. Therefore, the threshold pump intensity for Brillouin Stokes generation in a medium of length l is

$$I_p^{th}(0) \approx \frac{\ln G_B^{th}}{\tilde{g}_B l_{eff}}. \tag{9.209}$$

For example, $\ln G_B^{th} \approx 21$ for Brillouin Stokes generation in single-mode silica fibers.

The conversion efficiency of a Brillouin generator from the pump to the Stokes is measured in terms of an intensity reflectivity defined as

$$R_B = \frac{I_S(0)}{I_p(0)}. \tag{9.210}$$

In the case when $\alpha_S = \alpha_p = 0$, the value of R_B can be found from the following transcendental relation (see Problem 9.9.12):

$$R_B = \left(G_B^{th}\right)^{r(1-R_B)-1} \tag{9.211}$$

under the approximation that $\omega_p \approx \omega_S$ for Brillouin scattering in the optical region, where $r = I_p(0)/I_p^{th}(0)$. Because of the large value of G_B^{th}, the relation in (9.211) indicates a sharp threshold for Brillouin generation. Below the Brillouin threshold, $r < 1$, and R_B quickly approaches zero. Above the threshold, R_B varies with pump intensity approximately as

$$R_B \approx 1 - \frac{1}{r} = 1 - \frac{I_p^{th}(0)}{I_p(0)}, \qquad \text{for } r > 1. \tag{9.212}$$

This relation leads to the important conclusion that

$$I_p(l) = I_p^{th}(0) \quad \text{if} \quad I_p(0) > I_p^{th}(0). \tag{9.213}$$

Therefore, when the input pump intensity exceeds the threshold pump intensity of a lossless Brillouin generator, the transmitted intensity is clamped at the level of the threshold pump intensity. The excess above the threshold is converted to the Stokes frequency and is reflected back to the input end. This characteristic allows very efficient Brillouin generation, but it also sets a very important limitation on the level of optical power that can be transmitted through an optical system. In particular, in a fiber-optic transmission system, the generation of the Brillouin Stokes in the optical fiber severely limits the transmission power level of the system.

EXAMPLE 9.22 The fiber Brillouin amplifier described in Example 9.21 becomes a fiber Brillouin generator for a Stokes signal at $\lambda_S = 1.55$ μm without an input signal at this

wavelength if the pump power at $\lambda_p = 1.5499$ μm is raised above a threshold level. Find the threshold pump power for this fiber Brillouin generator if the linewidth of the pump is 100 MHz. What is the threshold pump power if the linewidth of the pump is only 1 MHz?

Solution Identifying $P_p^{th} = I_p^{th}(0)\mathcal{A}_{eff}$ and using $\ln G_B^{th} \approx 21$, we have, from (9.209),

$$P_p^{th} = \frac{\mathcal{A}_{eff}}{l_{eff}} \frac{\ln G_B^{th}}{\tilde{g}_B} = \frac{21\mathcal{A}_{eff}}{\tilde{g}_B l_{eff}} \tag{9.214}$$

for the threshold pump power of a fiber Brillouin generator. Using the parameters obtained in Example 9.21 with $\tilde{g}_B = 6.2 \times 10^{-12}$ m W^{-1} for a pump wave of 100 MHz linewidth, we find that

$$P_p^{th} = \frac{21 \times 5 \times 10^{-11}}{6.2 \times 10^{-12} \times 14.86 \times 10^3} \text{ W} = 11.4 \text{ mW}.$$

If the pump has a narrow linewidth of only 1 MHz, we have $\tilde{g}_B = 4.23 \times 10^{-11}$ m W^{-1} from Example 9.18. Then the threshold pump power is reduced to

$$P_p^{th} = \frac{21 \times 5 \times 10^{-11}}{4.23 \times 10^{-11} \times 14.86 \times 10^3} \text{ W} = 1.67 \text{ mW}.$$

The Brillouin threshold pump power can be increased substantially if the linewidth of the pump is large. Because the power that remains in the pump is clamped to the Brillouin threshold, with the rest reflected back to the input end, suppressing the Brillouin Stokes generation by sufficiently increasing the Brillouin threshold is essential for the operation of a Raman amplifier, as discussed in Example 9.19, as well as for the operation of a Raman generator.

Because Raman and Brillouin gains exist in the same medium and both Raman Stokes and Brillouin Stokes can grow from spontaneous emission, these two processes compete with each other for the same pump power source. The one that has a lower threshold pump intensity quickly monopolizes the pump power and prohibits the other from occuring. Because $\tilde{g}_{B0}$ is usually much larger than $\tilde{g}_{R0}$ in the same medium, Brillouin generation usually dominates although G_B^{th} is larger than G_R^{th}. However, because the Brillouin gain has a very narrow linewidth, the threshold pump intensity for Brillouin generation increases very quickly when the pump wave has a linewidth exceeding Δf_B. Therefore, Brillouin generation dominates only when the pump has a narrow linewidth, whereas Raman generation dominates when the linewidth of the pump is larger than the Brillouin linewidth. The pump power required for a Raman amplifier is generally lower than the threshold pump power of a Raman generator, and that for a Brillouin amplifier is lower than the threshold of a Brillouin generator. However, because the Brillouin gain factor can be a few orders of magnitude higher than the

Raman gain factor, stimulated Brillouin scattering can easily occur well below the power required for a Raman amplifier. The consequences of stimulated Brillouin scattering in a Raman amplifier include significant reduction of the Raman gain by depletion of the pump power, generation of noise, and distortion of the signal waveform. It is therefore necessary to suppress stimulated Brillouin scattering in a Raman amplifier by, for example, using a pump of a sufficiently broad linewidth.

Besides the amplifiers and the generators discussed above, the Raman gain can also be utilized to construct a Raman laser by placing such a gain medium in an optical oscillator that is in resonance with the Raman Stokes frequency. Similarly, a Brillouin laser can also be constructed by placing a Brillouin gain medium in an optical oscillator that is in resonance with the Brillouin Stokes frequency.

9.10 Nonlinear optical interactions in waveguides

As we have seen in preceding sections, the efficiency of a nonlinear optical interaction generally increases with the intensities of the interacting optical waves and the interaction length. In a homogeneous bulk medium, the intensity of an optical wave can be increased by tightening the focus of the beam to reduce its cross-sectional spot size, but often at the expense of reducing the effective interaction length due to an increase in the beam divergence as a result of the decrease in the beam spot size. In an optical waveguide, however, an optical wave is guided and confined to a small cross-sectional area for the entire length of the waveguide. Because of optical confinement, a guided optical wave can maintain a high intensity over a long distance that is practically limited only by the length and the attenuation coefficient of the waveguide. Therefore, both high intensity and long interaction length desired for efficient nonlinear optical interactions can be simultaneously fulfilled in an optical waveguide. For example, in a low-loss optical fiber, the effective interaction length is on the order of tens of kilometers and the optical intensity can be quite high at a modest power level because of the small core diameter of a typical single-mode fiber. This unique characteristic makes optical waveguides ideal media for efficient nonlinear optical devices.

Coupled-wave theory is used in the analysis of interactions among waves of different frequencies, including the acousto-optic interactions discussed in Chapter 8 and the nonlinear optical interactions discussed in preceding sections. In the analysis of the coupling of waveguide modes, however, coupled-mode theory has to be used. In general, both the interaction among different optical frequencies and the characteristics of the waveguide modes have to be considered for a nonlinear optical interaction in an optical waveguide. Therefore, a combination of coupled-wave and coupled-mode theories has to be employed in the analysis of such an interaction.

First, the total field of the interacting waves is expanded in terms of the fields of individual frequencies:

$$\mathbf{E}(\mathbf{r}, t) = \sum_q \mathbf{E}_q(\mathbf{r}) \exp(-\mathrm{i}\omega_q t), \tag{9.215}$$

where $\mathbf{E}_q(\mathbf{r})$ is the spatially dependent total field amplitude for the frequency ω_q. Then, instead of taking out a uniquely defined fast-varying spatial variation as done in (4.5) for the formulation of coupled-wave theory, we expand each field $\mathbf{E}_q(\mathbf{r})$ at a given frequency ω_q in terms of the waveguide modes:

$$\mathbf{E}_q(\mathbf{r}) = \sum_\nu A_{q,\nu}(z)\hat{\boldsymbol{\mathcal{E}}}_{q,\nu}(x, y) \exp(\mathrm{i}\beta_{q,\nu} z). \tag{9.216}$$

Note that the propagation constant $\beta_{q,\nu}$ is a function of both optical frequency ω_q and waveguide mode ν. Note also that this expansion is valid only when nonlinear polarization $\mathbf{P}^{(n)}$ is small compared to linear polarization $\mathbf{P}^{(1)}$ so that the waveguide modes defined by the linear optical properties of the medium remain a valid concept. Because $\hat{\boldsymbol{\mathcal{E}}}_{q,\nu}$ is the normalized mode field pattern defined in Section 2.4, the power contained in waveguide mode ν at frequency ω_q is simply given by

$$P_{q,\nu} = |A_{q,\nu}|^2 = A_{q,\nu} A^*_{q,\nu}. \tag{9.217}$$

Following the procedures used in formulating the coupled-mode equations discussed in Chapter 4 and allowing for any possible linear coupling besides nonlinear, we find the following coupled-mode equation that accounts for an nth-order nonlinear interaction in a waveguide structure:[6]

$$\pm\frac{\mathrm{d}A_{q,\nu}}{\mathrm{d}z} = \sum_\mu \mathrm{i}\kappa_{q,\nu\mu} A_{q,\mu} \mathrm{e}^{\mathrm{i}(\beta_{q,\mu}-\beta_{q,\nu})z} + \mathrm{i}\omega_q \mathrm{e}^{-\mathrm{i}\beta_{q,\nu} z} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}^*_{q,\nu} \cdot \mathbf{P}^{(n)}_q \mathrm{d}x\mathrm{d}y, \tag{9.218}$$

where the plus sign is taken for a forward-propagating mode with $\beta_{q,\nu} > 0$, and the minus sign is for a backward-propagating mode with $\beta_{q,\nu} < 0$. In summary, for nonlinear interactions in optical waveguides, the expansion in (9.216) replaces that in (9.53) and (9.218) replaces (9.55). Besides the nonlinear effect characterized by $\mathbf{P}^{(n)}$, linear effects such as a periodic grating or an externally applied voltage also modify the behavior of the waves in a waveguide and lead to coupling between different waveguide modes. Coupling between different waveguides in the presence of optical nonlinearity is also possible. The first term on the right-hand side of (9.218) accounts for the possibility

[6] To be precise, just like that of the linear coupling coefficient $\kappa_{\nu\mu}$, the form of the nonlinear term on the right-hand side of (9.218) also has to be modified when $\hat{\boldsymbol{\mathcal{E}}}_{q,\nu}$ represents nonorthogonal modes of different individual waveguides in a structure that consists of multiple waveguides. Such modification is normally not significant and is ignored here. No such approximation is incurred in the use of (9.218), however, if $\hat{\boldsymbol{\mathcal{E}}}_{q,\nu}$ represents the modes of the entire structure, which are orthogonal to each other. This is always true in the case of a single waveguide. It is also true if the supermodes of a structure that consists of multiple waveguides are used in the analysis.

of such linear coupling effects based on the coupled-mode formulation discussed in Section 4.2. Therefore, (9.218) can be viewed as an extension of (4.33) or (4.39) to include the nonlinear perturbation. In the case of coupling between different waveguides, $\kappa_{q,\nu\mu}$ still has to be evaluated using (4.40) due to the nonorthogonality between modes of different waveguides.

Many guided-wave nonlinear optical devices have direct bulk counterparts. The use of a waveguide for such a device offers the advantages of improved efficiency, phase matching, or miniaturization of the device but is not absolutely necessary for the device function. Guided-wave optical frequency converters typically fall into this category. Some nonlinear optical devices rely on the waveguide geometry for their functions and thus have no bulk counterparts. All-optical switches and modulators that use waveguide interferometers or waveguide couplers belong to this category. Sometimes, the use of an optical waveguide is necessary for the practical reason that only a waveguide can provide the long interaction length required for the function of a device though the basic function of the device does not depend on the waveguide geometry. Many nonlinear optical devices that use optical fibers belong to this category.

9.11 Guided-wave optical frequency converters

All of the optical frequency converters discussed in Section 9.6 can be made in waveguide structures. The basic principles and characteristics of these devices are the same as their bulk counterparts, except that the characteristics of the waveguide modes have to be considered. Though a guided-wave optical frequency converter generally takes the form of a single waveguide, there is often a possibility that multiple waveguide modes are involved in the frequency conversion process. Each individual frequency component can consist of multiple waveguide modes, as expressed by (9.216). Even when each frequency component is represented by only one waveguide mode, it is still possible for the different interacting frequency components to be in different waveguide modes.

For a parametric second-order process in a waveguide involving three different frequencies with $\omega_3 = \omega_1 + \omega_2$, we have

$$\hat{\boldsymbol{\mathcal{E}}}^*_{3,\nu} \cdot \mathbf{P}^{(2)}_3 = 2\epsilon_0 \sum_{\mu,\xi} \hat{\boldsymbol{\mathcal{E}}}^*_{3,\nu} \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{\boldsymbol{\mathcal{E}}}_{1,\mu} \hat{\boldsymbol{\mathcal{E}}}_{2,\xi} A_{1,\mu} A_{2,\xi} \mathrm{e}^{\mathrm{i}(\beta_{1,\mu}+\beta_{2,\xi})z} \tag{9.219}$$

to replace (9.56), and similar expressions for $\hat{\boldsymbol{\mathcal{E}}}^*_{1,\mu} \cdot \mathbf{P}^{(2)}_1$ and $\hat{\boldsymbol{\mathcal{E}}}^*_{2,\xi} \cdot \mathbf{P}^{(2)}_2$ to replace (9.57) and (9.58), respectively. The interacting waves in an efficient frequency converter normally propagate in the same direction though contradirectional geometry is also possible. Here we consider only codirectional geometry with all of the interacting waves

propagating in the forward direction. Then, using (9.218), we can write

$$\frac{dA_{3,\nu}}{dz} = i\omega_3 \sum_{\mu,\xi} C_{\nu\mu\xi} A_{1,\mu} A_{2,\xi} e^{i\Delta\beta_{\nu\mu\xi} z}, \tag{9.220}$$

$$\frac{dA_{1,\mu}}{dz} = i\omega_1 \sum_{\nu,\xi} C^*_{\nu\mu\xi} A_{3,\nu} A^*_{2,\xi} e^{-i\Delta\beta_{\nu\mu\xi} z}, \tag{9.221}$$

$$\frac{dA_{2,\xi}}{dz} = i\omega_2 \sum_{\nu,\mu} C^*_{\nu\mu\xi} A_{3,\nu} A^*_{1,\mu} e^{-i\Delta\beta_{\nu\mu\xi} z}, \tag{9.222}$$

where

$$\begin{aligned} C_{\nu\mu\xi} &= 2\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}^*_{3,\nu} \cdot \boldsymbol{\chi}^{(2)}(\omega_3 = \omega_1 + \omega_2) : \hat{\boldsymbol{\mathcal{E}}}_{1,\mu} \hat{\boldsymbol{\mathcal{E}}}_{2,\xi} dx dy \\ &= 4\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}^*_{3,\nu} \cdot \mathbf{d}(\omega_3 = \omega_1 + \omega_2) : \hat{\boldsymbol{\mathcal{E}}}_{1,\mu} \hat{\boldsymbol{\mathcal{E}}}_{2,\xi} dx dy \end{aligned} \tag{9.223}$$

is the *effective nonlinear coefficient* that accounts for the overlapping of the field distribution patterns of different waveguide modes, as well as for any possible spatial variations in $\boldsymbol{\chi}^{(2)}$ due to the waveguide structure, and

$$\Delta\beta_{\nu\mu\xi} = \beta_{1,\mu} + \beta_{2,\xi} - \beta_{3,\nu} \tag{9.224}$$

is the phase mismatch. In comparison to the effective nonlinear susceptibility, χ_{eff}, defined in (9.59) for the interaction of plane waves, the effective nonlinear coefficient defined above for the interaction of waveguide modes has the following relation:

$$|C_{\nu\mu\xi}|^2 = \frac{|\chi_{\text{eff}}|^2}{2c^3\epsilon_0 n_{3,\nu} n_{1,\mu} n_{2,\xi}} \frac{\Gamma_{\nu\mu\xi}}{\mathcal{A}} = \frac{2|d_{\text{eff}}|^2}{c^3\epsilon_0 n_{3,\nu} n_{1,\mu} n_{2,\xi}} \frac{\Gamma_{\nu\mu\xi}}{\mathcal{A}}, \tag{9.225}$$

where $n_{q,\nu} = c\beta_{q,\nu}/\omega_q$ is the effective refractive index of a waveguide mode, $\Gamma_{\nu\mu\xi}$ is the *overlap factor* for the interacting waveguide modes, and $\mathcal{A}$ is the cross-sectional area of the waveguide core. The overlap factor $\Gamma_{\nu\mu\xi}$ accounts for the differences in the mode field distributions among the interacting waves and any transverse spatial variations in the nonlinear susceptibility. An effective area for the interaction can be defined as $\mathcal{A}_{\text{eff}} = \mathcal{A}/\Gamma_{\nu\mu\xi}$.

In the case of second-harmonic generation in a waveguide, we have

$$\mathbf{P}^{(2)}_{2\omega} = \epsilon_0 \sum_{\mu,\xi} \boldsymbol{\chi}^{(2)}(2\omega = \omega + \omega) : \hat{\boldsymbol{\mathcal{E}}}_{\omega,\mu} \hat{\boldsymbol{\mathcal{E}}}_{\omega,\xi} A_{\omega,\mu} A_{\omega,\xi} e^{i(\beta_{\omega,\mu} + \beta_{\omega,\xi})z}, \tag{9.226}$$

$$\mathbf{P}^{(2)}_{\omega} = 2\epsilon_0 \sum_{\nu,\xi} \boldsymbol{\chi}^{(2)}(\omega = 2\omega - \omega) : \hat{\boldsymbol{\mathcal{E}}}_{2\omega,\nu} \hat{\boldsymbol{\mathcal{E}}}^*_{\omega,\xi} A_{2\omega,\nu} A^*_{\omega,\xi} e^{i(\beta_{2\omega,\nu} - \beta_{\omega,\xi})z}. \tag{9.227}$$

Therefore, the coupled equations for second-harmonic generation in a waveguide are

$$\frac{\mathrm{d}A_{2\omega,\nu}}{\mathrm{d}z} = \mathrm{i}\omega \sum_{\mu,\xi} C_{\nu\mu\xi} A_{\omega,\mu} A_{\omega,\xi} \mathrm{e}^{\mathrm{i}\Delta\beta_{\nu\mu\xi} z}, \tag{9.228}$$

$$\frac{\mathrm{d}A_{\omega,\mu}}{\mathrm{d}z} = \mathrm{i}\omega \sum_{\nu,\xi} C^*_{\nu\mu\xi} A_{2\omega,\nu} A^*_{\omega,\xi} \mathrm{e}^{-\mathrm{i}\Delta\beta_{\nu\mu\xi} z}, \tag{9.229}$$

where

$$\begin{aligned} C_{\nu\mu\xi} &= 2\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}^*_{2\omega,\nu} \cdot \boldsymbol{\chi}^{(2)}(2\omega = \omega + \omega) : \hat{\boldsymbol{\mathcal{E}}}_{\omega,\mu} \hat{\boldsymbol{\mathcal{E}}}_{\omega,\xi} \mathrm{d}x\mathrm{d}y \\ &= 4\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}^*_{2\omega,\nu} \cdot \mathbf{d}(2\omega = \omega + \omega) : \hat{\boldsymbol{\mathcal{E}}}_{\omega,\mu} \hat{\boldsymbol{\mathcal{E}}}_{\omega,\xi} \mathrm{d}x\mathrm{d}y \end{aligned} \tag{9.230}$$

and

$$\Delta\beta_{\nu\mu\xi} = \beta_{\omega,\mu} + \beta_{\omega,\xi} - \beta_{2\omega,\nu}. \tag{9.231}$$

All of the general concepts discussed in Section 9.4 for parametric second-order interactions are equally valid for such interactions in an optical waveguide, except that the form of each relation must be modified to factor in the characteristics of the waveguide modes. For instance, Manley–Rowe relations still exist but such relations have to be expressed in terms of optical powers in the waveguide modes (see Problem 9.11.1). Phase matching is also most important for an efficient interaction, but it is now determined by the propagation constants of the interacting waveguide modes. Therefore, the coherence length for the coupling of the mode fields $\boldsymbol{\mathcal{E}}_{3,\nu}$, $\boldsymbol{\mathcal{E}}_{1,\mu}$, and $\boldsymbol{\mathcal{E}}_{2,\xi}$ is

$$l_{\mathrm{coh}} = \frac{\pi}{|\Delta\beta_{\nu\mu\xi}|}. \tag{9.232}$$

Because the propagation constant $\beta_{q,\nu}$ for a given frequency ω_q is mode dependent due to modal dispersion, the phase mismatch and, consequently, the efficiency of an interaction are dependent on specific combination of the modes among the interacting frequency components. In a multimode waveguide, there can be many different mode combinations, as is indicated by the summation over the mode indices on the right-hand side of the coupled equations in (9.220)–(9.222) and those in (9.228) and (9.229). However, it is unlikely and undesirable, though not impossible, for multiple mode combinations to be simultaneously phase matched in a particular interaction. In a practical optical frequency converter, normally only one waveguide mode for each frequency component is efficiently coupled to other frequency components in the interaction.

When an interaction involves only one waveguide mode in each frequency component, the coupled equations have the form of those for the corresponding interaction in a bulk medium though the mode amplitudes are used instead of the field amplitudes and the coefficients in the equations look different. Then, the characteristics of

any guided-wave optical frequency converter can be obtained by converting those of its bulk counterpart described in Section 9.6 with the following modifications: (1) the mode power $P_{q,\nu}$ is used in place of $I_q \mathcal{A}_q$; (2) $\Delta\beta$ is used instead of Δk; (3) the nonlinear coefficient $C_{\nu\mu\xi}$ is used in place of χ_{eff} by replacing a compound coefficient of the form on the right-hand side of (9.225) in any relation for a bulk device with $|C_{\nu\mu\xi}|^2$ for a guided-wave device.

For example, consider a second-harmonic generator in which the fundamental and second-harmonic waves each contain only one mode. Then, the coupled equations in (9.228) and (9.229) are reduced to

$$\frac{\mathrm{d}A_{2\omega,\nu}}{\mathrm{d}z} = \mathrm{i}\omega C A_{\omega,\mu}^2 \mathrm{e}^{\mathrm{i}\Delta\beta z}, \tag{9.233}$$

$$\frac{\mathrm{d}A_{\omega,\mu}}{\mathrm{d}z} = \mathrm{i}\omega C^* A_{2\omega,\nu} A_{\omega,\mu}^* \mathrm{e}^{-\mathrm{i}\Delta\beta z}, \tag{9.234}$$

where $C = C_{\nu\mu\mu}$ and $\Delta\beta = 2\beta_{\omega,\mu} - \beta_{2\omega,\nu}$. In the low-efficiency limit when depletion of power in the fundamental wave is negligible, we can obtain, by integrating (9.233) or by converting (9.110), the following relation for a waveguide of length l:

$$P_{2\omega,\nu}(l) = \omega^2 |C|^2 P_{\omega,\mu}^2 l^2 \frac{\sin^2(\Delta\beta l/2)}{(\Delta\beta l/2)^2} = \frac{4\pi^2 c^2}{\lambda^2} |C|^2 P_{\omega,\mu}^2 l^2 \frac{\sin^2(\Delta\beta l/2)}{(\Delta\beta l/2)^2}. \tag{9.235}$$

In the high-efficiency limit with perfect phase matching, we have

$$P_{2\omega,\nu}(l) = P_{\omega,\mu}(0) \tanh^2 \kappa l, \tag{9.236}$$

$$P_{\omega,\mu}(l) = P_{\omega,\mu}(0) \operatorname{sech}^2 \kappa l, \tag{9.237}$$

with the coefficient κ given by

$$\kappa = \left[\omega^2 |C|^2 P_{\omega,\mu}(0)\right]^{1/2} = \left[\frac{4\pi^2 c^2}{\lambda^2} |C|^2 P_{\omega,\mu}(0)\right]^{1/2}. \tag{9.238}$$

The techniques for phase matching discussed in Section 9.5 are also applicable to guided-wave devices. Besides, the modal dispersion in a waveguide can also be used for phase matching if modes of different orders are involved in an interaction. Often, a combination of different techniques is employed. For example, a waveguide is fabricated along a certain direction in a crystal for collinear birefringent phase matching, but temperature is used for fine tuning once the wave propagation direction is fixed by the waveguide structure. Quasi-phase matching is particularly useful for guided-wave devices because of its advantages discussed in Section 9.5 and because of its compatibility with microfabrication technology. For a guided-wave device that is quasi-phase matched using a periodic structure with a duty factor ξ, the coupled equations can be transformed in a manner similar to the transformation shown in (9.96), resulting in a phase mismatch of $\Delta\beta_{\text{Q}} = \Delta\beta + qK$ that is minimized with a particular integer q

and a nonlinear coefficient C_Q given by

$$C_Q = 2C \frac{\sin \xi q\pi}{q\pi} e^{-i\xi q\pi} \tag{9.239}$$

according to (9.100). With quasi-phase matching, we have to replace the phase mismatch $\Delta\beta$ in (9.235) by $\Delta\beta_Q$, and the parameter C in (9.235) and (9.238) by C_Q. For a first-order structure with a 50% duty factor, $|C_Q| = 2|C|/\pi$. In a slab waveguide, the fanned structure shown in Fig. 9.14(c) can also be used for continuous wavelength tuning through quasi-phase matching.

EXAMPLE 9.23 A PPLN waveguide is used for second-harmonic generation of a fundamental wave at 1.10 μm wavelength. The waveguide is a diffused channel waveguide formed by Ti diffusion into a PPLN crystal similar to the one described in Example 9.11. It has a diffusion depth of $d = 2$ μm and a width of $w = 3$ μm, for an effective waveguide core area of $\mathcal{A} = wd = 6$ μm^2. It is a single-mode waveguide for the fundamental wavelength at 1.1 μm. The overlap factor for second-harmonic generation at this wavelength in this waveguide is $\Gamma = 0.4$. The grating period of the PPLN is properly chosen as a first-order grating with a 50% duty factor for quasi-phase matching of the interacting waveguide modes. For easy comparison to second-harmonic generation in the bulk PPLN crystal considered in Example 9.12(d), we take the waveguide length to be $l = 1$ cm. Find the normalized second-harmonic conversion efficiency for this device.

Solution By replacing $\Delta\beta$ with $\Delta\beta_Q$ and C with C_Q in (9.235), we have the following normalized efficiency in the low-efficiency limit for the PPLN waveguide:

$$\hat{\eta}_{SH} = \frac{P_{2\omega}(l)}{P_\omega^2} = \frac{4\pi^2 c^2}{\lambda^2} |C_Q|^2 l^2 \frac{\sin^2(\Delta\beta_Q l/2)}{(\Delta\beta_Q l/2)^2}. \tag{9.240}$$

For perfect quasi-phase matching with a first-order grating that has a 50% duty factor, we have

$$\hat{\eta}_{SH} = \frac{4\pi^2 c^2}{\lambda^2} |C_Q|^2 l^2 = \frac{32|d_{eff}|^2}{c\epsilon_0 n_\omega^2 n_{2\omega} \lambda^2} \frac{\Gamma}{\mathcal{A}} l^2, \tag{9.241}$$

where n_ω and $n_{2\omega}$ are the effective refractive index, n_β, of the waveguide modes at the fundamental and second-harmonic frequencies, respectively. Because the index change created by Ti diffusion is very small, typically on the order of 0.5%, we can simply take the refractive index of the bulk PPLN as a very good approximation for the effective refractive index of a waveguide mode. From Example 9.11, we have $d_{eff} = d_{33} = -25.2$ pm V^{-1}, $n_\omega = n_\omega^e = 2.1536$, and $n_{2\omega} = n_{2\omega}^e = 2.2260$. We then

find the following normalized conversion efficiency:

$$\hat{\eta}_{\text{SH}} = \frac{32 \times (25.2 \times 10^{-12})^2 \times 0.4 \times (1 \times 10^{-2})^2}{3 \times 10^8 \times 8.85 \times 10^{-12} \times (2.1536)^2 \times 2.226 \times (1.1 \times 10^{-6})^2 \times 6 \times 10^{-12}} \text{ W}^{-1}$$
$$= 409\% \text{ W}^{-1}.$$

This normalized conversion efficiency for the PPLN waveguide is 159 times that obtained in Example 9.12(d) for the bulk PPLN. Further increase in the efficiency is possible by increasing the length of the waveguide. In the waveguide device, the efficiency continues to increase quadratically with length l because the optical waves remain confined as the waveguide length is increased. In a bulk device, the best efficiency only increases linearly with length l, as seen in (9.121), because of the limitation imposed by diffraction (see Problem 9.11.2).

Note that (9.240) and (9.241) are valid only in the low-efficiency limit. Clearly, $\hat{\eta}_{\text{SH}} = 409\% \text{ W}^{-1}$ obtained in this example does not mean that it is possible to obtain an unphysical efficiency of 409% by launching a fundamental beam of 1 W power into the waveguide. Nor does it mean that a conversion efficiency of 100% is obtained by launching a fundamental beam of 244 mW into the waveguide. It only means that a very low input power of the fundamental wave is needed to obtain a decent conversion efficiency. For example, an input fundamental power of only $P_\omega = 24.4$ mW is required to have an output second-harmonic power of $P_{2\omega} = 2.44$ mW for a conversion efficiency of 10%. A conversion efficiency approaching 100% is theoretically possible, but with an input fundamental power found by using the relation in (9.236) for the high-efficiency limit (see Problem 9.11.3).

9.12 Guided-wave all-optical modulators and switches

As discussed in Section 9.7, an all-optical modulator can be either of refractive type, which utilizes $\chi^{(3)\prime}$, or of absorptive type, which utilizes $\chi^{(3)\prime\prime}$. For a guided-wave nonlinear optical device, however, any absorptive loss in the waveguide is detrimental to the device function due to the fact that the primary advantage of using an optical waveguide for the device is the long interaction length made possible by the waveguiding effect. Therefore, all practical guided-wave all-optical modulators and switches are of refractive type based on the optical Kerr effect. The majority of such devices require only one optical frequency for their operation though some involve two or more frequencies at a time. For a guided-wave all-optical modulator or switch that requires only one frequency at a time for its operation, we have

$$\mathbf{P}^{(3)} = 3\epsilon_0 \sum_{\mu,\xi,\zeta} \chi^{(3)}(\omega = \omega + \omega - \omega) \vdots \hat{\boldsymbol{\varepsilon}}_\mu \hat{\boldsymbol{\varepsilon}}_\xi \hat{\boldsymbol{\varepsilon}}_\zeta^* A_\mu A_\xi A_\zeta^* \mathrm{e}^{\mathrm{i}(\beta_\mu + \beta_\xi - \beta_\zeta)z}. \tag{9.242}$$

According to (9.218), we have the following general coupled-mode equation for such a device:

$$\pm\frac{\mathrm{d}A_\nu}{\mathrm{d}z} = \sum_\mu \mathrm{i}\kappa_{\nu\mu}A_\mu \mathrm{e}^{\mathrm{i}(\beta_\mu-\beta_\nu)z} + \sum_{\mu,\xi,\zeta} \mathrm{i}\omega C_{\nu\mu\xi\zeta}A_\mu A_\xi A_\zeta^* \mathrm{e}^{\mathrm{i}(\beta_\mu+\beta_\xi-\beta_\zeta-\beta_\nu)z}, \tag{9.243}$$

where $\kappa_{\nu\mu}$ is the linear coupling coefficient defined in Section 4.2, subject to any modifications such as those caused by the electro-optic, magneto-optic, or acousto-optic effects discussed in earlier chapters, and $C_{\nu\mu\xi\zeta}$ is the nonlinear coefficient given by

$$C_{\nu\mu\xi\zeta} = 3\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\nu^* \cdot \boldsymbol{\chi}^{(3)} \vdots \hat{\boldsymbol{\mathcal{E}}}_\mu \hat{\boldsymbol{\mathcal{E}}}_\xi \hat{\boldsymbol{\mathcal{E}}}_\zeta^* \mathrm{d}x\mathrm{d}y. \tag{9.244}$$

Self-phase modulation

In the simplest situation when a waveguide mode $\boldsymbol{\mathcal{E}}_\nu$ at a particular optical frequency ω is not coupled to any other frequencies or any other modes at the same frequency, (9.243) reduces to

$$\frac{\mathrm{d}A_\nu}{\mathrm{d}z} = \mathrm{i}\sigma_{\nu\nu}A_\nu|A_\nu|^2, \tag{9.245}$$

where

$$\sigma_{\nu\nu} = \omega C_{\nu\nu\nu\nu} = 3\omega\epsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\boldsymbol{\mathcal{E}}}_\nu^* \cdot \boldsymbol{\chi}^{(3)} \vdots \hat{\boldsymbol{\mathcal{E}}}_\nu \hat{\boldsymbol{\mathcal{E}}}_\nu \hat{\boldsymbol{\mathcal{E}}}_\nu^* \mathrm{d}x\mathrm{d}y. \tag{9.246}$$

Because $\boldsymbol{\chi}^{(3)}$ is real for a device based on the purely refractive optical Kerr effect, the nonlinear coefficient $\sigma_{\nu\nu}$ is also a real quantity. It is then clear from (9.245) that only the phase, but not the magnitude, of A_ν varies with z. Therefore, the mode power $P_\nu = |A_\nu|^2$ is a constant that is independent of z. The solution of (9.245) can be easily obtained:

$$A_\nu(z) = A_\nu(0)\exp(\mathrm{i}\sigma_{\nu\nu}P_\nu z) = A_\nu(0)\exp(\mathrm{i}\beta_\nu^{\mathrm{NL}} z), \tag{9.247}$$

where $\beta_\nu^{\mathrm{NL}} = \sigma_{\nu\nu}P_\nu$ is a power-dependent modification on the propagation constant. Clearly, the consequence of the optical Kerr effect on an individual waveguide mode is an effective propagation constant that is a function of the mode power:

$$\beta_\nu^{\mathrm{eff}} = \beta_\nu + \beta_\nu^{\mathrm{NL}} = \beta_\nu + \sigma_{\nu\nu}P_\nu, \tag{9.248}$$

where β_ν is the power-independent linear propagation constant of the mode. This effect leads to the following self-phase modulation for the mode field over a distance l in the

waveguide:

$$\varphi_\nu^{\mathrm{NL}} = \beta_\nu^{\mathrm{NL}} l = \sigma_{\nu\nu} P_\nu l = \sigma_{\nu\nu} |A_\nu|^2 l, \tag{9.249}$$

which is linearly dependent on the mode power.

For a waveguide that is fabricated in an isotropic medium, such as silica glass, $\chi^{(3)}_{xxxx} = \chi^{(3)}_{yyyy} = \chi^{(3)}_{zzzz} = \chi^{(3)}_{1111}$. Then, $\sigma_{\nu\nu}$ defined in (9.246) becomes

$$\sigma_{\nu\nu} = 3\omega\epsilon_0 \chi^{(3)}_{1111} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{\boldsymbol{\mathcal{E}}}_\nu|^4 \mathrm{d}x\mathrm{d}y. \tag{9.250}$$

It is then convenient to define an effective area for an individual waveguide mode in a third-order nonlinear process as

$$\mathcal{A}_\nu^{\mathrm{eff}} = \frac{\left[\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{\boldsymbol{\mathcal{E}}}_\nu|^2 \mathrm{d}x\mathrm{d}y\right]^2}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{\boldsymbol{\mathcal{E}}}_\nu|^4 \mathrm{d}x\mathrm{d}y} = \left(\frac{\omega\mu_0}{2\beta_\nu}\right)^2 \frac{1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{\boldsymbol{\mathcal{E}}}_\nu|^4 \mathrm{d}x\mathrm{d}y}, \tag{9.251}$$

where the orthonormality relation given in (2.44) is used.[7] Then, by using (9.248), (9.250), and (9.251), we find that the power-dependent effective index of the waveguide mode can be written

$$n_\nu^{\mathrm{eff}} = n_\nu + n_{2\nu} \frac{P_\nu}{\mathcal{A}_\nu^{\mathrm{eff}}}, \tag{9.252}$$

where $n_\nu^{\mathrm{eff}} = c\beta_\nu^{\mathrm{eff}}/\omega$, $n_\nu = c\beta_\nu/\omega$, and

$$n_{2\nu} = \frac{3\chi^{(3)}_{1111}}{4c\epsilon_0 n_\nu^2}. \tag{9.253}$$

Clearly, (9.252) has the form of (9.49), and (9.253) has the form of (9.50). Then self-phase modulation for the mode field can be expressed in the form of (9.143) as

$$\varphi_\nu^{\mathrm{NL}} = \beta_\nu^{\mathrm{NL}} l = \frac{\omega}{c} n_{2\nu} \frac{P_\nu}{\mathcal{A}_\nu^{\mathrm{eff}}} l = \frac{2\pi n_{2\nu}}{\lambda} \frac{P_\nu}{\mathcal{A}_\nu^{\mathrm{eff}}} l. \tag{9.254}$$

Note that (9.253) is valid only for a mode in a waveguide fabricated in a noncrystalline isotropic medium. For a mode in a waveguide based on a crystalline material, such as GaAs or $LiNbO_3$, (9.254) can still be used, but (9.253) is generally not valid because the nonlinear refractive index $n_{2\nu}$ in this situation is a function of the mode field polarization direction with respect to the principal axes of the crystal.

[7] The orthonormality relation in (2.44) is strictly accurate for TE modes only. It is used here as an approximation for other types of modes. In a weakly guiding waveguide, it is a good approximation.

Two-mode interaction

Guided-wave all-optical modulators and switches function on the same basic principle as guided-wave electro-optic modulators and switches by transforming a differential phase shift between two waveguide modes into an amplitude modulation, except that the required phase shift is controlled by the optical power in the waveguide structure rather than by an externally applied voltage. There are two basic approaches to transforming a differential phase shift into an amplitude modulation: by interference or by phase-sensitive coupling.

The operation of most devices involves only two modes of either the same waveguide or two separate waveguides. For a device that functions on two waveguide modes a and b at the same frequency ω, the total field is simply $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}) \exp(-\mathrm{i}\omega t)$ with

$$\mathbf{E}(\mathbf{r}) = A(z)\hat{\boldsymbol{\mathcal{E}}}_a(x, y)\mathrm{e}^{\mathrm{i}\beta_a z} + B(z)\hat{\boldsymbol{\mathcal{E}}}_b(x, y)\mathrm{e}^{\mathrm{i}\beta_b z}, \tag{9.255}$$

which has the same form as that given in (4.25). By using (9.243) and by keeping only the major nonlinear terms, the coupled-mode equations for such a two-mode device can be written as

$$\pm\frac{\mathrm{d}A}{\mathrm{d}z} = \mathrm{i}\kappa_{aa}A + \mathrm{i}\kappa_{ab}B\mathrm{e}^{\mathrm{i}(\beta_b-\beta_a)z} + \mathrm{i}\sigma_{aa}|A|^2A + \text{nonlinear cross terms}, \tag{9.256}$$

$$\pm\frac{\mathrm{d}B}{\mathrm{d}z} = \mathrm{i}\kappa_{bb}B + \mathrm{i}\kappa_{ba}A\mathrm{e}^{\mathrm{i}(\beta_a-\beta_b)z} + \mathrm{i}\sigma_{bb}|B|^2B + \text{nonlinear cross terms}, \tag{9.257}$$

where $\sigma_{aa} = \omega C_{aaaa}$ and $\sigma_{bb} = \omega C_{bbbb}$, as defined in (9.246). In general, $\sigma_{aa} \neq \sigma_{bb}$. The nonlinear cross terms are those that represent direct nonlinear coupling between the two modes with the nonlinear coefficients C_{aaab}, C_{aaba}, C_{abaa}, C_{aabb}, C_{abab}, C_{abba}, and C_{abbb} for (9.256) and C_{bbba}, C_{bbab}, C_{babb}, C_{bbaa}, C_{baba}, C_{baab}, and C_{baaa} for (9.257). Such nonlinear cross terms are generally much smaller than the direct nonlinear terms characterized by σ_{aa} and σ_{bb}, which are explicitly expressed in the above coupled equations.

In a device that is based solely on interference, $\kappa_{ab} = \kappa_{ba} = 0$, and the nonlinear cross terms vanish also. Therefore, there is generally no direct power exchange between the two modes. The nonlinear differential phase shift between the two modes controls the interference condition, thus turning an optical-power-dependent phase change into an amplitude modulation or switching. In a device that is based on coupling, $\kappa_{ab} \neq 0$ and $\kappa_{ba} \neq 0$. The function of modulation or switching is then a result of direct exchange of power between the two modes. In such a device, the power-dependent differential phase shift controls the effective coupling coefficient between the two modes through its influence on the phase matching between them.

Nonlinear optical mode mixers

A nonlinear mode mixer is a simple all-optical switch based on the power-dependent interference effect between two modes in a multimode waveguide, such as the TE_0 and TE_1 modes of a slab waveguide. Two different modes in an unperturbed waveguide are orthogonal to each other in the absence of nonlinear effects. Even when the optical Kerr effect is present, significant direct coupling between them occurs only when the power in the waveguide reaches a critical level. Below this critical power level, the optical Kerr effect leads to sufficient self-phase modulation in each individual mode but no significant cross-phase modulation or power exchange between the mutually orthogonal modes. For a nonlinear two-mode mixer operating in this regime, $\kappa_{aa} = \kappa_{bb} = \kappa_{ab} = \kappa_{ba} = 0$, and the nonlinear cross interaction between the two modes can also be neglected. Consequently, both (9.256) and (9.257) reduce to the form of (9.245) with the solution given in (9.247). Therefore, the total field in the two-mode mixer is

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= A(0)\hat{\boldsymbol{\mathcal{E}}}_a(x, y)\mathrm{e}^{\mathrm{i}\beta_a^{\mathrm{eff}}z} + B(0)\hat{\boldsymbol{\mathcal{E}}}_b(x, y)\mathrm{e}^{\mathrm{i}\beta_b^{\mathrm{eff}}z} \\ &= \left[A(0)\hat{\boldsymbol{\mathcal{E}}}_a(x, y) + B(0)\hat{\boldsymbol{\mathcal{E}}}_b(x, y)\mathrm{e}^{\mathrm{i}(\beta_b^{\mathrm{eff}}-\beta_a^{\mathrm{eff}})z}\right]\mathrm{e}^{\mathrm{i}\beta_a^{\mathrm{eff}}z}. \end{aligned} \tag{9.258}$$

For a mode mixer of a length l, the total differential phase shift between the two modes over the length of the device is

$$\Delta\varphi = (\beta_b^{\mathrm{eff}} - \beta_a^{\mathrm{eff}})l = \Delta\varphi_{\mathrm{L}} + \Delta\varphi_{\mathrm{NL}}, \tag{9.259}$$

where $\Delta\varphi_{\mathrm{L}} = (\beta_b - \beta_a)l$ is the linear differential phase shift due to modal dispersion and $\Delta\varphi_{\mathrm{NL}} = (\beta_b^{\mathrm{NL}} - \beta_a^{\mathrm{NL}})l$ is the nonlinear differential phase shift due to the difference in the self-phase modulation of the two different modes. For a given device of a fixed length, the value of $\Delta\varphi_{\mathrm{L}}$ is fixed, but that of $\Delta\varphi_{\mathrm{NL}}$ varies with the powers in the modes. Therefore, $\Delta\varphi$ can be controlled by the power coupled into the waveguide. Even when that power is evenly divided between the two modes, there is still a power-dependent differential phase shift between the two modes because the self-phase modulation expressed by (9.254) for a waveguide mode is also a function of the mode-dependent effective area $\mathcal{A}_v^{\mathrm{eff}}$.

Figure 9.33 illustrates the principle of a two-mode mixer. In this example, the power launched into the waveguide is equally divided between the two modes so that $P_a = P_b = P/2$ and $A(0) = B(0)$ at the input end. Therefore, the total field is asymmetrically distributed with its peak on one side of the waveguide. The linear differential phase shift in this example is $\Delta\varphi_{\mathrm{L}} = 2n\pi$, where n is an integer. At low power levels when the power-dependent nonlinear differential phase shift $\Delta\varphi_{\mathrm{NL}}$ is negligibly small, the field distribution at the output end is the same as that at the input end, as shown in Fig. 9.33(a). At a power level of P_π when $\Delta\varphi_{\mathrm{NL}} = \pi$, the total differential phase shift is $\Delta\varphi = (2n + 1)\pi$. Then, at the output end the peak of the total field is switched to

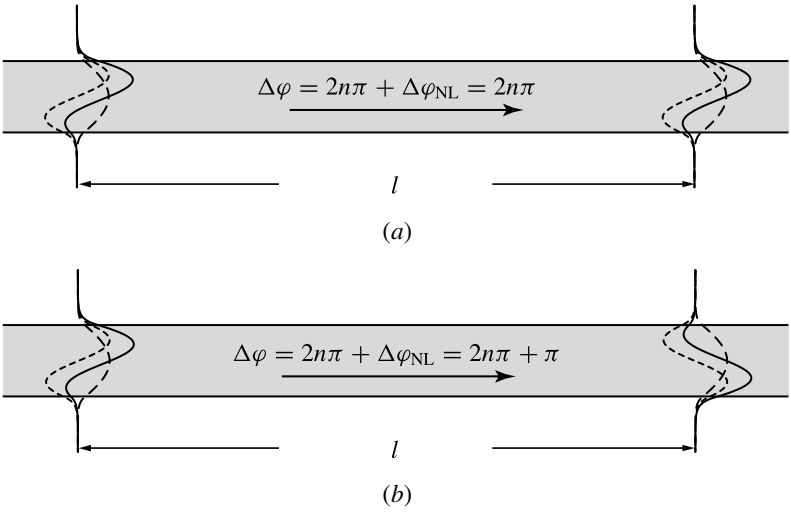


Figure 9.33 Power-dependent field distribution characteristics of a nonlinear mode mixer with a linear differential phase shift of $2n\pi$ (a) at low power levels when the nonlinear phase shift is negligible and (b) at a power level P_π when the nonlinear differential phase shift is π. The long dashed and short dashed curves respectively show fields of two individual modes, and the solid curve represents the total field of the two modes.

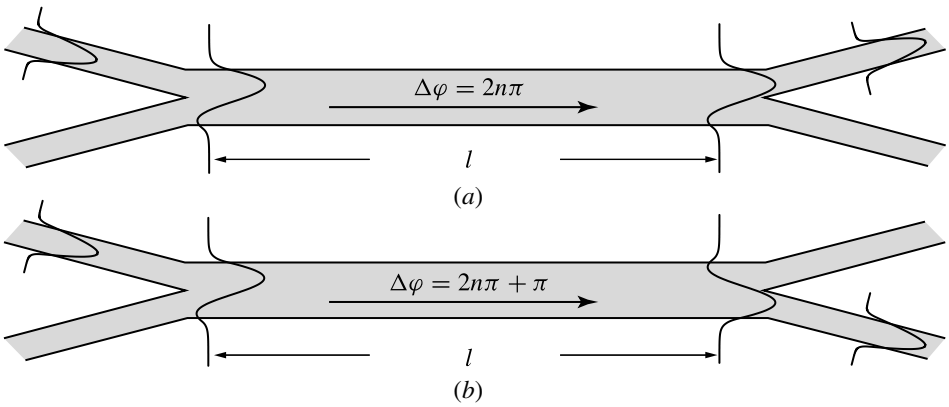


Figure 9.34 Mode mixer with Y-junction waveguides at its input and output ends for all-optical switching between separate waveguides.

the other side of the waveguide, as shown in Fig. 9.33(b). In this manner, a nonlinear mode mixer functions as a power-dependent all-optical switch.

A nonlinear mode mixer can take the form of a two-mode slab waveguide or that of a two-mode channel waveguide. In the latter case, both input and output ends of the mode mixer can be connected to Y-junction waveguides for all-optical switching of optical power between separate waveguides, as shown in Fig. 9.34. Such a device also functions as a *nonlinear mode sorter*.

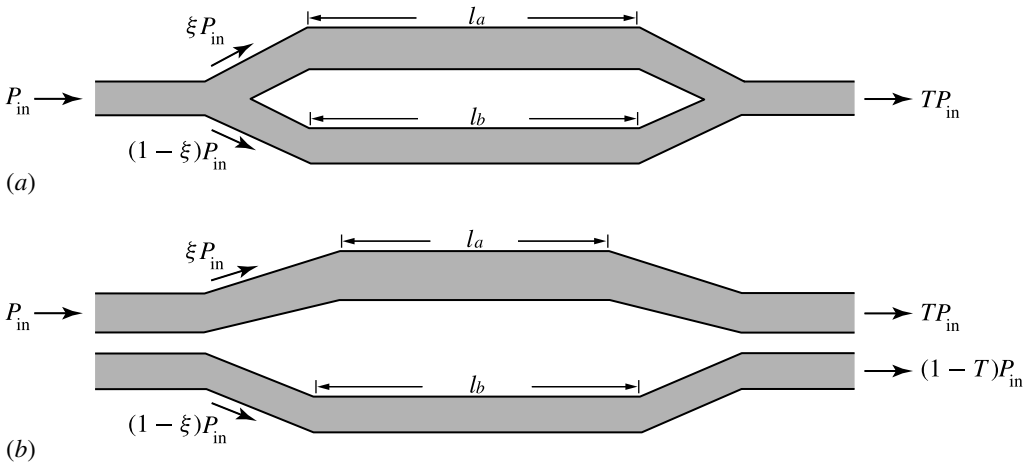


Figure 9.35 Single-input, all-optical Mach–Zehnder interferometers (*a*) using two Y-junction waveguides and (*b*) using two directional couplers for beam splitting at the input end and beam combining at the output end. The two arms of an all-optical interferometer are generally not balanced.

All-optical Mach–Zehnder interferometers

A nonlinear mode mixer functions as a nonlinear interferometric device only when the optical power in the waveguide is kept below the critical power level to prevent direct coupling of the modes. This limitation is caused by the fact that the two modes overlap in space while propagating codirectionally. It can be avoided in a nonlinear interferometer that consists of two separate arms such as one in the form of a Mach–Zehnder interferometer as shown in Fig. 9.35. There are a few significant differences between a nonlinear Mach–Zehnder interferometer and a nonlinear mode mixer: (1) both arms of the interferometer are generally single-mode waveguides; (2) at any power level, there is no cross modulation between the fields in the two separate arms of the interferometer; (3) the two fields that are combined at the output end of the interferometer can experience different propagation distances because the lengths of the two arms do not have to be the same.

An all-optical Mach–Zehnder interferometer is based on the same principle as the electro-optic Mach–Zehnder interferometer discussed in Section 6.4 except that the differential phase shift $\Delta\varphi$ between its two arms is controlled by the optical power rather than by an applied electric field. An all-optical Mach–Zehnder interferometer can have a single input channel, as shown in Fig. 9.35, or three input channels, as shown in Fig. 9.36.

Figure 9.35 shows two possible structures of single-input, all-optical Mach–Zehnder interferometers. The beam-splitting and beam-combining couplers at the input and output ends, respectively, of an all-optical Mach–Zehnder interferometer can be either Y-junction waveguides, as shown in Fig. 9.35(*a*), or directional couplers, as shown

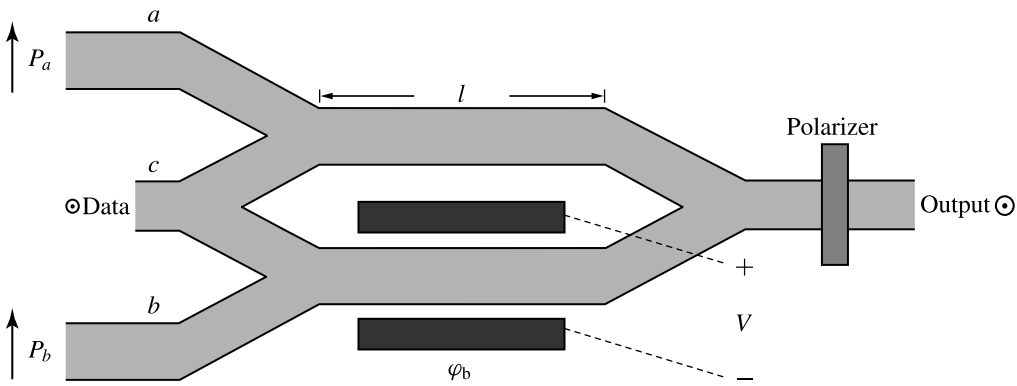


Figure 9.36 Three-input, symmetric all-optical Mach–Zehnder interferometer using Y-junction waveguides. The bias voltage V can provide a bias phase difference of φ_b between the two arms.

in Fig. 9.35(b). The two arms of an all-optical interferometer are not required to be identical. Therefore, in general, the linear differential phase shift is $\Delta\varphi_L = \beta_b l_b - \beta_a l_a$, and the nonlinear differential phase shift can be expressed as

$$\Delta\varphi_{NL} = \beta_b^{NL} l_b - \beta_a^{NL} l_a = \frac{2\pi}{\lambda}\left(n_{2b}\frac{P_b}{\mathcal{A}_b^{eff}} l_b - n_{2a}\frac{P_a}{\mathcal{A}_a^{eff}} l_a\right), \tag{9.260}$$

where l_a and l_b are the lengths of the two separate arms, respectively. We see that a power-dependent nonlinear differential phase shift can be obtained only when the two arms are not balanced, due to unbalanced excitation or physical asymmetry between them. With unbalanced excitation, $P_a \neq P_b$. Physical asymmetry exists when the waveguides that form the two arms have different lengths, $l_a \neq l_b$, or different effective areas, $\mathcal{A}_a^{eff} \neq \mathcal{A}_b^{eff}$, or different values of nonlinearity, $n_{2a} \neq n_{2b}$, or any combination of them.

To facilitate the possibility of unbalanced excitation, the Y-junction waveguides or directional couplers used in an all-optical Mach–Zehnder interferometer are not necessarily 3-dB couplers. For a given device, however, the beam-splitting coupler at the input end and the beam-combining coupler at the output end are usually identical couplers with a fixed power-splitting ratio of $\xi : (1 - \xi)$ between the two arms, as also shown in Fig. 9.35. For an all-optical Mach–Zehnder interferometer that uses such Y-junction waveguides as input and output couplers, the power transmittance is (see Problem 9.12.1)

$$T = 1 - 2\xi(1 - \xi)(1 - \cos\Delta\varphi), \tag{9.261}$$

where $\Delta\varphi = \Delta\varphi_L + \Delta\varphi_{NL}$ is the total differential phase shift. For one that uses such directional couplers as input and output couplers, the power transmittance through the same channel is (see Problem 9.12.1)

$$T = 1 - 2\xi(1 - \xi)(1 + \cos\Delta\varphi). \tag{9.262}$$

Note that (9.261) reduces to the form of (6.77), and (9.262) reduces to that of (6.79), if the two arms of the interferometer are equally excited so that $\xi = 1/2$. As discussed above, such balanced excitation is feasible only when the two arms of the interferometer are physically asymmetric. Such physical asymmetry leads to a nonvanishing linear differential phase shift, $\Delta\varphi_{\mathrm{L}} \neq 0$, which acts as a bias phase shift. By properly adjusting the asymmetry between the two arms, the value of this bias phase shift can be chosen for a desired operating point of the device.

Figure 9.36 shows the structure of a three-input, symmetric all-optical Mach–Zehnder interferometer using Y-junction waveguides. This device consists of three input channels that are fed into a symmetric Mach–Zehnder interferometer with arms of equal lengths. The data signal is sent through the central channel c, and the control signals are fed into either channel a or b or both. The data signal wave is orthogonally polarized with respect to the control signals to avoid interference between them. A polarizer at the output end allows only the polarization of the data signal to pass. Interaction between the data signal and the control signals is through cross-phase modulation only. Because a data signal sent through channel c is equally split between the two arms of the interferometer, nonlinear phase shifts caused by self-phase modulation of the data signal in the two arms cancel. The net differential nonlinear phase shift is caused by the cross-phase modulation imposed by any control signals on the data signal. This differential nonlinear phase shift has exactly the form of (9.260) with $n_{2a} = n_{2b} = n_2$, $l_a = l_b = l$, and $\mathcal{A}_a^{\mathrm{eff}} = \mathcal{A}_b^{\mathrm{eff}} = \mathcal{A}_{\mathrm{eff}}$ for a symmetric Mach–Zehnder interferometer:

$$\Delta\varphi_{\mathrm{NL}} = \frac{2\pi}{\lambda}\frac{n_2 l}{\mathcal{A}_{\mathrm{eff}}}(P_b - P_a), \tag{9.263}$$

where n_2 is the nonlinear refractive index due to cross-phase modulation between orthogonally polarized waves. Though the two arms of a symmetric Mach–Zehnder interferometer are equal in length, it is still possible to introduce a linear phase difference between them by a bias voltage if the device is fabricated on an electro-optic material. For a symmetric Mach–Zehnder interferometer using Y-junction waveguides, the transmittance of the data signal is that given in (9.261) with $\xi = 1/2$, which is reduced to the following simple form:

$$T = \cos^2\frac{\Delta\varphi}{2}. \tag{9.264}$$

An all-optical interferometer has many useful applications. Like an electro-optic interferometer, it can be used as an amplitude modulator or, when accompanied by a directional coupler instead of a Y-junction waveguide at the output end, as a switch. Unlike an electro-optic interferometer, however, its function is completely controlled by the input optical power alone. Therefore, there are some unique applications of an all-optical interferometer that are not possible with an electro-optic interferometer. For instance, with unbalanced excitation in an all-optical interferometer with symmetric arms, it is possible to shape an optical pulse by taking advantage of the fact that the

power-dependent transmittance of the device now varies across the envelope of the pulse. Pulse shortening can be achieved if the maximum transmittance occurs at the peak of the pulse while the wings of the pulse have very low transmittance. All-optical Mach–Zehnder interferometers can be made to perform certain unique functions, such as optical logic, optical sampling, and optical ON–OFF switching.

EXAMPLE 9.24 A three-input, symmetric all-optical Mach–Zehnder interferometer as shown in Fig. 9.36 consists of AlGaAs channel waveguides fabricated on a GaAs substrate along the [110] crystal axis on the (001) plane. The data signal launched into channel c is a TM-like mode polarized in the [001] direction. A control signal is launched into channel a as a TE-like mode polarized in the $[1\bar{1}0]$ direction. No control signal is launched into channel b. Both data and control signals are at $\lambda = 1.55$ μm wavelength. The length of both arms of the interferometer is $l = 2$ cm, and the effective area of the channel waveguide is $\mathcal{A}_{\text{eff}} = 6\ \mu\text{m}^2 = 6 \times 10^{-12}\ \text{m}^2$. The nonlinear refractive index characterizing cross-phase modulation between TE-like and TM-like modes in this AlGaAs waveguide at $\lambda = 1.55$ μm is $n_2 = 1.3 \times 10^{-17}\ \text{m}^2\ \text{W}^{-1}$. No linear bias phase is applied to either arm of the device. Ignoring all possible linear and nonlinear losses, find the power of the control signal needed for this device to function as an all-optical ON–OFF switch. If the control signal is in the form of an optical pulse of $\Delta t_{\text{ps}} = 1$ ps pulsewidth, what is the switching energy of the control pulse?

Solution For the device to function as an all-optical ON–OFF switch, both the ON state with a transmittance of $T = 1$ and the OFF state with a transmittance of $T = 0$ have to be accessible by varying the power of the control signal. Because there is no linear phase bias, the total differential phase shift of the device is contributed solely by the nonlinear effect; thus $\Delta\varphi = \Delta\varphi_{\text{NL}}$. Because no control signal is launched into channel b, $P_b = 0$. From (9.264), we then find that the minimum nonlinear differential phase shift required for $T = 1$ is $\Delta\varphi_{\text{NL}} = 0$ and that required for $T = 0$ is $\Delta\varphi_{\text{NL}} = -\pi$. Therefore, we find from (9.263) that the ON state can be reached by simply making $P_a = 0$ so that $\Delta\varphi_{\text{NL}} = 0$. By setting $\Delta\varphi_{\text{NL}} = -\pi$ in (9.263), we find the following control signal power required to reach the OFF state:

$$P_a = \frac{\lambda \mathcal{A}_{\text{eff}}}{2n_2 l} = \frac{1.55 \times 10^{-6} \times 6 \times 10^{-12}}{2 \times 1.3 \times 10^{-17} \times 2 \times 10^{-2}}\ \text{W} = 17.88\ \text{W}.$$

If the control signal is in the form of an optical pulse of $\Delta t_{\text{ps}} = 1$ ps pulsewidth, the device is in the ON state with $T = 1$ in the absence of a control pulse. The device can be switched to the OFF state with a control pulse of a switching energy of $U_{\text{ps}} = P_a \Delta t_{\text{ps}} = 17.88$ pJ.

Clearly, a waveguide Mach–Zehnder interferometer operated with a CW beam at 17.88 W is not practical, but it is practical with an ultrashort pulse of 17.88 W peak power such as the one of 1 ps pulsewidth considered here. For this reason, the control

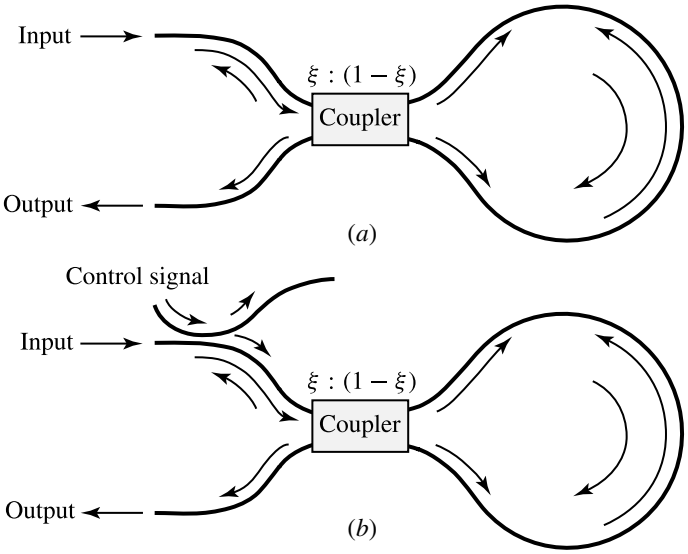


Figure 9.37 Nonlinear optical loop mirrors: (*a*) single-input configuration without a control signal and (*b*) two-input configuration with a control signal.

signal for this device is generally in the form of ultrashort laser pulses though the data signal can be of any waveform.

Here we have ignored the losses and dispersion of the waveguide. In reality, the waveguide has both linear losses, mainly from scattering and impurity absorption, and nonlinear losses, from both two-photon and three-photon absorption processes. These losses will increase the switching power of the device while reducing its extinction ratio between ON and OFF states. When the device is operated with short optical pulses, the dispersion of the waveguide can broaden the pulses and introduce an additional linear phase shift in the pulses. The consequences are also an increase in the switching energy and a reduction in the extinction ratio.

Nonlinear optical loop mirrors

A nonlinear optical loop mirror,[8] is a folded Mach–Zehnder interferometer in the so-called *Sagnac configuration*, as shown in Fig. 9.37. The basic device shown in Fig. 9.37(*a*) consists of a single-mode waveguide loop, such as a single-mode fiber or a single-mode semiconductor waveguide, that is closed with a four-port directional coupler. The two paths of opposite propagation directions in the loop are equivalent to the two arms of an interferometer. The single coupler, which has a power-splitting ratio of $\xi : (1 - \xi)$, serves as both the power-splitting input coupler and the power-combining output coupler. An input field is split into two contrapropagating

[8] Doran, N.J. and Wood, D., "Nonlinear-optical loop mirror," *Optics Letters* **13**: 56–58, 1988.

fields that travel through exactly the same loop path but in opposite directions before recombining at the coupler to form the output of the device.

The optical field launched into the device can be a short pulse that has a spatial span much shorter than the loop length. Then interaction between contradirectionally propagating pulses in the loop is negligible so that only the self-phase modulation of each individual pulse needs to be considered. It can also be a very long pulse or a CW wave that fills up the entire loop. Then the cross-phase modulation between contradirectionally propagating waves needs to be considered as well. Because of the exact symmetry between the two contradirectional paths, $\Delta\varphi_{\mathrm{L}} = 0$ irrespective of the operating condition. It can be shown that for both cases discussed here, we have (see Problem 9.12.4)

$$\Delta\varphi = \Delta\varphi_{\mathrm{NL}} = (1 - 2\xi)\frac{2\pi n_2}{\lambda}\frac{P_{\mathrm{in}}}{\mathcal{A}_{\mathrm{eff}}}l, \tag{9.265}$$

where P_{in} is the input power launched into the device and l is the length of the loop. The transmittance of the device is that given in (9.262) with $\Delta\varphi = \Delta\varphi_{\mathrm{NL}}$ given above. The device also has a reflectance $R = 1 - T$ back to the original input port. In the linear regime at low power levels, the device functions as a mirror with $R = 4\xi(1 - \xi)$ and $T = 1 - 4\xi(1 - \xi)$. In the nonlinear regime at high power levels, the device functions as a nonlinear mirror with power-dependent reflectance and transmittance due to the dependence of $\Delta\varphi_{\mathrm{NL}}$ on the input power.

Similarly to the Mach–Zehnder interferometer, a nonlinear optical loop mirror can also accept a control signal to switch the data signal. Figure 9.37(*b*) shows a two-input configuration for such a purpose. More sophisticated configurations are also possible. With a control signal, a nonlinear optical loop mirror can perform such functions as optical switching, sampling, multiplexing, and demultiplexing.

There are several advantages of using the nonlinear optical loop mirror as an all-optical interferometric device over the conventional all-optical Mach–Zehnder interferometer with two separate arms. Because the two contrapropagating fields in a nonlinear optical loop mirror travel over exactly the same path in opposite directions, they experience exactly the same linear effects, which cancel out when the two fields are combined in returning to the coupler. Therefore, the device is stable against external perturbations and does not require interferometric alignment. This unique characteristic allows a very long fiber on the order of kilometers to be used for a nonlinear optical loop mirror to function at a low optical power level with sufficient self-phase modulation, making it a truly practical all-optical device. Because response and relaxation of Kerr nonlinearity in silica fibers are nearly instantaneous, a nonlinear fiber loop mirror is also ideal for many applications that use ultrashort optical pulses. The precise match in length of the contradirectional paths in this device ensures precise coincidence of the returning pulses, which is a daunting task with a conventional interferometer with separate arms

considering the fact that the path length can be as long as a few kilometers while the pulses can be shorter than 1 ps.

EXAMPLE 9.25 A single-input nonlinear optical fiber loop mirror of the configuration shown in Fig. 9.37(a) consists of a single-mode fiber that has a loop length of $l = 100$ m and an effective cross-sectional area of $\mathcal{A}_{\text{eff}} = 3 \times 10^{-11}$ m^2 for an optical wave at $\lambda = 1.55$ μm. The self-phase modulation nonlinear refractive index of this fiber is $n_2 = 3.2 \times 10^{-20}$ m^2 W^{-1}. At low input power levels, this loop mirror has a transmittance of $T = 25\%$. Find the lowest input power that is required for it to have a transmittance of 100%.

Solution With a low-power transmittance of $T = 25\% = 1/4$, we find by solving $T = 1 - 4\xi(1 - \xi) = 1/4$ that $\xi = 1/4$ for the power-splitting ratio of the coupler in the device. By plugging $\xi = 1/4$ and $T = 1$ into (9.262), we find that the nonlinear phase shift required for $T = 1$ at a high power level is a solution of the following condition: $1 + \cos \Delta\varphi = 0$. Therefore, $\Delta\varphi = (2n + 1)\pi$ for any integer n. From (9.265), we see that $P_{\text{in}} \propto \Delta\varphi$. The lowest required power for $T = 100\%$ can be obtained by plugging $\Delta\varphi = \pi$ and $\xi = 1/4$ into (9.265) to find that

$$P_{\text{in}} = \frac{\lambda \mathcal{A}_{\text{eff}}}{n_2 l} = \frac{1.55 \times 10^{-6} \times 3 \times 10^{-11}}{3.2 \times 10^{-20} \times 100} \text{ W} = 14.53 \text{ W}.$$

This power is too high for this fiber device to be practical if the input is a CW signal. It is not a problem if the input signal consists of very short pulses. For instance, an average power of only 1.453 mW is required if the input signal is made up with pulses of 1 ps pulsewidth at a repetition rate of 100 MHz. For this reason, nonlinear optical loop mirrors are generally operated with very short laser pulses.

Nonlinear directional couplers

The coupling efficiency of a directional coupler can be varied by varying the phase mismatch or the coupling coefficients between the two waveguides that form the directional coupler. For an electrically modulated directional coupler discussed in Section 6.4, the coupling coefficient is a function of an externally applied voltage that induces changes in the refractive index of the waveguide material through the Pockels effect. For an all-optical nonlinear directional coupler based on the optical Kerr effect, the coupling coefficient can be varied by varying the value or the distribution of the optical power launched into the device. A nonlinear directional coupler can be formed using two parallel waveguides fabricated in such nonlinear crystals as GaAs or $LiNbO_3$, as shown in Fig. 9.38(a). It can also be formed using a dual-core optical fiber, as shown in Fig. 9.38(b). The advantage of using a dual-core fiber is that a coupler of a very long interaction length on the order of kilometers can be easily realized to make practical use

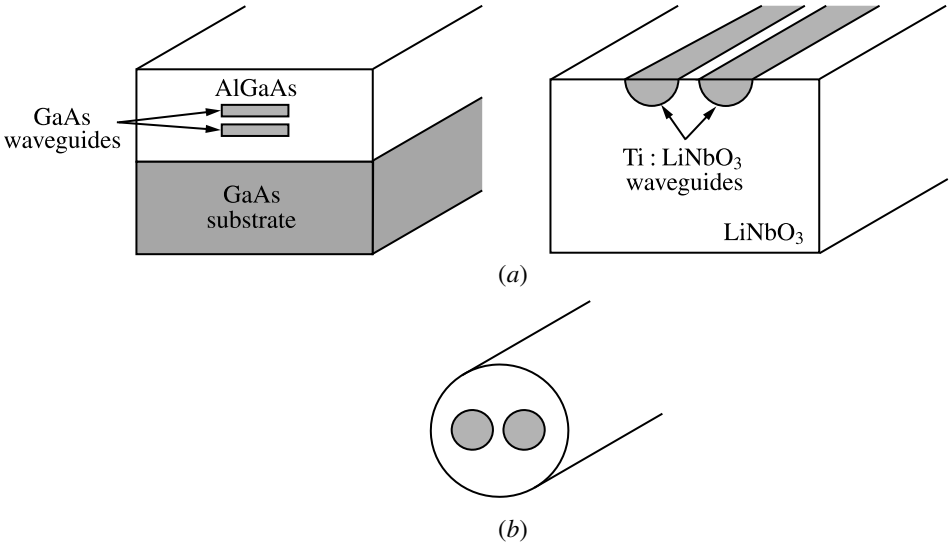


Figure 9.38 Nonlinear directional couplers formed by (*a*) two parallel waveguides and (*b*) a dual-core optical fiber.

of the small optical nonlinearity in a fiber. In the following, we consider for simplicity only symmetric directional couplers in which the two waveguide channels are identical. Asymmetric nonlinear directional couplers have similar characteristics.

For a symmetric nonlinear directional coupler that is formed by two identical single-mode waveguides, we have $\beta_a = \beta_b$ and $\kappa_{aa} = \kappa_{bb}$. Therefore, the effective linear propagation constant for each individual waveguide mode is $\beta = \beta_a + \kappa_{aa} = \beta_b + \kappa_{bb}$. In addition, $\kappa_{ab} = \kappa_{ba}^* \equiv \kappa$, which is real and positive, and $\sigma_{aa} = \sigma_{bb} \equiv \sigma$, which is real but can be either positive or negative depending on the sign of $\chi^{(3)}$ of the Kerr medium. The coupled equations for a symmetric nonlinear directional coupler can then be written as

$$\frac{\mathrm{d}\tilde{A}}{\mathrm{d}z} = \mathrm{i}\kappa\tilde{B} + \mathrm{i}\sigma|\tilde{A}|^2\tilde{A} + \text{nonlinear cross terms}, \tag{9.266}$$

$$\frac{\mathrm{d}\tilde{B}}{\mathrm{d}z} = \mathrm{i}\kappa\tilde{A} + \mathrm{i}\sigma|\tilde{B}|^2\tilde{B} + \text{nonlinear cross terms}, \tag{9.267}$$

where $\tilde{A} = A\mathrm{e}^{-\mathrm{i}\kappa_{aa}z}$ and $\tilde{B} = B\mathrm{e}^{-\mathrm{i}\kappa_{bb}z}$ as defined in (4.52). The terms characterized by the coupling coefficient κ represent linear coupling between the two modes. The terms characterized by σ contribute to the self-phase modulation of each individual mode. The nonlinear cross terms, which are not explicitly spelled out because of their complexity, contribute to direct nonlinear coupling between the two modes. In general, the nonlinear cross terms, though not completely negligible, are much smaller than the terms that represent linear coupling and self-phase modulation in each equation. Indeed, the direct nonlinear coupling contributed by the nonlinear cross terms is not

necessary for the functioning of a nonlinear directional coupler. The basic operation principle of a nonlinear directional coupler is that the self-phase modulation of each individual mode creates a power-dependent differential phase shift that leads to a power-dependent phase mismatch between the two modes. As a consequence, the coupling coefficient would become power dependent even if the only coupling were the linear coupling characterized by the linear coefficient κ. The direct nonlinear coupling contributed by the nonlinear cross terms acts as an additional perturbation, which changes the detailed quantitative characteristics of a nonlinear directional coupler. The general characteristics of a nonlinear directional coupler can be fully understood without considering the nonlinear cross terms.

We consider only the simple case when the nonlinear cross terms are neglected. We also assume that the input optical power is initially launched into only waveguide a so that $P_a(0) = P_{\text{in}}$ and $P_b(0) = 0$. Under these assumptions, the coupling efficiency of a nonlinear coupler that has an interaction length l is found to be

$$\eta = \frac{P_b(l)}{P_{\text{in}}} = \frac{1}{2}\left[1 - \text{cn}(2\kappa l, m)\right] = \frac{1}{2}\left[1 - \text{cn}\left(2\kappa l, \frac{\sigma}{4\kappa}P_{\text{in}}\right)\right], \tag{9.268}$$

where

$$m = \frac{\sigma}{4\kappa}P_{\text{in}} = \frac{P_{\text{in}}}{P_{\text{c}}} \tag{9.269}$$

is an index that characterizes the level of the input power with respect to a critical power level P_{c} that is defined as

$$P_{\text{c}} = \frac{4\kappa}{\sigma} = \frac{2\kappa\lambda\mathcal{A}_{\text{eff}}}{\pi n_2}, \tag{9.270}$$

and $\text{cn}(z, m)$ is a Jacobi elliptic function defined by

$$z = \int_x^1 \frac{\text{d}t}{(1-t^2)^{1/2}(1-m^2+m^2t^2)^{1/2}} = \text{cn}^{-1}(x, m). \tag{9.271}$$

For the symmetric coupler under consideration, $P_a(l) + P_b(l) = P_{\text{in}}$, and the power transmittance through the input channel is

$$T = \frac{P_a(l)}{P_{\text{in}}} = 1 - \eta. \tag{9.272}$$

It can be clearly seen from (9.268) that the coupling efficiency of a nonlinear coupler is a function of the input power to the device. Figure 9.39 shows the coupling efficiency as a function of interaction length l, normalized to the linear coupling length $l_{\text{c}}^{\text{PM}} = \pi/2\kappa$, at various input power levels that are characterized by different values of the index m. In the limit of very low input powers, $P_{\text{in}} \ll P_{\text{c}}$ and $m \approx 0$, the coupling efficiency

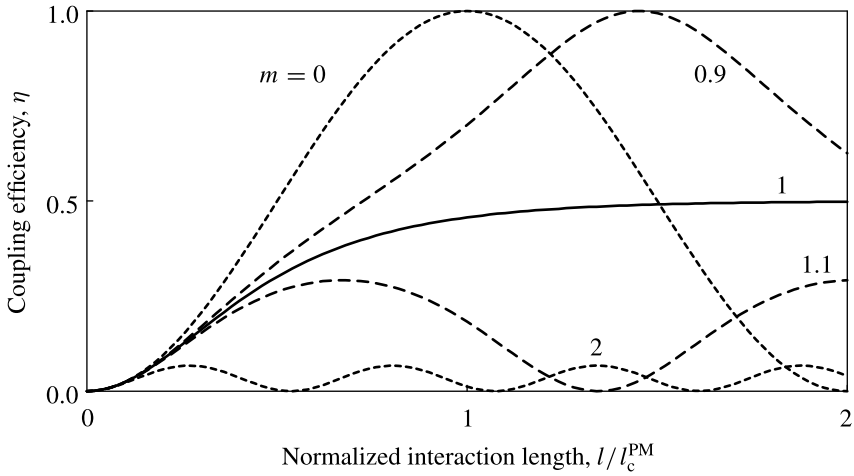


Figure 9.39 Coupling efficiency of a symmetric nonlinear directional coupler as a function of interaction length l, normalized to the linear coupling length $l_c^{PM} = \pi/2\kappa$, at various input power levels that are characterized by different values of the index $m = P_{in}/P_c$.

reduces to that of the phase-matched linear directional coupler, $\eta = (1 - \cos 2\kappa l)/2 = \sin^2 \kappa l$ given in (4.85), because $\mathrm{cn}(2\kappa l, 0) = \cos(2\kappa l)$, and the coupling length is just l_c^{PM}.

As the input power increases, a power-dependent phase mismatch between the two waveguide channels is generated by the power-dependent differential phase shift. At relatively low input powers, $P_{in} < P_c$ and $m < 1$, this power-dependent phase mismatch has the effect of slowing down the power transfer between the two channels. This phase mismatch is reduced as more power is transferred and is later even reversed as more than 50% of the input power is transferred. The nonlinear directional coupler thus acts like a reversed-$\Delta\beta$ coupler. Complete switching of power with $\eta = 1$ to reach the cross state still occurs, but the coupling length is longer than the linear coupling length and it increases as the input power increases. These effects can be observed from the curve for $m = 0.9$ in Fig. 9.39.

At high input powers, $P_{in} > P_c$ and $m > 1$, the initial phase mismatch is so large that the power transfer never reaches the 50% point for the phase mismatch to be reversed. Therefore, the coupling efficiency oscillates, but $\eta < 1/2$ for any device length. The cross state cannot be reached at such high power levels, as can be seen in Fig. 9.39 from the curves for $m = 1.1$ and $m = 2$. At the critical power level, $P_{in} = P_c$ and $m = 1$, the coupling efficiency stays at $\eta = 1/2$ indefinitely after 50% of the input power is transferred. This state is unstable as any perturbation caused by noise or fluctuations in the input power can tip this balance between the two channels.

Figure 9.40 shows the coupling efficiency as a function of input power P_{in}, normalized to the critical power P_c, for a symmetric nonlinear directional coupler with a fixed length $l = l_c^{PM}$, known as the *half-beat-length coupler*, and another with $l = 2l_c^{PM}$, known as

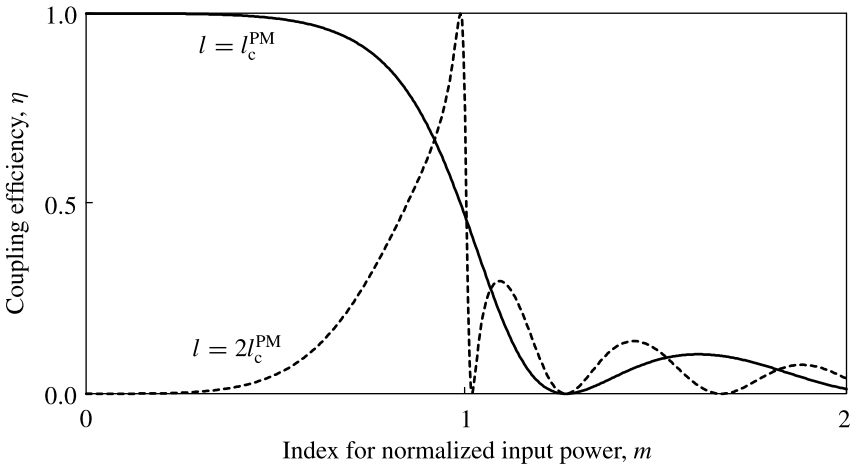


Figure 9.40 Coupling efficiency as a function of index m, which represents the input power P_{in} normalized to the critical power P_c, for two symmetric nonlinear directional couplers of fixed lengths $l = l_c^{PM}$ and $l = 2l_c^{PM}$.

the *beat-length coupler*. For the half-beat-length coupler, which starts with a linear coupling efficiency of $\eta = 1$ at a very low input power, the coupling efficiency remains high until the input power approaches the level of P_c when it drops and remains low for all high power levels above P_c. For the beat-length coupler, which starts with a linear coupling efficiency of $\eta = 0$ at a very low power level, only a very narrow power range exists for high coupling efficiencies with a peak value of $\eta = 1$.

In the above discussions, only symmetric directional couplers are considered, and the effect of nonlinear cross terms in (9.266) and (9.267) are ignored. There exists a general analytical solution in the form of elliptic functions for the coupled nonlinear differential equations even when the structure of the coupler is asymmetric and the nonlinear cross terms are considered. The primary effect of the nonlinear cross terms is to cause a change in the value of P_c depending on the strength of the nonlinear cross coupling between the two waveguide modes. For an asymmetric coupler, the initial linear phase mismatch leads to power-dependent characteristics that are nonreciprocal with respect to detuning between the two channels.

EXAMPLE 9.26 A half-beat-length nonlinear directional coupler of the structure shown in Fig. 9.38(*a*) for a TE-like mode has a length of $l = 1.5$ cm. It consists of two parallel AlGaAs channel waveguides on a GaAs substrate with the same structural parameters as the AlGaAs waveguides described in Example 9.24. At $\lambda = 1.55$ μm wavelength, the nonlinear refractive index characterizing self-phase modulation for the TE-like mode in the waveguide is $n_2 = 1.5 \times 10^{-17}$ m^2 W^{-1}. Find the critical power of the device.

Solution Because the device is a half-beat-length coupler, we have $l_c^{PM} = l = 1.5$ cm. Therefore, the coupling coefficient is $\kappa = \pi/2l_c^{PM} = \pi/2l$. By plugging this relation into (9.270), we find the following critical power:

$$P_c = \frac{\lambda \mathcal{A}_{eff}}{n_2 l} = \frac{1.55 \times 10^{-6} \times 6 \times 10^{-12}}{1.5 \times 10^{-17} \times 1.5 \times 10^{-2}} \text{ W} = 41.3 \text{ W}.$$

If the device is operated with a short pulse of $\Delta t_{ps} = 1$ ps like that considered in Example 9.24, then the critical pulse energy is 41.3 pJ.

PROBLEMS

9.1.1 Three optical fields at wavelengths of $\lambda_1 = 750$ nm, $\lambda_2 = 600$ nm, and $\lambda_3 = 500$ nm, corresponding to frequencies of $\omega_1 = 2\pi c/\lambda_1$, $\omega_2 = 2\pi c/\lambda_2$, and $\omega_3 = 2\pi c/\lambda_3$, respectively, are involved in second-order nonlinear optical interactions. The optical fields at the three frequencies are $\mathbf{E}(\omega_1) = E_1\hat{x}$, $\mathbf{E}(\omega_2) = E_2(\hat{y} + \hat{z})/\sqrt{2}$, and $\mathbf{E}(\omega_3) = E_3\hat{z}$, where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are aligned with the principal x, y, and z axes of the nonlinear crystal.

a. Find the nonlinear polarization $\mathbf{P}^{(2)}$ at the frequency of $\omega_5 = 2\pi c/\lambda_5$ where $\lambda_5 = 3$ μm. Express each of the components of $\mathbf{P}^{(2)}(\omega_5)$ explicitly in terms of the elements of $\boldsymbol{\chi}^{(2)}$ and the given magnitudes, E_1, E_2, and E_3, of the three optical fields.

b. If nonlinear interaction takes place in a $LiNbO_3$ crystal, which belongs to the $3m$ point group, what are the expressions of the components of $\mathbf{P}^{(2)}(\omega_5)$ in terms of the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$?

9.1.2 Answer the questions in Problem 9.1.1 for the nonlinear polarization $\mathbf{P}^{(2)}$ at the frequency of $\omega_6 = 2\pi c/\lambda_6$ for $\lambda_6 = 250$ nm.

9.1.3 Answer the questions in Problem 9.1.1 for the nonlinear polarization $\mathbf{P}^{(2)}$ at the frequency of $\omega_7 = 2\pi c/\lambda_7$ for $\lambda_7 = 1.5$ μm.

9.2.1 Verify the reality condition given in (9.23) for nonlinear susceptibilities.

9.2.2 Spell out explicitly the relations among the frequency-dependent elements of the $\boldsymbol{\chi}^{(3)}$ tensor that characterize the interaction of four frequencies ω_1, ω_2, ω_3, and ω_4 for $\omega_4 = \omega_1 + \omega_2 + \omega_3$ under (a) intrinsic permutation symmetry, (b) full permutation symmetry, and (c) Kleiman's symmetry condition, respectively.

9.2.3 Show that $\boldsymbol{\chi}^{(2)}$ contributed by electric-dipole interaction is identically zero in a centrosymmetric material, whereas a nonzero $\boldsymbol{\chi}^{(3)}$ exists in any material.

9.2.4 Verify the relation between the Pockels coefficients and the $\boldsymbol{\chi}^{(2)}$ elements expressed in (9.31) and that between the electro-optic Kerr coefficients and the $\boldsymbol{\chi}^{(3)}$ elements expressed in (9.32).

9.2.5 In this problem, we calculate the linear and nonlinear susceptibilities of a material containing N valence electrons per unit volume using a one-dimensional

anharmonic oscillator model. Each of these electrons, with a charge $q = -e$ and a mass m_0, oscillates in an anharmonic potential of the form

$$V(x) = \frac{1}{2}m_0\omega_0^2 x^2 + \frac{1}{3}m_0 a x^3 \tag{9.273}$$

with a damping constant γ so that its motion in response to externally applied optical fields can be described by

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x + ax^2 = \frac{F}{m_0}, \tag{9.274}$$

where $F = -e(E_m\mathrm{e}^{-\mathrm{i}\omega_m t} + E_n\mathrm{e}^{-\mathrm{i}\omega_n t} + E_p\mathrm{e}^{-\mathrm{i}\omega_p t} + \cdots) + \text{c.c.}$ We are interested in linear and nonlinear susceptibilities contributed by electric-dipole interactions. The electric-dipole polarization is defined as

$$P(t) = -Nex(t). \tag{9.275}$$

For the material response at a particular frequency ω_q, $P(t) = P(\omega_q)\mathrm{e}^{-\mathrm{i}\omega_q t} + \text{c.c.}$ and $x(t) = x(\omega_q)\mathrm{e}^{-\mathrm{i}\omega_q t} + \text{c.c.}$ so that $P(\omega_q) = -Nex(\omega_q)$. The linear and nonlinear susceptibilities can be found by solving (9.274) using the classical perturbation method. In this approach, $x(t)$ is expanded in a perturbation series as $x = x^{(1)} + x^{(2)} + x^{(3)} + \cdots$. Each order of $x^{(n)}$ is solved successively from (9.274). The polarizations of different orders are then defined as $P^{(n)}(\omega_q) = -Nex^{(n)}(\omega_q)$ for $n = 1, 2, 3, \ldots$

a. Is this material centrosymmetric or noncentrosymmetric?
b. Find the linear susceptibility $\chi^{(1)}(\omega)$.
c. Find the second-order susceptibilities: $\chi^{(2)}(\omega = \omega_1 + \omega_2)$, $\chi^{(2)}(2\omega_1 = \omega_1 + \omega_1)$, and $\chi^{(2)}(0 = \omega_1 - \omega_1)$, where $\omega_1 \neq \omega_2$.
d. Find the third-order susceptibilities: $\chi^{(3)}(\omega = \omega_1 + \omega_2 + \omega_3)$, $\chi^{(3)}(3\omega_1 = \omega_1 + \omega_1 + \omega_1)$, and $\chi^{(3)}(\omega_1 = \omega_1 + \omega_1 - \omega_1)$, where $\omega_1 \neq \omega_2 \neq \omega_3$.

9.2.6 There is a relationship between the second-order susceptibility $\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2)$ and the linear susceptibilities $\chi_{ii}^{(1)}(\omega_3)$, $\chi_{jj}^{(1)}(\omega_1)$, and $\chi_{kk}^{(1)}(\omega_2)$. It is known as *Miller's rule* and states that the ratio

$$\Delta_{ijk}^{(2)} = \frac{\chi_{ijk}^{(2)}(\omega_3 = \omega_1 + \omega_2)}{\chi_{ii}^{(1)}(\omega_3)\chi_{jj}^{(1)}(\omega_1)\chi_{kk}^{(1)}(\omega_2)} \tag{9.276}$$

is nearly a constant for all noncentrosymmetric crystals.

a. Use the results in Problem 9.2.5 to find the constant $\Delta^{(2)}$ for a one-dimensional case.
b. Estimate the value of the constant $\Delta^{(2)}$ for typical noncentrosymmetric solid crystals by taking the following typical values: $N \approx 10^{29}$ m^{-3} and $\omega_0 \approx 10^{16}$ rad s^{-1}. We also take $|a|x^2 \approx \omega_0^2 x$ and $x \approx N^{-1/3}$ so that $|a| \approx \omega_0^2 N^{1/3}$.

Find the typical range of $\chi^{(2)}$ by considering the fact that $\chi^{(1)}$ falls in the range of 1–10 for most crystals.

c. Find a similar rule for the third-order susceptibility $\chi^{(3)}_{ijkl}(\omega_4 = \omega_1 + \omega_2 + \omega_3)$. What is the constant $\Delta^{(3)}$ for a one-dimensional case in this situation?

d. Estimate the value of the constant $\Delta^{(3)}$ for typical noncentrosymmetric solid crystals by taking the parameters used in (b). Find the typical range of $\chi^{(3)}$ by considering the fact that $\chi^{(1)}$ falls in the range of 1–10 for most crystals.

e. What are the physical implications of Miller's rule and the similar rule for $\chi^{(3)}$?

9.2.7 Use the results from Problem 9.2.5 to answer the following questions.

a. Show the permutation symmetry of $\chi^{(2)}(\omega_3 = \omega_1 + \omega_2)$ for lossless media when $\omega_1 \neq \omega_2$.

b. What is the permutation symmetry relation of $\chi^{(2)}$ for lossless media in the case of frequency degeneracy, $\omega_1 = \omega_2$?

c. Without calculation, can you write down similar permutation symmetry relations of $\chi^{(3)}(\omega_4 = \omega_1 + \omega_2 + \omega_3)$ for lossless media in the cases of no frequency degeneracy, two-frequency degeneracy, and three-frequency degeneracy, respectively?

9.2.8 In this problem, we calculate the linear and nonlinear susceptibilities of a material containing N valence electrons per unit volume using a one-dimensional anharmonic oscillator model that is different from the one considered in Problem 9.2.5. Each of these electrons, with a charge $q = -e$ and a mass m_0, oscillates in an anharmonic potential of the form

$$V(x) = \frac{1}{2}m_0\omega_0^2 x^2 + \frac{1}{4}m_0 b x^4 \tag{9.277}$$

with a damping constant γ so that its motion in response to externally applied optical fields can be described by

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + 2\gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x + bx^3 = \frac{F}{m_0}, \tag{9.278}$$

where $F = -e(E_m \mathrm{e}^{-\mathrm{i}\omega_m t} + E_n \mathrm{e}^{-\mathrm{i}\omega_n t} + E_p \mathrm{e}^{-\mathrm{i}\omega_p t} + \cdots) + \text{c.c.}$

a. Is this a centrosymmetric or noncentrosymmetric material?

b. Use the perturbation method and the definition for the electric-dipole polarization given in Problem 9.2.5 to find $\chi^{(1)}(\omega)$, $\chi^{(2)}(2\omega = \omega + \omega)$, and $\chi^{(3)}(3\omega = \omega + \omega + \omega)$ contributed by electric-dipole interactions.

c. Explain why $\chi^{(2)} = 0$ in this problem.

9.2.9 For centrosymmetric materials, there is a relation between $\chi^{(3)}$ and $\chi^{(1)}$ similar to Miller's rule discussed in Problem 9.2.6.

a. Use the results obtained in Problem 9.2.8(b) to show that there is a constant $\Delta^{(3)}$ relating $\chi^{(3)}(3\omega)$ and the linear susceptibilities $\chi^{(1)}(\omega)$ and $\chi^{(1)}(3\omega)$ for centrosymmetric materials. Find this constant. Compare it with that found in Problem 9.2.6(c) for noncentrosymmetric crystals.

b. Estimate the value of the constant $\Delta^{(3)}$ for typical centrosymmetric solids by taking the following typical values: $N \approx 10^{29}\,\mathrm{m}^{-3}$ and $\omega_0 \approx 10^{16}\,\mathrm{rad\,s}^{-1}$. We also take $|b|x^3 \approx \omega_0^2 x$ and $x \approx N^{-1/3}$ so that $|b| \approx \omega_0^2 N^{2/3}$. Find the typical range of $\chi^{(3)}$ by considering the fact that $\chi^{(1)}$ falls in the range of 1–10 for most solids. Compare the numerical results obtained here with those found in Problem 9.2.6(d).

9.2.10 Consider an isotropic two-level absorbing medium with a single resonance frequency at ω_0. Follow the procedure used in Problem 9.2.8 to find $\chi^{(2)}(0 = \omega - \omega)$ and $\chi^{(3)}(\omega = \omega + \omega - \omega)$.

9.2.11 Answer the following questions regarding nonlinear optical susceptibilities.

a. Does a second-order nonlinear optical effect exist on the surface of a centrosymmetric material? Why?

b. In general, would you expect a highly refractory material to have larger or smaller nonlinear optical susceptibilities than a less refractory material? Explain.

c. Given ten nonlinear crystals without any knowledge of their $\boldsymbol{\chi}^{(2)}$, but with the refractive indices checked out from a handbook, how do you make an intelligent guess at which ones are likely to have a large $\boldsymbol{\chi}^{(2)}$ before taking any measurements? What do you base your guess on?

d. How does the effect of quantum confinement, such as that in the quantum-well structures of a semiconductor, enhance nonlinear susceptibilities?

9.3.1 What are the two unique features that set nonlinear optical processes apart from linear optical processes? Identify which one, or both, of these unique features each of the second-order processes listed in Table 9.4 has. Identify this also for each of the third-order processes listed in Table 9.5.

9.3.2 GaAs is a cubic crystal of $\bar{4}3m$ symmetry. Its nonlinear susceptibilities have relatively large values.

a. We know that many useful and efficient electro-optic devices, such as electro-optic modulators and demodulators, can be fabricated with GaAs. However, bulk GaAs is not used for efficient second-harmonic generation in any optical spectral region ranging from the ultraviolet to the infrared. Because both the Pockels effect and second-harmonic generation use $\boldsymbol{\chi}^{(2)}$, there must be some fundamental reasons for this. What are the reasons?

b. GaAs also has a large $\boldsymbol{\chi}^{(3)}$. Among the third-order nonlinear optical processes listed in Table 9.5, which ones do you think could easily take place in GaAs? Which ones do you think would be unlikely to happen? Explain.

9.3.3 A GaAs semiconductor laser has the structure shown in Fig. 9.41 where the active waveguide core layer has junction planes perpendicular to the [001] crystal axis, which is taken to be the z axis. The laser light propagates in the [110] direction with $\hat{k} = (\hat{x} + \hat{y})/\sqrt{2}$. It is found that second-harmonic emission at a frequency that is twice the laser frequency can be observed at the laser facets when the laser field is a TE mode of the waveguide, but it disappears when the laser field is a TM mode of the waveguide.

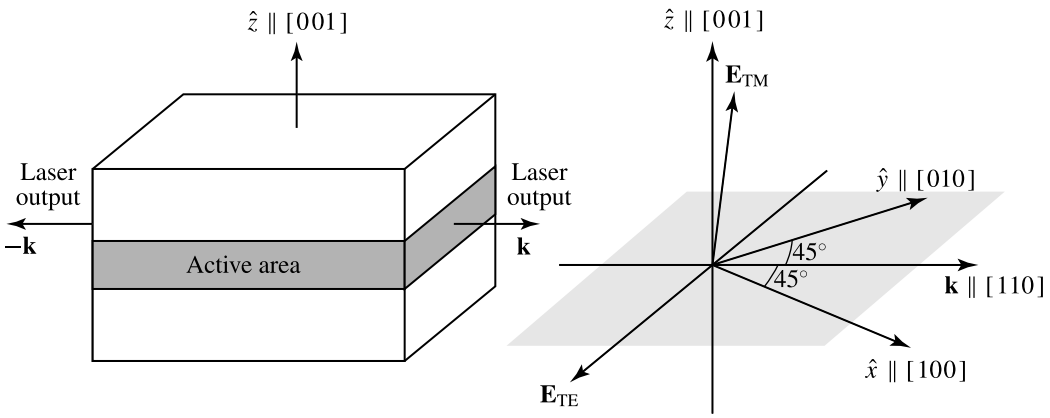


Figure 9.41 Crystal axes and field directions in a GaAs laser structure.

a. Explain why a TE laser mode can generate second-harmonic emission while a TM laser mode cannot.
b. What is the polarization of the second-harmonic field generated by the TE-polarized laser mode?
c. Explain why second-harmonic emission can be generated at the laser facets. Is there any contradiction to the discussions in Problem 9.3.2(a)?

9.3.4 The only nonvanishing elements of the $\boldsymbol{\chi}^{(2)}$ tensor for GaAs are $\chi_{14}^{(2)} = \chi_{25}^{(2)} = \chi_{36}^{(2)}$. A GaAs crystal is cleaved at the (011) surface so that the cleaved surface is normal to the [011] direction with the surface normal $\hat{n} = (\hat{y} + \hat{z})/\sqrt{2}$. The crystal has a thickness of a few millimeters. It is placed in a nondispersive medium, such as a vacuum.

a. A linearly polarized optical wave at $\lambda = 1$ μm is normally incident on the crystal surface with its polarization direction, $\hat{e}$, making an angle ϕ with the [100] crystal axis, $\hat{x}$. Is it possible to generate substantial second-harmonic radiation in transmission? Why?
b. It is possible to generate second-harmonic radiation in reflection. With the arrangement in (a), what is the dependence of the second-harmonic intensity in reflection on the polarization direction $\hat{e}$ as a function of angle ϕ?

Discuss the relation between the polarization directions of the fundamental and second-harmonic fields.

c. Now, the crystal is instead cleaved at the (001) surface with $\hat{n} = \hat{z}$, and the fundamental wave is incident at an angle θ with respect to the surface normal and is TE polarized. If we rotate the crystal with respect to its normal axis $\hat{n}$ but with the incident fundamental beam fixed in space, how does the second-harmonic intensity in reflection vary with the rotation angle?

d. In (c), what is the relation between the polarization directions of the fundamental field and the second-harmonic field in reflection? Does it vary with the rotation angle of the crystal?

e. At a fixed rotational position of the crystal, what is the dependence of the second-harmonic intensity in reflection on the incident angle θ?

9.3.5 Answer the following questions regarding the interaction of an optical field at a frequency ω with a low-frequency RF field at a frequency Ω or with a DC field at $\Omega = 0$. Briefly explain your answer to each question by showing how a process being considered can or cannot happen. Assume that there is no spontaneous emission of any sort involved.

a. Can a low-frequency field at $\Omega \neq 0$ be generated by a single-frequency optical field through a second-order nonlinearity characterized by $\boldsymbol{\chi}^{(2)}$? Is this possible if it is a DC field with $\Omega = 0$?

b. Answer the questions in (a) if a third-order nonlinearity characterized by $\boldsymbol{\chi}^{(3)}$ is considered instead of the second-order nonlinearity.

c. Can the RF field be phase modulated by a single-frequency optical field in a $\boldsymbol{\chi}^{(2)}$ material without generating an optical field at another frequency?

d. Answer the question in (c) for a $\boldsymbol{\chi}^{(3)}$ material.

e. Can the RF field be amplified in magnitude through a $\boldsymbol{\chi}^{(2)}$ or $\boldsymbol{\chi}^{(3)}$ process?

9.3.6 Show that a circularly polarized optical wave cannot produce third-harmonic radiation directly through a third-order nonlinear optical process in an isotropic medium.

9.3.7 Consider the optical-field-induced birefringence with circularly polarized optical waves in an isotropic medium.

a. Show that the nonlinear polarization generated by a circularly polarized optical field, $\mathbf{E} = \hat{e}_{\pm} E(\omega)$, is circularly polarized with the same helicity as that of the optical field:

$$\mathbf{P}^{(3)}(\omega) = \hat{e}_{\pm} 3\epsilon_0 \left(\chi^{(3)}_{1122} + \chi^{(3)}_{1212} \right) |E(\omega)|^2 E(\omega), \tag{9.279}$$

where $\hat{e}_+$ and $\hat{e}_-$ are the eigenvectors of the left- and right-circular polarizations, respectively.

b. What is the field-induced nonlinear index of refraction for a circularly polarized wave in an isotropic medium?

9.3.8 Consider the optical-field-induced birefringence with linearly polarized optical waves in an isotropic medium.
   a. What is the nonlinear polarization generated by a field that is linearly polarized along $\hat{x}$? What if the field is linearly polarized at an angle of 45° with respect to the x and y coordinate axes?
   b. Use the results obtained in (a) to show that $\chi^{(3)}_{1111} = \chi^{(3)}_{1122} + \chi^{(3)}_{1212} + \chi^{(3)}_{1221}$ for an isotropic medium. Show also that $\chi^{(3)}_{1111} = 3\chi^{(3)}_{1122}$ if Kleiman's symmetry condition is valid.
   c. What is the field-induced nonlinear index of refraction for a linearly polarized wave in an isotropic medium?

9.3.9 GaAs is a cubic crystal of $\bar{4}3m$ symmetry, which has the isotropic linear optical property that $n_x = n_y = n_z = n$. However, it is not centrosymmetric. Therefore, its nonlinear optical properties are different from those of isotropic media. Consider a linearly polarized optical wave at a frequency ω propagating along the z axis of the crystal. The optical field is polarized in a direction on the xy plane that makes an angle ϕ with respect to the x axis.
   a. The optical wave may change the index of refraction of the medium through optical-field-induced birefringence. Show that the direction of the optical field polarization is not changed by this effect only when the incident optical wave is polarized at $\phi = m\pi/4$, where $m = 0, \pm 1, \pm 2, \ldots$
   b. What is the third-harmonic nonlinear polarization? For efficient harmonic generation, the optical frequencies involved have to be far away from the GaAs bandgap to avoid absorption. Is phase matching possible under this condition? Explain.

9.3.10 The only nonvanishing elements of the third-order nonlinear susceptibility tensor of an isotropic medium are those of the following forms: $\chi^{(3)}_{1111}$, $\chi^{(3)}_{1122}$, $\chi^{(3)}_{1212}$, and $\chi^{(3)}_{1221}$, with $\chi^{(3)}_{1111} = \chi^{(3)}_{1122} + \chi^{(3)}_{1212} + \chi^{(3)}_{1221}$.
   a. Show that by applying a DC electric field, it is possible to generate second-harmonic radiation of a linearly polarized fundamental optical wave in this medium.
   b. How do you apply this DC field and choose the polarization direction for the fundamental wave so that the second-harmonic field is polarized in a direction parallel to the fundamental field polarization?
   c. With an applied DC field, is it possible to generate the second harmonic of a circularly polarized fundamental wave? What is the polarization of this second-harmonic field if it is possible?
   d. Without the applied DC field, is it possible to generate second-harmonic emission in this medium? Explain.

9.3.11 Answer the following questions regarding harmonic generation by considering nonlinear susceptibilities and phase matching carefully.

a. Is it possible to generate a second-harmonic signal by the incidence of a laser beam from free space on a liquid solution if only the electric-dipole interaction is important? Consider all possibilities and explain your answer.
b. Silicon is a centrosymmetric crystal and absorbs very little infrared at wavelengths longer than 1 μm. However, when an intense beam of infrared is incident upon a polished silicon wafer, we can still detect second-harmonic radiation in reflection. What do you think is happening?
c. Silver is a very good metal that strongly reflects light in the visible region. However, when a shiny surface of silver is illuminated with a laser beam in the visible, we can also detect second-harmonic radiation in reflection. What happens? Do you think there is anything different in this case from what happens in (b)?
d. $\chi^{(3)}$ exists in any medium including optical fibers. Why are optical fibers not used for third-harmonic generation in order to take advantage of their long interaction length and strong optical confinement?

9.3.12 Write down the nonlinear susceptibilities that are responsible for the following nonlinear optical processes: stimulated Raman scattering, Stokes–anti-Stokes coupling, stimulated Raman anti-Stokes scattering, optical rectification, and optical-field-induced birefringence. Clearly identify the relations among the component frequencies involved in each process. Also discuss the conditions for each process to take place by considering the following questions.
a. Is phase matching required?
b. Is the process associated with the real or the imaginary part of the responsible nonlinear susceptibility?
c. Is material excitation or de-excitation involved?

9.3.13 Answer the following general questions.
a. The third-order nonlinear susceptibility in a semiconductor such as AlGaAs/GaAs is substantially enhanced by growing quantum-well structures in the material. What is the primary mechanism of this enhancement?
b. Give two reasons why it is necessary to use very thin crystals when doing second-harmonic generation with femtosecond pulses.
c. How does the efficiency of phase-matched third-harmonic generation depend on the length of nonlinear interaction?
d. Name three nonlinear optical effects that may complicate the propagation of an intense optical pulse through a centrosymmetric medium that is not phase-matchable. What are these effects on the characteristics of the optical pulse?
e. Which nonlinear optical processes do you expect to see on sending an intense laser pulse into an isotropic medium? Why do you pick them over other processes?

f. Which nonlinear optical process surely occurs when a short optical pulse propagates in an optical fiber? What other processes might occur under certain conditions?

g. Under what conditions can the Stokes and anti-Stokes frequencies be simultaneously observed in a spontaneous Raman process? Under what conditions can they be simultaneously observed in a stimulated Raman process?

9.3.14 Consider phase-matched degenerate four-wave mixing of optical waves at an optical frequency ω and a corresponding wavelength λ in an isotropic medium. The arrangement involves two contrapropagating pump waves of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_1' = -\mathbf{k}_1$ and a probe wave of wavevector $\mathbf{k}_\mathrm{i}$ for the generation of an output wave of a wavevector $\mathbf{k}_\mathrm{s} = -\mathbf{k}_\mathrm{i}$.

a. Show that the nonlinear polarization is

$$\mathbf{P}_\mathrm{s}^{(3)}(\mathbf{k}_\mathrm{s} = -\mathbf{k}_\mathrm{i}, \omega) = A(\mathbf{E}_1 \cdot \mathbf{E}_\mathrm{i}^*)\mathbf{E}_1' + B(\mathbf{E}_1' \cdot \mathbf{E}_\mathrm{i}^*)\mathbf{E}_1 + C(\mathbf{E}_1 \cdot \mathbf{E}_1')\mathbf{E}_\mathrm{i}^*. \tag{9.280}$$

b. If the angle between $\mathbf{k}_1$ and $\mathbf{k}_\mathrm{i}$ is θ, what are the periods of the static gratings corresponding to the A and B terms, respectively, each in terms of the optical wavelength λ?

c. If both pump waves are s polarized (normal to the plane formed by the wavevectors), what is the polarization of the output wave generated by an s-polarized probe wave? What is the polarization if the probe wave is p polarized (parallel to the plane formed by the wavevectors)? Indicate the contribution from each term in (9.280).

d. Answer the questions in (c) for the situation where one pump wave is s polarized but the other is p polarized.

e. In a gaseous medium, the static gratings may degrade with time because of atomic thermal motion and time-dependent interactions if such gratings are created by short optical pulses. In such a situation, the degenerate four-wave mixing signal then depends on a parameter $a = \tau/\Delta t_\mathrm{ps}$, where τ is the atomic relaxation time constant and Δt_ps is the pulse duration. Based on this fact and the results obtained in (b)–(d), discuss an experimental approach that allows the deduction of information on the constant τ with variable pulse durations.

9.4.1 Consider a $4mm$ crystal, such as $BaTiO_3$, which has the following nonvanishing elements of the $\chi^{(2)}$ tensor: $\chi_{15}^{(2)} = \chi_{24}^{(2)}$, $\chi_{31}^{(2)} = \chi_{32}^{(2)}$, and $\chi_{33}^{(2)}$. The uniaxial crystal axis is the z axis.

a. An optical wave at a frequency ω propagates through the crystal along the x axis with its electric field $\mathbf{E}(\omega)$ polarized in the yz plane making an angle ϕ with respect to the y axis. Write down the second-harmonic nonlinear

polarization $\mathbf{P}^{(2)}(2\omega)$ as a function of the angle ϕ and the nonvanishing elements of $\chi^{(2)}$. What is χ_{eff} if $\mathbf{E}\,(2\omega)$ is polarized at an angle θ with respect to the y axis?

b. What should the direction of $\mathbf{E}(\omega)$ be so that $\mathbf{P}^{(2)}(2\omega) \parallel \mathbf{E}(\omega)$? What should it be so that $\mathbf{P}^{(2)}(2\omega) \perp \mathbf{E}(\omega)$?

c. How do you arrange the directions of beam propagation $\mathbf{k}$ and field polarization $\mathbf{E}(\omega)$ so that an optical rectification field can be generated along the direction of beam propagation? How do you arrange them so that an optical rectification field can be generated in a direction perpendicular to the direction of beam propagation?

9.4.2 In the process of second-harmonic generation, the fundamental field at frequency ω induces a nonlinear polarization $\mathbf{P}^{(2)}(2\omega)$ in the nonlinear medium at the second-harmonic frequency 2ω. This nonlinear polarization acts as a source to generate the second-harmonic radiation field $\mathbf{E}(2\omega)$.

a. Consider the interaction between $\mathbf{P}^{(2)}(2\omega)$ and $\mathbf{E}(2\omega)$. Show that power is converted from $\mathbf{E}(\omega)$ through $\mathbf{P}^{(2)}(2\omega)$ to $\mathbf{E}(2\omega)$ if $\mathbf{P}^{(2)}(2\omega)$ lags $\mathbf{E}(2\omega)$ in phase. Show also that this process is most efficient when the phase lag is $\pi/2$.

b. If there is a phase mismatch, the power shuttles back and forth between the second-harmonic and the fundamental waves. Show that after every coherence length, the power flow changes direction. In other words, if the power is converted from the fundamental to the second harmonic at $z = 0$, it will be converted from the second harmonic back to the fundamental at $z = l_{\text{coh}}$, and so on.

c. Suppose that at the beginning location $z = 0$ only the fundamental wave exists. What happens if the interaction length is one coherence length long? What happens if it is exactly two coherence lengths long?

9.4.3 A GaAs crystal has $\bar{4}3m$ symmetry. It may be used for second-harmonic generation in the infrared region because it is transparent in the 1- to 10-μm spectral range. Because it is a cubic crystal, however, phase matching is not possible in its transparent spectral region. Consider second-harmonic generation of a fundamental wave at a frequency ω propagating over a distance l with its $\mathbf{k}$ vector in the $[1\bar{1}0]$ direction of a GaAs crystal. Calculate the intensity of the second-harmonic wave if the fundamental wave is linearly polarized with (a) $\hat{e} \parallel [001]$ crystal direction, (b) $\hat{e} \parallel [110]$ crystal direction, and (c) $\hat{e} \parallel [111]$ crystal direction, respectively.

9.4.4 A plane wave at frequency ω, linearly polarized in the x direction, traverses an isotropic fluid. It produces a nonlinear polarization at the third-harmonic frequency 3ω.

a. What is the direction of this third-harmonic polarization?

b. Express the polarization at 3ω in terms of an element of the nonlinear susceptibility tensor $\boldsymbol{\chi}^{(3)}$.
c. Calculate the third-harmonic field generated by this polarization, assuming that $\mathcal{E}_{3\omega}(z) = 0$ at $z = 0$. The indices of refraction at the two frequencies are $n(\omega)$ and $n(3\omega)$, respectively. Is phase matching possible?

9.4.5 Consider the attenuation of an optical beam at frequency ω_1 by one- and two-photon absorption processes. The evolution of its intensity as it propagates along the z direction can be described as

$$\frac{\mathrm{d}I}{\mathrm{d}z} = -\alpha I - \alpha_2 I^2, \tag{9.281}$$

where α and α_2 are one- and two-photon absorption coefficients, respectively.
a. Calculate $I(z)$ for $I = I_0$ at $z = 0$.
b. Relate α and α_2 to their relevant linear or nonlinear susceptibilities, respectively.
c. Consider also the attenuation of a weak beam at ω_1 due to two-photon absorption at a combined photon energy of $\hbar\omega_1 + \hbar\omega_2$ in the presence of a strong beam at ω_2 that has an intensity of $I(\omega_2)$. Calculate $I(\omega_1)$ as a function of z in terms of nonlinear susceptibility $\chi^{(3)}(\omega_1 = \omega_1 + \omega_2 - \omega_2)$.

9.5.1 Show that in uniaxial crystals of symmetry classes 4, 6, 422, 622, $4mm$, and $6mm$ the effective second-order susceptibility, χ_{eff}, for collinear interaction of optical waves propagating in a direction $\hat{k}$ that makes an angle θ with the optical axis $\hat{z}$ and an angle ϕ with the principal axis $\hat{x}$ is independent of the angle ϕ.

9.5.2 Efficient second-order nonlinear optical frequency conversion using temperature tuning with 90° phase matching in a uniaxial crystal is not always possible. The reason is that the effective nonlinear susceptibility may be zero when phase matching is accomplished in this manner.
a. Show that 90° phase matching in a negative uniaxial crystal is not possible for type II phase matching but is possible for type I phase matching only in crystals of symmetry classes 3, 4, 6, $4mm$, $6mm$, $3m$, $\overline{4}$, and $\overline{4}2m$.
b. Show that 90° phase matching in a positive uniaxial crystal is not possible for type I phase matching but is possible for type II phase matching only in crystals of symmetry classes 3, 4, 6, $4mm$, $6mm$, $3m$, $\overline{4}$, and $\overline{4}2m$.

9.5.3 In this problem, we consider the angular tolerance for 90° phase matching in uniaxial crystals.
a. According to Problem 9.5.2, 90° phase matching in a negative uniaxial crystal is possible only for type I phase matching in certain crystals. Find the angular tolerance for this case.
b. Also according to Problem 9.5.2, 90° phase matching in a positive uniaxial crystal is possible only for type II phase matching in certain crystals. Find the angular tolerance for this case.

c. Because the efficiency of second-harmonic generation is proportional to the square of the fundamental intensity, the efficiency can usually be improved by simply focusing the beam. Assuming a Gaussian beam profile, how much can the fundamental beam be focused while maintaining phase matching through the entire crystal length?

9.5.4 Consider second-harmonic generation with a fundamental wave at $\lambda = 1.06\ \mu\text{m}$ in $LiNbO_3$, which is a negative uniaxial crystal. Phase matching is accomplished by 90° type I temperature tuning at 300 K with $n_o(\omega) = n_e(2\omega) = 2.234$. The dispersion of these indices with respect to wavelength can be expressed as

$$\left.\frac{dn_o}{d\lambda}\right|_{\lambda} = -6.17 \times 10^4\ \text{m}^{-1}, \quad \left.\frac{dn_e}{d\lambda}\right|_{\lambda/2} = -4.2 \times 10^5\ \text{m}^{-1}. \tag{9.282}$$

Assume that the crystal length is $l = 5$ cm.

a. What is the FWHM phase-matching spectral range?

b. If the laser generating the fundamental wave has a cavity length of 1 m, how many longitudinal laser modes can be simultaneously phase matched within the FWHM phase-matching range?

c. What is the implication of this result for the second-harmonic generation of ultrashort mode-locked laser pulses?

9.5.5 For the 90° type I phase matching in $LiNbO_3$ discussed in Problem 9.5.4, what is the FWHM phase-matching angle at the central phase-matched wavelength of $\lambda = 1.06\ \mu\text{m}$? Use the result obtained in Problem 9.5.3(a) to solve this.

9.5.6 $LiNbO_3$ is a uniaxial crystal of $3m$ symmetry. Its ordinary and extraordinary indices of refraction are both highly dependent on temperature. At room temperature, it has $n_o = 2.238$ and $n_e = 2.159$ at 1 μm wavelength and $n_o = 2.343$ and $n_e = 2.248$ at 500 nm wavelength. A laser beam at the fundamental wavelength of 1 μm propagates through the crystal at an angle θ with respect to the optical axis $\hat{z}$ in a plane at an angle ϕ with respect to the x axis to generate the second harmonic at 500 nm.

a. How should the polarizations of the fundamental and second-harmonic waves be chosen, respectively, for type I phase matching? What is χ_{eff} as a function of θ, ϕ, and the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$?

b. Answer the questions in (a) for type II phase matching.

c. If $\chi_{31}^{(2)} < 0$ and $\chi_{22}^{(2)} > 0$, what are the best choices for the value of ϕ in (a) and (b), respectively?

d. Show that angle tuning at room temperature is not possible for both type I and type II phase matching. What can be done for phase matching in order to carry out this second-harmonic generation process efficiently?

9.5.7 $LiNbO_3$ is a negative uniaxial crystal of $3m$ symmetry that has the following nonvanishing elements of $\boldsymbol{\chi}^{(2)}$: $\chi_{15}^{(2)} = \chi_{24}^{(2)}$, $\chi_{31}^{(2)} = \chi_{32}^{(2)}$, $\chi_{33}^{(2)}$, $\chi_{22}^{(2)} = -\chi_{21}^{(2)} = -\chi_{16}^{(2)}$. The ordinary and extraordinary refractive indices as functions

of wavelength and temperature are given in (9.92) and (9.93), respectively. The optical axis is the z axis. A laser beam at the fundamental wavelength of 1.064 μm propagates through the crystal at an angle θ with respect to $\hat{z}$ in a plane at an angle ϕ with respect to $\hat{x}$ to generate the second harmonic at 532 nm wavelength. It can be shown that angle tuning for phase matching is not possible at room temperature for either type of phase matching in this case. Therefore, we only consider temperature tuning for phase matching.

a. Find χ_{eff} for type I phase matching as a function of θ, ϕ, and the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$.
b. Find χ_{eff} for type II phase matching as a function of θ, ϕ, and the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$.
c. Find the phase-matching temperature for 90° type I phase matching.
d. In (c), how are the fundamental and second-harmonic fields polarized with respect to each other?
e. Can second-harmonic emission be generated with 90° type II phase matching? If your answer is yes, describe how it can be done. If your answer is no, explain why it cannot be done.

9.5.8 Consider nondegenerate sum-frequency generation, $\omega_3 = \omega_1 + \omega_2$ with $\omega_1 \neq \omega_2$, in ADP or KDP, which are negative uniaxial crystals of $\overline{4}2m$ symmetry. The optical axis is the z axis. The wave propagates at an angle θ with respect to $\hat{z}$ in a plane at an angle ϕ with respect to the x axis.

a. What is the condition for collinear phase matching? Very briefly describe how this can be accomplished.
b. For type I phase matching, how should the polarizations, in terms of ordinary or extraordinary waves, of the three waves be chosen respectively? What is the effective nonlinear susceptibility, χ_{eff}, as a function of θ, ϕ, and the nonvanishing elements of $\boldsymbol{\chi}^{(2)}$?
c. Answer the questions in (b) for the type II phase-matching condition.
d. What determines the optimum choice for the value of θ? Should the same value of θ be chosen for the cases in (b) and (c)?
e. What determines the optimum choice for the value of ϕ? Should the same value of ϕ be chosen for the cases in (b) and (c)? Write down the respective values chosen for ϕ in the two cases.

9.5.9 The uniaxial nonlinear crystal KDP has $\overline{4}2m$ symmetry with the only nonvanishing $\boldsymbol{\chi}^{(2)}$ elements being $\chi_{14}^{(2)} = \chi_{25}^{(2)} \neq \chi_{36}^{(2)}$. Consider second-harmonic generation in KDP with a fundamental wave at 1.064 μm to emit a second-harmonic wave at 532 nm. At room temperature, the ordinary and extraordinary indices of refraction for KDP at these two frequencies are

$$n_{\text{o}}(\omega) = 1.506\,617, \quad n_{\text{o}}(2\omega) = 1.527\,838,$$
$$n_{\text{e}}(\omega) = 1.468\,102, \quad n_{\text{e}}(2\omega) = 1.481\,803.$$

These indices have the following temperature dependencies:

$$n_{\mathrm{o}}(T) = n_{\mathrm{o}} + 4.02 \times 10^{-5}(n_{\mathrm{o}}^2 - 1.432)(298 - T), \tag{9.283}$$

$$n_{\mathrm{e}}(T) = n_{\mathrm{e}} + 2.21 \times 10^{-5}(n_{\mathrm{e}}^2 - 1.105)(298 - T), \tag{9.284}$$

where T is the absolute temperature in K, and n_0 and n_{e} are the values at 298 K.

a. Find the phase-matching angle for type I phase matching at room temperature.
b. What is the maximum value of $|\chi_{\mathrm{eff}}|$ for the case in (a)? How should the incident fundamental beam be arranged to obtain this maximum nonlinear coefficient for second-harmonic generation?
c. Is it possible to obtain 90° type I phase matching by temperature tuning? Explain.
d. Is it possible to obtain 90° type II phase matching by temperature tuning? Explain.
e. Is it possible to obtain type II phase matching by angle tuning at room temperature? Explain.

9.5.10 BBO is a negative uniaxial crystal. It is phase matchable for second-harmonic and sum-frequency generation in the spectral range from near infrared to near ultraviolet. Its ordinary and extraordinary indices at room temperature as a function of optical wavelength are given by the following Sellmeier equations:

$$n_{\mathrm{o}}^2 = 2.7359 + \frac{0.018\,78}{\lambda^2 - 0.018\,22} - 0.013\,54\lambda^2, \tag{9.285}$$

$$n_{\mathrm{e}}^2 = 2.3753 + \frac{0.012\,24}{\lambda^2 - 0.016\,67} - 0.015\,16\lambda^2, \tag{9.286}$$

where λ is in micrometers. It is desired to generate the second and third harmonics of the fundamental wavelength at 1.064 μm of a Nd : YAG laser using BBO crystals. The second harmonic will be generated by doubling the fundamental frequency, while the third harmonic will be generated by summing the fundamental and the second-harmonic frequencies. Collinear phase matching with the waves propagating in a direction making an angle θ with the unique z axis and an angle ϕ with the x axis is considered.

a. For the SHG to generate light at 532 nm, write down the equations required to be solved for the phase-matching angles for type I and type II phase-matching conditions, respectively. Calculate the phase-matching angle for type I phase matching.
b. Calculate the walk-off angle for the SHG under type I phase matching in (a). Suppose that the fundamental beam is a Gaussian beam that is focused to a beam waist size of $w_0 = 100$ μm. What is the aperture distance? Under what conditions will the SHG efficiency increase quadratically with the length of the nonlinear crystal?

c. For the SFG using 1.064 μm and 532 nm input waves to generate a third-harmonic wave at 354.7 nm, calculate the type I phase-matching angle.
d. For the SFG to generate a third-harmonic wave at 354.7 nm, what are the possibilities for type II phase matching? Find the phase-matching angle.

9.5.11 KTP is a biaxial crystal of $mm2$ symmetry. It is a very popular crystal for nonlinear optical frequency conversion applications. Birefringent phase matching in such a biaxial crystal is normally accomplished in a plane normal to one of the three principal axes. For a propagation vector $\mathbf{k}$, the angle θ is that between $\hat{k}$ and $\hat{z}$, and the angle ϕ is that between $\hat{k}$ and $\hat{x}$, both defined in the same manner as those for a uniaxial crystal. Collinear phase matching for parametric generation in KTP with a pump wavelength at 527 nm can be achieved in any of the three principal planes: xy, yz, and zx. Use the data for KTP listed in Table 9.3 to plot the angle-tuning curves in the form of parametric wavelengths, within the transparency window of KTP up to 4.5 μm, as a function of the phase-matching angle for collinear phase matching in each of the three planes. Note that the tuning angle is ϕ for phase matching in the xy plane, but it is θ for phase matching in the yz or zx plane.

9.5.12 What is the grating period required for second-order quasi-phase matching? What duty factor has to be chosen for such a second-order grating in order to maximize the value of $|\chi_Q|$? What is this $|\chi_Q|$? How does it compare with that of the optimized first-order grating?

9.5.13 A PPKTP crystal is used for frequency doubling of an optical wave at 1.064 μm wavelength to its second harmonic at 532 nm. The grating period of this PPKTP is chosen for quasi-phase matching of this process. The properties of KTP are listed in Table 9.3. How should the waves be polarized for the largest nonlinear susceptibility under quasi-phase matching? What is the required first-order grating period at room temperature?

9.5.14 A first-order PPKTP crystal is pumped at $\lambda_3 = 532$ nm to generate a parametric signal at $\lambda_1 = 1.3$ μm. The properties of KTP are listed in Table 9.3.
a. What is the idler wavelength λ_2?
b. How should the pump wave be polarized for the largest nonlinear susceptibility under quasi-phase matching? What are the polarizations of the signal and idler waves, respectively?
c. What is the grating period required for quasi-phase matching at room temperature?
d. If the pump wavelength is at $\lambda_3 = 860$ nm instead, what is the idler wavelength λ_2? What is the required grating period? Compare this grating period with that obtained in (c).

9.5.15 Answer briefly the following questions regarding phase matching.
a. What are collinear and noncollinear phase-matching processes?

b. Discuss the physical mechanisms that can be utilized for collinear phase matching.
c. What is the best way to minimize the walk-off effect in a second-order nonlinear frequency converter?
d. Why is phase matching by temperature tuning also called noncritical phase matching?
e. What methods can be used for phase matching in a nonbirefringent material within a spectral region that is far away from resonances?
f. If phase matching is not possible, what is the longest crystal you should use for a second-harmonic generation process?

9.6.1 Consider sum-frequency generation in the low-efficiency limit.
a. Show that the optimum crystal length for the highest efficiency is equal to the coherence length of the interaction if there is a finite phase mismatch of Δk.
b. For a given interaction length l, what is the maximum acceptable phase mismatch for maintaining an efficiency higher than 50% of the peak efficiency at perfect phase matching?

9.6.2 Consider SFG with $\omega_3 = \omega_1 + \omega_2$ and DFG with $\omega_2 = \omega_3 - \omega_1$. Answer the following questions briefly without unnecessary elaboration.
a. In the low-efficiency limit, what is $I(\omega_3)$ in SFG as a function of ω_3, χ_{eff}, $I(\omega_1)$, $I(\omega_2)$, the interaction length l, and the phase mismatch Δk?
b. Sketch the dependence of $I(\omega_3)$ as a function of phase mismatch Δk. Identify the zeros.
c. With a given type of crystal and given input optical powers at ω_1 and ω_2, what can be done to maximize the output power at ω_3? Write down four steps that can be taken to accomplish this task.
d. How is the answer in (a) modified for DFG?
e. In the high-efficiency limit with a very strong pump beam at one frequency interacting with a weak pump beam at another frequency, sketch the evolution of the intensities of the pump beams and the signal beam as a function of the interaction length l for SFG.
f. Consider the same situation in (e) for DFG. What is the difference in the evolution of the beam intensities between the cases of DFG and SFG?

9.6.3 Three optical waves of the same intensity at the frequencies of ω_1, ω_2, and ω_3 with $\omega_1 + \omega_2 = \omega_3$ are sent together into a nonlinear optical crystal. If they propagate with $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3$, what decides or controls whether sum-frequency generation with the optical power being converted from ω_1 and ω_2 to ω_3 or difference-frequency generation with the optical power being converted in the opposite direction will take place?

9.6.4 Answer the following questions regarding the characteristics of a second-harmonic generator.

a. In the case of perfect phase matching, sketch the evolution of the fundamental and second-harmonic intensities as a function of distance in the low-efficiency limit. What is the dependence of the second-harmonic intensity on the distance?

b. Answer the questions in (a) for the high-efficiency limit.

c. If phase matching is not perfect with $\Delta k \neq 0$, how are the answers in (a) changed? Show the sketch.

d. In the presence of phase mismatch, how should the crystal length be chosen for the maximum second-harmonic output?

9.6.5 Very often, we use two nonlinear crystals in tandem as shown in Fig. 9.42 to increase the overall nonlinear conversion efficiency because of the limitation in the length of available crystals. Assume that both crystals are cut and oriented for perfect phase matching. Consider for simplicity the case of second-harmonic generation although your answers to the following questions can be easily extended to other second-order frequency generation processes. Consider also only the low-efficiency limit, which is usually the case when there is a need to use two crystals.

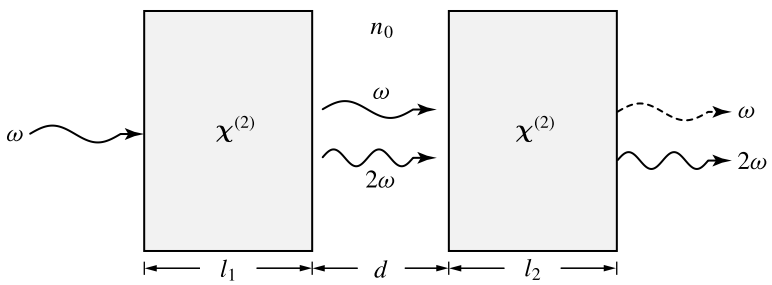


Figure 9.42 Second-harmonic generation with two nonlinear crystals in tandem.

a. If the medium between the crystals is isotropic and nondispersive, what is the effect of the distance d between the two crystals on the overall nonlinear conversion efficiency? Explain.

b. What is the dependence of the overall conversion efficiency on the lengths, l_1 and l_2, of the crystals and the distance d between them?

c. Discuss what can happen if the medium between the crystals is birefringent but nondispersive. What should you do in this case to ensure maximum conversion efficiency?

d. What can happen if the medium between the crystals is isotropic but is dispersive? What has to be done to ensure maximum conversion efficiency?

e. Answer the questions in (d) if the medium is nondispersive but is nonlinear with an intensity-dependent refractive index.

9.6.6 A mode-locked short optical pulse usually contains many equally spaced longitudinal mode frequencies. Consider a transform-limited Gaussian pulse that contains 100 longitudinal modes with the central mode at frequency ω_0 and the spacing between two adjacent modes being $\Delta\omega_L$. The Gaussian pulse has a FWHM pulsewidth Δt_{ps} in the time domain and a spectral width $\Delta\omega_{ps}$ in the frequency domain so that its time-domain intensity $I(t)$ and frequency-domain spectral intensity $I(\omega)$ are, respectively,

$$I(t) = I_0 \exp\left[-\left(\frac{t}{\Delta t_{ps}}\right)^2\right] \quad \text{and} \quad I(\omega) = I(\omega_0) \exp\left[-\left(\frac{\omega - \omega_0}{\Delta\omega}\right)^2\right]. \tag{9.287}$$

a. In a phase-matched second-harmonic generation process with this Gaussian pulse, how many longitudinal mode frequencies are found in the second-harmonic pulse? What is the mode spacing? Sketch the spectrum and identify the frequencies of these modes.
b. If the phase-matching spectral range is much larger than $\Delta\omega_{ps}$, what are the spectral width and the pulsewidth of the second-harmonic pulse?
c. If the phase-matching spectral range were smaller than $\Delta\omega_{ps}$, what would happen to the spectral width and the pulsewidth of the second-harmonic pulse?

9.6.7 $LiIO_3$ is an attractive material for second-harmonic generation because of its high nonlinear coefficients and its resistance to damage. Because its index of refraction changes little with temperature, phase matching is normally accomplished by angle tuning. It is a uniaxial crystal with its ordinary and extraordinary indices of refraction as a function of wavelength given by the following Sellmeier equations:

$$n_o^2 = 1 + \frac{2.40\lambda^2}{\lambda^2 - 0.022}, \tag{9.288}$$

$$n_e^2 = 1 + \frac{1.91\lambda^2}{\lambda^2 - 0.019}, \tag{9.289}$$

where λ is in micrometers. The crystal has group 6 symmetry, and its principal second-order nonlinear susceptibility is $\chi_{31}^{(2)} = \chi_{zxx}^{(2)}(2\omega = \omega + \omega) = 20$ pm V^{-1}. We want to generate second harmonics at 532 nm wavelength from a fundamental plane wave at 1.064 μm wavelength using a $LiIO_3$ crystal. Collinear phase matching is considered.

a. Sketch the proper crystal orientation, direction of beam propagation, and polarization directions of the fundamental and second-harmonic fields for phase-matched conversion. Calculate the phase-matching angle θ_{PM} and the walk-off angle ρ.
b. Calculate the value of χ_{eff} under the phase-matched condition.

c. For a crystal of $l = 5$ mm length, calculate the input fundamental intensity $I(\omega)$ at 1.064 μm that is required for 50% power conversion to the second harmonic at 532 nm.

d. Assume that the laser beams are circular Gaussian beams. If 50% power conversion efficiency is to be accomplished in such a manner that the walk-off in space at the output end is less than one-tenth of the fundamental beam diameter, how much total power is needed for the input fundamental beam?

9.6.8 Examine the effects of phase mismatch for optical parametric frequency converters. Compare them to the effects of phase mismatch for sum-frequency generators. What is the reason for the differences?

9.6.9 Compare the differences between an optical parametric frequency converter and an OPA in terms of their basic principles, operating characteristics, and applications.

9.6.10 Find the intensity gain of an OPA with a phase mismatch of Δk and compare the result to (9.137). Show that in the case when $\Delta k/2 > \kappa$, the parametric gain is approximately

$$G = 1 + \kappa^2 l^2 \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2}. \tag{9.290}$$

9.6.11 A wavelength-tunable OPA or OPO is normally operated with a pump at a fixed wavelength of λ_3 and a fixed input intensity of $I_3(0)$ while the signal and idler wavelengths, λ_1 and λ_2, respectively, are tuned together in opposite directions. In this process, other characteristic parameters, such as d_{eff} and the refractive indices of the crystal at the signal and idler wavelengths, of the device also vary accordingly. If such variations can be ignored, at what signal and idler wavelengths does a given wavelength-tunable OPA have the largest gain? What is the implication of this wavelength dependence of the parametric gain on the practical design and applications of OPAs and OPOs?

9.6.12 A crystal that has a large $\chi^{(2)}$ is used for nonlinear frequency conversion.

a. When two laser beams at frequencies ω_1 and ω_2 are sent together into the crystal, what determines whether you will see $\omega_1 + \omega_2$, or $\omega_1 - \omega_2$, or $2\omega_1$, or $2\omega_2$ at the output?

b. When three beams at three different frequencies, ω_1, ω_2, and ω_3, with $\omega_3 = \omega_1 + \omega_2$, propagate simultaneously in the crystal, what determines whether the sum-frequency process of $\omega_1 + \omega_2 \rightarrow \omega_3$ or the difference-frequency process of $\omega_3 - \omega_2 \rightarrow \omega_1$ will occur?

9.6.13 The nonlinear crystal $AgGaS_2$ is a good material for parametric generation of infrared frequencies. It has $\bar{4}2m$ symmetry and is transparent in the infrared spectral range. We can assume Kleiman's symmetry so that all of its nonvanishing $\chi^{(2)}$ elements have equal magnitude. Its refractive indices can be described

by the following Sellmeier equations:

$$n_o^2 = 5.728 + \frac{0.241\,07}{\lambda^2 - 0.087\,03} - 0.002\,10\lambda^2, \tag{9.291}$$

$$n_e^2 = 5.497 + \frac{0.202\,59}{\lambda^2 - 0.130\,70} - 0.002\,33\lambda^2, \tag{9.292}$$

where λ is in micrometers. An OPO can be constructed using this crystal and pumped with a laser beam at 1.064 μm to generate wavelength-tunable infrared light in the longer-wavelength range. The pump beam propagates through the crystal at an angle θ with respect to the crystal z axis in a plane at an angle ϕ with respect to the x axis.

a. Find χ_{eff} for type I phase matching as a function of θ, ϕ, and the nonvanishing elements of $\chi^{(2)}$.
b. Find χ_{eff} for type II phase matching as a function of θ, ϕ, and the nonvanishing elements of $\chi^{(2)}$.
c. If the signal is at 4 μm wavelength, what is the wavelength of the idler?
d. Find the phase-matching angle for type I phase matching to generate the 4-μm signal.
e. If we rotate the crystal by $\pm 1^\circ$ around the phase-matching angle for the 4-μm signal, what are the wavelength tuning ranges for the signal and the idler, respectively?

9.6.14 An OPO is constructed using the uniaxial nonlinear crystal $AgGaS_2$ and pumped with a laser beam at 1.064 μm to generate wavelength-tunable infrared light in the wavelength range between 3 and 5 μm. A linearly polarized pump beam propagates through the crystal at an angle θ with respect to the crystal z axis in a plane at an angle ϕ with respect to the x axis. Consider using only type I phase matching. Use the data for $AgGaS_2$ given in Problem 9.6.13 to answer the following questions.

a. What is the wavelength range of the idler?
b. How do you orient the direction of polarization of your pump beam for maximum efficiency?
c. It is desired that only one crystal be used for the entire signal wavelength range of 3–5 μm. Wavelength tuning in this range is to be accomplished by tuning the angle of the crystal. How do you cut your crystal so that the length of the crystal needed is minimized for a given pump beam cross section? Sketch for clarity.
d. Following (c), what is the range of tuning angle for the desired wavelength-tuning range?
e. What is the change in nonlinear conversion efficiency across the signal wavelength range?

9.6.15 An $AgGaS_2$ OPO with a pump beam at $\lambda_3 = 1.053$ μm can cover a wavelength range from 1 to 12 μm by angle tuning. The internal small-signal gain of the OPO is determined by $\kappa^2 l^2$ for a crystal of length l, where κ is given in (9.133). The effective nonlinear susceptibility d_{eff} depends on the signal and idler wavelengths through its dependence on the tuning angle. For a given crystal length and a given pump intensity, the gain of the OPO thus varies with the signal and idler wavelengths through its dependence on d_{eff} and refractive indices in addition to its explicit dependence on λ_1 and λ_2 as seen in (9.133). Collinear phase matching is considered. Use the data given in Table 9.3 for $AgGaS_2$.

a. Plot the angle-tuning curves in the form of parametric wavelengths versus phase-matching angle for both type I and type II phase matching.
b. Find d_{eff} as a function of angles θ and ϕ for both types of phase matching, where θ is the angle between the propagation vector $\mathbf{k}$ and the crystal optical axis $\hat{z}$ and ϕ is that between $\mathbf{k}$ and the principal axis $\hat{x}$. For each type of phase matching, maximize the value of $|d_{\text{eff}}|$ by properly choosing the value for ϕ.
c. Plot the maximum value of $|d_{\text{eff}}|$ as a function of signal and idler wavelengths for both types of phase matching.
d. Plot the small-signal gain $\kappa^2 l^2$, normalized to its peak value, for the OPO as a function of signal and idler wavelengths for both types of phase matching. Compare these curves with those obtained in (c).
e. Compare type I and type II phase matching in terms of wavelength coverage, tuning, and efficiency.

9.6.16 The angle-tuning curves for a $LiNbO_3$ OPO pumped at $\lambda_3 = 527$ nm are shown in Fig. 9.11. The effective nonlinear susceptibilities are given respectively in (9.86) and (9.88) for type I and type II phase matching. As described in Problem 9.6.15, the internal small-signal gain of the OPO is $\kappa^2 l^2$ for a crystal of length l, where κ is given in (9.133). Answer the following questions for an angle-tuned $LiNbO_3$ OPO pumped at 527 nm with collinear phase matching. Use the data given in Table 9.3 for $LiNbO_3$.

a. Plot the maximum value of $|d_{\text{eff}}|$ as a function of signal and idler wavelengths for both types of phase matching.
b. Plot the small-signal gain $\kappa^2 l^2$, normalized to its peak value, for the OPO as a function of signal and idler wavelengths for both types of phase matching. Compare these curves with those obtained in (a).
c. Compare type I and type II phase matching in terms of wavelength coverage, tuning, and efficiency.

9.6.17 Answer the questions in Problem 9.6.16 for a temperature-tuned $LiNbO_3$ OPO pumped at $\lambda_3 = 527$ nm. The temperature-tuning curves are shown in Fig. 9.13.

9.6.18 In this problem, we consider a $LiNbO_3$ OPO pumped at 527 nm such as those considered in Problems 9.6.16 and 9.6.17, but with quasi-phase matching with

a PPLN crystal that has a first-order grating of 50% duty factor. Tuning of the parametric wavelengths is accomplished by varying the grating period in a fanned structure as shown in Fig. 9.14(*c*). Use the data given in Table 9.3 for $LiNbO_3$ to answer the following questions.

a. Plot the tuning curve in the form of parametric wavelengths versus the phase-matching grating period.
b. Plot the maximum value of $|d_{\text{eff}}|$ as a function of signal and idler wavelengths.
c. Plot the small-signal gain $\kappa^2 l^2$, normalized to its peak value, for the OPO as a function of signal and idler wavelengths. Compare these curves with those obtained in (b).
d. Compare this OPO with quasi-phase matching to the OPOs with angle tuning and temperature tuning that are considered in Problems 9.6.16 and 9.6.17, respectively, in terms of wavelength coverage, tuning, and efficiency.

9.6.19 Answer the following questions regarding frequency converters.

a. What are the three most important factors to consider for efficient operation of second-order nonlinear optical frequency converters? List them in descending order according to their relative importance.
b. For a sum-frequency generator operating at the low-efficiency limit, how does the sum-frequency signal depend on the pump beam intensities, the crystal length, the effective nonlinear susceptibility, and the signal wavelength, respectively? Assume perfect phase matching.
c. What are the usual arrangements employed for efficient generation of third-harmonic and fourth-harmonic frequencies from a fundamental frequency?
d. What are the advantages and disadvantages of a singly resonant OPO in comparison to a doubly resonant OPO? Are most practical OPOs of singly resonant type or doubly resonant type?
e. How are the signal and idler frequencies of an OPO tuned in practice?

9.7.1 Consider a one-beam optical Kerr effect in an isotropic medium.

a. Show that if an optical wave is linearly or circularly polarized, its polarization state remains unchanged under the optical Kerr effect. For the intensity-dependent index of refraction expressed in the form of $n = n_0 + n_2 I$, find n_2 for a circularly polarized wave and compare it to that of a linearly polarized wave of the same intensity.
b. The field of an elliptically polarized wave can be expressed as a linear superposition of two circularly polarized components of unequal magnitudes as $\boldsymbol{\mathcal{E}} = \mathcal{E}_+ \hat{e}_+ + \mathcal{E}_- \hat{e}_-$ with $|\mathcal{E}_+| \neq |\mathcal{E}_-|$. Show that the two circularly polarized components of unequal magnitudes experience different intensity dependencies due to a field-dependent circular birefringence of

$$\Delta n_{\text{c}} = \Delta n_+ - \Delta n_- = -\frac{3\chi^{(3)}_{1221}}{n_0}\left(|\mathcal{E}_+|^2 - |\mathcal{E}_-|^2\right). \tag{9.293}$$

What is the angle of rotation of the elliptical polarization after the wave propagates through the medium over a length l?

9.7.2 Consider a cubic crystal with principal axes $\hat{x}$, $\hat{y}$, and $\hat{z}$. It has a length l in the z direction. The optical waves being considered propagate in the z direction so that they are polarized in the xy plane. The nonvanishing elements of the third-order nonlinear susceptibility of the cubic crystal have the following forms: $\chi^{(3)}_{1111} > \chi^{(3)}_{1122}, \chi^{(3)}_{1212}, \chi^{(3)}_{1221} > 0$. The crystal is slightly stressed to become slightly birefringent in its linear optical property so that $\delta n = n_{0y} - n_{0x} > 0$ and $n_0 \gg \delta n$, but its nonlinear optical susceptibilities are not changed by stress.

a. A beam at a frequency ω_0 that is linearly polarized at an angle θ with respect to the x axis at $z = 0$ is launched into the crystal. Its output polarization state at $z = l$ can be used to deduce stress-induced birefringence in the crystal. If we take $\theta = 45^\circ$ and find that the beam is circularly polarized at $z = l$, what is the minimum value of δn?

b. Because of the stress-induced birefringence, the polarization state of the output beam depends sensitively on the angle θ. It is possible to get rid of this problem by an optical Kerr effect induced by another strong pump beam at a frequency ω that is different from ω_0. If such a beam that is linearly polarized in the x direction is launched, what is the minimum required intensity of this beam for the optical field at frequency ω_0 to be always linearly polarized at $z = l$ no matter what value θ has?

c. If the pump field at ω is linearly polarized in the y direction instead, what is the minimum intensity required to accomplish the effect in (b)?

9.7.3 The optical Kerr effect changes the divergence of a Gaussian optical beam significantly when the effective focal length of a Kerr lens is on the order of the Rayleigh range of the beam. Show that by making the length l of the Kerr medium the largest allowed under the thin-lens condition, the minimum power required for $f_K = z_R$ is a constant that is independent of the focusing condition of the beam. For the silica Kerr lens described in Example 9.15, what is the minimum peak power of the pulses to reach this condition? What is the corresponding average power? Note that the beam waist is located at the incident surface of the silica plate.

9.7.4 For a Gaussian beam at a given power with a given beam waist radius w_0, the effective focal length of a Kerr lens depends on the location z of the Kerr lens with respect to the beam waist, which is located at the origin of the z axis. As a consequence, both the beam waist radius w_{0K} and the divergence of the beam after the Kerr lens vary with the location of the Kerr lens. Use the relations for Gaussian beam focusing in Problem 1.7.5 to answer the following questions.

a. Find $f_K(z)$ and $w_{0K}(z)$ as a function of location z of the Kerr lens.
b. Plot $f_K(z)/f_K(0)$ and $w_{0K}(z)/w_0$ as a function of z/z_R for the three values of $f_K(0)/z_R = 0.1$, 1, and 10.

9.7.5 Using an isotropic nonlinear medium such as a cell containing liquid CS_2, design a Kerr shutter to gate the signal carried by an optical wave at a wavelength λ. The CS_2 cell has a length l and is pumped by a short gating pulse at a wavelength λ'.
a. It is desired that the signal be gated by the gating pulse such that its transmission is synchronized to the arrival of the gating pulse and lasts only within the duration of the pulse. Sketch your setup and describe how you arrange the polarization of the beams for your setup to work.
b. Find the peak intensity of your gating pulse in terms of the linear refractive index, n_0, and the nonlinear refractive index, n_2, of CS_2.

9.7.6 It is possible to use $\chi^{(3)}$ processes to shorten optical pulses. This objective can be accomplished through processes that involve either $\chi^{(3)'}$ or $\chi^{(3)''}$. Because the temporal pulsewidth is related to its spectral width through the Fourier transform, it is generally necessary to broaden the pulse spectrum when we start with a transform-limited pulse. However, broadening the spectrum does not automatically result in a shortened pulse before chirping in the pulse is removed.
a. Show that a $\chi^{(3)'}$ process, such as self-phase modulation, broadens the pulse spectrum but does not by itself shorten the pulsewidth.
b. What else is needed for shortening the pulse by using a $\chi^{(3)'}$ process?
c. Show that a $\chi^{(3)''}$ process, such as absorption saturation, can reduce the pulsewidth without the help of other processes.
d. Does a $\chi^{(3)''}$ process broaden the pulse spectrum? Discuss mathematically.

9.7.7 An optical wave at a frequency ω propagates in an isotropic medium that has a third-order nonlinear susceptibility $\chi^{(3)}(\omega = \omega + \omega - \omega)$. If the material has a bandgap larger than $\hbar\omega$ but smaller than $2\hbar\omega$, two-photon absorption is possible. Discuss the nonlinear susceptibility conditions for the beam to propagate through the medium with the possibilities of (a) self-focusing, (b) self-defocusing, and (c) attenuation due to two-photon absorption, respectively.

9.7.8 Answer the following questions regarding nonlinear optical phase modulation.
a. What is the major problem preventing the scaling up of power of a solid-state laser such as Nd : YAG by continuously increasing the length of the laser crystal?
b. What is the solution to the problem in (a)?
c. Describe the principle of an optical power limiter that utilizes a nonlinear optical effect to protect a detector from excessively high input light intensities in a Gaussian beam.

d. Sketch the distribution of the new frequencies generated by self-phase modulation as a function of time when a short optical pulse propagates through a medium with $n_2 > 0$. What is the distribution if $n_2 < 0$?

9.8.1 A dispersive bistable optical device consists of a Fabry–Perot cavity filled with a purely dispersive optical Kerr medium. Consider its operation under the condition that $|\varphi| < 1$.

a. Show that the conditions for bistability in such a device are those given in (9.172) and (9.173).

b. Show that the threshold input intensity, I_{th}, required for this device to be able to reach its bistability is that given in (9.174).

9.8.2 From the characteristics of a dispersive bistable optical device as shown in Fig. 9.28, we find that when operating in its bistable state with $a = F^2\varphi_0^2/\pi^2 > 3$, the device has an up-transition point at an input intensity $I_{\text{in}}^{\text{up}}$ and a down-transition point at $I_{\text{in}}^{\text{down}}$. The bistability range is $\Delta I_{\text{in}} = I_{\text{in}}^{\text{up}} - I_{\text{in}}^{\text{down}}$. Consider its operation under the condition that $|\varphi| < 1$ and $a \gg 3$.

a. Show that the up-transition point occurs at

$$I_{\text{in}}^{\text{up}} \approx \left(\frac{4a}{27} + \frac{1}{3} + \frac{1}{3a}\right)\frac{3\sqrt{3a}}{8} I_{\text{th}}, \tag{9.294}$$

and the down-transition point occurs at

$$I_{\text{in}}^{\text{down}} \approx \left(1 - \frac{1}{3a}\right)\frac{3\sqrt{3a}}{8} I_{\text{th}}. \tag{9.295}$$

b. Show that at the up-transition point, the output intensity of the device makes the following jump from a low level to a high level:

$$\left(I_{\text{out}}^{\text{up}}\right)_{\text{low}} \approx \left(\frac{1}{3} + \frac{1}{2a}\right)\frac{F^2}{F_0^2}\frac{3\sqrt{3a}}{8} I_{\text{th}} \Rightarrow \left(I_{\text{out}}^{\text{up}}\right)_{\text{high}} \approx \left(\frac{4}{3} - \frac{1}{a}\right)\frac{F^2}{F_0^2}\frac{3\sqrt{3a}}{8} I_{\text{th}}. \tag{9.296}$$

c. Show that at the down-transition point, the output intensity of the device makes the following jump from a high level to a low level:

$$\left(I_{\text{out}}^{\text{down}}\right)_{\text{high}} \approx \left(1 - \frac{1}{2a}\right)\frac{F^2}{F_0^2}\frac{3\sqrt{3a}}{8} I_{\text{th}} \quad \Rightarrow \quad \left(I_{\text{out}}^{\text{down}}\right)_{\text{low}} \approx \frac{1}{a}\frac{F^2}{F_0^2}\frac{3\sqrt{3a}}{8} I_{\text{th}}. \tag{9.297}$$

9.8.3 It is desired that the dispersive bistable device described in Example 9.16 be operated with a bistability input power range of $\Delta P_{\text{in}} = 1$ mW between the up-transition and down-transition points.

a. What value of the biased phase φ_0 should be chosen?

b. What are the required input powers at the two transition points?

c. What are the output powers at the two transition points?

9.8.4 Find the threshold intensity for the dispersive bistable device described in Example 9.16 if the reflectivities of both mirrors are increased to $R = 99.5\%$. What is the threshold power if the Gaussian input beam is still focused to a spot size of $w_0 = 20$ μm? Compare the results with those found in Example 9.16.

9.8.5 Show that the condition for bistability in a Fabry–Perot cavity filled with a saturable absorber is $C_0 > 8$ as given in (9.178) under the assumptions that $2kl = 2m\pi$ and $\alpha l \ll 1$.

9.8.6 An absorptive bistable device has the structure of a vertical cavity that consists of a doped GaAs semiconductor saturable absorber layer between two symmetric GaAs/AlGaAs DBR mirrors of reflectivity R. The unsaturated absorption coefficient is $\alpha_0 = 8 \times 10^4$ m^{-1} at $\lambda = 850$ nm.

a. If the length of the absorber layer is $l = 0.5$ μm, what is the required reflectivity R for bistability?

b. If the reflectivity of the DBR mirrors is limited to $R = 99\%$, what should the length of the absorber layer be in order for the device to function bistably?

9.9.1 Stimulated Raman scattering and stimulated Brillouin scattering have some common features but also many fundamental and practical differences. Discuss the similarities and differences between stimulated Raman scattering and stimulated Brillouin scattering in terms of their physical mechanisms, geometric characteristics, spectral characteristics, and gain.

9.9.2 Consider stimulated Raman scattering in a very long interaction medium such as an optical fiber of hundreds of meters in length. Discuss the fundamental differences among pumping with (a) a CW laser beam, (b) a nanosecond laser pulse, (c) a picosecond laser pulse, and (d) a femtosecond laser pulse. Assume that the CW laser intensity is strong enough to generate at least the first-order Stokes signal and that the peak intensities of the pulses are at least as strong as that of the CW beam.

9.9.3 In this problem, we consider Raman amplification in a codirectional configuration in the case when $\alpha_\mathrm{p} = \alpha_\mathrm{S} = \alpha$.

a. When the medium has a finite absorption coefficient of $\alpha \neq 0$, the Manley–Rowe relation given in (9.81) is no longer valid. Instead, show, using the coupled equations given in (9.188) and (9.189), that we have

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_\mathrm{S}}{\omega_\mathrm{S}}\mathrm{e}^{\alpha z}\right) = -\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{I_\mathrm{p}}{\omega_\mathrm{p}}\mathrm{e}^{\alpha z}\right). \tag{9.298}$$

Therefore,

$$\frac{I_\mathrm{S}(z)}{\omega_\mathrm{S}}\mathrm{e}^{\alpha z} + \frac{I_\mathrm{p}(z)}{\omega_\mathrm{p}}\mathrm{e}^{\alpha z} = \frac{I_\mathrm{S}(0)}{\omega_\mathrm{S}} + \frac{I_\mathrm{p}(0)}{\omega_\mathrm{p}}, \tag{9.299}$$

which is a constant independent of z.

b. Using the relation in (9.299), show that the coupled equations in (9.188) and (9.189) have the following exact analytical solution:

$$I_S(l) = \frac{\omega_p I_S(0) + \omega_S I_p(0)}{\omega_p I_S(0) + \omega_S I_p(0) \exp(-g l_{eff})} I_S(0) e^{-\alpha l}, \tag{9.300}$$

$$I_p(l) = \frac{\omega_S I_p(0) + \omega_p I_S(0)}{\omega_S I_p(0) + \omega_p I_S(0) \exp(g l_{eff})} I_p(0) e^{-\alpha l}, \tag{9.301}$$

where l_{eff} is that given in (9.192) with $\alpha_p = \alpha$ and

$$g = \tilde{g}_R \left[I_p(0) + \frac{\omega_p}{\omega_S} I_S(0) \right]. \tag{9.302}$$

c. Show that, for any value of α, the pump and the Stokes signal intensities have the following relation:

$$\frac{I_S(l)}{I_p(l)} = \frac{I_S(0)}{I_p(0)} \exp(g l_{eff}). \tag{9.303}$$

9.9.4 By using the relations in (9.196), (9.197), and (9.300) and by taking the realistic assumption that $I_p(0) \gg I_S^{eff}(0) = I_{S,th}^{eff}(0)$, show that the efficiency of a Raman Stokes generator of length l in the case when $\alpha_p = \alpha_S = \alpha$ is that given by (9.199).

9.9.5 A low-loss single-mode silica fiber has a short length such that $\alpha l \ll 1$. Therefore, the linear absorption loss in the fiber can be neglected in this problem. The fiber is pumped with an optical beam at $\lambda_p = 1$ μm. The Raman gain peak of this fiber appears at a frequency shift of 460 cm^{-1}. Plot the Raman conversion efficiency η_R defined in (9.199) as a function of the pump ratio r for Raman Stokes generation in the fiber. Examine the behavior of η_R around the threshold.

9.9.6 Because of the typical long interaction length in an optical fiber, a nonlinear optical phenomenon such as stimulated Raman scattering can become important even though the nonlinear susceptibility might not be very large. The Raman spectra of oxide glasses, such as various silica, germania, and phosphorous glasses, that are used in the fabrication of optical fibers show a broad band of frequencies rather than discrete Raman lines because of the amorphous nature of glasses. Consider a germania-doped silica fiber that has a Raman spectral peak at a frequency shift of 440 cm^{-1}.

a. A high-power pulsed laser beam at $\lambda = 1$ μm wavelength is sent through such a fiber that is long enough to generate up to the fifth-order Stokes signal without completely depleting the pump laser power. Sketch the expected spectrum of the output at the exit end of the fiber. Identify the wavelengths of the peaks in the spectrum.

b. Would you expect anti-Stokes lines to be seen? Explain.

c. Consider only the coupling of the pump beam and the first Stokes signal. Write down the equations in the slowly varying amplitude approximation

to describe the wave propagation while ignoring the effects of optical-field-induced birefringence and absorption losses. Does phase matching need to be considered for this coupling? Why?

d. If depletion of the pump beam is negligible throughout the entire fiber, show that Raman amplification of the Stokes signal is

$$\frac{P_S(l)}{P_S(0)} = \exp\left(\frac{\tilde{g}_R P_p l}{\mathcal{A}_{eff}}\right), \tag{9.304}$$

where P_p is the input pump power, l is the length of the fiber, $\mathcal{A}_{eff}$ is the effective cross-sectional area of the fiber core, and $\tilde{g}_R$ is the Raman gain factor.

e. If the pump power is subject to attenuation due to linear absorption in the fiber but is not subject to appreciable depletion due to stimulated Raman scattering, how should the expression in (9.304) be modified for Raman amplification in this situation?

9.9.7 With the same input Stokes signal power and the same parameters for the fiber Raman amplifier described in Example 9.19, what is the output signal power if the pump power is doubled to $P_p = 480$ mW? What is it if the pump power is cut in half to $P_p = 120$ mW?

9.9.8 In this problem, we consider the effect of the changes in some key parameters of the fiber Raman amplifier considered in Example 9.19. In each case, only the single indicated parameter is changed while all other parameters are kept the same as those described in Example 9.19. Find the required pump power for each of the following cases: (a) the fiber absorption coefficient is reduced by half to $\alpha = 0.1$ dB km^{-1}, (b) the length of the fiber is doubled to $l = 50$ km, (c) the Raman gain factor is doubled to $\tilde{g}_R = 1.29 \times 10^{-13}$ m W^{-1}. Compare the results to the pump power found in Example 9.19. Which parameter change has the most significant effect?

9.9.9 Nonlinear optical effects, such as stimulated Raman and Brillouin scattering, can be troublesome problems that limit the capability of fiber-optic communication systems. However, they can also be used in certain situations to our advantage. For example, optical amplifiers based on stimulated Raman gain have been developed for amplifying optical signals in fiber communication systems. We consider this application in this problem. The amplifier consists of an optical fiber of length l, pumped by an optical beam at frequency ω_p from the left input end at $z = 0$ as shown in Fig. 9.43. The optical signal has a carrier frequency at ω_S that matches the first Raman Stokes frequency down-shifted from ω_p for $\omega_S = \omega_p - \Omega_R$, where Ω_R is the peak Raman resonance frequency. It has been shown that if the pump is a CW beam, bidirectional amplification of the signal is possible. Assume that the attenuation coefficient α of the fiber is the same for pump and signal frequencies.

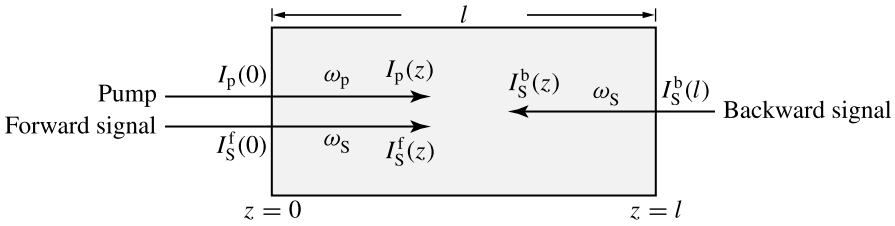


Figure 9.43 Bidirectional Raman amplification.

a. Consider the pump and the forward-propagating signal only. Ignore the dispersion in the refractive index between ω_p and ω_S. Show that the Stokes signal and the pump intensities are described by the coupled differential equations given in (9.188) and (9.189).
b. Assume that the fiber is lossless with $\alpha = 0$ but take pump depletion into consideration. Find the total Stokes signal gain $G_R^f = I_S^f(l)/I_S^f(0)$.
c. Assume no pump depletion due to the Raman effect but take fiber attenuation for both pump and signal into consideration. Find the total Stokes signal gain defined in (b).
d. Now consider the amplification of the backward-propagating signal. The total Stokes signal gain for this backward-propagating signal is defined as $G_R^b = I_S^b(0)/I_S^b(l)$, where $I_S^b(l)$ is the input intensity of the backward-propagating signal and $I_S^b(0)$ is its output intensity. It can be easily seen that if fiber loss is ignored, G_R^b for backward amplification is the same as G_R^f for forward amplification found in (b). The situation is less clear when fiber loss is considered. Find the Stokes signal gain G_R^b for backward amplification under the same assumptions taken in (c) and compare it to the forward Stokes signal gain obtained in (c).

9.9.10 A germania-doped silica fiber has an effective core area of $\mathcal{A}_{eff} = 2.8 \times 10^{-11}$ m^2. Its loss at the optical wavelength of 1.064 μm is about 1 dB km^{-1}, and its group-velocity dispersion at this wavelength is about 40 ps km^{-1} nm^{-1}. It has a strong Raman gain peak at a frequency shift of 460 cm^{-1}. A train of optical pulses at 1.064 μm wavelength at a repetition rate of 76 MHz is sent into the fiber. Stimulated Raman Stokes signals are observed. The Raman gain coefficient under these conditions is $\tilde{g}_R \approx 1 \times 10^{-13}$ m W^{-1}. The threshold for stimulated Raman scattering is at about

$$\frac{\tilde{g}_R P_p l_{eff}}{\mathcal{A}_{eff}} = 16, \tag{9.305}$$

where P_p is the peak power of the pulses and l_{eff} is the effective interaction length of the stimulated Raman scattering process.

a. What is the wavelength of the first Raman Stokes signal?
b. Show that for pulses in the range of about 10 ps to 1 ns at the same 76-MHz repetition rate, the average power $\overline{P}$ of the pulse train for the Raman threshold is independent of pulse duration. What is the value of this threshold average power, P_{th}?
c. It is experimentally observed that P_{th} changes when the pulses become as short as 3 ps. Do you think it increases or decreases? Give a possible explanation.
d. What do you expect to happen when the pulses get substantially shorter, say, down to about 100 fs?
e. What is P_{th} when the input is a CW beam instead of a pulse train? What is the minimum length of fiber for the stimulated Raman scattering to reach threshold in this case?

9.9.11 In this problem, we consider Brillouin amplification that occurs only in a contradirectional configuration. The results obtained in the following apply to contradirectional Raman amplification as well if $\tilde{g}_{\text{B}}$ is replaced by $\tilde{g}_{\text{R}}$. We consider only the case when $\alpha_{\text{p}} = \alpha_{\text{S}} = 0$ so that exact analytical solution can be found for the coupled equations given in (9.201) and (9.202).

a. Show that

$$\frac{I_{\text{p}}(z)}{\omega_{\text{p}}} - \frac{I_{\text{S}}(z)}{\omega_{\text{S}}} = \frac{I_{\text{p}}(0)}{\omega_{\text{p}}} - \frac{I_{\text{S}}(0)}{\omega_{\text{S}}}, \tag{9.306}$$

which is a constant independent of z.

b. By using the relation in (9.306), show that

$$\frac{I_{\text{S}}(0)}{I_{\text{p}}(0)} = \frac{I_{\text{S}}(l)}{I_{\text{p}}(l)} \exp(gl), \tag{9.307}$$

where

$$g = \tilde{g}_{\text{B}} \left[I_{\text{p}}(0) - \frac{\omega_{\text{p}}}{\omega_{\text{S}}} I_{\text{S}}(0) \right] = \tilde{g}_{\text{B}} \left[I_{\text{p}}(l) - \frac{\omega_{\text{p}}}{\omega_{\text{S}}} I_{\text{S}}(l) \right]. \tag{9.308}$$

With given values of $I_{\text{p}}(0)$ and $I_{\text{S}}(l)$ as the boundary conditions for a contradirectional amplifier, the solutions for $I_{\text{S}}(0)$ and $I_{\text{p}}(l)$ cannot be explicitly expressed in terms of $I_{\text{p}}(0)$ and $I_{\text{S}}(l)$ but can be found from the relations in (9.307) and (9.308).

9.9.12 By using the relations in (9.208), (9.307), and (9.308), show that the reflectivity, R_{B}, of a Brillouin generator defined in (9.210) can be found from the relation in (9.211) in the case when $\alpha_{\text{p}} = \alpha_{\text{S}} = 0$.

9.9.13 A low-loss single-mode silica fiber has a short length such that $\alpha l \ll 1$. Therefore, linear absorption loss in the fiber can be neglected in this problem. The

fiber is pumped with an optical beam at $\lambda_p = 1$ μm. Plot the Brillouin conversion efficiency R_B given in (9.211) as a function of the pump ratio r for Brillouin Stokes generation in the fiber. Examine the behavior of R_B around the threshold. Discuss the applicability of the approximate relation given in (9.212) by comparing it to the accurate plot.

9.9.14 A single-mode silica fiber has an attenuation coefficient of 0.3 dB km^{-1} and an effective cross-sectional area of 50 μm^2 at 1.3 μm optical wavelength. A CW optical beam at this wavelength is launched into the fiber. The Raman gain peak of this fiber appears at a frequency shift of 460 cm^{-1}. The Raman and Brillouin gain factors have the characteristics described in Section 9.9 for silica fibers.

a. If the fiber has a length of 100 km and the optical beam has a linewidth of 10 MHz, what are the critical powers of the beam that reach the Raman and Brillouin thresholds, respectively, in this fiber? What is the maximum power of the beam that can be transmitted through this fiber?

b. How do the answers of the questions in (a) vary if the fiber length varies between 1 and 100 km but the linewidth of the beam remains at 10 MHz? Plot them as functions of the fiber length.

c. How do the answers of the questions in (a) vary if the fiber length is fixed at 100 km but the linewidth of the optical beam varies between 1 MHz and 100 GHz? Plot them as functions of the linewidth of the optical beam.

9.9.15 Suppression of stimulated Brillouin scattering in a Raman amplifier or generator can be accomplished by using a pump that has a sufficiently broad linewidth to raise the Brillouin threshold pump power. Find the respective pump linewidths required to suppress the competition from stimulated Brillouin scattering for (a) the fiber Raman amplifier described in Example 9.19 and (b) the fiber Raman generator described in Example 9.20.

9.10.1 Discuss the advantages of using waveguides for nonlinear optical devices.

9.11.1 Manley–Rowe relations exist for nonlinear processes that take place in waveguide structures.

a. Find the general form of the Manley–Rowe relations that are equivalent to (9.68) and (9.69) for parametric second-order interactions in a multimode waveguide. What is the physical meaning of such relations? Write down the form of such relations in the special case when each frequency component consists of only one waveguide mode.

b. Answer the questions in (a) for second-harmonic generation in a waveguide.

9.11.2 The conversion efficiency of the PPLN waveguide second-harmonic generator described in Example 9.23 is compared to that of the bulk PPLN second-harmonic generator described in Example 9.12(d). An output second-harmonic power of $P_{2\omega} = 1$ mW is desired.

a. What are the required input fundamental power P_ω and the conversion efficiency η_{SH} if a PPLN waveguide of length $l = 1$ cm is used? What are they if a bulk PPLN crystal of the same length is used?
b. What are the required input fundamental power P_ω and the conversion efficiency η_{SH} for the PPLN waveguide second-harmonic generator and for the bulk PPLN second-harmonic generator, respectively, if the lengths of both devices are doubled to $l = 2$ cm?

9.11.3 Find the input power of the fundamental wave required for the PPLN second-harmonic generator described in Example 9.23 to have a conversion efficiency of $\eta_{SH} = 99\%$. By doubling the length of the PPLN waveguide, the efficiency can be increased at the same input power, or the input power can be reduced for the same efficiency. What is the increased efficiency if the input power is kept unchanged while the waveguide length is doubled? What is the reduced input power for an efficiency of 99% if the waveguide length is doubled?

9.11.4 In this problem, we consider the possibility of SFG and SHG in a nonbirefringent optical fiber.
a. How does the efficiency of SFG and SHG depend on the interaction length and the intensities of the component frequencies?
b. From (a), what potential advantages does an optical fiber offer to SFG in comparison to a bulk material?
c. It might be difficult, if not impossible, to carry out SFG or SHG in a nonbirefringent optical fiber. What are the fundamental difficulties with nonbirefringent glass fibers that could prevent the realization of the great advantages found in (b)?
d. However, contrary to the common expectations considered in (c), SHG with high efficiencies have been observed in glass fibers under certain conditions. What are the possible mechanisms that might be responsible for this effect?

9.11.5 In a single-mode birefringent glass optical fiber, it is possible to phase match for parametric generation with a single pump frequency ω_p to generate parametric Stokes and anti-Stokes frequencies, ω_S and ω_{AS}, respectively, with $\Omega = \omega_{AS} - \omega_p = \omega_p - \omega_S$. This is accomplished by compensating material dispersion with birefringence. Assume that x and y are the slow and fast axes, respectively, of the birefringent fiber, i.e., $n_x > n_y$, where n_x and n_y are the effective refractive indices including the waveguide geometry and the material property. The group index of refraction and the dispersion are defined, respectively, as

$$N_i = \frac{\mathrm{d}}{\mathrm{d}\omega}(\omega n_i) = n_i - \lambda \frac{\mathrm{d}n_i}{\mathrm{d}\lambda} \tag{9.309}$$

and

$$D = \omega \frac{\mathrm{d}^2}{\mathrm{d}\omega^2}(\omega n_i) = \lambda^2 \frac{\mathrm{d}^2 n_i}{\mathrm{d}\lambda^2}, \tag{9.310}$$

where i is x or y. The birefringence of the fiber is defined as $B = n_x - n_y$. Assume that $B \approx N_x - N_y$ and $D = D_x = D_y$ for simplicity, which are good approximations.

a. What is the nonlinear susceptibility that is responsible for this parametric process? What are its nonvanishing elements?
b. Write down the general expression for the phase mismatch in this process.
c. Based on your answer to (a), how many different parametric processes are possible through different combinations of the polarization directions of the participating fields? Write them down.
d. Among the processes you identified in (c), which have the pump polarized along a birefringent axis and rely on the birefringence for phase matching? Derive the frequency shift Ω in terms of B and D.
e. Among the processes identified in (c), there are also processes in which the pump has components along both birefringent axes. Derive the frequency shift Ω for these processes in terms of B and D.

(See Morgan, P. N. and Liu, J. M., "Parametric four-photon mixing followed by Raman scattering with optical pulses in birefringent optical fibers," *IEEE Journal of Quantum Electronics* **27**(4): 1011–1021, Apr. 1991.)

9.12.1 Show that the transmittance of an all-optical Mach–Zehnder interferometer using two Y-junction waveguides as shown in Fig. 9.35(*a*) is that given in (9.261). Show also that the transmittance of an all-optical Mach–Zehnder interferometer using two directional couplers as shown in Fig. 9.35(*b*) is that given in (9.262).

9.12.2 An ultrafast all-optical gate based on the three-input, symmetric Mach–Zehnder interferometer shown in Fig. 9.36 is considered for optical logic operations. The device is fabricated on $LiNbO_3$. A continuous stream of data pulses of a peak power P_c is incident in the central waveguide c. This beam is split equally between the two arms of the Mach–Zehnder interferometer and is recombined at the output. A DC bias voltage is applied to the lower arm of the interferometer so that a relative phase shift of φ_b is produced between the two arms. Control pulses of peak powers $P_a = P_b = P$ are fed into either waveguide a or b, or both. Each control pulse propagates only in one arm of the interferometer to change the refractive index seen by the data pulses through cross-phase modulation. The interaction length of the interferometer arms is l. To avoid interference between the control and data pulses, different polarizations at the same wavelength λ can be used. The TM-like mode is excited in the signal channel c, and the

TE-like mode is excited in the control channels a and b. A polarizer at the output transmits only TM-polarized signal pulses. For simplicity, assume that the pulses are all square pulses. The total differential phase shift between the two arms is $\Delta\varphi = \varphi_b + \Delta\varphi_{NL}$. What is the required power P of the control pulses as a function of λ, l, A_{eff}, and n_2 for the operation of this device as a logic gate? Describe the required bias phase and the input conditions for the device to function as (a) an inverter, (b) an AND gate, and (c) an exclusive OR gate, respectively. (See Lattes, A., Haus, H. A., Leonberger, F. J., and Ippen, E. P., "An ultrafast all-optical gate," *IEEE Journal of Quantum Electronics* **QE-19** (11): 1718–1723, Nov. 1983.)

9.12.3 The AlGaAs/GaAs symmetric all-optical Mach–Zehnder interferometer described in Example 9.24 can be used for all-optical sampling to digitize a continuous optical waveform if both input and output Y junctions are replaced by 3-dB directional couplers as shown in Fig. 9.44. The optical waveform to be digitized through pulse sampling is fed into one port of the input directional coupler as a TE-like mode. The sampling pulses are the control pulses fed into one arm of the interferometer through the control port as a TM-like mode. Describe the sampling operation of this device. The pulsewidth of the sampling pulses is $\Delta t_{ps} = 1$ ps. The sampling rate is 1 GHz. With the parameters for the AlGaAs waveguide given in Example 9.24, what is the required average power of the sampling pulse stream?

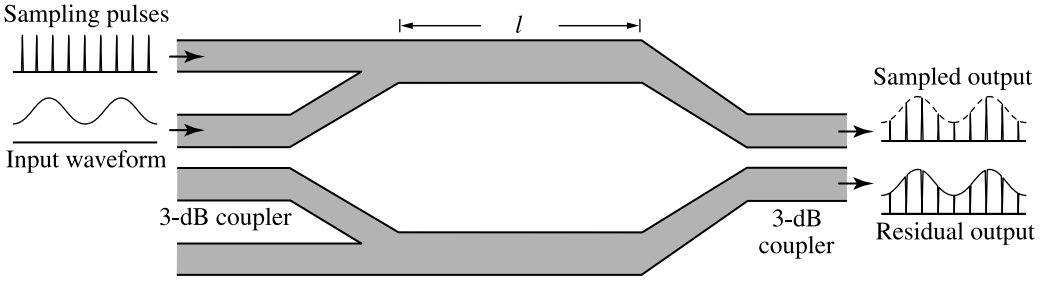


Figure 9.44 All-optical sampling device.

9.12.4 The transmittance of a nonlinear optical loop mirror shown in Fig. 9.37(a) is that given in (9.262).

a. In case the optical field consists of very short pulses such that interaction between contradirectionally propagating pulses is negligible, show that $\Delta\varphi$ is that given by (9.265).

b. In case the optical field consists of very long pulses or CW waves such that the interaction between contradirectionally propagating fields cannot be

ignored but there are no polarization changes in the fields along the loop path, show that $\Delta\varphi$ given by (9.265) is still valid when used in conjunction with (9.262).

c. Discuss the validity of (9.265) for the nonlinear optical loop mirror in the case when there are polarization changes in the fields along the loop path due to various reasons.

9.12.5 A nonlinear optical loop mirror as shown in Fig. 9.37(*b*) is used as an all-optical demultiplexer for switching a 20 Gbits s^{-1} data signal at $\lambda_d = 1.55$ μm with switching pulses of $\Delta t_{ps} = 2$ ps at $\lambda_s = 1.53$ μm at a repetition rate of 2.5 GHz to obtain a demultiplexed 2.5 Gbits s^{-1} data signal at λ_d. The switching pulses act as the control signal and are launched into the fiber loop through the control input port. The coupler of the loop mirror is designed such that the switching pulses propagate only in one direction in the loop but each data pulse is split into two contrapropagating pulses in the loop. It is desired that the transmittance be $T = 0$ for the data signal in the absence of a switching pulse while $T = 1$ for the data signal in the presence of a switching pulse. For this function to be achieved, what are the required power-splitting ratios for λ_d and λ_s, respectively, at the couplers? If the fiber has an effective cross-sectional area of $\mathcal{A}_{eff} = 4 \times 10^{-11}$ m^2, a nonlinear refractive index of $n_2 = 3.2 \times 10^{-20}$ m^2 W^{-1}, and a loop length of $l = 500$ m, what is the required average power of the switching pulse train for this demultiplexing function?

SELECT BIBLIOGRAPHY

Agrawal, G. P., *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press, 2001.

Bloembergen, N., *Nonlinear Optics*, 4th edn. Singapore: World Scientific, 1996.

Boyd, R. W., *Nonlinear Optics*. Boston, MA: Academic Press, 1992.

Butcher, P. N. and Cotter, D., *The Elements of Nonlinear Optics*. Cambridge: Cambridge University Press, 1990.

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge University Press, 1996.

Dmitriev, V. G., Gurzadyan, G. G., and Nikogosyan, D. N., *Handbook of Nonlinear Optical Crystals*, 3rd edn. Berlin: Springer, 1999.

Guo, Y., *Nonlinear Photonics*: *Nonlinearities in Optics, Optoelectronics, and Fiber Communications*. Berlin: Springer, 2002.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Nye, J. F., *Physical Properties of Crystals*. London: Oxford University Press, 1957.

Ostrowsky, D. B. and Reinisch, R., eds., *Guided Wave Nonlinear Optics*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1992.

Sauter, E. G., *Nonlinear Optics*. New York: Wiley, 1996.

Schwartz, K., *The Physics of Optical Recording*. New York: Springer-Verlag, 1993.

Shen, Y. R., *The Principles of Nonlinear Optics*. New York: Wiley, 1984.

Sirotin, Yu. I. and Shaskolskaya, M. P., *Fundamentals of Crystal Physics*. Moscow: Mir Publishers, 1982.

Sutherland, R. L., *Handbook of Nonlinear Optics*. New York: Marcel Dekker, 1996.

Zernike, F. and Midwinter, J. E., *Applied Nonlinear Optics*. New York: Wiley, 1973.

ADVANCED READING LIST

Adair, R., Chase, L. L., and Payne, S. A., "Nonlinear refractive index of optical crystals," *Physical Review B* **39**(5): 3337–3350, Feb. 1989.

Armstrong, J. A., Bloembergen, N., Ducuing, J., and Pershan, P. S., "Interaction between light waves in a nonlinear dielectric," *Physical Review* **127**(6): 1918–1939, Sep. 1962.

Bloembergen, N., "Nonlinear optics and spectroscopy," *Reviews of Modern Physics* **54**(3): 685–695, July 1982.

"Nonlinear optics: past, present, and future," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 876–880, Nov.–Dec. 2000.

Brabec, T. and Krausz, F., "Intense few-cycle laser fields: frontiers of nonlinear optics," *Reviews of Modern Physics* **72**(2): 545–591, Apr. 2000.

Byer, R. L., "Nonlinear optics and solid-state lasers: 2000," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 911–930, Nov.–Dec. 2000.

"Quasi-phasematched nonlinear interactions and devices," *Journal of Nonlinear Optical Physics and Materials* **6**(4): 549–592, Dec. 1997.

Dianov, E. M., "Advances in Raman fibers," *Journal of Lightwave Technology* **20**(8): 1457–1462, Aug. 2002.

Dimitrov, V. and Sakka, S., "Linear and nonlinear optical properties of simple oxides. II," *Journal of Applied Physics* **79**(3): 1741–1745, Feb. 1996.

Doran, N. J. and Wood, D., "Nonlinear-optical loop mirror," *Optics Letters* **13**(1): 56–58, Jan. 1988.

Dorn, R., Baums, D., Kersten, P., and Regener, R., "Nonlinear optical materials for integrated optics: telecommunications and sensors," *Advanced Materials* **4**(7): 464–473, July–Aug. 1992.

Eaton, D. F., "Nonlinear optical materials," *Science* **253**(5017): 281–287, July 1991.

Fainman, Y., Ma, J., and Lee, S. H., "Non-linear optical materials and applications," *Material Science Reports* **9**(2–3): 53–139, Jan. 1993.

Garmire, E., "Resonant optical nonlinearities in semiconductors," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1094–1110, Nov.–Dec. 2000.

Hansryd, J., Andrekson, P. A., Westlund, M., Li, J., and Hedekvist, P. O., "Fiber-based optical parametric amplifiers and their applications," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(3): 506–520, May–June 2002.

Henry, W. M., "Fibre acoustic modes and stimulated Brillouin scattering," *International Journal of Optoelectronics* **7**(4): 453–478, July–Aug. 1992.

Houe, M. and Townsend, P. D., "An introduction to methods of periodic poling for second-harmonic generation," *Journal of Physics D* **28**(9): 1747–1763, Sep. 1995.

Ishizuki, H., Suhara, T., Fujimura, M., and Nishihara, H., "Wavelength-conversion type picosecond optical switching using a waveguide QPM-SHG/DFG device," *Optical and Quantum Electronics* **33**: 953–961, 2001.

Islam, M. N., "Raman amplifiers for telecommunications," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(3): 548–559, May–June 2002.

Jeannes, F., Lugagne-Delpon, E., Tanguy, C., and Oudar, J. L., "Nonlinear optical and bistable properties of a wafer-fused vertical-cavity device based on InGaAsP," *Optics Communications* **134**: 607–614, Jan. 1997.

Jensen, S. M., "The nonlinear coherent coupler," *IEEE Journal of Quantum Electronics* **QE-18**(10): 1580–1583, Oct. 1982.

Klingshirn, C., "Non-linear optical properties of semiconductors," *Semiconductor Science and Technology* **5**(6): 457–469, June 1990.

Lattes, A., Haus, H. A., Leonberger, F. J., and Ippen, E. P., "An ultrafast all-optical gate," *IEEE Journal of Quantum Electronics* **QE-19**(11): 1718–1723, Nov. 1983.

Lin, C., "Nonlinear optics in fibers for fiber measurements and special device functions," *Journal of Lightwave Technology* **LT-4**(8): 1103–1115, Aug. 1986.

Moloney, J. V. and Newell, A. C., "Nonlinear optics," *Physica D* **44**(1–2): 1–37, Aug. 1990.

Myers, L. E. and Bosenberg, W. R., "Periodically poled lithium niobate and quasi-phase-matched optical parametric oscillators," *IEEE Journal of Quantum Electronics* **33**(10): 1663–1672, Oct. 1997.

Nikogosyan, D. N. and Gurzadyan, G. G., "Crystals for nonlinear optics," *Soviet Journal of Quantum Electronics* **17**(8): 970–977, Aug. 1987.

Reinisch, R. and Vitrant, R., "Optical bistability," *Progress in Quantum Electronics* **18**(1): 1–38, 1994.

Robert, D. A., "Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions," *IEEE Journal of Quantum Electronics* **28**(10): 2057–2074, Oct. 1992.

Ryvkin, B. S., "Optical bistability in semiconductors (review)," *Soviet Physics–Semiconductors* **19**(1): 1–15, Jan. 1985.

Shen, Y. R., "Surface nonlinear optics: a historical perspective," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1375–1379, Nov.–Dec. 2000.

Shiraki, K., Ohashi, M., and Tateda, M., "SBS threshold of a fiber with a Brillouin frequency shift distribution," *Journal of Lightwave Technology* **14**(1): 50–57, Jan. 1996.

Shoji, I., Kondo, T., and Ito, R., "Second-order nonlinear susceptibilities of various dielectric and semiconductor materials," *Optical and Quantum Electronics* **34**(8): 797–833, Aug. 2002.

Sibbett, W., Grant, R. S., and Spence, D. E., "Broadly tunable femtosecond solid-state laser sources," *Applied Physics B* **B58**(3): 171–181, Mar. 1994.

Stegeman, G. I. and Seaton, C. T., "Nonlinear integrated optics," *Journal of Applied Physics* **58**(12): R57–R78, Dec. 1985.

Stegeman, G. I. and Wright, E. M., "All-optical waveguide switching," *Optical and Quantum Electronics* **22**(2): 95–122, Mar. 1990.

Stegeman, G. I., Wright, E. M., Flinlayson, N., Zanoi, R., and Seaton, C. T., "Third order nonlinear integrated optics," *Journal of Lightwave Technology* **6**(6): 953–970, June 1988.

Steinmeyer, G., Sutter, D. H., Gallmann, L., Matuschek, N., and Keller, U., "Frontiers in ultrashort pulse generation: pushing the limits in linear and nonlinear optics," *Science* **286**(5444): 1507–1512, Nov. 1999.

Stolen, R. H., "Nonlinearity in fiber transmission," *Proceedings of the IEEE* **68**(10): 1232–1236, Oct. 1980.

Stolen, R. H., Gordon, J. P., Tomlinson, W. J., and Haus, H. A., "Raman response function of silica-core fibers," *Journal of the Optical Society of America B* **6**(6): 1159–1166, June 1989.

Sudo, S. and Itoh, H., "Efficient non-linear optical fibres and their applications," *Optical and Quantum Electronics* **22**(3): 187–212, May 1990.

Tang, C. L., "Tutorial on optical parametric processes and devices," *Journal of Nonlinear Optical Physics and Materials* **6**(4): 535–547, Dec. 1997.

Toliver, P., Runser, R. J., Glesk, I., and Prucnal, P. R., "Comparison of three nonlinear interferometric optical switch geometries," *Optics Communications* **175**: 365–373, 2000.

Part IV

Lasers

10 Laser amplifiers

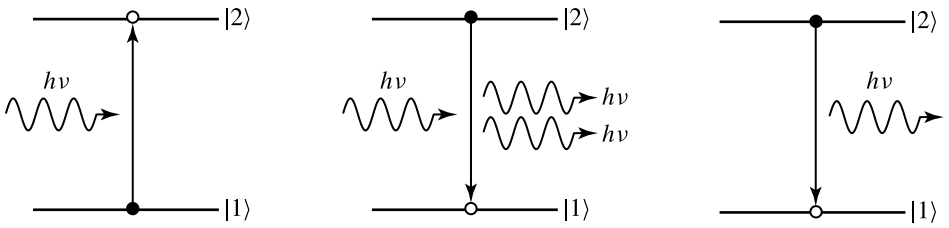
The word laser is an acronym for *light amplification by stimulated emission of radiation*. However, the term laser generally refers to a *laser oscillator*, which generates laser light without an input light wave. A device that amplifies a laser beam by stimulated emission is called a *laser amplifier*. Laser light is generally highly collimated with a very small divergence and highly coherent in time and space. It also has a relatively narrow spectral linewidth and a high intensity in comparison with light generated from ordinary sources. Due to the process of stimulated emission, an optical wave amplified by a laser amplifier preserves most of the characteristics, including the frequency spectrum, the coherence, the polarization, the divergence, and the direction of propagation, of the input wave. In this chapter, we discuss the characteristics of laser amplifiers. Laser oscillators are discussed in Chapter 11. Optical fiber amplifiers are of particular interest in photonics applications. They are specifically discussed in Section 10.5. Semiconductor laser amplifiers are discussed in Chapter 13.

10.1 Optical transitions

Optical absorption and emission occur through the interaction of optical radiation with electrons in a material system that defines the energy levels of the electrons. Depending on the properties of a given material, electrons that interact with optical radiation can be either those bound to individual atoms or those residing in the energy-band structures of a material such as a semiconductor. In any event, the absorption or emission of a photon by an electron is associated with a resonant transition of the electron between a lower energy level $|1\rangle$ of energy E_1 and an upper energy level $|2\rangle$ of energy E_2, as illustrated in Fig. 10.1. The resonance frequency, ν_{21}, of the transition is determined by the separation between the energy levels:

$$\nu_{21} = \frac{E_2 - E_1}{h}. \tag{10.1}$$

In an atomic or molecular system, a given energy level usually consists of a number of *degenerate* quantum-mechanical states, which have the same energy. Therefore, the



(*a*) (*b*) (*c*)

Figure 10.1 (*a*) Absorption, (*b*) stimulated emission, and (*c*) spontaneous emission of photons and resonant transitions in a material.

energy levels $|1\rangle$ and $|2\rangle$ are generally characterized by *degeneracy factors* g_1 and g_2, respectively.

There are basically three types of processes associated with resonant optical transitions between two energy levels in a system: *absorption*, *stimulated emission*, and *spontaneous emission*, which are illustrated in Figs. 10.1(*a*), (*b*), and (*c*), respectively. Absorption and stimulated emission of photons are both associated with *induced transitions* between the energy levels caused by interaction of an electron with the existing optical radiation. If an electron is initially in the lower level $|1\rangle$, it can absorb a photon to make a transition to the upper level $|2\rangle$. If an electron is initially in the upper level $|2\rangle$, the optical radiation can stimulate it to emit a photon by making a downward transition to the lower level $|1\rangle$. Irrespective of the presence or absence of any existing optical radiation, an electron initially in the upper level $|2\rangle$ can also spontaneously relax to the lower level $|1\rangle$ by emitting a spontaneous photon.

A photon emitted by stimulated emission has the same frequency, phase, polarization, and propagation direction as the optical radiation that induces the process. In contrast, spontaneously emitted photons are random in phase and polarization and are emitted in all directions, though their frequencies are still dictated by the separation between the two energy levels, subject to a degree of uncertainty determined by the linewidth of the transition. Therefore, stimulated emission results in the amplification of an optical signal, whereas spontaneous emission merely adds noise to an optical signal. Absorption simply leads to the attenuation of an optical signal.

Spectral lineshape

A resonant transition is selective of the frequency of the interacting optical field because the process is associated with absorption or emission of a photon whose frequency is determined by the energy change of the transition indicated in (10.1). The spectral characteristic of a resonant transition is never infinitely sharp, however. The finite spectral width of a resonant transition is dictated by the uncertainty principle of quantum mechanics, but it can be understood intuitively without the details of quantum mechanics

by following the line of reasoning in Section 1.10. One important conclusion learned from these discussions is that any response that has a finite relaxation time in the time domain must have a finite spectral width in the frequency domain. As we shall see below, the rate of the induced transitions between two energy levels in a given system is directly proportional to the spontaneous emission rate from the upper to the lower level in that system. Therefore, it is a basic law of physics that any allowed resonant transition between two energy levels has a finite relaxation time constant because at least the upper level has a finite lifetime due to spontaneous emission. Consequently, for each particular resonant transition between two energy levels, there is a characteristic *lineshape function*, $\hat{g}(\nu)$, of finite linewidth that characterizes the optical processes associated with the transition. The lineshape function is generally normalized as

$$\int_0^\infty \hat{g}(\nu)\mathrm{d}\nu = \int_0^\infty \hat{g}(\omega)\mathrm{d}\omega = 1, \tag{10.2}$$

where $\hat{g}(\nu) = 2\pi\hat{g}(\omega)$.

Homogeneous broadening

If all of the atoms in a material that participate in a resonant interaction associated with the energy levels $|1\rangle$ and $|2\rangle$ are indistinguishable, their responses to an electromagnetic field are characterized by the same resonance frequency ν_{21} and the same relaxation constant γ_{21}. In such a homogeneous system, the physical mechanisms that contribute to the linewidth of the transition affect all atoms equally. Spectral broadening due to such mechanisms is called *homogeneous broadening*.

From the discussions in Section 1.10, the spectral characteristics of a damped response characterized by a single resonance frequency and a single relaxation constant, such as that of a resonant interaction in a homogeneously broadened system, are described by the functions given in (1.176). As we shall see in Section 10.2, in the interaction of a material with an optical field, the absorption and emission of optical energy are characterized by the imaginary part χ'' of the susceptibility of the material. Therefore, the spectral characteristics of optical absorption and emission due to a resonant transition in a homogeneously broadened medium are described by the Lorentzian lineshape function of $\chi''(\omega)$ given in (1.176). Using the normalization condition in (10.2), we find that the resonant transition between $|1\rangle$ and $|2\rangle$ has the following normalized Lorentzian lineshape function:

$$\hat{g}(\omega) = \frac{1}{\pi}\frac{\gamma_{21}}{(\omega-\omega_{21})^2+\gamma_{21}^2}, \tag{10.3}$$

which has a FWHM of $\Delta\omega_\mathrm{h} = 2\gamma_{21}$, or

$$\hat{g}(\nu) = \frac{\Delta\nu_\mathrm{h}}{2\pi[(\nu-\nu_{21})^2+(\Delta\nu_\mathrm{h}/2)^2]}, \tag{10.4}$$

where $\Delta\nu_h = \gamma_{21}/\pi$ is the FWHM of $\hat{g}(\nu)$. We see that the spectrum has a finite width that is determined by the relaxation constant γ_{21}.

The fundamental mechanism for homogeneous broadening is *lifetime broadening* due to the finite lifetimes, τ_1 and τ_2, of the energy levels, $|1\rangle$ and $|2\rangle$, respectively, that are involved in the resonant transition. The population in an energy level can relax through both radiative transitions and nonradiative transitions to lower levels. Radiative relaxation is associated with population relaxation through spontaneous emission of radiation. The radiative relaxation rate of the transition from level $|2\rangle$ to level $|1\rangle$ is characterized by a constant A_{21}, known as the *Einstein A coefficient*, or a time constant $\tau_{sp} = 1/A_{21}$, known as the *spontaneous radiative lifetime* between $|2\rangle$ and $|1\rangle$. Both A_{21} and τ_{sp} are discussed in further detail later. The total radiative relaxation rate, γ_2^{rad}, of level $|2\rangle$ is the sum of all radiative spontaneous transition rates from $|2\rangle$ to other levels: $\gamma_2^{rad} = \sum_i A_{2i}$. The nonradiative relaxation rate accounts for all other population relaxation mechanisms that do not result in the emission of photons. Therefore, the total relaxation rate is the sum of the radiative and nonradiative relaxation rates, and the lifetime of an energy level has both radiative and nonradiative contributions:

$$\gamma_2 = \gamma_2^{rad} + \gamma_2^{nonrad}, \qquad \frac{1}{\tau_2} = \frac{1}{\tau_2^{rad}} + \frac{1}{\tau_2^{nonrad}}, \tag{10.5}$$

where $\tau_2 = 1/\gamma_2$, $\tau_2^{rad} = 1/\gamma_2^{rad}$, and $\tau_2^{nonrad} = 1/\gamma_2^{nonrad}$. The same concept can be applied to level $|1\rangle$ to obtain similar relations for γ_1 and τ_1. Even though τ_2 is contributed by both radiative and nonradiative decay from level $|2\rangle$, fluorescent emission from level $|2\rangle$ decays at the total relaxation rate γ_2 of the population in level $|2\rangle$. Therefore, the decay time constant of the fluorescent emission associated with population relaxation from $|2\rangle$ is τ_2, not τ_2^{rad}. For this reason, the total lifetimes τ_1 and τ_2 are known as the *fluorescence lifetimes* of energy levels $|1\rangle$ and $|2\rangle$, respectively. The contributions of various relaxation rates to the radiative and nonradiative lifetimes, and to the fluorescence lifetimes, of the upper and lower laser levels are summarized in Fig. 10.2.

The nonradiative relaxation rate of an energy level is a function of external perturbations such as collisions and thermal vibrations. It can therefore be changed by varying the conditions of the surrounding environment. The minimum broadening is called *natural broadening* and is caused only by radiative relaxation when the nonradiative processes are eliminated. The linewidth due to natural broadening alone is

$$\gamma_{21}^{natural} = \frac{1}{2}\left(\gamma_1^{rad} + \gamma_2^{rad}\right) = \frac{1}{2}\left(\frac{1}{\tau_1^{rad}} + \frac{1}{\tau_2^{rad}}\right). \tag{10.6}$$

The total contribution of lifetime broadening to the linewidth due to both radiative and nonradiative relaxation processes is

$$\gamma_{21}^{life} = \frac{1}{2}(\gamma_1 + \gamma_2) = \frac{1}{2}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) \geq \gamma_{21}^{natural}. \tag{10.7}$$

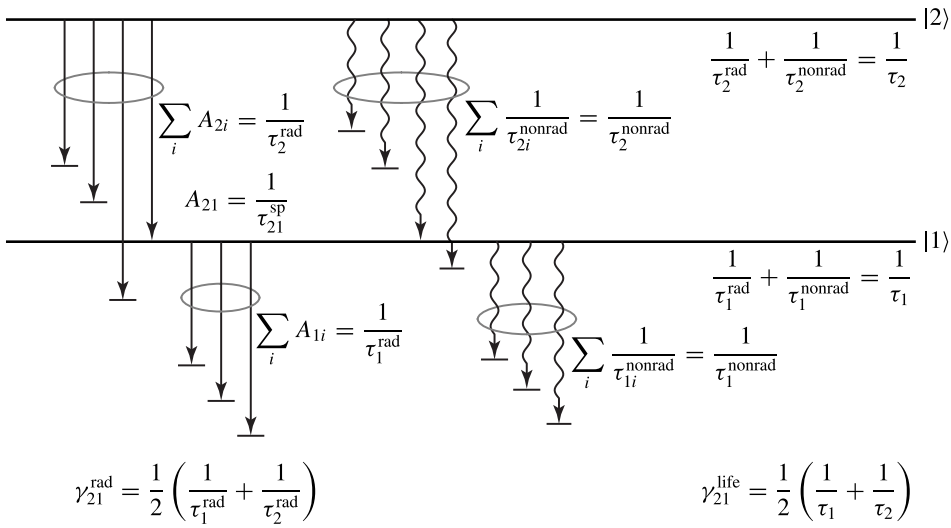


Figure 10.2 Contributions of various relaxation rates to the radiative and nonradiative lifetimes, and to the fluorescence lifetimes, of the upper and lower laser levels. The homogeneous natural linewidth is determined by radiative lifetimes, whereas the lifetime-broadened linewidth is determined by fluorescence lifetimes.

These contributions to $\gamma_{21}^{\text{natural}}$ and $\gamma_{21}^{\text{life}}$ are also summarized in Fig. 10.2. Note that the linewidth is determined by the lifetimes of both upper and lower laser levels. In the case when the lower laser level $|1\rangle$ is the ground state of an atomic system, such as in the situation of the ruby emission line at 694.3 nm, we have $\gamma_1 = 0$ and $\tau_1 = \infty$. Then, the linewidth due to lifetime broadening is solely determined by the lifetime of the upper laser level, τ_2.

Other mechanisms that affect all atoms equally can further increase the homogeneous linewidth without changing the fluorescence lifetimes τ_2 and τ_1 of the upper and lower laser levels. One important mechanism is collision-induced phase randomization of the emitted radiation. Collisions among atoms in a gas or liquid and collisions of atoms with phonons in a solid normally have two possible effects. One is reduction of the fluorescence lifetimes of the upper and lower laser levels by increasing the nonradiative relaxation rates. Such a process increases lifetime broadening; its effect is included in $\gamma_{21}^{\text{life}}$ through the dependence of $\gamma_{21}^{\text{life}}$ on γ_1^{nonrad} and γ_2^{nonrad} contained in γ_1 and γ_2, respectively. Collisions can also increase a homogeneous linewidth without reducing the fluorescence lifetimes by simply interrupting the phase of the radiation emitted through radiative relaxation. This dephasing process, quantified by a linewidth-broadening factor $\gamma_{21}^{\text{dephase}}$, is often more important than the lifetime-reduction process, resulting in a homogeneous linewidth that is significantly broader than the linewidth contributed by lifetime broadening. Therefore, the homogeneous linewidth can increase both with pressure and with temperature in a gas medium, and with active-ion concentration and temperature in a liquid or solid medium. In general, the homogeneous linewidth,

including the contributions of such external mechanisms, is a function of pressure, P, active-ion concentration, N, and temperature, T:

$$\gamma_{21}(P, N, T) = \gamma_{21}^{\text{life}} + \gamma_{21}^{\text{dephase}} \geq \gamma_{21}^{\text{life}} \geq \gamma_{21}^{\text{natural}}. \tag{10.8}$$

EXAMPLE 10.1 The energy levels of laser transitions, along with radiative transition rates[1] and emission wavelengths, of Nd : YAG are shown in Fig. 10.3. The upper level $^4F_{3/2}$ relaxes radiatively to four lower levels. The lowest level $^4I_{9/2}$ is the ground level of the system. In this example, we consider the dominant transition that takes place between the upper level $^4F_{3/2}$, labeled level $|2\rangle$, and the lower level $^4I_{11/2}$, labeled level $|1\rangle$, for the well-known Nd : YAG emission wavelength of $\lambda = 1.064$ μm. The relaxation of the upper level $^4F_{3/2}$ is predominantly radiative with a fluorescence lifetime of $\tau_2 = 240$ μs. The relaxation of the lower level $^4I_{11/2}$ is nonradiative with a fluorescence lifetime[2] of $\tau_1 = 200$ ps. (a) Find the spontaneous radiative lifetime τ_{sp} between $|2\rangle$ and $|1\rangle$. (b) Find the radiative and nonradiative relaxation rates, γ_2^{rad} and γ_2^{nonrad}, and the corresponding lifetimes, τ_2^{rad} and τ_2^{nonrad}, for the upper level $|2\rangle$. (c) Find the natural linewidth, $\Delta\nu_{\text{natural}}$, and the lifetime-broadened homogeneous linewidth, $\Delta\nu_{\text{life}}$. (d) If the measured linewidth at room temperature is $\Delta\nu = 150$ GHz with a homogeneously broadened component of $\Delta\nu_{\text{h}} = 120$ GHz, what is the linewidth-broadening factor $\gamma_{21}^{\text{dephase}}$ due to dephasing through phonon collisions?

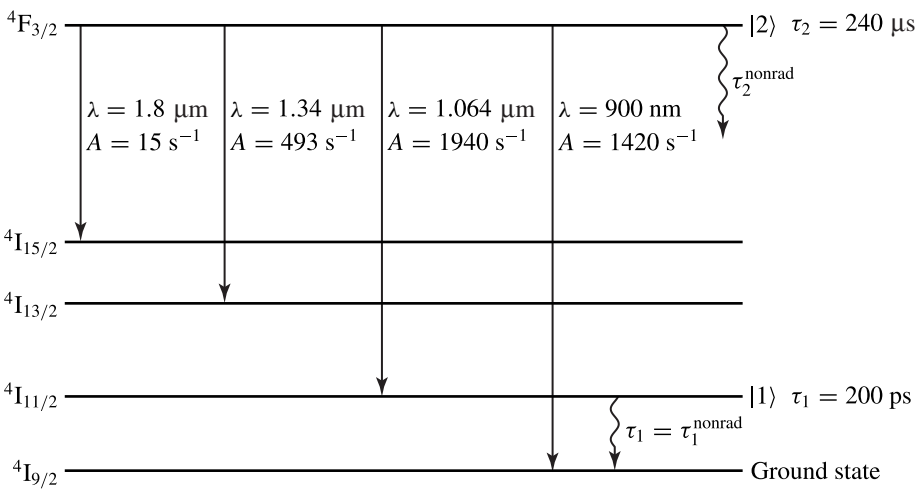


Figure 10.3 Energy levels of Nd : YAG.

[1] Krupke, W. F., "Radiative transition probabilities within the $4f^3$ ground configuration of Nd : YAG," *IEEE Journal of Quantum Electronics* **QE-7**: 153–159, 1971.

[2] Payne, S. A. and Bibeau, C., "Picosecond nonradiative processes in neodymium-doped crystals and glasses: mechanisms for the energy gap law," *Journal of Luminescence* **79**: 143–159, 1998.

Solution (a) Using the radiative transition rate between $^4F_{3/2}$ and $^4I_{11/2}$, we find that

$$\tau_{sp} = \frac{1}{A_{21}} = \frac{1}{1940}\ \text{s} = 515\ \mu\text{s}.$$

(b) The total radiative relaxation rate is the sum of the radiative transition rates from $^4F_{3/2}$ to all four lower levels. Therefore, the radiative relaxation rate and the radiative lifetime are, respectively,

$$\gamma_2^{rad} = \sum_i A_{2i} = 3868\ \text{s}^{-1}, \quad \tau_2^{rad} = \frac{1}{\gamma_2^{rad}} = 259\ \mu\text{s}.$$

Note that $\tau_2^{rad} > \tau_2 = 240\ \mu$s, as expected. Because the total relaxation rate of the upper level is $\gamma_2 = 1/\tau_2 = 4167\ \text{s}^{-1}$, the nonradiative relaxation rate and the nonradiative lifetime are, respectively,

$$\gamma_2^{nonrad} = \gamma_2 - \gamma_2^{rad} = 299\ \text{s}^{-1}, \quad \tau_2^{nonrad} = \frac{1}{\gamma_2^{nonrad}} = 3.34\ \text{ms}.$$

(c) Because level $|1\rangle$ relaxes only nonradiatively, $\gamma_1^{rad} = 0$ and $\tau_1^{rad} = \infty$. Therefore,

$$\gamma_{21}^{natural} = \frac{1}{2}\left(\gamma_1^{rad} + \gamma_2^{rad}\right) = \frac{1}{2}(0 + 3868)\ \text{s}^{-1} = 1.93 \times 10^3\ \text{s}^{-1}.$$

Using (10.7), we find that

$$\gamma_{21}^{life} = \frac{1}{2}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) = \frac{1}{2}\left(\frac{1}{200 \times 10^{-12}} + \frac{1}{240 \times 10^{-6}}\right)\ \text{s}^{-1} = 2.5 \times 10^9\ \text{s}^{-1}.$$

From (10.3) and (10.4), we know that $\Delta\nu_h = \gamma_{21}/\pi$. Using a similar relation, we find that

$$\Delta\nu_{natural} = \frac{\gamma_{21}^{natural}}{\pi} = 616\ \text{Hz}, \quad \Delta\nu_{life} = \frac{\gamma_{21}^{life}}{\pi} = 796\ \text{MHz}.$$

(d) For $\Delta\nu_h = 120$ GHz, we have $\gamma_{21} = \pi\Delta\nu_h = 3.77 \times 10^{11}\ \text{s}^{-1}$. Therefore,

$$\gamma_{21}^{dephase} = \gamma_{21} - \gamma_{21}^{life} = 3.745 \times 10^{11}\ \text{s}^{-1}.$$

Clearly, $\gamma_{21} \approx \gamma_{21}^{dephase} \gg \gamma_{21}^{life}$ in this example.

Inhomogeneous broadening

A resonant transition can be further broadened by inhomogeneous broadening if certain physical mechanisms exist that do not affect all atoms equally, causing energy levels $|1\rangle$ and/or $|2\rangle$ to shift differently among different groups of atoms. The resulting inhomogeneous shifts of the resonance frequency contribute to inhomogeneous broadening of the transition spectrum on top of the original homogeneous broadening. If we express the homogeneous lineshape function given in (10.4) as $\hat{g}_h(\nu, \nu_{21})$ to indicate explicitly that its resonance frequency is at ν_{21}, the homogeneously broadened spectrum of a

group of atoms whose resonance frequency is shifted from ν_{21} to ν_k is $\hat{g}_{\mathrm{h}}(\nu, \nu_k)$. The distribution of atoms in the system can be described by a probability density function $p(\nu_k)$ with

$$\int_0^\infty p(\nu_k)\mathrm{d}\nu_k = 1. \tag{10.9}$$

The probability that the resonance frequency of a given atom falls in the range between ν_k and $\nu_k + \mathrm{d}\nu_k$ is $p(\nu_k)\mathrm{d}\nu_k$. Then, the overall spectral lineshape of the inhomogeneously broadened transition is

$$\hat{g}(\nu) = \int_0^\infty p(\nu_k)\hat{g}_{\mathrm{h}}(\nu, \nu_k)\mathrm{d}\nu_k. \tag{10.10}$$

The overall lineshape function obtained from (10.10) depends on the degree of inhomogeneous broadening in comparison to the homogeneous broadening of the atoms. Mathematically, it depends on the spread of the distribution $p(\nu_k)$ in comparison to the homogeneous linewidth.

One possibility for inhomogeneous broadening is the existence of different isotopes, which have slightly different resonance frequencies for a given resonant transition. In this situation, $p(\nu_k)\mathrm{d}\nu_k$ represents the percentage of each isotope group among all atoms and (10.10) becomes simply the weighted sum of the isotope groups.

Other mechanisms for inhomogeneous broadening include the Doppler effect in a gaseous medium at a low pressure and the random distribution of active impurity atoms doped in a solid host. The inhomogeneous frequency shifts caused by these mechanisms are usually randomly distributed, resulting in a Gaussian functional distribution for $p(\nu_k)$. In an extremely inhomogeneously broadened system, the spread of this distribution dominates the homogeneous linewidth. Then, the transition is characterized by a normalized *Gaussian lineshape*:

$$\hat{g}(\nu) = \frac{2(\ln 2)^{1/2}}{\pi^{1/2}\Delta\nu_{\mathrm{inh}}} \exp\left[-4\ln 2\frac{(\nu - \nu_0)^2}{\Delta\nu_{\mathrm{inh}}^2}\right], \tag{10.11}$$

where ν_0 is the center frequency and $\Delta\nu_{\mathrm{inh}}$ is the FWHM of the inhomogeneously broadened spectral distribution. In terms of the angular frequency, the normalized Gaussian lineshape is

$$\hat{g}(\omega) = \frac{2(\ln 2)^{1/2}}{\pi^{1/2}\Delta\omega_{\mathrm{inh}}} \exp\left[-4\ln 2\frac{(\omega - \omega_0)^2}{\Delta\omega_{\mathrm{inh}}^2}\right], \tag{10.12}$$

where $\omega_0 = 2\pi\nu_0$ and $\Delta\omega_{\mathrm{inh}} = 2\pi\Delta\nu_{\mathrm{inh}}$. Figure 10.4 compares the normalized Lorentzian lineshape function and the normalized Gaussian lineshape function of the same FWHM. In Fig. 10.4(*a*), we show $\hat{g}(\nu)$ as expressed in (10.4) for the Lorentzian lineshape and in (10.11) for the Gaussian lineshape, both with a normalized area as

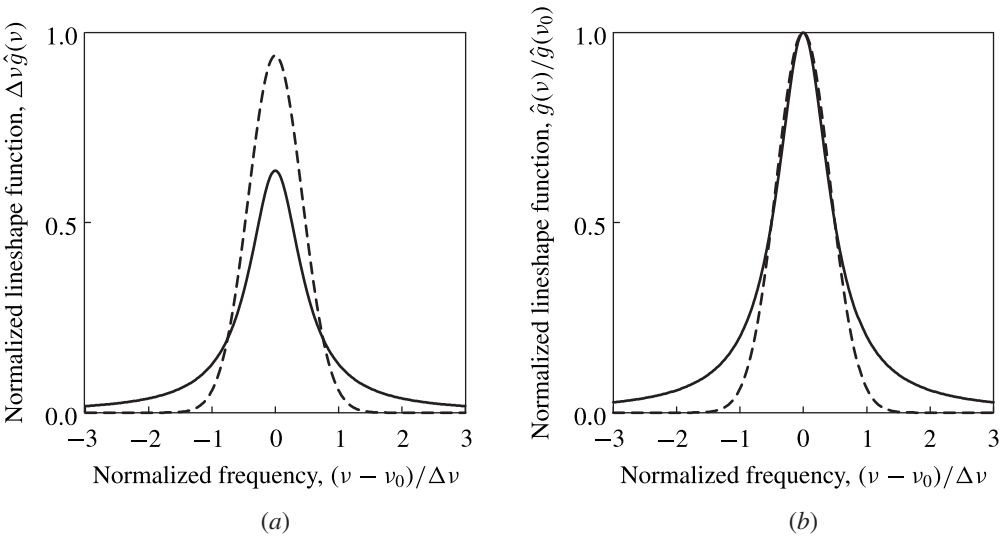


Figure 10.4 Normalized Lorentzian (solid curves) and Gaussian (dashed curves) lineshape functions of the same FWHM with (*a*) a normalized area as $\hat{g}(\nu)$ is defined and (*b*) a normalized peak value. For the Lorentzian lineshape, $\nu_0 = \nu_{21}$ and $\Delta\nu = \Delta\nu_h$. For the Gaussian lineshape, $\Delta\nu = \Delta\nu_{inh}$.

defined in (10.2). In Fig. 10.4(*b*), the lineshapes that are normalized to the same peak value are shown.

Whether a medium is homogeneously or inhomogeneously broadened is often a function of pressure and temperature. In a gas at low pressure, the velocity distribution of the gas molecules in thermal equilibrium is characterized by the Maxwellian velocity distribution, which is a Gaussian function. This velocity distribution leads to a Gaussian distribution of the Doppler frequency shifts with a linewidth $\Delta\nu_D$ given by

$$\Delta\nu_D = 2\nu\left(\frac{2(\ln 2)k_B T}{Mc^2}\right)^{1/2} = \frac{2^{3/2}(\ln 2)^{1/2}}{\lambda}\left(\frac{k_B T}{M}\right)^{1/2}, \tag{10.13}$$

where λ is the emission wavelength, k_B is the Boltzmann constant, T is the temperature in kelvins, and M is the mass of the atom or molecule that emits the radiation. When this Doppler-broadening effect dominates, the Gaussian lineshape has an inhomogeneous linewidth of $\Delta\nu_{inh} = \Delta\nu_D$. When the pressure is increased, frequent collisions among the gas molecules cause the homogeneous linewidth to increase. At a certain pressure, the homogeneous linewidth $\Delta\nu_h$ finally dominates the Doppler linewidth $\Delta\nu_D$. Then the medium becomes predominantly homogeneously broadened.

Another good example is the linewidth associated with the impurity ions doped in a solid host, such as Nd : YAG or Nd : glass. At low temperatures, the homogeneous linewidth of the Nd^{3+} ions is narrow. The lineshape is dominated by inhomogeneous shifts of the resonance frequency due to variations in the local environment of individual Nd^{3+} ions. As a result, the lineshape function is inhomogeneously broadened. As the temperature increases, the homogeneous linewidth increases because of increased

collisions of phonons with the ions. At room temperature, the spectral line of Nd : YAG at 1.064 μm has a total linewidth of $\Delta\nu \approx 120$–180 GHz with an inhomogeneous component of only about 6–30 GHz. Therefore, Nd : YAG becomes pretty much homogeneously broadened at room temperature. In comparison, Nd : glass has a much larger inhomogeneous linewidth than Nd : YAG because the glass host provides a larger range of local variations than the YAG crystal. At room temperature, the same spectral line of Nd : glass appears at 1.054 μm with a total linewidth of $\Delta\nu \approx 5$–7 THz, which is predominantly inhomogeneously broadened.

EXAMPLE 10.2 The emission at 632.8 nm wavelength of the HeNe laser is caused by radiative transitions in the Ne atoms. The linewidth of this emission is inhomogeneously broadened due to Doppler broadening. The atomic mass number of Ne is 20, and the typical gas temperature of a HeNe laser is about 400 K. Find the emission linewidth.

Solution The mass of a Ne atom of mass number 20 is $M = 20 \times 1.67 \times 10^{-27}$ kg. Using (10.13), we find that the inhomogeneously broadened linewidth due to Doppler broadening is

$$\Delta\nu_{\mathrm{D}} = \frac{2^{3/2}(\ln 2)^{1/2}}{632.8 \times 10^{-9}} \times \left(\frac{1.38 \times 10^{-23} \times 400}{20 \times 1.67 \times 10^{-27}}\right)^{1/2} \text{Hz} = 1.5 \text{ GHz}.$$

Transition rates

The probability per unit time for a resonant optical process to occur is measured by the transition rate of the process. Because of the resonant nature of the interaction, the transition rate of an induced process is a function of the spectral distribution of the optical radiation and the spectral characteristics of the resonant transition.

The spectral distribution of an optical field is characterized by its spectral energy density, $u(\nu)$, which is the energy density of the optical radiation per unit frequency interval at the optical frequency ν. The total energy density of the radiation is

$$u = \int_0^\infty u(\nu)\mathrm{d}\nu. \tag{10.14}$$

The spectral intensity distribution, $I(\nu)$, of the radiation is related to $u(\nu)$ by the relation

$$I(\nu) = \frac{c}{n}u(\nu), \tag{10.15}$$

where n is the refractive index of the medium, and the total intensity is simply

$$I = \int_0^\infty I(\nu)\mathrm{d}\nu. \tag{10.16}$$

Because an induced transition is stimulated by optical radiation, its transition rate is proportional to the energy density of the optical radiation within the spectral response range of the transition. The transition rate for the upward transition from $|1\rangle$ to $|2\rangle$ associated with absorption in the frequency range between ν and $\nu + \mathrm{d}\nu$ is

$$W_{12}(\nu)\mathrm{d}\nu = B_{12}u(\nu)\hat{g}(\nu)\mathrm{d}\nu \qquad (\mathrm{s}^{-1}), \tag{10.17}$$

whereas that for the downward transition from $|2\rangle$ to $|1\rangle$ associated with stimulated emission in the frequency range between ν and $\nu + \mathrm{d}\nu$ is

$$W_{21}(\nu)\mathrm{d}\nu = B_{21}u(\nu)\hat{g}(\nu)\mathrm{d}\nu \qquad (\mathrm{s}^{-1}). \tag{10.18}$$

The spontaneous emission rate is independent of the energy density of the radiation. The spontaneous emission spectrum associated with a particular resonant transition is determined solely by the lineshape function of the transition:

$$W_{\mathrm{sp}}(\nu)\mathrm{d}\nu = A_{21}\hat{g}(\nu)\mathrm{d}\nu \qquad (\mathrm{s}^{-1}). \tag{10.19}$$

The A and B constants defined above are known as the *Einstein A and B coefficients*, respectively.

The induced and the spontaneous transition rates for a given system are not independent of each other but are directly proportional to one another. Such a relationship was first obtained by Einstein by considering the interaction of blackbody radiation with an ensemble of identical atomic systems in thermal equilibrium.

The spectral energy density of blackbody radiation at a temperature T is given by Planck's formula:

$$u(\nu) = \frac{8\pi n^3 h\nu^3}{c^3}\frac{1}{\mathrm{e}^{h\nu/k_{\mathrm{B}}T} - 1}, \tag{10.20}$$

where k_{B} is the Boltzmann constant. In thermal equilibrium with blackbody radiation, the total induced transition rates are

$$W_{12} = \int_0^\infty W_{12}(\nu)\mathrm{d}\nu = B_{12}\int_0^\infty u(\nu)\hat{g}(\nu)\mathrm{d}\nu \tag{10.21}$$

and

$$W_{21} = \int_0^\infty W_{21}(\nu)\mathrm{d}\nu = B_{21}\int_0^\infty u(\nu)\hat{g}(\nu)\mathrm{d}\nu. \tag{10.22}$$

The total spontaneous emission rate is

$$W_{\mathrm{sp}} = \int_0^\infty W_{\mathrm{sp}}(\nu)\mathrm{d}\nu = A_{21}. \tag{10.23}$$

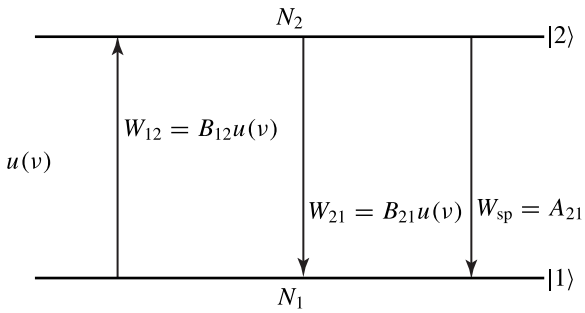


Figure 10.5 Resonant transitions in the interaction of a radiation field with two atomic levels $|1\rangle$ and $|2\rangle$ of population densities N_1 and N_2, respectively.

The rates associated with resonant transitions between two atomic levels $|1\rangle$ and $|2\rangle$ in the interaction with a radiation field of energy density $u(\nu)$ are summarized in Fig. 10.5. If N_2 and N_1 are the population densities per unit volume of the atoms in levels $|2\rangle$ and $|1\rangle$, respectively, the number of atoms per unit volume making the downward transition per unit time accompanied by the emission of radiation in a frequency range from ν to $\nu + \mathrm{d}\nu$ is $N_2(W_{21}(\nu) + W_{sp}(\nu))\mathrm{d}\nu$, and the number of atoms per unit volume making the upward transition per unit time assisted by the absorption of radiation in the same frequency range is $N_1 W_{12}(\nu)\mathrm{d}\nu$. In equilibrium, both the blackbody radiation spectral density and the atomic population density in each energy level should reach a steady state, meaning that

$$N_2[W_{21}(\nu) + W_{sp}(\nu)] = N_1 W_{12}(\nu). \tag{10.24}$$

This relation spells out the *principle of detailed balance* in thermal equilibrium. Therefore, the steady-state population distribution in thermal equilibrium satisfies

$$\frac{N_2}{N_1} = \frac{W_{12}(\nu)}{W_{21}(\nu) + W_{sp}(\nu)} = \frac{B_{12}u(\nu)}{B_{21}u(\nu) + A_{21}}. \tag{10.25}$$

In thermal equilibrium at temperature T, however, the population ratio of the atoms in the upper and the lower levels follows the Boltzmann distribution. Taking into account the degeneracy factors, g_2 and g_1, of these energy levels, we have

$$\frac{N_2}{N_1} = \frac{g_2}{g_1}\exp(-h\nu/k_B T) \tag{10.26}$$

for the population densities associated with a transition energy of $h\nu$. Combining (10.25) and (10.26), we have

$$u(\nu) = \frac{A_{21}/B_{21}}{(g_1 B_{12}/g_2 B_{21})\mathrm{e}^{h\nu/k_B T} - 1}. \tag{10.27}$$

Identifying (10.27) with (10.20), we find that

$$\frac{A_{21}}{B_{21}} = \frac{8\pi n^3 h\nu^3}{c^3} \tag{10.28}$$

and

$$g_1 B_{12} = g_2 B_{21}. \tag{10.29}$$

The *spontaneous radiative lifetime* of the atoms in level $|2\rangle$ associated with the radiative spontaneous transition from $|2\rangle$ to $|1\rangle$ is

$$\tau_{sp} = \frac{1}{W_{sp}} = \frac{1}{A_{21}}. \tag{10.30}$$

Therefore, the spectral dependence of the spontaneous emission rate can be expressed as

$$W_{sp}(\nu) = \frac{1}{\tau_{sp}} \hat{g}(\nu). \tag{10.31}$$

According to the relations in (10.28) and (10.29), *the transition rates of both of the induced processes of absorption and stimulated emission are directly proportional to the spontaneous emission rate.* In terms of τ_{sp}, the spectral dependence of the induced transition rates between energy levels $|1\rangle$ and $|2\rangle$ can be generally expressed as

$$W_{21}(\nu) = \frac{c^3}{8\pi n^3 h\nu^3 \tau_{sp}} u(\nu)\hat{g}(\nu) = \frac{c^2}{8\pi n^2 h\nu^3 \tau_{sp}} I(\nu)\hat{g}(\nu) \tag{10.32}$$

for the transition from $|2\rangle$ to $|1\rangle$ associated with stimulated emission and

$$W_{12}(\nu) = \frac{g_2}{g_1} W_{21}(\nu) \tag{10.33}$$

for the transition from $|1\rangle$ to $|2\rangle$ associated with absorption.

Because $W(\nu)$ is the transition rate per unit frequency according to the definition in (10.17)–(10.19), we have $W(\nu)\mathrm{d}\nu = W(\omega)\mathrm{d}\omega$. Therefore, $W_{sp}(\nu) = 2\pi W_{sp}(\omega)$, $W_{21}(\nu) = 2\pi W_{21}(\omega)$, and $W_{12}(\nu) = 2\pi W_{12}(\omega)$.

Transition cross section

It is often useful to express the transition probability of an atom in its interaction with optical radiation at a frequency ν in terms of the *transition cross section*, $\sigma(\nu)$. For transitions between energy levels $|1\rangle$ and $|2\rangle$, the transition cross sections $\sigma_{21}(\nu)$ and $\sigma_{12}(\nu)$ are defined through the following relations to the transition rates:

$$W_{21}(\nu) = \frac{I(\nu)}{h\nu} \sigma_{21}(\nu) \tag{10.34}$$

and

$$W_{12}(\nu) = \frac{I(\nu)}{h\nu} \sigma_{12}(\nu). \tag{10.35}$$

The transition cross section $\sigma_{21}(\nu)$, which is associated with stimulated emission, is also called the *emission cross section*, $\sigma_e(\nu)$, whereas $\sigma_{12}(\nu)$, which is associated with

absorption, is also called the *absorption cross section*, $\sigma_a(\nu)$. From (10.32), we find that

$$\sigma_e(\nu) = \sigma_{21}(\nu) = \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}} \hat{g}(\nu). \tag{10.36}$$

According to (10.29) and (10.33), $g_1\sigma_{12} = g_2\sigma_{21}$. Therefore,

$$\sigma_a(\nu) = \sigma_{12}(\nu) = \frac{g_2}{g_1}\sigma_{21}(\nu) = \frac{g_2}{g_1}\sigma_e(\nu). \tag{10.37}$$

The transition cross sections have the unit of area in square meters but are often quoted in square centimeters.

Note that $\sigma(\nu) = \sigma(\omega)$ because $\sigma(\nu)$ is simply defined as the value of the transition cross section *at the frequency* ν rather than as that per unit frequency, but $W(\nu) = 2\pi W(\omega)$ and $\hat{g}(\nu) = 2\pi\hat{g}(\omega)$. Therefore, in terms of ω,

$$\sigma_e(\omega) = \sigma_{21}(\omega) = \frac{\pi^2 c^2}{n^2\omega^2\tau_{sp}}\hat{g}(\omega) \quad \text{and} \quad \sigma_a(\omega) = \frac{g_2}{g_1}\sigma_e(\omega). \tag{10.38}$$

For the ideal Lorentzian and Gaussian lineshapes expressed in (10.4) and (10.11), respectively, the peak value of $\hat{g}(\nu)$ occurs at the center of the spectrum and is a function of linewidth $\Delta\nu$ only. By applying this fact to (10.36), the peak value of the emission cross section at the center wavelength λ of the spectrum can be expressed as

$$\sigma_e^h = \frac{\lambda^2}{4\pi^2 n^2 \Delta\nu_h \tau_{sp}} \tag{10.39}$$

for a homogeneously broadened medium with an ideal Lorentzian lineshape, and as

$$\sigma_e^{inh} = \frac{(\ln 2)^{1/2}\lambda^2}{4\pi^{3/2} n^2 \Delta\nu_{inh}\tau_{sp}} \tag{10.40}$$

for an inhomogeneously broadened medium with an ideal Gaussian lineshape. In practice, the experimentally measured peak emission cross section usually differs from that calculated using these formulas because the spectral lineshape of a realistic laser gain medium is generally determined by a combination of many different mechanisms and, consequently, is rarely ideal Lorentzian or ideal Gaussian. Nevertheless, these formulas provide a very good estimate for the peak value of the emission cross section. They also clearly indicate that *the emission cross section varies quadratically with the emission wavelength but is inversely proportional to both the emission linewidth and the spontaneous radiative lifetime of the laser transition*.

The characteristics of some representative laser materials are listed in Table 10.1. As seen from Table 10.1, the parameters vary over a wide range among different types of laser gain media. For example, the peak value of the emission cross section

Table 10.1 *Characteristics of some laser materials*

Gain medium	Wavelength λ (μm)	System[a]	Cross section[b] σ_e (m^2)	Spontaneous linewidth[c]		Lifetimes[d]		Index
				$\Delta\nu$	$\Delta\lambda$ (nm)	τ_{sp}	τ_2	n
HeNe	0.6328	I,4	3.0×10^{-17}	1.5 GHz	0.002	300 ns	30 ns	1
Ar ion	0.488	I,4	2.5×10^{-16}	2.7 GHz	0.004	13 ns	10 ns	1
CO_2	10.6	I,4	3.0×10^{-22}	60 MHz	0.02	4 s	1 μs	1
Copper vapor	0.5105	I,3	8.6×10^{-18}	2.3 GHz	0.002	500 ns	500 ns	1
KrF excimer	0.248	H,3	2.6×10^{-20}	10 THz	2	10 ns	8 ns	1
R6G dye	0.57–0.65	H/I,Q2	2.3×10^{-20}	30 THz	33	6 ns	4 ns	1.4
Ruby[e]	0.6943	H,3	$1.25–2.5 \times 10^{-24}$	330 GHz	0.53	3 ms	3 ms	1.76
Nd : YAG	1.064	H,4	$2–10 \times 10^{-23}$	150 GHz	0.56	515 μs	240 μs	1.82
Nd : glass	1.054	I,4	4.0×10^{-24}	6 THz	22	330 μs	330 μs	1.53
Er : fiber	1.53	H/I,3	6.0×10^{-25}	5 THz	40	10 ms	10 ms	1.46
Ti : sapphire[f]	0.66–1.1	H,Q2	3.4×10^{-23}	100 THz	180	3.9 μs	3.2 μs	1.76
Cr : LiSAF	0.78–1.01	H,Q2	4.8×10^{-24}	83 THz	200	67 μs	67 μs	1.4
Semiconductor	0.37–1.65	H/I,Q2	$1–5 \times 10^{-20}$	10–20 THz	20–100	~1 ns	~1 ns	3–4

[a] H, homogeneously broadened; I, inhomogeneously broadened; Q2, quasi-two-level system; 3, three-level system; 4, four-level system.

[b] Both the absorption and emission cross sections depend on the optical frequency. The absorption and emission cross sections generally have different peak values and different spectral dependencies. Listed is the peak value of the emission cross section.

[c] The spontaneous linewidth determines the gain bandwidth of a medium when population inversion is achieved.

[d] The spontaneous lifetime τ_{sp} is related to the transition rate, whereas the fluorescence lifetime τ_2 is related to the upper-level population relaxation. The fluorescence lifetime of a gaseous medium varies with temperature and pressure; that of a liquid or solid medium varies with temperature, the host material, and the concentration of the active ions or molecules. For example, τ_2 of CO_2 varies from 100 ns to 1 ms depending on temperature and pressure.

[e] Ruby is sapphire (Al_2O_3) doped with Cr^{3+} ions. The sapphire crystal is uniaxial. For ruby, the value of σ_e for emission with $\mathbf{E} \perp c$, which is listed, is larger than that for $\mathbf{E} \parallel c$.

[f] For Ti : sapphire, the value of σ_e for $\mathbf{E} \parallel c$, which is listed, is larger than that for $\mathbf{E} \perp c$.

varies from 6×10^{-25} m^2 for Er : fiber to 2.5×10^{-16} m^2 for the Ar-ion laser, whereas the spontaneous emission linewidth varies from 60 MHz for CO_2 to 100 THz for Ti : sapphire. The fluorescence lifetime varies from the order of 1 ns for semiconductor gain media to the order of 10 ms for Er : fiber.

EXAMPLE 10.3 The emission at 632.8 nm wavelength of a HeNe laser is inhomogeneously broadened due to Doppler broadening with a linewidth of $\Delta\nu \approx \Delta\nu_{\text{inh}} = \Delta\nu_{\text{D}} = 1.5$ GHz. The spontaneous radiative lifetime is $\tau_{\text{sp}} = 300$ ns. Being a gas laser, the refractive index of the medium is $n \approx 1$. Find the peak emission cross section of the HeNe laser at this wavelength.

Solution Using (10.40), we find the following peak emission cross section at $\lambda = 632.8$ nm for the HeNe laser:

$$\sigma_{\text{e}} = \sigma_{\text{e}}^{\text{inh}} = \frac{(\ln 2)^{1/2} \times (632.8 \times 10^{-9})^2}{4 \times \pi^{3/2} \times 1^2 \times 1.5 \times 10^9 \times 300 \times 10^{-9}} \text{ m}^2 = 3.3 \times 10^{-17} \text{ m}^2.$$

This calculated result agrees, by a small difference of only 10%, with the value of 3.0×10^{-17} m^2 quoted in the literature, which is listed in Table 10.1.

10.2 Optical absorption and amplification

Optical absorption results in attenuation of an optical field, while stimulated emission leads to amplification of an optical field. To quantify the net effect of a resonant transition on the attenuation or amplification of an optical field, we consider the interaction of a monochromatic plane optical field at a frequency ν with a material that consists of electronic or atomic systems with population densities N_1 and N_2 in energy levels $|1\rangle$ and $|2\rangle$, respectively. Because the spectral intensity distribution of the monochromatic plane optical field that has an intensity I is simply $I(\nu) = I\delta(\nu' - \nu)$, the induced transition rates between energy levels $|1\rangle$ and $|2\rangle$ in this interaction are

$$W_{21} = \frac{I}{h\nu}\sigma_{\text{e}}(\nu) \qquad \text{and} \qquad W_{12} = \frac{I}{h\nu}\sigma_{\text{a}}(\nu). \tag{10.41}$$

The net power that is transferred from the optical field to the material is the difference between that absorbed by the material and that emitted due to stimulated emission:

$$\begin{aligned} \overline{W}_{\text{p}} &= h\nu W_{12} N_1 - h\nu W_{21} N_2 \\ &= [N_1\sigma_{\text{a}}(\nu) - N_2\sigma_{\text{e}}(\nu)]\, I. \end{aligned} \tag{10.42}$$

In the case when $\overline{W}_{\text{p}} > 0$, there is net power absorption from the optical field by the medium due to resonant transitions between energy levels $|1\rangle$ and $|2\rangle$. The absorption

coefficient is

$$\alpha(\nu) = N_1\sigma_a(\nu) - N_2\sigma_e(\nu) = \left(N_1 - \frac{g_1}{g_2}N_2\right)\sigma_a(\nu). \tag{10.43}$$

In the case when $\overline{W}_p < 0$, net power flows from the medium to the optical field, resulting in an amplification to the optical field with a gain coefficient given by

$$g(\nu) = N_2\sigma_e(\nu) - N_1\sigma_a(\nu) = \left(N_2 - \frac{g_2}{g_1}N_1\right)\sigma_e(\nu). \tag{10.44}$$

The coefficients α and g are quoted per meter, and are also often quoted per centimeter. Note that $\alpha(\nu) = \alpha(\omega)$ and $g(\nu) = g(\omega)$ because $\sigma(\nu) = \sigma(\omega)$.

According to (10.41), both $\sigma_a(\nu)$ and $\sigma_e(\nu)$ are positive because $W_{21} \geq 0$ and $W_{12} \geq 0$ by definition. We then find that $\alpha(\nu) > 0$ and $g(\nu) < 0$ if $N_1 > (g_1/g_2)N_2$, whereas $g(\nu) > 0$ and $\alpha(\nu) < 0$ if $N_2 > (g_2/g_1)N_1$. Therefore, a material absorbs optical energy in its normal state of thermal equilibrium when the lower energy level is more populated than the upper energy level. In order to provide a net optical gain to the optical field, a material has to be in a nonequilibrium state of *population inversion* with the upper energy level more populated than the lower energy level.

EXAMPLE 10.4 The upper and lower laser levels of the ruby laser are shown in Fig. 10.6. The lower laser level $|1\rangle$ of the ruby laser is the ground state $^4A_2$, which has a degeneracy factor of $g_1 = 4$. The upper laser level $|2\rangle$ is the ^2E state, which consists of two closely spaced $2\overline{A}$ and $\overline{E}$ sublevels, each with a degeneracy factor of 2. The 694.3 nm ruby laser transition takes place between the $\overline{E}$ sublevel, which has a degeneracy factor of $g(\overline{E}) = 2$, and the ground state $^4A_2$ with an emission cross section of $\sigma_e^{line} = 2.5 \times 10^{-24}$ m^2 for the $\mathbf{E} \perp c$ polarization. Find the peak value of the absorption cross section for an optical wave at 694.3 nm polarized with $\mathbf{E} \perp c$. At room temperature without pumping, what is the absorption coefficient at 694.3 nm of a ruby crystal that is doped with a Cr concentration of 1.58×10^{25} m^{-3}?

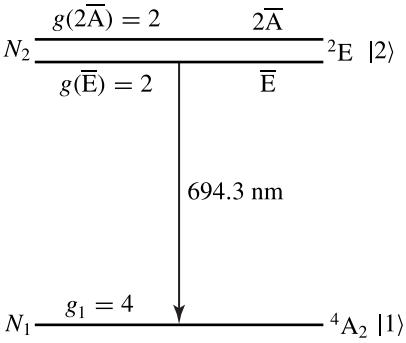


Figure 10.6 Upper and lower laser levels of the ruby laser.

Solution To find σ_a, we use (10.37) by taking $\sigma_e = \sigma_e^{line}$ for the $\overline{E}$ sublevel. Thus, we have

$$\sigma_a = \frac{g(\overline{E})}{g_1}\sigma_e = \frac{2}{4} \times 2.5 \times 10^{-24}\ m^2 = 1.25 \times 10^{-24}\ m^2.$$

At room temperature without pumping, the upper laser level is almost totally unpopulated because it is 1.786 eV above ground level. Virtually all of the Cr ions are in the ground level. Therefore, $N_1 = 1.58 \times 10^{25}\ m^{-3}$ and $N_2 = 0$. Then, the absorption coefficient is

$$\alpha = N_1\sigma_a = 1.58 \times 10^{25} \times 1.25 \times 10^{-24}\ m^{-1} = 19.75\ m^{-1}.$$

Laser level splitting

In Example 10.4, we see that the upper laser level of the ruby laser consists of two closely spaced but clearly separate sublevels, corresponding to laser lines at 692.9 and 694.3 nm, respectively. The population N_2 in the upper laser level is split between these two sublevels. Such laser level splitting also occurs in most other lasers. As a consequence, only a fractional population of ξN_2 that resides in a particular sublevel of the upper laser level is directly responsible for a particular laser transition, whereas the remaining population of $(1 - \xi)N_2$ that resides in other sublevels does not contribute to this transition. While taking N_2 to be the total population of the upper laser level including all sublevels in a situation like this, the emission cross section σ_e used in (10.43) and (10.44) is a cross section that is weighted as $\sigma_e = \xi\sigma_e^{line}$, where σ_e^{line} is the emission cross section of the specific transition line for the population ξN_2 in its sublevel. The parameter ξ of a given laser medium varies with many factors, including temperature, crystal quality, doping concentration, and the presence of codopants. This explains the variations in the measured values for the emission cross section of a laser medium, as seen in the values of σ_e for ruby and Nd : YAG listed in Table 10.1. A similar effect also exists for the absorption cross section.

EXAMPLE 10.5 For the ruby laser, the $2\overline{A}$ and $\overline{E}$ sublevels within upper laser level $|2\rangle$ of the ^2E state have an energy separation of $\Delta E = 29\ cm^{-1}$, which is $\Delta E = 3.6$ meV ($1\ cm^{-1} \equiv 30$ GHz $\equiv 124\ \mu eV$). As discussed in Example 10.4, the laser transition between sublevel $\overline{E}$ and the ground state $^4A_2$ is the 694.3-nm line with $\sigma_e^{line} = 2.5 \times 10^{-24}\ m^2$ for $\mathbf{E} \perp c$ polarization. The 692.9-nm transition between sublevel $2\overline{A}$ and the ground state $^4A_2$ has a similar cross section. (a) What is the population distribution at 300 K between the two sublevels in the upper laser level? What is the weighted emission cross section σ_e for the 694.3-nm transition? (b) What is σ_e for the 694.3-nm transition at 77 K?

Solution (a) At $T = 300$ K, $k_B T = 25.9$ meV. Because the $2\overline{A}$ state lies above the $\overline{E}$ state by an energy difference of $\Delta E = 3.6$ meV and the two states have the same degeneracy factor of 2, we have

$$\frac{N(2\overline{A})}{N(\overline{E})} = \frac{g(2\overline{A})}{g(\overline{E})} \exp(-\Delta E/k_B T) = \frac{2}{2} \times e^{-3.6/25.9} = 0.87.$$

Therefore, the fraction of the N_2 population in sublevel $\overline{E}$ is

$$\xi = \frac{N(\overline{E})}{N(2\overline{A}) + N(\overline{E})} = \frac{1}{0.87 + 1} = 0.535.$$

This means that only 53.5% of population N_2 in the upper laser level ^{2}E state contributes directly to the 694.3-nm transition. Thus the weighted emission cross section for the 694.3-nm transition is

$$\sigma_e = \xi \sigma_e^{\text{line}} = 0.535 \times 2.5 \times 10^{-24}\ \text{m}^2 = 1.34 \times 10^{-24}\ \text{m}^2.$$

(b) At $T = 77$ K, $k_B T = 6.64$ meV. Then,

$$\frac{N(2\overline{A})}{N(\overline{E})} = \frac{g(2\overline{A})}{g(\overline{E})} \exp(-\Delta E/k_B T) = \frac{2}{2} \times e^{-3.6/6.64} = 0.58,$$

and $\xi = 1/(0.58 + 1) = 0.632$. Therefore, 63.2% of population N_2 in the upper laser level now contributes to the 694.3-nm transition with a weighted emission cross section of

$$\sigma_e = \xi \sigma_e^{\text{line}} = 0.632 \times 2.5 \times 10^{-24}\ \text{m}^2 = 1.58 \times 10^{-24}\ \text{m}^2.$$

If the temperature is further lowered, σ_e for the 694.3-nm transition will further increase toward its maximum value of 2.5×10^{-24} m^2 as the $\overline{E}$ sublevel takes up a larger fraction of the total population in the upper laser level.

In many systems, the degenerate states in each laser level are split not into clearly separate sublevels but into very closely spaced sublevels that form a small quasi-continuous energy band, as shown in Fig. 10.7. In a molecular gas medium such as CO_2, for example, transition levels $|1\rangle$ and $|2\rangle$ are defined by the vibrational states of the CO_2 molecule, each of which consists of many closely spaced rotational sublevels. In laser dyes, transition levels $|1\rangle$ and $|2\rangle$ are electronic states. Due to the vibrational and rotational motions of the dye molecules, the electronic states are split into vibrational sublevels, which are further split into finer structures of rotational sublevels. In dielectric solid-state media doped with transition-metal or rare-earth ions, such as Ti : sapphire, Nd : glass, and Er-doped glass fiber, transition levels $|1\rangle$ and $|2\rangle$ are the electronic energy levels of the dopant ions. The degeneracies in such energy levels are contributed by the angular momentum states of the dopant ion. Because a dopant ion is embedded in a host solid-state medium, the electric fields of the neighboring atoms in the host medium cause

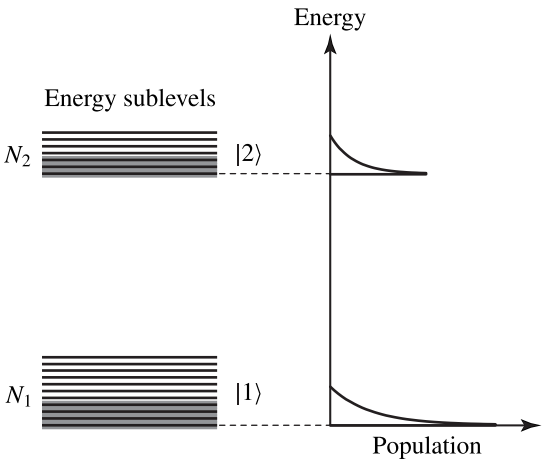


Figure 10.7 Splitting of the upper and lower transition levels into respective quasi-continuous bands of sublevels.

some or all of its degenerate angular momentum states within a given energy level to split into a band of sublevels due to the Stark effect. Interaction with phonons in the lattice of the host medium can further broaden the energy band in each level.

Within the band of a transition level, the population at a higher sublevel can relax to a lower sublevel very quickly through nonradiative processes. In CO_2 and laser dyes, such relaxation takes place through collisions among the molecules. In a solid-state medium that is doped with transition-metal or rare-earth ions, relaxation takes place through interaction of the ions with the phonons, i.e., the lattice vibrations, of the host material. These are thermal processes whose efficiency depends on temperature. Because the sublevels within the band of a transition level are very closely spaced in energy, at room temperature relaxation among them takes place in a time much shorter than that between different transition levels. As a result, before any optical transition begins, the sublevels within each transition level are generally thermalized to be in thermal equilibrium with the medium. This thermalization leads to a Boltzmann population distribution among the sublevels within the band of each transition level. Within a band, the lower sublevels are more populated than the higher sublevels. Consequently, N_1 is not evenly distributed among the g_1 states of level $|1\rangle$, and N_2 is not evenly distributed among the g_2 states of level $|2\rangle$, as illustrated also in Fig. 10.7. Because of this nonuniform population distribution, absorption occurs with a higher probability from a lower sublevel in the band of level $|1\rangle$ to a higher sublevel in the band of level $|2\rangle$, whereas emission is more likely to take place from a low-lying sublevel in level $|2\rangle$ to a high-lying sublevel in level $|1\rangle$. The consequences are:

1. The absorption and emission spectra associated with the same pair of transition levels $|1\rangle$ and $|2\rangle$ that consist of split sublevels are generally not identical. They have

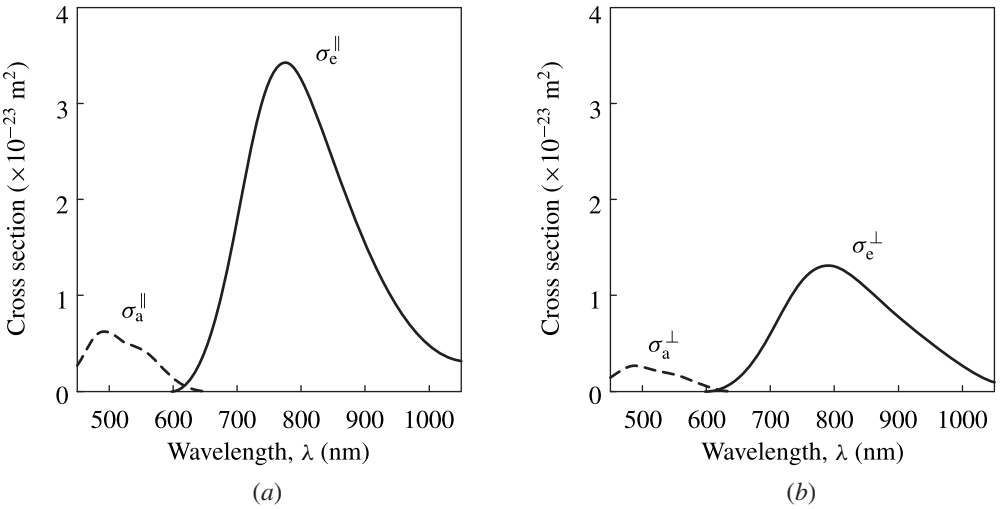


Figure 10.8 Spectra of the absorption and emission cross sections of Ti : sapphire at room temperature for (*a*) $\mathbf{E} \parallel c$ and (*b*) $\mathbf{E} \perp c$. Sapphire is a uniaxial crystal. The absorption and emission spectra depend on the polarization of the radiation with respect to the unique crystal axis c, which defines the optical axis. The extraordinary polarization with $\mathbf{E} \parallel c$ has larger absorption and emission cross sections than the ordinary polarization with $\mathbf{E} \perp c$. Note that the curves for σ_a and σ_e for both polarizations are properly scaled according to (10.46) with a degeneracy ratio of $g_1/g_2 = 3/2$ for Ti : Al_2O_3. (Adapted from Moulton, P. F., "Spectroscopic and laser characteristics of Ti : Al_2O_3," *Journal of the Optical Society of America B* **3**(1): 125–133, Jan. 1986.)

different shapes and widths, and both vary with temperature. The absorption spectrum is generally shifted to the side of shorter wavelengths, corresponding to higher photon energies, with respect to the emission spectrum. As an example, Fig. 10.8 shows the spectra of the absorption and emission cross sections of Ti : sapphire at room temperature.

2. The relation in (10.36) is still valid but $\hat{g}(\nu)$ now represents the normalized emission spectral lineshape, which is different from the absorption lineshape. Therefore, the relation in (10.37) is no longer valid. Instead, $\sigma_a(\nu)$ and $\sigma_e(\nu)$ satisfy the following general relation (see Problem 10.2.4(b)):

$$\frac{1}{\tau_{sp}} = \frac{8\pi n^2}{c^2} \int_0^\infty \nu^2 \sigma_e(\nu) d\nu = \frac{8\pi n^2}{c^2} \frac{g_1}{g_2} \int_0^\infty \nu^2 \sigma_a(\nu) d\nu. \tag{10.45}$$

The validity of this relation is based on the assumption that all components in either of the two levels are equally populated or all transitions between the two levels have equal probability regardless of the components involved. Because the experimentally measured spectra of emission and absorption cross sections are normally expressed as functions of wavelength rather than frequency, we can convert the relation in

(10.45) into the following useful relation in terms of wavelength:

$$\frac{1}{\tau_{sp}} = 8\pi n^2 c \int_0^\infty \frac{\sigma_e(\lambda)}{\lambda^4} d\lambda = 8\pi n^2 c \frac{g_1}{g_2} \int_0^\infty \frac{\sigma_a(\lambda)}{\lambda^4} d\lambda. \tag{10.46}$$

This relation can be used to determine the spontaneous lifetime τ_{sp} by integrating the experimentally measured absorption or emission cross section as a function of wavelength. Note that $\sigma(\lambda) = \sigma(\nu)$ for $\lambda = c/\nu$.

3. A detailed relation between $\sigma_e(\nu)$ and $\sigma_a(\nu)$ is known as the *McCumber relation*:[3]

$$\sigma_e(\nu) = \sigma_a(\nu) \exp\left(\frac{h\nu_c - h\nu}{k_B T}\right), \tag{10.47}$$

where ν_c is the optical frequency at which the absorption and emission cross sections are equal: $\sigma_e(\nu_c) = \sigma_a(\nu_c)$. In terms of wavelength, the McCumber relation can be expressed as

$$\sigma_e(\lambda) = \sigma_a(\lambda) \exp\left[\frac{hc}{k_B T}\left(\frac{1}{\lambda_c} - \frac{1}{\lambda}\right)\right], \tag{10.48}$$

where $\lambda_c = c/\nu_c$ for which $\sigma_e(\lambda_c) = \sigma_a(\lambda_c)$. The photon energy $h\nu_c$ corresponds to the temperature-dependent excitation energy that is equivalent to the free energy required to move one atom from the lower level $|1\rangle$ to the upper level $|2\rangle$. According to (10.47), the spectra of $\sigma_e(\nu)$ and $\sigma_a(\nu)$ associated with the transition between two energy levels $|1\rangle$ and $|2\rangle$ cross at only one frequency ν_c. The McCumber relation is generally applicable because it does not depend on the assumption that is required for the validity of (10.45). The only assumption needed is that the sublevels within either level $|1\rangle$ or $|2\rangle$ reach thermal equilibrium in a time shorter than the lifetime of each energy level.

4. The first parts of the relations in (10.43) and (10.44) for the absorption and emission coefficients, respectively, are still valid, but not the second parts:

$$\alpha(\nu) = N_1\sigma_a(\nu) - N_2\sigma_e(\nu) \neq \left(N_1 - \frac{g_1}{g_2}N_2\right)\sigma_a(\nu) \tag{10.49}$$

and

$$g(\nu) = N_2\sigma_e(\nu) - N_1\sigma_a(\nu) \neq \left(N_2 - \frac{g_2}{g_1}N_1\right)\sigma_e(\nu). \tag{10.50}$$

EXAMPLE 10.6 Find the spontaneous lifetime τ_{sp} for the laser transition of Ti : sapphire from the spectrum of its emission cross section shown in Fig. 10.8.

[3] This relation is based on the principle of detailed balance and is a generalization of (10.24). For details, see McCumber, D. E., "Theory of phonon-terminated optical masers," *Physical Review* **134**: A299–A306, 1964.

Solution According to (10.46), τ_{sp} can be found from $\sigma_e(\lambda)$ by integrating $\sigma_e(\lambda)/\lambda^4$ over the entire spectrum. In applying (10.46) to the spectra shown in Fig. 10.8, however, we have to account for the difference between the emission spectra for different polarizations. With respect to the unique crystal axis c of the uniaxial sapphire, there are three polarization modes, one parallel to c and two perpendicular to c. As a consequence, the spontaneous emission resulting from the radiative transition that defines τ_{sp} for Ti : sapphire has a 1 : 2 ratio between the $\mathbf{E} \parallel c$ and $\mathbf{E} \perp c$ polarized emissions. Therefore, for a uniaxial crystal such as Ti : sapphire, (10.46) has to be modified as

$$\int_0^\infty \frac{\sigma_e^{\parallel}(\lambda) + 2\sigma_e^{\perp}(\lambda)}{3\lambda^4} d\lambda = \frac{g_1}{g_2} \int_0^\infty \frac{\sigma_a^{\parallel}(\lambda) + 2\sigma_a^{\perp}(\lambda)}{3\lambda^4} d\lambda = \frac{1}{8\pi n^2 c \tau_{sp}}. \tag{10.51}$$

Using the $\sigma_e^{\parallel}(\lambda)$ and $\sigma_e^{\perp}(\lambda)$ spectra shown in Fig. 10.8, we find that

$$\int_0^\infty \frac{\sigma_e^{\parallel}(\lambda) + 2\sigma_e^{\perp}(\lambda)}{3\lambda^4} d\lambda = 1.1 \times 10^{-5}\ \mathrm{m}^{-1}.$$

Using (10.51) and $n = 1.76$ for Ti : sapphire, we find that

$$\tau_{sp} = \frac{1}{8\pi \times 1.76^2 \times 3 \times 10^8 \times 1.1 \times 10^{-5}}\ \mathrm{s} = 3.89\ \mu\mathrm{s}.$$

Resonant optical susceptibility

The macroscopic optical properties of a medium are characterized by its electric susceptibility. As seen in Section 1.10, resonances in a medium contribute to the dispersion in the susceptibility of the medium. Clearly, the optical properties of a material are functions of the resonant optical transitions between the energy levels of the electrons in the material.

From the viewpoint of the macroscopic optical properties of a medium, interaction between an optical field and a medium is characterized by the polarization induced by the optical field in the medium. The power exchange between the optical field and the medium is given by (1.30). For resonant interaction of an isotropic medium with a monochromatic plane optical field at a frequency $\omega = 2\pi\nu$, we have $\boldsymbol{E}(t) = \mathbf{E}e^{-i\omega t} + \mathbf{E}^* e^{i\omega t}$ and $\boldsymbol{P}_{res}(t) = \epsilon_0(\chi_{res}(\omega)\mathbf{E}e^{-i\omega t} + \chi_{res}^*(\omega)\mathbf{E}^* e^{i\omega t})$, where $\boldsymbol{P}_{res}$ is the polarization contributed by the resonant transitions and χ_{res} is the resonant susceptibility. Using (1.30), we find that the time-averaged power density absorbed by the medium is

$$\overline{W}_p = 2\omega\epsilon_0 \chi_{res}''(\omega)|\mathbf{E}|^2 = \frac{\omega}{nc}\chi_{res}''(\omega) I. \tag{10.52}$$

By identifying (10.52) with (10.42), we find that the imaginary part of the susceptibility contributed by the resonant transitions between energy levels $|1\rangle$ and $|2\rangle$ is

$$\chi''_{\text{res}}(\omega) = \frac{nc}{\omega}[N_1\sigma_{\text{a}}(\omega) - N_2\sigma_{\text{e}}(\omega)]. \tag{10.53}$$

The real part $\chi'_{\text{res}}(\omega)$ of the resonant susceptibility can be found through the Kramers–Kronig relations given in (1.177).

As discussed in Sections 1.5 and 1.10, a medium has an optical loss if $\chi'' > 0$, and it has an optical gain if $\chi'' < 0$. It is also clear from (10.52) that there is a net power loss from the optical field to the medium if $\chi''_{\text{res}} > 0$, but there is a net power gain for the optical field if $\chi''_{\text{res}} < 0$. By comparing (10.53) with (10.43) and (10.44), we find that the medium has an absorption coefficient given by

$$\alpha(\omega) = \frac{\omega}{nc}\chi''_{\text{res}}(\omega) \tag{10.54}$$

in the case of normal population distribution when $\chi''_{\text{res}} > 0$, whereas it has a gain coefficient given by

$$g(\omega) = -\frac{\omega}{nc}\chi''_{\text{res}}(\omega) \tag{10.55}$$

in the case of population inversion when $\chi''_{\text{res}} < 0$.

Note that the material susceptibility characterizes the response of a material to the excitation of an electromagnetic field. Therefore, the resonant susceptibility χ_{res} accounts for only the contributions from the induced processes of absorption and stimulated emission, but not that from the process of spontaneous emission. The resonant susceptibility contributed by the induced transitions between two energy levels is proportional to the population difference between the two levels, but the power density of the optical radiation due to spontaneous emission is a function of the population density in the upper energy level alone.

By taking $\Delta\mathbf{P} = \mathbf{P}_{\text{res}}$, the behavior of an optical field propagating in the presence of resonant transitions can be formulated with the coupled-wave theory discussed in Section 4.1, if the medium is spatially homogeneous, or with the coupled-mode theory discussed in Section 4.2, if the medium has waveguiding structures. Here we consider the simplest situation involving a monochromatic wave at a frequency ω that propagates along the z direction in a spatially homogeneous, isotropic medium with a resonant susceptibility χ_{res}. Then, the index q in the coupled-wave equation expressed in (4.13) can be dropped:

$$\frac{\text{d}\mathcal{E}(z)}{\text{d}z} = \frac{\text{i}\omega^2\mu_0}{2k}\mathbf{P}_{\text{res}}(z)\text{e}^{-\text{i}kz}, \tag{10.56}$$

where

$$\mathbf{P}_{\text{res}}(z) = \epsilon_0\chi_{\text{res}}(\omega)\mathbf{E}(z) = \epsilon_0(\chi'_{\text{res}} + \text{i}\chi''_{\text{res}})\mathcal{E}(z)\text{e}^{\text{i}kz}. \tag{10.57}$$

Substitution of (10.57) in (10.56) yields

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}z} = \mathrm{i}\frac{\omega}{2nc}\chi'_{\mathrm{res}}\mathcal{E} - \frac{\omega}{2nc}\chi''_{\mathrm{res}}\mathcal{E}. \tag{10.58}$$

We see from this equation that, as the optical field propagates, not only is its amplitude varied by the resonant susceptibility, but its phase is modified as well.

When the phase information of the optical wave is of no interest, we can take $\mathcal{E}^* \cdot (10.58) + \text{c.c.}$ to find the evolution of the intensity of the optical wave as it propagates through the medium. Using the relations in (10.54) and (10.55), we find that

$$\frac{\mathrm{d}I}{\mathrm{d}z} = -\alpha I \tag{10.59}$$

in the case of optical attenuation when $\chi''_{\mathrm{res}} > 0$, and

$$\frac{\mathrm{d}I}{\mathrm{d}z} = gI \tag{10.60}$$

in the case of optical amplification when $\chi''_{\mathrm{res}} < 0$. Clearly, the coefficients α and g respectively characterize the attenuation and growth of the optical intensity per unit length traveled by the optical wave in a medium.

10.3 Population inversion and optical gain

From the discussions in the preceding section, it is clear that population inversion is the basic condition for the presence of an optical gain. In the normal state of any system in thermal equilibrium, a low-energy state is always more populated than a high-energy state, hence no population inversion. Population inversion in a system can only be accomplished through a process called *pumping* by actively exciting the atoms in a low-energy state to a high-energy state. If left alone, the atoms in a system will relax to thermal equilibrium. Therefore, population inversion is a nonequilibrium state that cannot be sustained without active pumping. To maintain a constant optical gain, continuous pumping is required to keep the population inversion at a constant level. This condition is clearly consistent with the law of conservation of energy: amplification of an optical wave leads to an increase in optical energy, which is possible only if there is a source supplying the energy.

Pumping is the process that supplies the energy to the gain medium for the amplification of an optical wave. There are many different pumping techniques, including optical excitation, electric current injection, electric discharge, chemical reaction, and excitation with particle beams. The use of a particular pumping technique depends on the properties of the gain medium being pumped. The lasers and optical amplifiers of particular interest in photonic systems are made of either dielectric solid-state media doped with active ions, such as Nd : YAG and Er : glass fiber, or direct-gap semiconductors, such as GaAs and InP. For dielectric media, the most commonly used pumping

technique is optical pumping either with incoherent light sources, such as flashlamps and light-emitting diodes, or with coherent light sources from other lasers. Semiconductor gain media can also be optically pumped, but they are usually pumped with electric current injection. In this section, we consider the general conditions for pumping to achieve population inversion for an optical gain. Detailed pumping mechanisms and physical setups are not addressed here because they depend on the specific gain medium used in a given application.

Rate equations

The net rate of increase of population density in a given energy level is described by a rate equation. As we shall see below, pumping for population inversion in any practical gain medium always requires the participation of more than two energy levels. In general, a rate equation has to be written for each energy level that is involved in the process. For simplicity but without loss of validity, however, we shall explicitly write down only the rate equations for the two energy levels, $|2\rangle$ and $|1\rangle$, that are directly associated with the resonant transition of interest. We are not interested in the population densities of other energy levels but only in how those levels affect N_2 and N_1.

In the presence of a monochromatic, coherent optical wave of intensity I at a frequency ν, the rate equations for N_2 and N_1 are

$$\frac{\mathrm{d}N_2}{\mathrm{d}t} = R_2 - \frac{N_2}{\tau_2} - \frac{I}{h\nu}(N_2\sigma_\mathrm{e} - N_1\sigma_\mathrm{a}), \tag{10.61}$$

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = R_1 - \frac{N_1}{\tau_1} + \frac{N_2}{\tau_{21}} + \frac{I}{h\nu}(N_2\sigma_\mathrm{e} - N_1\sigma_\mathrm{a}), \tag{10.62}$$

where R_2 and R_1 are the total rates of pumping into energy levels $|2\rangle$ and $|1\rangle$, respectively, and τ_2 and τ_1 are the fluorescence lifetimes of levels $|2\rangle$ and $|1\rangle$, respectively. The rate of population decay, including radiative and nonradiative spontaneous relaxation, from level $|2\rangle$ to level $|1\rangle$ is τ_{21}^{-1}. Because it is possible for the population in level $|2\rangle$ to relax to other energy levels also, the total population decay rate of level $|2\rangle$ is $\tau_2^{-1} \geq \tau_{21}^{-1}$. Therefore, in general, we have

$$\tau_2 \leq \tau_{21} \leq \tau_\mathrm{sp}. \tag{10.63}$$

Note that τ_{21}^{-1} is not the same as γ_{21} defined in (10.8): τ_{21}^{-1} is purely the rate of *population relaxation* from level $|2\rangle$ to level $|1\rangle$, whereas γ_{21} is the rate of *phase relaxation* of the polarization associated with the transition between these two levels.

In an optical gain medium, level $|2\rangle$ is known as the *upper laser level* and level $|1\rangle$ is known as the *lower laser level*. The fluorescence lifetime τ_2 of the upper laser level is an important parameter that determines the effectiveness of a gain medium. Generally speaking, the upper laser level has to be a metastable state with a relatively large τ_2 for a gain medium to be useful.

Population inversion

Population inversion in a medium is generally defined as

$$N_2 > \frac{g_2}{g_1} N_1. \tag{10.64}$$

According to (10.50), however, this condition does not guarantee an optical gain at a particular optical frequency ν if $\sigma_a(\nu) \neq (g_2/g_1)\sigma_e(\nu)$ when the population in each level, $|1\rangle$ or $|2\rangle$, is distributed unevenly among its sublevels. For this reason, when the condition for population inversion given in (10.64) is achieved in a medium, we may find an optical gain at an optical frequency ν where $\sigma_a(\nu) \leq (g_2/g_1)\sigma_e(\nu)$, but at the same time find an optical loss at another frequency ν' where $\sigma_a(\nu') > (g_2/g_1)\sigma_e(\nu')$. What really matters for an optical wave at a given frequency is the optical gain at that particular frequency. Therefore, in the following discussions, we shall consider, instead of the condition in (10.64), the following condition:

$$N_2\sigma_e(\nu) - N_1\sigma_a(\nu) > 0, \tag{10.65}$$

which guarantees an optical gain at frequency ν, as the effective condition of population inversion as far as an optical signal at frequency ν is concerned.

The pumping requirement for the condition in (10.65) to be satisfied depends on the properties of a medium. For atomic and molecular media, there are three different basic systems. Each has a different pumping requirement to reach effective population inversion for an optical gain. The pumping requirement can be found by solving the coupled rate equations in (10.61) and (10.62).

Two-level system

When the only energy levels involved in the pumping and the relaxation processes are the upper and the lower laser levels $|2\rangle$ and $|1\rangle$, the system can be considered as a two-level system. In such a system, level $|1\rangle$ is the ground state with $\tau_1 = \infty$, and level $|2\rangle$ relaxes only to level $|1\rangle$ so that $\tau_{21} = \tau_2$. The total population density is $N_t = N_1 + N_2$. While a pumping mechanism excites atoms from the lower laser level to the upper laser level, the same pump also stimulates atoms in the upper laser level to relax to the lower laser level. Therefore, irrespective of the specific pumping technique used, $R_2 = -R_1 = W_{12}^p N_1 - W_{21}^p N_2$, where W_{12}^p and W_{21}^p are the *pumping transition probability rates*, or simply the *pumping rates*, from $|1\rangle$ to $|2\rangle$ and from $|2\rangle$ to $|1\rangle$, respectively. Under these conditions, (10.61) and (10.62) are equivalent to each other. The upward and downward pumping transition rates are not independent of each other but are directly proportional to each other because both are associated with the interaction of the same pump source with a given set of energy levels. We take the upward pumping rate to be $W_{12}^p = W_p$ and the downward pumping rate to be $W_{21}^p = pW_p$, where

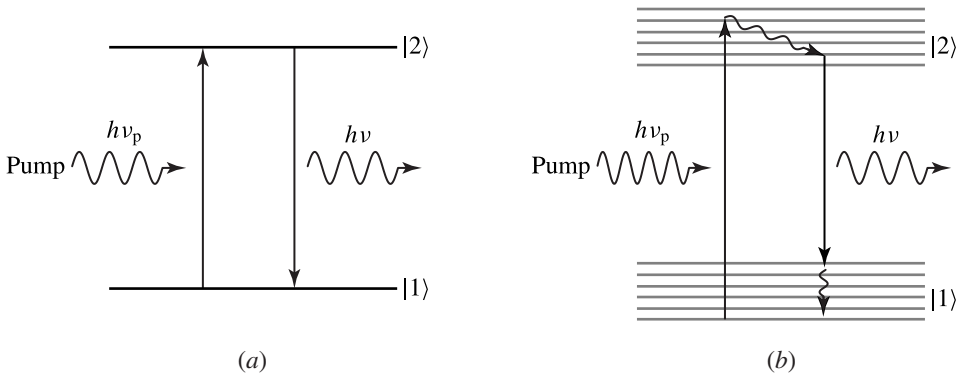


Figure 10.9 (*a*) Pumping scheme of a true two-level system. (*b*) Pumping scheme of a quasi-two-level system.

p is a constant that depends on the detailed characteristics of the two-level atomic system and the pump source. In the steady state when $dN_2/dt = dN_1/dt = 0$, we then find that

$$N_2\sigma_e - N_1\sigma_a = \frac{W_p\tau_2(\sigma_e - p\sigma_a) - \sigma_a}{1 + (1+p)W_p\tau_2 + (I\tau_2/h\nu)(\sigma_e + \sigma_a)}N_t. \tag{10.66}$$

Using the relation in (10.41), we find that, for optical pumping,

$$p = \frac{\sigma_e^p}{\sigma_a^p} = \frac{\sigma_e(\lambda_p)}{\sigma_a(\lambda_p)}, \tag{10.67}$$

where σ_a^p and σ_e^p are the absorption and emission cross sections, respectively, at the pump wavelength.

In a true two-level system, shown in Fig. 10.9(*a*), the energy levels $|2\rangle$ and $|1\rangle$ can each be degenerate with degeneracies g_2 and g_1, respectively, but the population densities in both levels are evenly distributed among the respective degenerate states. In this situation, $p = \sigma_e^p/\sigma_a^p = g_1/g_2 = \sigma_e/\sigma_a$. Then, we find from (10.66) that

$$N_2\sigma_e - N_1\sigma_a = \frac{-\sigma_a}{1 + (\sigma_e + \sigma_a)(I/h\nu + W_p/\sigma_a)\tau_2}N_t < 0. \tag{10.68}$$

No matter how a true two-level system is pumped, it is clearly not possible to achieve population inversion for an optical gain in the steady state. This situation can be understood by considering the fact that the pump for a two-level system has to be in resonance with the transition between the two levels, thus inducing downward transitions as well as upward transitions. In the steady state, the two-level system would reach thermal equilibrium with the pump at a finite temperature T, resulting in a Boltzmann population distribution of the form given in (10.26) without population inversion.

As discussed in the preceding section and illustrated in Fig. 10.7, however, in many cases an energy level is actually split into a band of closely spaced, but not exactly degenerate, sublevels with its population density unevenly distributed among these sublevels. A system is not a true two-level system, but is known as a *quasi-two-level*

system, if either or both of the two levels involved are split in such a manner. By pumping such a quasi-two-level system properly, it is possible to reach the needed population inversion in the steady state for an optical gain at a particular laser frequency ν because the ratio $p = \sigma_e^p/\sigma_a^p$ at the pump frequency ν_p can now be made different from the ratio σ_e/σ_a at the laser frequency ν due to the uneven population distribution among the sublevels within an energy level. From (10.66), we find that the pumping requirements for a steady-state optical gain from a quasi-two-level system are

$$p = \frac{\sigma_e^p}{\sigma_a^p} < \frac{\sigma_e}{\sigma_a} \quad \text{and} \quad W_p > \frac{1}{\tau_2}\frac{\sigma_a}{\sigma_e - p\sigma_a}. \tag{10.69}$$

Because the absorption spectrum is generally shifted to the short-wavelength side of the emission spectrum, as discussed in the preceding section and demonstrated in Fig. 10.8, these conditions can be satisfied by pumping sufficiently strongly at a higher transition energy than the photon energy corresponding to the peak of the emission spectrum. In the case of optical pumping, this condition means that the pump wavelength has to be shorter than the emission wavelength. Figure 10.9(*b*) illustrates such a pumping scheme of a quasi-two-level system. Indeed, many laser gain media, including laser dyes, semiconductor gain media, and vibronic solid-state gain media, are often pumped as a quasi-two-level system.

Three-level system

Population inversion in steady state is possible for a system that has three energy levels involved in the process. Figure 10.10 shows the energy-level diagram of an idealized three-level system. The lower laser level $|1\rangle$ is the ground state, $E_1 = E_0$, or is very close to the ground state, within an energy separation of $\Delta E_{10} \ll k_B T$ from the ground state, so that it is initially populated. The atoms are pumped to an energy level $|3\rangle$ above the upper laser level $|2\rangle$.

An effective three-level system satisfies the following conditions: (1) population relaxation from level $|3\rangle$ to level $|2\rangle$ is very fast and efficient, ideally $\tau_2 \gg \tau_{32} \approx \tau_3$,

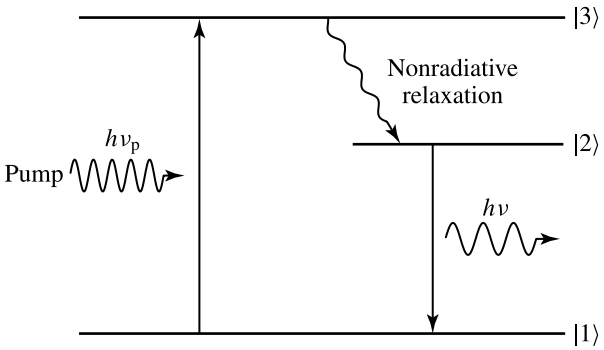


Figure 10.10 Energy levels of a three-level system.

so that the atoms excited by the pump quickly end up in level $|2\rangle$; (2) level $|3\rangle$ lies sufficiently high above level $|2\rangle$ with $\Delta E_{32} \gg k_B T$ so that the population in level $|2\rangle$ cannot be thermally excited back to level $|3\rangle$; (3) the lower laser level $|1\rangle$ is the ground state, or its population relaxes very slowly if it is not the ground state. Under these conditions, $R_2 \approx W_p N_1$, $R_1 \approx -W_p N_1$, and $N_1 + N_2 \approx N_t$. Furthermore, $\tau_1 \approx \infty$ and $\tau_{21} \approx \tau_2$. The parameter W_p is the effective pumping transition probability rate for exciting an atom in the ground state to eventually reach the upper laser level. It is proportional to the power of the pump. In the steady state with a constant pump, W_p is a constant and $dN_2/dt = dN_1/dt = 0$. With these conditions, we find that

$$N_2\sigma_e - N_1\sigma_a = \frac{W_p\tau_2\sigma_e - \sigma_a}{1 + W_p\tau_2 + (I\tau_2/h\nu)(\sigma_e + \sigma_a)} N_t. \tag{10.70}$$

Therefore, the pumping condition for a constant optical gain under steady-state population inversion is

$$W_p > \frac{\sigma_a}{\tau_2\sigma_e}. \tag{10.71}$$

This condition sets the *minimum pumping requirement* for effective population inversion to reach an optical gain in a three-level system. This requirement can be understood by considering the fact that almost all of the population initially resides in the lower laser level $|1\rangle$. To achieve effective population inversion, the pump has to be strong enough to depopulate sufficient population density from level $|1\rangle$, while the system has to be able to keep it in level $|2\rangle$. In the case when $\sigma_a = \sigma_e$, no population inversion occurs before at least one-half of the total population is transferred from level $|1\rangle$ to level $|2\rangle$.

Four-level system

A four-level system, shown schematically in Fig. 10.11, is more efficient than a three-level system. A four-level system differs from a three-level system in that the lower laser level $|1\rangle$ lies sufficiently high above the ground level $|0\rangle$, with $\Delta E_{10} \gg k_B T$.

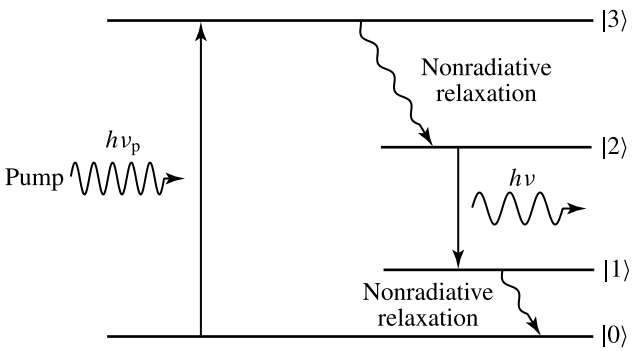


Figure 10.11 Energy levels of a four-level system.

Therefore, in thermal equilibrium, the population in $|1\rangle$ is negligibly small compared with that in $|0\rangle$. Pumping takes place from level $|0\rangle$ to level $|3\rangle$.

An effective four-level system also has to satisfy the conditions concerning levels $|3\rangle$ and $|2\rangle$ discussed above for an effective three-level system. In addition, it has to satisfy the condition that the population in level $|1\rangle$ relaxes very quickly back to the ground level, ideally $\tau_1 \approx \tau_{10} \ll \tau_2$, so that level $|1\rangle$ remains relatively unpopulated in comparison with level $|2\rangle$ when the system is pumped. Under these conditions, $N_1 \approx 0$ and $R_2 \approx W_p(N_t - N_2)$, where the effective pumping transition probability rate W_p is again proportional to the pump power. Then, (10.62) can be ignored because $N_1 \approx 0$. In the steady state when W_p is held constant, by taking $dN_2/dt = 0$ for (10.61), we find that

$$N_2\sigma_e - N_1\sigma_a \approx N_2\sigma_e = \frac{W_p\tau_2\sigma_e}{1 + W_p\tau_2 + (I\tau_2/h\nu)\sigma_e}N_t. \tag{10.72}$$

This result indicates that there is *no minimum pumping requirement* for an ideal four-level system that satisfies the conditions discussed above. Real systems are rarely ideal, but a practical four-level system is still much more efficient than a three-level system. There is no minimum pumping requirement for population inversion in a four-level system because level $|1\rangle$ is initially empty in such a system.

Optical gain

When the condition in (10.65) is satisfied for a given system, an *optical gain coefficient* at a given optical frequency ν can be evaluated with $g = N_2\sigma_e - N_1\sigma_a$ according to (10.50). The optical gain coefficient is a function of the optical signal intensity, I, as a result of the dependence of N_2 and N_1 on I due to stimulated emission that changes the population densities by causing downward transitions from level $|2\rangle$ to level $|1\rangle$. This effect causes saturation of the optical gain coefficient by the optical signal. For all three basic systems discussed above, the optical gain coefficient can be expressed as a function of the optical signal intensity, I (see Problem 10.3.1(a)):

$$g = \frac{g_0}{1 + I/I_{sat}}, \tag{10.73}$$

where g_0 is the *unsaturated gain coefficient*, which is independent of the optical signal intensity, and I_{sat} is the *saturation intensity* of a medium, which can be generally expressed as

$$I_{sat} = \frac{h\nu}{\tau_s\sigma_e}. \tag{10.74}$$

The time constant τ_s is an effective *saturation lifetime* of the effective population inversion. It can be considered as an effective decay time constant for the optical gain coefficient through the relaxation of the effective population inversion. Both g_0 and τ_s

are functions of the intrinsic properties of a gain medium, as well as of the pumping rate. They can be found from (10.66), (10.70), and (10.72) for the quasi-two-level, three-level, and four-level systems, respectively. The results are summarized below (see Problem 10.3.1(b)).

Quasi-two-level system:

$$g_0 = (W_p \tau_s \sigma_e - \sigma_a) N_t, \tag{10.75}$$

$$\tau_s = \tau_2 \frac{1 + \sigma_a/\sigma_e}{1 + (1 + p) W_p \tau_2}. \tag{10.76}$$

Three-level system:

$$g_0 = (W_p \tau_s \sigma_e - \sigma_a) N_t, \tag{10.77}$$

$$\tau_s = \tau_2 \frac{1 + \sigma_a/\sigma_e}{1 + W_p \tau_2}. \tag{10.78}$$

Four-level system:

$$g_0 = W_p \tau_s \sigma_e N_t, \tag{10.79}$$

$$\tau_s = \frac{\tau_2}{1 + W_p \tau_2}. \tag{10.80}$$

The minimum pumping requirement for a medium to have an optical gain is clearly $g_0 > 0$. It can be shown that the minimum pumping requirements obtained by applying this condition to (10.75) and (10.77) are identical to those given in (10.69) and (10.71) for the quasi-two-level and the three-level systems, respectively (see Problem 10.3.1(c)). As for the four-level system, both (10.72) and (10.79) clearly indicate that it has no minimum pumping requirement.

For a desired unsaturated gain coefficient of g_0, the required pumping rate can be found by solving (10.75) and (10.76) for a quasi-two-level system, (10.77) and (10.78) for a three-level system, and (10.79) and (10.80) for a four-level system. The results are summarized below (see Problem 10.3.2).

Quasi-two-level system:

$$W_p = \frac{1}{\tau_2} \frac{\sigma_a N_t + g_0}{(\sigma_e - p\sigma_a) N_t - (1 + p) g_0}. \tag{10.81}$$

Three-level system:

$$W_p = \frac{1}{\tau_2} \frac{\sigma_a N_t + g_0}{\sigma_e N_t - g_0}. \tag{10.82}$$

Four-level system:

$$W_p = \frac{1}{\tau_2} \frac{g_0}{\sigma_e N_t - g_0}. \tag{10.83}$$

In the limit when $p \to 0$, a quasi-two-level system is identical to a three-level system. In the limit when $p \to 0$ and $\sigma_a \to 0$, a quasi-two-level system behaves like a four-level system. In the limit when $\sigma_a \to 0$, a three-level system behaves like a four-level system. For a quasi-two-level system, it is clearly desirable to choose a pump wavelength for which the value of p is as small as possible.

Unsaturated gain coefficient

The unsaturated gain coefficient is also known as the *small-signal gain coefficient* because it is the gain coefficient of a weak optical field that does not saturate the gain medium. In the case of optical pumping with a pump quantum efficiency η_p, the pump intensity required for a desired pumping transition probability rate can be found by using (10.41) as

$$I_p = \frac{1}{\eta_p} \frac{h\nu_p}{\sigma_a^p} W_p, \tag{10.84}$$

where $h\nu_p$ is the energy of the pump photon. The pump quantum efficiency η_p is the net probability of exciting an atom to the upper laser level by each absorbed pump photon. In general, $\eta_p \leq 1$.

It is convenient to define a *saturation pump intensity*, I_p^{sat}, for a laser amplifier for which $W_p \tau_2 = 1$ as

$$I_p^{sat} = \frac{h\nu_p}{\eta_p \tau_2 \sigma_a^p}. \tag{10.85}$$

This is the pump intensity that pumps one-half of the population in a three- or four-level system, and about one-half in a quasi-two-level system, to the upper laser level. *At this level and above, absorption of the pump power is significantly saturated due to depletion of the ground-state population by pumping.* For a pump intensity of I_p, we have $W_p \tau_2 = I_p / I_p^{sat}$.

For a four-level system, we have $g > 0$ as long as the medium is pumped because there is no minimum pumping requirement. For a quasi-two-level or three-level system, we find that $g > 0$ only when the pumping level exceeds its minimum pumping requirement; below that, the medium has absorption for $g < 0$.

When the unsaturated gain coefficient is zero, the medium becomes *transparent*, or *bleached*, to the optical signal, neither absorbing it nor amplifying it. A quasi-two-level or three-level system reaches *transparency*, or the *bleached condition*, at the following *transparency pumping rate*:

$$W_p^{tr} = \frac{1}{\tau_2} \frac{\sigma_a}{\sigma_e - p\sigma_a}. \tag{10.86}$$

The pump intensity corresponding to the transparency pumping rate is the *transparency pump intensity*, I_p^{tr}, which can be expressed as

$$I_p^{tr} = \frac{1}{\eta_p}\frac{h\nu_p}{\tau_2\sigma_a^p}\frac{\sigma_a}{\sigma_e - p\sigma_a} = \frac{\sigma_a}{\sigma_e - p\sigma_a} I_p^{sat}. \tag{10.87}$$

For a quasi-two-level system, $p \neq 0$ in general. For a three-level system, we take $p = 0$ for (10.86) and (10.87). For a four-level system, $I_p^{tr} = 0$ because a four-level system has no minimum pumping requirement and is thus transparent without pumping.

It can be seen from (10.75)–(10.80) that for any system, g_0 increases with pump intensity less than linearly because τ_s decreases with pump intensity though W_p is linearly proportional to the pump intensity. This dependence of τ_s on the pump intensity is caused by the fact that as the pump excites atoms from the ground state to any excited state to eventually reach the upper laser level, it depletes the population in the ground state. Consequently, as the pump intensity increases, fewer atoms remain available for excitation in the ground state, thus reducing the differential increase of the effective population inversion with respect to the increase of the pump intensity.

It can be shown by using the relations in (10.75), (10.77), and (10.79) that the unsaturated gain coefficient of any system can be expressed as a function of pump intensity in the following general form:

$$g_0 = \frac{(\sigma_e - p\sigma_a)N_t}{1 + (1+p)I_p/I_p^{sat}}\left(\frac{I_p}{I_p^{sat}} - \frac{I_p^{tr}}{I_p^{sat}}\right) = \frac{(\sigma_e + \sigma_a)N_t}{1 + (1+p)I_p/I_p^{sat}}\frac{I_p}{I_p^{sat}} - \sigma_a N_t. \tag{10.88}$$

For a quasi-two-level system, $p \neq 0$ and $I_p^{tr} \neq 0$. For a three-level system, $p = 0$ but $I_p^{tr} \neq 0$. For a four-level system, $p = 0$ and $I_p^{tr} = 0$. Note that for a quasi-two-level system or a three-level system, (10.88) is also valid when $I_p < I_p^{tr}$ for $g_0 < 0$. In this situation, the medium has an absorption coefficient of $\alpha = -g_0$ at the laser transition frequency.

As can be seen in (10.88), g_0 varies with I_p sublinearly at high pumping levels due to the dependence of τ_s on I_p as discussed above. For a four-level system, however, the unsaturated gain coefficient varies approximately linearly with I_p at a low pumping level such that $I_p/I_p^{sat} \ll 1$. For a quasi-two-level system or a three-level system, significant pumping is needed just to reach transparency, but the unsaturated gain coefficient also varies approximately linearly with I_p for small variations of the pump intensity near the transparency point.

Gain saturation

The optical gain coefficient is a function of the intensity of the optical wave traveling in the gain medium; it decreases as the optical signal intensity increases. According to (10.73), the optical gain coefficient is reduced to one-half that of the unsaturated gain coefficient g_0 when the optical signal intensity reaches the saturation intensity I_{sat}. The

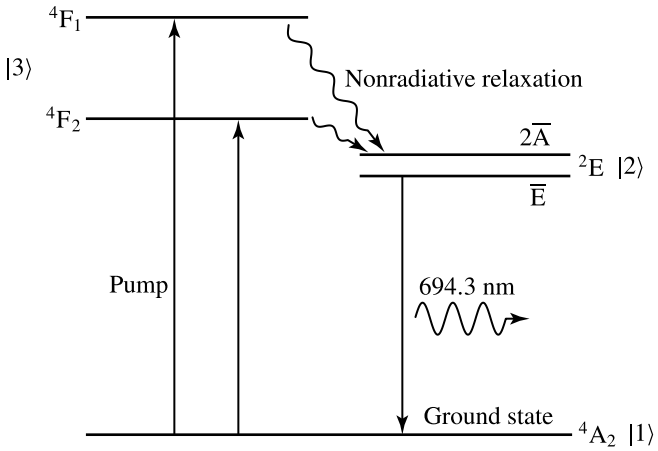


Figure 10.12 Energy levels of the three-level ruby laser.

smaller the value of I_{sat}, the easier it is for the gain to become saturated. For a quasi-two-level system, $\tau_s = \tau_2(1 - p\sigma_a/\sigma_e)$ at transparency. For three-level and four-level systems, $\tau_s = \tau_2$ at transparency. For all three systems, $\tau_s < \tau_2$ as the gain medium is pumped above transparency for a positive gain coefficient. Therefore, I_{sat} increases as the gain medium is pumped harder for a larger unsaturated gain coefficient.

EXAMPLE 10.7 The ruby laser is a three-level system. As shown in Fig. 10.12, it has two primary pump bands at 404 and 554 nm wavelengths, from the $^4A_2$ ground state to the $^4F_1$ and $^4F_2$ states, respectively, both of which relax quickly to the ^2E state so that $\tau_{32} \ll \tau_2 = 3$ ms. The absorption cross sections for $\mathbf{E} \perp c$ polarization at 404 and 554 nm pump wavelengths are both $\sigma_a^p = 2 \times 10^{-23}$ m^{-1}. Assume a 100% pump quantum efficiency of $\eta_p = 1$. (a) Find the transparency pumping rate of a ruby crystal for the 694.3 nm transition with $\mathbf{E} \perp c$ polarization. Find the transparency pump intensity for each pump band. What is the saturation intensity at transparency? (b) A ruby crystal rod doped with 0.05 wt. % Cr_2O_3 has a Cr concentration of 1.58×10^{25} m^{-3}. It is pumped for an unsaturated gain coefficient of 5 m^{-1}. Find the required pumping rate, the saturation intensity at this pumping rate, and the required pump intensity for each pump band.

Solution (a) Because $\sigma_a = 1.25 \times 10^{-24}$ m^2 from Example 10.4, the transparency pumping rate is

$$W_p^{tr} = \frac{\sigma_a}{\tau_2 \sigma_e} = \frac{1.25 \times 10^{-24}}{3 \times 10^{-3} \times 1.34 \times 10^{-24}} \text{ s}^{-1} = 311 \text{ s}^{-1}.$$

The pump photons at 404 and 554 nm wavelengths have photon energies of $h\nu_{p1} = (1.2398/0.404)$ eV $= 3.069$ eV and $h\nu_{p2} = (1.2398/0.554)$ eV $= 2.238$ eV,

respectively. The transparency pump intensity for the 404 nm pump band is

$$I_{\rm p}^{\rm tr} = \frac{h\nu_{\rm p1}}{\sigma_{\rm a}^{\rm p}} W_{\rm p}^{\rm tr} = \frac{3.069 \times 1.6 \times 10^{-19} \times 311}{2 \times 10^{-23}} \text{ W m}^{-2} = 7.64 \text{ MW m}^{-2},$$

and that for the 554 nm pump band is

$$I_{\rm p}^{\rm tr} = \frac{h\nu_{\rm p2}}{\sigma_{\rm a}^{\rm p}} W_{\rm p}^{\rm tr} = \frac{2.238 \times 1.6 \times 10^{-19} \times 311}{2 \times 10^{-23}} \text{ W m}^{-2} = 5.57 \text{ MW m}^{-2},$$

For the three-level ruby, $\tau_{\rm s} = \tau_2 = 3$ ms at transparency. The photon energy for $\lambda = 694.3$ nm is $h\nu = (1.2398/0.6943)$ eV $= 1.786$ eV $= 1.786 \times 1.6 \times 10^{-19}$ J. Therefore, the saturation intensity at transparency is

$$I_{\rm sat} = \frac{h\nu}{\tau_{\rm s}\sigma_{\rm e}} = \frac{1.786 \times 1.6 \times 10^{-19}}{3 \times 10^{-3} \times 1.34 \times 10^{-24}} \text{ W m}^{-2} = 71.1 \text{ MW m}^{-2}.$$

(b) For $N_{\rm t} = 1.58 \times 10^{25}$ m^{-3}, we find that $\sigma_{\rm e}N_{\rm t} = 1.34 \times 10^{-24} \times 1.58 \times 10^{25}$ m^{-1} $= 21.17$ m^{-1} and $\sigma_{\rm a}N_{\rm t} = 1.25 \times 10^{-24} \times 1.58 \times 10^{25}$ m^{-1} $= 19.75$ m^{-1}. Thus, using (10.82), we find that, for $g_0 = 5$ m^{-1}, the required pumping rate is

$$W_{\rm p} = \frac{1}{3 \times 10^{-3}} \times \frac{19.75 + 5}{21.17 - 5} \text{ s}^{-1} = 510 \text{ s}^{-1},$$

which is 1.64 times the transparency pumping rate of $W_{\rm p}^{\rm tr} = 311$ s^{-1}. Therefore, the required pump power is 1.64 times the transparency pump power: $I_{\rm p} = 1.64 I_{\rm p}^{\rm tr} = 12.53$ MW m^{-2} for the 404 nm pump and $I_{\rm p} = 1.64 I_{\rm p}^{\rm tr} = 9.13$ MW m^{-2} for the 554 nm pump. At this pumping level, $W_{\rm p}\tau_2 = 1.53$. Therefore, from (10.78), $\tau_{\rm s} = \tau_2(1 + 1.25/1.34)/(1 + 1.53) = 2.29$ ms. Then, the saturation intensity is

$$I_{\rm sat} = \frac{h\nu}{\tau_{\rm s}\sigma_{\rm e}} = \frac{1.786 \times 1.6 \times 10^{-19}}{2.29 \times 10^{-3} \times 1.34 \times 10^{-24}} \text{ W m}^{-2} = 93.1 \text{ MW m}^{-2}.$$

It is clear from this example that a very high pump power is required just to bring a ruby crystal to transparency because of the fact that it is a three-level system. For this reason, it is only feasible to pump a ruby laser with a pulsed pump. As a consequence, CW operation is never realized for the ruby laser. Ruby lasers are always operated in the pulsed mode, most notably in the Q-switched mode for the generation of giant pulses. The situation is very different for four-level systems, such as Nd : YAG, or quasi-two-level systems, such as Ti : sapphire and Cr : LiSAF. See Problem 10.3.5(a) for a comparison with Nd : YAG and Problem 10.3.6(a) for a comparison with Ti : sapphire.

Spontaneous emission power

When the upper laser level of a gain medium is populated, there is spontaneous emission. The upper laser level population can be found by solving $N_1 + N_2 = N_{\rm t}$ and

$N_2\sigma_e - N_1\sigma_a = g$ simultaneously to have

$$N_2 = \frac{\sigma_a N_t + g}{\sigma_e + \sigma_a}. \tag{10.89}$$

This relation is valid for all systems, including the four-level system. Though we have used $N_1 + N_2 = N_t$, which is not valid for a four-level system, to obtain this relation, (10.89) reduces to $N_2 = g/\sigma_e$ in the case of a four-level system, for which $\sigma_a = 0$. Note that, in the case of a quasi-two-level or a three-level system, $g = -\alpha$ when the medium is not sufficiently pumped to reach transparency. Because the maximum value of the absorption coefficient is $\alpha_0 = \sigma_a N_t$, we find that $N_2 \geq 0$ for any positive or negative values of g. Note also that g appearing in (10.89) is the saturated gain coefficient if stimulated emission exists in the medium. According to the discussions in Section 10.1, the spontaneous emission power is proportional to N_2 only and is independent of N_1. Therefore, regardless of whether the medium has a gain or a loss, the *spontaneous emission power density*, which is defined as the spontaneous emission power per unit volume of the medium in watts per cubic meter, is

$$\hat{P}_{sp} = \frac{h\nu}{\tau_{sp}} N_2 = \frac{h\nu}{\tau_{sp}} \frac{\sigma_a N_t + g}{\sigma_e + \sigma_a}, \tag{10.90}$$

where g can be positive, for a medium pumped above transparency, or negative, for a medium below transparency. For a gain volume of $\mathcal{V}$, the spontaneous emission power is $P_{sp} = \hat{P}_{sp}\mathcal{V}$.

In the case when the gain is not saturated so that $g = g_0$, we find from (10.75), (10.77), and (10.79) that $\sigma_a N_t + g = W_p \tau_s \sigma_e N_t$. For a medium that is optically pumped with a pump intensity I_p, we then have

$$N_2 = \frac{W_p \tau_2}{1 + (1 + p) W_p \tau_2} N_t = \frac{I_p / I_p^{sat}}{1 + (1 + p) I_p / I_p^{sat}} N_t. \tag{10.91}$$

Then, the spontaneous emission power density in the absence of gain saturation can be expressed as

$$\hat{P}_{sp} = \frac{h\nu}{\tau_{sp}} \frac{I_p / I_p^{sat}}{1 + (1 + p) I_p / I_p^{sat}} N_t. \tag{10.92}$$

At transparency, $g = g_0 = 0$. The spontaneous emission power density at transparency, which is known as the *critical fluorescence power density*, is

$$\hat{P}_{sp}^{tr} = \frac{h\nu}{\tau_{sp}} \frac{\sigma_a}{\sigma_e + \sigma_a} N_t = \frac{h\nu}{\tau_{sp}} \frac{I_p^{tr} / I_p^{sat}}{1 + (1 + p) I_p^{tr} / I_p^{sat}} N_t. \tag{10.93}$$

For a gain volume of $\mathcal{V}$, the *critical fluorescence power* is $P_{sp}^{tr} = \hat{P}_{sp}^{tr}\mathcal{V}$.

EXAMPLE 10.8 A ruby crystal doped with 0.05 wt. % Cr_2O_3 for a Cr concentration of 1.58×10^{25} m^{-3} as discussed in Example 10.7 is considered. Almost all of the

population in the upper laser level of a ruby laser crystal relaxes radiatively by to the ground state so that $\tau_{sp} = \tau_{21} = \tau_2 = 3$ ms. Find the critical fluorescence power density corresponding to transparency for the 694.3 nm line at 300 K. What is the spontaneous emission power density if the ruby crystal is pumped above transparency for a gain coefficient of 5 m^{-1} for the 694.3 nm line? What is it if the crystal is insufficiently pumped so that it has an absorption coefficient of 5 m^{-1} for the 694.3 nm line? If a ruby laser rod of 6 cm length and 4 mm cross-sectional diameter is uniformly pumped, what are the spontaneous emission powers in the three cases considered here?

Solution When we consider transparency for the 694.3 nm transition, we take $\sigma_a = 1.25 \times 10^{-24}$ m^2 and $\sigma_e = 1.34 \times 10^{-24}$ m^2 for this transition at 300 K, which are obtained in Examples 10.4 and 10.5, respectively. However, the spontaneous emission is broadband covering both emission lines at 692.9 and 694.3 nm. Therefore, we take an average photon energy of the two for $h\nu = 1.787$ eV. Then, we find from (10.93) the following critical fluorescence power density at the transparency point for the 694.3 nm line:

$$\hat{P}_{sp}^{tr} = \frac{1.787 \times 1.6 \times 10^{-19} \times 1.25 \times 10^{-24} \times 1.58 \times 10^{25}}{3 \times 10^{-3} \times (1.34 \times 10^{-24} + 1.25 \times 10^{-24})} \text{ W m}^{-3} = 727 \text{ MW m}^{-3}.$$

When pumped for a gain coefficient of 5 m^{-1} for the 694.3 nm line, we have

$$\hat{P}_{sp} = \frac{1.787 \times 1.6 \times 10^{-19} \times (1.25 \times 10^{-24} \times 1.58 \times 10^{25} + 5)}{3 \times 10^{-3} \times (1.34 \times 10^{-24} + 1.25 \times 10^{-24})} \text{ W m}^{-3}$$
$$= 911 \text{ MW m}^{-3}.$$

When the crystal is insufficiently pumped so that there is an absorption coefficient of 5 m^{-1}, $g = -5$ m^{-1}. Then,

$$\hat{P}_{sp} = \frac{1.787 \times 1.6 \times 10^{-19} \times (1.25 \times 10^{-24} \times 1.58 \times 10^{25} - 5)}{3 \times 10^{-3} \times (1.34 \times 10^{-24} + 1.25 \times 10^{-24})} \text{ W m}^{-3}$$
$$= 543 \text{ MW m}^{-3}.$$

For a rod of 6 cm length and 4 mm cross-sectional diameter, the volume is $\mathcal{V} = \pi \times (4 \times 10^{-3}/2)^2 \times 6 \times 10^{-2}$ $m^3 = 7.54 \times 10^{-7}$ m^3. Therefore, the critical fluorescence power is $P_{sp}^{tr} = \hat{P}_{sp}^{tr}\mathcal{V} = 548$ W. The total spontaneous emission power is $P_{sp} = 687$ W for $g = 5$ m^{-1} and $P_{sp} = 409$ W for $g = -5$ m^{-1}. From the consideration of energy conservation, it is clear that the power required to pump the crystal to a particular state has to be at least, and most often far exceed, that emitted spontaneously by the crystal. Therefore, these numbers again show the high power required to pump a ruby laser crystal just to its transparency point. For example, if the pumping efficiency is 10%, the pump power required to pump this crystal to transparency is as high as 5.48 kW. See Problem 10.3.5(b) for a comparison with Nd : YAG and Problem 10.3.6(b) for a comparison with Ti : sapphire.

10.4 Laser amplifiers

Any medium that has an optical gain can be used to amplify an optical signal. Depending on the physical mechanism responsible for the optical gain, there are two different categories of optical amplifiers: the *nonlinear optical amplifiers* and the *laser amplifiers*. The optical gain of a nonlinear optical amplifier is associated with a nonlinear optical process in a nonlinear medium, whereas the gain of a laser amplifier results from the population inversion in a medium. Important nonlinear optical amplifiers include the OPAs, discussed in Section 9.6, and the Raman and Brillouin amplifiers, discussed in Section 9.9. In this section, the general characteristics of laser amplifiers are addressed. We consider only continously pumped laser amplifiers operating in the steady state. Not considered here are pulsed laser amplifiers that require transient dynamical analysis, including those for regenerative amplification of ultrashort laser pulses and those using transient pumping for high-power amplification of giant laser pulses.

We consider single-pass, traveling-wave laser amplifiers, as shown in Fig. 10.13. Such a laser amplifier does not form a resonant optical cavity; therefore, the optical signal being amplified passes through it only once as a traveling wave. A laser amplifier can be pumped in many different ways, but the most commonly employed techniques are *electrical pumping* and *optical pumping*. For electrical pumping, a *transverse pumping* arrangement is more convenient and is most often used though a *longitudinal pumping* arrangement is also possible. For optical pumping, both transverse and longitudinal pumping arrangements can be easily implemented. However, for an optically pumped amplifier that has a long length but a relatively small absorption coefficient at the pump frequency, such as the fiber amplifier discussed in the following section, the longitudinal pumping arrangement is much more efficient than the transverse pumping arrangement. Longitudinal optical pumping can be arranged as unidirectional forward, unidirectional backward, or bidirectional. The concepts of these different pumping schemes are also illustrated in Fig. 10.13.

The most important characteristics of a laser amplifier are *gain, efficiency, bandwidth, and noise*. These four characteristics are addressed in the following discussions.

Amplifier gain

Ignoring the contribution of noise, the amplification of the intensity, I_s, of an optical signal propagating through a laser amplifier can be described by

$$\frac{dI_s}{dz} = gI_s = \frac{g_0(z)}{1 + I_s/I_{sat}} I_s, \tag{10.94}$$

where $g_0(z)$ is the unsaturated gain coefficient, which can be spatially varying in the longitudinal direction, and I_{sat} is the saturation intensity of the gain medium, both

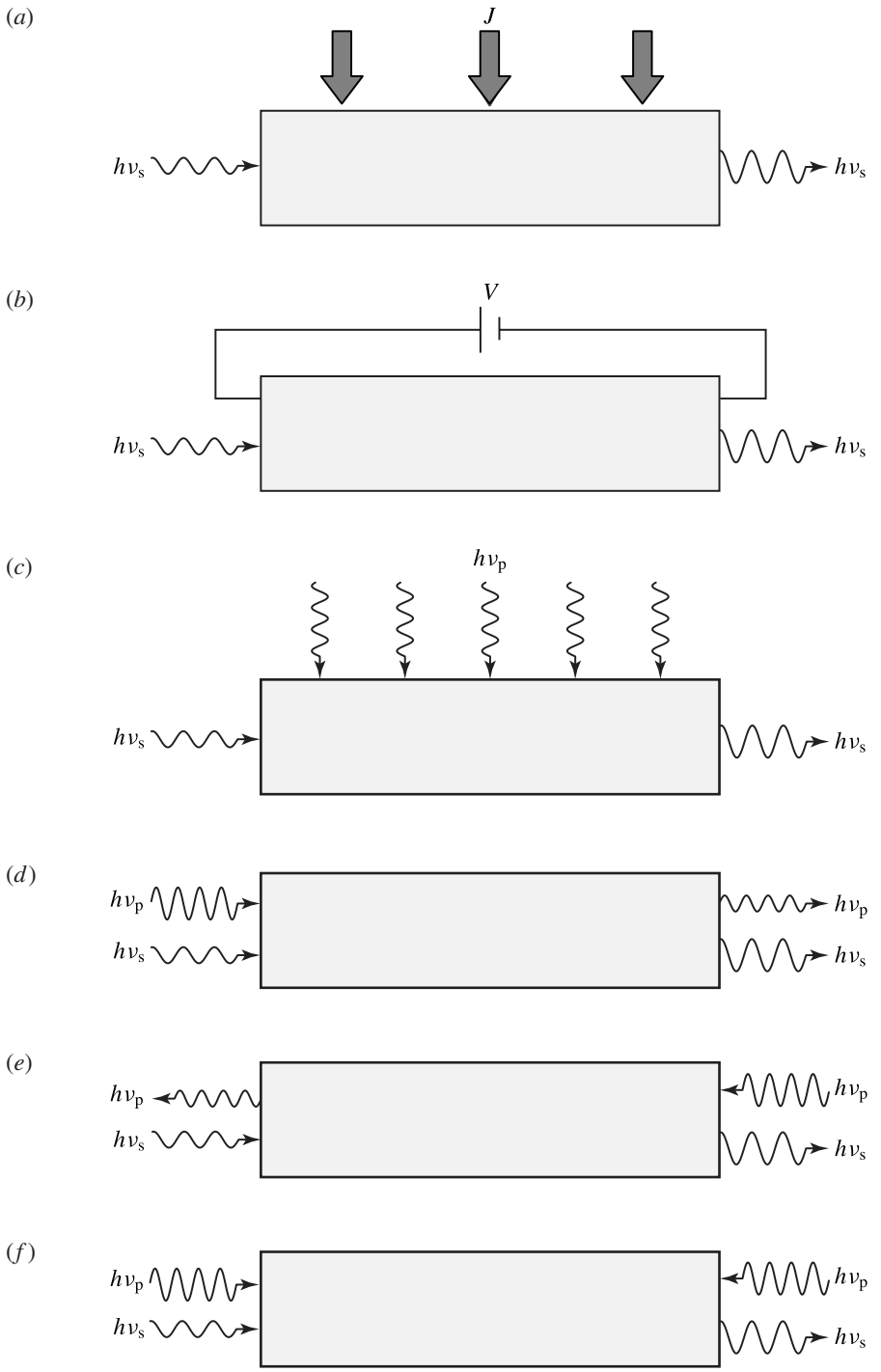


Figure 10.13 Single-pass, traveling-wave laser amplifiers with various pumping arrangements: (*a*) transverse electrical pumping; (*b*) longitudinal electrical pumping; (*c*) transverse optical pumping; (*d*) unidirectional forward, longitudinal optical pumping; (*e*) unidirectional backward, longitudinal optical pumping; (*f*) bidirectional, longitudinal optical pumping.

defined in the preceding section. Here we assume transverse uniformity but consider the possibility of longitudinal nonuniformity by taking the unsaturated gain coefficient $g_0(z)$ to be a function of z.

Such a longitudinally nonuniform gain distribution is a common scenario in an amplifier under longitudinal optical pumping because of pump absorption by the gain medium. In the following discussions, we assume for simplicity that the signal beam is collimated throughout the length of the amplifier such that divergence of the beam is negligible. This assumption allows us to consider the power, P_s, of the optical signal and to convert (10.94) into

$$\frac{dP_s}{dz} = gP_s = \frac{g_0(z)}{1 + P_s/P_{sat}} P_s, \tag{10.95}$$

where P_{sat} is the *saturation power* obtained by integrating I_{sat} over the cross-sectional area of the signal beam. By integrating (10.95), the following relation is obtained:

$$\frac{P_s(z)}{P_s(0)} \exp\left[\frac{P_s(z) - P_s(0)}{P_{sat}}\right] = \exp \int_0^z g_0(z)dz, \tag{10.96}$$

where $P_s(0)$ is the power of the signal beam at $z = 0$. When $P_s \ll P_{sat}$, the power of the optical signal grows exponentially with distance. As P_s approaches the value of P_{sat}, the growth slows down. Eventually, the signal grows only linearly with distance when $P_s \gg P_{sat}$.

The *power gain* of a signal amplified by a laser amplifier is defined as

$$G = \frac{P_s^{out}}{P_s^{in}}, \tag{10.97}$$

where P_s^{in} and P_s^{out} are the input and output powers of the signal, respectively. By using the relation in (10.96) while identifying P_s^{out} and P_s^{in} with $P_s(l)$ and $P_s(0)$, respectively, for an amplifier of a length l, the following relation for the power gain of the signal is found:

$$G = G_0 e^{(1-G)P_s^{in}/P_{sat}}, \tag{10.98}$$

where G_0 is the *unsaturated power gain*, or the *small-signal power gain*. For a single pass through the amplifier, G_0 is given by

$$G_0 = \exp \int_0^l g_0(z)dz. \tag{10.99}$$

Note that, according to (10.98), $G_0 \geq G > 1$ because $g_0 > 0$ for an amplifier. For a weak optical signal such that $P_s^{in} < P_s^{out} \ll P_{sat}$, the power gain is simply the small-signal power gain, $G = G_0$. If the signal power approaches or even exceeds the saturation power of the amplifier, the relation in (10.98) clearly indicates that $G < G_0$ because of gain saturation. In this situation, the overall gain, G, can be found by solving (10.98)

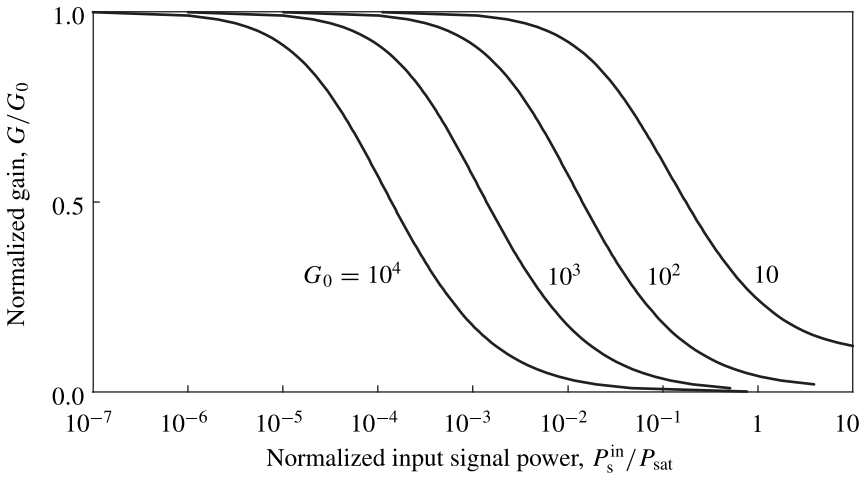


Figure 10.14 Gain, normalized to the unsaturated gain as G/G_0, of a laser amplifier as a function of input signal power, normalized to the saturation power as P_s^{in}/P_{sat}, for a few different values of the unsaturated power gain.

when the values of P_s^{in} and P_{sat}, as well as that of G_0, are given. Figure 10.14 shows the amplifier gain as a function of input signal power for a few different values of the unsaturated power gain.

The unsaturated gain coefficient g_0 of an optically pumped laser amplifier depends on the pump intensity according to (10.88). For both longitudinal and transverse pumping, the pump intensity normally varies in space because of absorption and diffraction of the pump beam. In the case of longitudinal optical pumping, the pump intensity is still a function of distance from the input end even when transverse uniformity is assumed, as is done in the above discussion. In general, $I_p(z)$ and $g_0(z)$ depend on many geometric parameters of the amplifier and have to be found numerically for each particular case.

A special situation of interest is when transverse divergence of the pump beam is nonexistent, such as in the case of an optical fiber amplifier, or can be ignored, such as in a short bulk laser amplifier with a highly collimated pump beam. In this situation, we can express $g_0(z)$ in terms of the pump power $P_p(z)$ instead of the pump intensity by integrating $I_p(z)$ over the transverse cross section. A *saturation pump power*, P_p^{sat}, can be defined by integrating I_p^{sat} over the transverse cross section. Then, by replacing I_p/I_p^{sat} with P_p/P_p^{sat} in (10.91) for N_2, we find that the absorption coefficient of the pump beam as a function of distance in the amplifier can be expressed as

$$\alpha_p(z) = \sigma_a^p[N_t - N_2(z)] = \alpha_p \frac{1 + pP_p(z)/P_p^{sat}}{1 + (1+p)P_p(z)/P_p^{sat}}, \tag{10.100}$$

where $\alpha_p = \sigma_a^p N_t$ is the intrinsic absorption coefficient at the pump wavelength in the absence of a strong pump beam so that pump depletion of the ground-state population is

negligible. Here we have assumed a low-loss medium where the pumping mechanism fully accounts for absorption of the pump beam. With this spatially varying pump absorption coefficient, we can write the following equation for the spatial evolution of the pump power:

$$\frac{\mathrm{d}P_\mathrm{p}}{\mathrm{d}z} = -\alpha_\mathrm{p}(z)P_\mathrm{p} = -\alpha_\mathrm{p}\frac{1 + pP_\mathrm{p}/P_\mathrm{p}^\mathrm{sat}}{1 + (1+p)P_\mathrm{p}/P_\mathrm{p}^\mathrm{sat}}P_\mathrm{p}. \tag{10.101}$$

This equation can be integrated to find the following solutions:

$$\frac{P_\mathrm{p}(z)}{P_\mathrm{p}(0)}\left[\frac{P_\mathrm{p}^\mathrm{sat} + pP_\mathrm{p}(z)}{P_\mathrm{p}^\mathrm{sat} + pP_\mathrm{p}(0)}\right]^{1/p} = \mathrm{e}^{-\alpha_\mathrm{p}z}, \quad \text{for } p \neq 0, \tag{10.102}$$

and

$$\frac{P_\mathrm{p}(z)}{P_\mathrm{p}(0)}\exp\left[\frac{P_\mathrm{p}(z) - P_\mathrm{p}(0)}{P_\mathrm{p}^\mathrm{sat}}\right] = \mathrm{e}^{-\alpha_\mathrm{p}z}, \quad \text{for } p = 0. \tag{10.103}$$

It can be seen from the relations in (10.102) and (10.103) that besides the absorption coefficient α_p, longitudinal variations of pump power in the amplifier strongly depend on the *pumping ratio* defined as

$$s = \frac{P_\mathrm{p}(0)}{P_\mathrm{p}^\mathrm{sat}} = \frac{P_\mathrm{p}^\mathrm{in}}{P_\mathrm{p}^\mathrm{sat}}, \tag{10.104}$$

where $P_\mathrm{p}^\mathrm{in} = P_\mathrm{p}(0)$ is the input pump power.

Once the pump power distribution, $P_\mathrm{p}(z)$, is found from the implicit solutions given in (10.102) and (10.103), the distribution of the small-signal gain coefficient $g_0(z)$ can be found from (10.88) as

$$\begin{aligned} g_0(z) &= \frac{(\sigma_\mathrm{e} - p\sigma_\mathrm{a})N_\mathrm{t}}{1 + (1+p)P_\mathrm{p}(z)/P_\mathrm{p}^\mathrm{sat}}\left[\frac{P_\mathrm{p}(z)}{P_\mathrm{p}^\mathrm{sat}} - \frac{P_\mathrm{p}^\mathrm{tr}}{P_\mathrm{p}^\mathrm{sat}}\right] \\ &= \frac{(\sigma_\mathrm{e} + \sigma_\mathrm{a})N_\mathrm{t}}{1 + (1+p)P_\mathrm{p}(z)/P_\mathrm{p}^\mathrm{sat}}\frac{P_\mathrm{p}(z)}{P_\mathrm{p}^\mathrm{sat}} - \sigma_\mathrm{a}N_\mathrm{t}. \end{aligned} \tag{10.105}$$

In general, numerical solution is required to find $P_\mathrm{p}(z)$ from the implicit solutions given in (10.102) and (10.103) in order to find $g_0(z)$. However, what really matters for an amplifier is the integral of $g_0(z)$ over the entire length of the amplifier, which gives the value of G_0 in (10.99). Closed-form solutions for both cases of $p \neq 0$ and $p = 0$ can be found by using (10.101) to integrate $g_0(z)$ in (10.105). For an amplifier of a length l, the integral can be expressed conveniently in terms of the input pump power $P_\mathrm{p}^\mathrm{in} = P_\mathrm{p}(0)$ launched at the input end and the remaining pump power $P_\mathrm{p}^\mathrm{out} = P_\mathrm{p}(l)$ at

the output end of the amplifier as (see Problem 10.4.2)

$$\int_0^l g_0(z)\mathrm{d}z = \sigma_\mathrm{e} N_\mathrm{t} l + \frac{(\sigma_\mathrm{e}+\sigma_\mathrm{a})N_\mathrm{t}}{\alpha_\mathrm{p}} \ln \frac{P_\mathrm{p}^\mathrm{out}}{P_\mathrm{p}^\mathrm{in}}$$
$$= \sigma_\mathrm{e} N_\mathrm{t} l + \frac{(\sigma_\mathrm{e}+\sigma_\mathrm{a})N_\mathrm{t}}{\alpha_\mathrm{p}} \ln(1-\zeta_\mathrm{p}), \tag{10.106}$$

where ζ_p is the *pump power utilization factor* that accounts for the pump power absorbed by the gain medium and is defined as

$$\zeta_\mathrm{p} = \frac{P_\mathrm{p}^\mathrm{in} - P_\mathrm{p}^\mathrm{out}}{P_\mathrm{p}^\mathrm{in}}. \tag{10.107}$$

The relation given in (10.106) is valid for both $p \neq 0$ and $p = 0$. It is a convenient form because all that is needed to evaluate the value of the gain integral is the value of ζ_p besides the basic parameters of the amplifier.

In theory, $P_\mathrm{p}^\mathrm{out}$ never completely vanishes and ζ_p never reaches the value of unity no matter how long the amplifier is because the pump power can only continue to decay. Therefore, there is no problem in utilizing (10.106) in principle. In practice, however, great uncertainty arises in using (10.106) when $P_\mathrm{p}^\mathrm{out}$ becomes very small. In an experimental setting, (10.106) yields no meaningful result when $P_\mathrm{p}^\mathrm{out}$ approaches the detection limit. To avoid such a limitation, (10.102) for $p \neq 0$ and (10.103) for $p = 0$ can be used to transform (10.106) into (see Problem 10.4.2)

$$\int_0^l g_0(z)\mathrm{d}z = \begin{cases} \dfrac{(\sigma_\mathrm{e}+\sigma_\mathrm{a})N_\mathrm{t}}{p\alpha_\mathrm{p}} \ln \dfrac{P_\mathrm{p}^\mathrm{sat} + pP_\mathrm{p}^\mathrm{in}}{P_\mathrm{p}^\mathrm{sat} + pP_\mathrm{p}^\mathrm{out}} - \sigma_\mathrm{a} N_\mathrm{t} l, & \text{for } p \neq 0, \\ \dfrac{(\sigma_\mathrm{e}+\sigma_\mathrm{a})N_\mathrm{t}}{\alpha_\mathrm{p}} \left(\dfrac{P_\mathrm{p}^\mathrm{in} - P_\mathrm{p}^\mathrm{out}}{P_\mathrm{p}^\mathrm{sat}} \right) - \sigma_\mathrm{a} N_\mathrm{t} l, & \text{for } p = 0. \end{cases} \tag{10.108}$$

Given an input pump power of $P_\mathrm{p}^\mathrm{in} = P_\mathrm{p}(0)$ at $z = 0$, the remaining pump power, $P_\mathrm{p}^\mathrm{out} = P_\mathrm{p}(l)$, at the output end $z = l$ of the amplifier can be found from (10.102) for $p \neq 0$ or from (10.103) for $p = 0$. In an experimental setting, both P_p^in and $P_\mathrm{p}^\mathrm{out}$ can be measured directly. Once the integral of $g_0(z)$ is found, G_0 can be found through (10.99).

EXAMPLE 10.9 A CW Nd : YAG laser amplifier for an optical signal at $\lambda_\mathrm{s} = 1.064$ μm is pumped with the output of a high-power semiconductor laser at $\lambda_\mathrm{p} = 808$ nm. The Nd : YAG crystal, which is doped with 1.1 at. % Nd for a concentration of $N_\mathrm{t} = 1.52 \times 10^{26}$ m^{-3}, has a length of 5 mm and a cross-sectional diameter of 5 mm. The pump optical beam is delivered through a multimode optical fiber of a 200-μm core diameter and is collimated to define a circular pumping spot of $w = 100$ μm radius throughout the length of the crystal. The signal beam is collimated to a spot size that matches the pumping area exactly. As shown in Fig. 10.15, the crystal surfaces are coated for a single

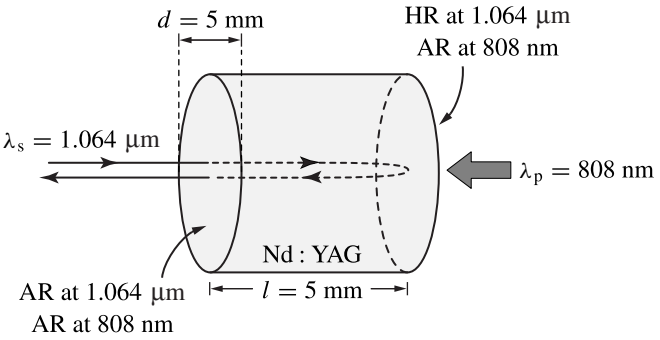


Figure 10.15 Schematics of a double-pass end-pumped Nd : YAG amplifier. AR means antireflection, and HR means high reflection.

pass of the pump beam through the crystal, but for double passes of the signal beam. At the 808 nm pump wavelength, the peak absorption cross section is 5.6×10^{-24} m^2. However, because the emission of the pump semiconductor laser has a broad spectral width of $\Delta\lambda_p = 3.5$ nm, the effective absorption cross section for the pump beam is reduced to $\sigma_a^p = 3.0 \times 10^{-24}$ m^2. The emission cross section accounting for all effects including the population ratio in the upper laser level is $\sigma_e = 3.1 \times 10^{-23}$ m^2 at the signal wavelength. The fluorescence lifetime is $\tau_2 = 240$ μs. The pump quantum efficiency is found to be $\eta_p = 0.8$. The amplifier is pumped with a pump power of $P_p = 2$ W. Assume that there is no additional attenuation to the pump beam besides the absorption for the pumping transition with the cross section σ_a^p. (a) Find the double-pass unsaturated power gain of the amplifier. (b) Find the gain and the output power for a signal with an input power of $P_s^{in} = 20$ mW.

Solution (a) The pump photon energy is $h\nu_p = (1.2398/0.808)$ eV $= 1.534$ eV. The signal photon energy is $h\nu_s = (1.2398/1.064)$ eV $= 1.165$ eV. Because the pump and signal beams are all well collimated, we can consider the pump and signal powers directly instead of the pump and signal intensities. We first find that

$$P_p^{sat} = \pi w^2 I_p^{sat} = \pi w^2 \frac{h\nu_p}{\eta_p \tau_2 \sigma_a^p} = \frac{\pi \times (100 \times 10^{-6})^2 \times 1.534 \times 1.6 \times 10^{-19}}{0.8 \times 240 \times 10^{-6} \times 3.0 \times 10^{-24}} \text{ W}$$
$$= 13.4 \text{ W}.$$

Then, with an input pump power of $P_p^{in} = 2$ W, the pumping ratio defined in (10.104) is

$$s = \frac{P_p(0)}{P_p^{sat}} = \frac{2}{13.4} = 0.149.$$

The pump absorption coefficient is $\alpha_p = \sigma_a^p N_t = 456$ m^{-1}. With an amplifier length of $l = 5$ mm, we find that $\alpha_p l = 2.28$. Because Nd : YAG is a four-level system for

the signal at 1.064 μm wavelength, we have $p = 0$ and $\sigma_a = 0$. We can then find $P_p^{out} = P_p(l)$ from the implicit solution in (10.103). By taking $z = l$ and defining the variable $x = P_p(l)/P_p^{sat}$, (10.103) can be expressed as

$$x e^x = s e^s e^{-\alpha_p l} = 0.149 \times e^{0.149-2.28},$$

which has a solution of $x = 1.74 \times 10^{-2}$. Thus,

$$x = \frac{P_p(l)}{P_p^{sat}} = 1.74 \times 10^{-2},$$

and $P_p^{out} = P_p(l) = 1.74 \times 10^{-2} \times P_p^{sat} = 1.74 \times 10^{-2} \times 13.4\ \text{W} = 0.233\ \text{W}$. With $p = 0$ and $\sigma_a = 0$, we then obtain from (10.108) the following integral:

$$\int_0^l g_0(z)\mathrm{d}z = \frac{\sigma_e N_t}{\alpha_p}\left(\frac{P_p^{in} - P_p^{out}}{P_p^{sat}}\right) = \frac{3.1 \times 10^{-23} \times 1.52 \times 10^{26}}{456} \times \frac{2 - 0.233}{13.4} = 1.36.$$

The double-pass unsaturated power gain can now be found as

$$G_0 = \exp\left[2\int_0^l g_0(z)\mathrm{d}z\right] = \exp(2 \times 1.36) = e^{2.72} = 15.2.$$

This unsaturated gain is about 11.8 dB.

(b) To find the power gain for the signal, we need to consider the gain saturation effect by first finding the saturation power of the amplifier. At the given pumping level, $W_p\tau_2 = s = 0.149$ at $z = 0$. Taking this value as an approximation throughout the gain medium, we have

$$\tau_s = \frac{\tau_2}{1 + W_p\tau_2} = \frac{240}{1.149}\ \mu\text{s} = 209\ \mu\text{s}.$$

Then the saturation intensity is

$$I_{sat} = \frac{h\nu_s}{\tau_s\sigma_e} = \frac{1.165 \times 1.6 \times 10^{-19}}{209 \times 10^{-6} \times 3.1 \times 10^{-23}}\ \text{W m}^{-2} = 28.8\ \text{MW m}^{-2},$$

and the saturation power is

$$P_{sat} = \pi w^2 I_{sat} = \pi \times (100 \times 10^{-6})^2 \times 28.8 \times 10^6\ \text{W} = 905\ \text{mW}.$$

For an input signal power of $P_s^{in} = 20$ mW, $P_s^{in}/P_{sat} = 0.022$. Therefore, according to (10.98), the gain can be found by solving

$$G = G_0 e^{(1-G)P_s^{in}/P_{sat}} = e^{2.72+0.022(1-G)}.$$

By solving this relation iteratively, we find that the signal power gain is $G = 11.9$,

which is 10.8 dB. The output signal power is

$$P_s^{out} = G P_s^{in} = 11.9 \times 20 \text{ mW} = 238 \text{ mW}.$$

Amplifier efficiency

The efficiency of a laser amplifier can be measured either as power efficiency or as quantum efficiency.

The *power conversion efficiency*, η_c, of a laser amplifier is defined as

$$\eta_c = \frac{P_s^{out} - P_s^{in}}{P_p}. \tag{10.109}$$

Another useful concept is the *differential power conversion efficiency*, also known as the *slope efficiency*, η_s, of an amplifier, which is defined as

$$\eta_s = \frac{dP_s^{out}}{dP_p}. \tag{10.110}$$

The differential power conversion efficiency measures the increase of the output signal power as the pump power increases. It is generally somewhat larger than the total power conversion efficiency measured by η_c.

The *quantum efficiency*, η_q, of a laser amplifier is defined as the number of signal photons generated per pump photon, in the case of optical pumping, or per pump electron, in the case of electrical pumping, that is absorbed by the gain medium. It can be expressed as

$$\eta_q = \frac{\Phi_s^{out} - \Phi_s^{in}}{\zeta_p \Phi_p}, \tag{10.111}$$

where Φ_s^{in} and Φ_s^{out} are the input and output photon fluxes, respectively, and Φ_p is the pump photon or electron flux. The maximum possible value of η_q is unity.

The power conversion efficiency is always less than the quantum efficiency. For the case of optical pumping, they have the following relationship:

$$\eta_q = \frac{\nu_p}{\nu_s}\frac{\eta_c}{\zeta_p} = \frac{\lambda_s}{\lambda_p}\frac{\eta_c}{\zeta_p}, \tag{10.112}$$

where ν_s and ν_p are the signal and pump frequencies, respectively, and λ_s and λ_p are the free-space signal and pump wavelengths, respectively. Because the maximum value of η_q is unity, the maximum possible power conversion efficiency of an optically pumped laser amplifier is λ_p/λ_s.

EXAMPLE 10.10 Find the power conversion efficiency and the quantum efficiency of the Nd : YAG laser amplifier described in Example 10.9 operated with a pump power of 2 W and an input signal power of 20 mW.

Solution From Example 10.9, the output signal power is $P_s^{out} = 238$ mW when the amplifier is operated with $P_p^{in} = 2$ W and $P_s^{in} = 20$ mW. Therefore, according to (10.109), the power conversion efficiency is

$$\eta_c = \frac{238 \times 10^{-3} - 20 \times 10^{-3}}{2} = 10.9\%.$$

For this amplifier, we have, from Example 10.9, $P_p^{out} = 0.233$ W. The pump power utilization factor is

$$\zeta_p = \frac{P_p^{in} - P_p^{out}}{P_p^{in}} = \frac{2 - 0.233}{2} = 0.884.$$

Using (10.112), we find that the quantum efficiency is

$$\eta_q = \frac{\lambda_s}{\lambda_p}\frac{\eta_c}{\zeta_p} = \frac{1.064 \times 10^{-6}}{808 \times 10^{-9}} \times \frac{10.9\%}{0.884} = 16.2\%.$$

Amplifier bandwidth

The optical bandwidth, B_o, of a laser amplifier is determined by the spectral width, $\Delta\nu_g$, of the gain coefficient $g(\nu)$ and any optical filter that might be incorporated into the device. In the case when there is no additional optical filter, $B_o = \Delta\nu_g$. On the other hand, if a narrow-band optical filter with a bandwidth much smaller than $\Delta\nu_g$ is used at the output of the amplifier, then B_o is simply that of the filter. From the results obtained in the preceding section regarding the gain coefficient, $g(\nu)$ is a function of $\sigma_e(\nu)$, $\sigma_a(\nu)$, and the pumping rate W_p. For a gain medium whose laser transition levels are narrow enough so that $\sigma_e(\nu)$ and $\sigma_a(\nu)$ have the same spectral distribution, or for a four-level system whose lower laser level is empty so that $g(\nu)$ is independent of $\sigma_a(\nu)$, the spectral distribution of $g(\nu)$ is simply that of $\sigma_e(\nu)$. However, as discussed in Section 10.2, the spectral distribution of $\sigma_e(\nu)$ can be very different from that of $\sigma_a(\nu)$ for many practical laser materials. For a quasi-two-level or a three-level system whose $\sigma_e(\nu)$ and $\sigma_a(\nu)$ have different spectral distributions, the spectral distribution of $g(\nu)$ not only depends on both $\sigma_e(\nu)$ and $\sigma_a(\nu)$ but also varies as the pumping rate W_p is varied, as can easily be observed by examining (10.75) and (10.77). Consequently, the optical bandwidth of a laser amplifier that consists of a quasi-two-level or a three-level gain medium is generally a function of the pumping rate, but that of a four-level amplifier is less sensitive to the pumping rate. Because of the resonant nature of the laser transition that is responsible for the gain of a laser amplifier, the optical bandwidth of a laser amplifier is generally quite small in the sense that $B_o \ll \nu_s$, where ν_s is the frequency of an optical signal being amplified.

EXAMPLE 10.11 Find the optical bandwidth of the Nd : YAG laser amplifier described in Example 10.9.

Solution The bandwidth of a laser amplifier is determined by the spectral width of its optical gain, which in turn is largely determined by the spontaneous emission linewidth of the gain medium. The linewidth of Nd : YAG varies with temperature, doping concentration, and crystal quality, but it typically falls between 120 and 180 GHz at room temperature. According to Table 10.1, we find that the typical spontaneous linewidth of Nd : YAG is $\Delta\nu = 150$ GHz. Therefore, we can expect that $B_o = \Delta\nu_g \approx \Delta\nu = 150$ GHz for the amplifier.

Amplifier noise

There are two intrinsic noise sources in a laser amplifier: quantum noise due to spontaneous emission and thermal noise associated with blackbody radiation. At room temperature, these two noise sources have the same magnitude at an electromagnetic wavelength of $\lambda = 44$ μm. Thermal noise dominates at long wavelengths, whereas quantum noise dominates at short wavelengths. Therefore, thermal noise in a laser amplifier that operates in the optical region at room temperature is negligible in the presence of quantum noise caused by spontaneous emission.

The spontaneous emission noise power at the output of a laser amplifier is the result of the *amplified spontaneous emission* (ASE) in the amplifier. It is a function of the gain and bandwidth of the amplifier and is given by

$$P_{sp} = N_{sp} h\nu_s B_o (G - 1), \tag{10.113}$$

where

$$N_{sp} = \frac{\sigma_e N_2}{\sigma_e N_2 - \sigma_a N_1} \tag{10.114}$$

is the amplifier *spontaneous emission factor* that measures the degree of population inversion in the amplifier. In a given amplifier, the value of N_{sp} varies with pumping rate and signal wavelength. It can be seen from (10.114) that $N_{sp} \geq 1$ for an amplifier with $\sigma_e N_2 > \sigma_a N_1$ so that $G > 1$. The minimum value, $N_{sp} = 1$, corresponds to complete population inversion with $N_1 = 0$. In the case of insufficient pumping with $\sigma_e N_2 < \sigma_a N_1$ so that $G < 1$, the amplifier actually attenuates the optical signal rather than amplifying it. Then, N_{sp} has a negative value, but the noise power P_{sp} is still positive because $G < 1$. Therefore, *an ideal amplifier has the minimum noise factor when the amplifying medium has complete population inversion so that* $N_{sp} = 1$.

By ignoring the effect of gain saturation, the spontaneous emission factor defined in (10.114) can be expressed in the following form in terms of pump intensity for an optically pumped system (see Problem 10.4.3):

$$N_{sp} = \left(1 - p\frac{\sigma_a}{\sigma_e} - \frac{I_p^{sat}}{I_p}\frac{\sigma_a}{\sigma_e}\right)^{-1} = \left(1 - p\frac{\sigma_a}{\sigma_e} - \frac{P_p^{sat}}{P_p}\frac{\sigma_a}{\sigma_e}\right)^{-1}, \tag{10.115}$$

where $p \neq 0$ and $\sigma_a \neq 0$ for a quasi-two-level system, $p = 0$ but $\sigma_a \neq 0$ for a three-level system, and $p = 0$ and $\sigma_a = 0$ for a four-level system. It can be seen from (10.115) that a four-level amplifier is normally less noisy than a quasi-two-level or a three-level amplifier because the lower laser level of a four-level system is normally not populated, so that $N_{sp} = 1$. Everything else being equal, a quasi-two-level system is expected to be noisier than a three-level system. We also see from (10.115) that the spontaneous emission factor is reduced toward its minimum value of unity when the pump intensity is increased. For a three-level erbium-doped fiber amplifier, an N_{sp} approaching the ideal minimum value of 1 can be obtained near the peak of the emission spectrum at a high pumping level. For a semiconductor laser amplifier, the value of N_{sp} typically ranges from 1.4 to more than 4, depending on the operating condition.

From (10.113), we see that the ASE of a laser amplifier is directly proportional to the optical bandwidth B_o of the amplifier. To increase the *signal-to-noise ratio* (SNR) at the amplifier output, the total noise power can be reduced to a minimum by placing at the output end of the amplifier an optical filter that has a narrow bandwidth matching the bandwidth of the optical signal.

Because of the spontaneous emission noise, the SNR of an optical signal always degrades after the optical signal passes through an amplifier. The degradation of the SNR of the optical signal passing through an amplifier is measured by the *optical noise figure* of the amplifier defined as

$$F_o = \frac{\text{SNR}_{in}}{\text{SNR}_{out}}, \tag{10.116}$$

where SNR_{in} and SNR_{out} represent the values of the optical SNR at the input and output ends of the amplifier, respectively.

The optical noise figure of an amplifier is a function of the gain and the spontaneous emission factor of the amplifier. It also depends on the photon statistics of the optical signal. For an optical signal that is characterized by a classical electromagnetic field with a large number of photons, the photon statistics can be described by the Poisson statistics. Then, the optical noise figure can be approximated by

$$F_o \approx \frac{1 + 2N_{sp}(G-1)}{G} = 2N_{sp} + \frac{1 - 2N_{sp}}{G}. \tag{10.117}$$

If the amplifying medium has complete population inversion so that $N_{sp} = 1$, then $F_o = 2 - 1/G$ and $2 > F_o > 1$ for $G > 1$. For a high-gain amplifier, $G \gg 1$. Then (10.117) can be further approximated by

$$F_o(G \gg 1) \approx 2N_{sp} \geq 2 \tag{10.118}$$

because $N_{sp} \geq 1$ for an amplifier of $G > 1$. Therefore, unless complete population inversion is achieved in the amplifying medium, the optical noise figure of a high-gain laser amplifier is subject to the *quantum limit* of $F_o \geq 2$, or $F_o \geq 3$ dB. For a low-gain amplifier with sufficient population inversion, it is possible to have a noise

figure less than 2. However, the relation in (10.117) does not imply that the value of F_o can be less than unity if $G < 1$ because N_{sp} has a negative value when $G < 1$. Therefore, an optical amplifier can never improve the SNR of an optical signal. No matter how a laser amplifier is pumped or operated and no matter whether the optical signal is amplified or attenuated, the optical noise figure is always larger than unity, $F_o > 1$.

EXAMPLE 10.12 Find the ASE noise power and the optical noise figure of the Nd : YAG laser amplifier described in Example 10.9 operated with a pump power of 2 W and an input signal power of 20 mW.

Solution From Example 10.9, we find that the power gain is $G = 11.9$ when the amplifier is operated with $P_p = 2$ W and $P_s^{in} = 20$ mW. From Example 10.11, we have $B_o = 150$ GHz. We also have $h\nu_s = 1.165$ eV for the signal photon energy. Because Nd : YAG operating at the signal wavelength of 1.064 μm is a four-level system with $N_1 = 0$, we have $N_{sp} = 1$. Therefore, according to the relation in (10.113), the ASE noise power for this amplifier under the specified operating condition is

$$P_{sp} = 1 \times 1.165 \times 1.6 \times 10^{-19} \times 150 \times 10^9 \times (11.9 - 1) \text{ W} = 305 \text{ nW}.$$

This is the power of the ASE noise at the output of the amplifier. It is only 1.28×10^{-6} of the output signal power of 238 mW. This noise power is small primarily because of the narrow bandwidth of the Nd : YAG laser amplifier.

The optical noise figure of the amplifier can be found by using (10.117) to be

$$F_o = \frac{1 + 2 \times 1 \times (11.9 - 1)}{11.9} = 1.92.$$

This amplifier has a noise figure of $F_o < 2$ because it is a four-level system with $N_{sp} = 1$.

Laser amplifiers and nonlinear optical amplifiers find their primary applications as *power amplifiers* for laser beams, as *optical repeaters* in long-distance optical transmission systems, and as *optical preamplifiers* to optical receivers. In addition, some laser amplifiers, particularly the semiconductor laser amplifiers, can also be used as *optical switches* and nonlinear optical processing devices. Basically all laser amplifiers and nonlinear optical amplifiers can be used as power amplifiers. In fact, because of their large noise figures and narrow bandwidths, many laser amplifiers and nonlinear optical amplifiers are unsuitable for other applications. As power amplifiers, they are used to amplify high-quality but low-power laser beams for the generation of high-power laser beams of good spatial and temporal qualities that cannot be easily generated directly from laser oscillators. The primary consideration in this application is the power conversion efficiency. Therefore, a laser power amplifier is normally operated at saturation level. Not all laser amplifiers and nonlinear optical amplifiers are suitable for the other

applications mentioned above. The suitability for each application varies from one type of amplifier to another. Some laser amplifiers, however, are suitable for multiple applications. For example, the erbium-doped fiber amplifiers, which are discussed in the following section, have found very important applications as power amplifiers, optical repeaters, and preamplifiers in optical communication systems.

10.5 Rare-earth ion-doped fiber amplifiers

Among the many different types of optical amplifiers, those that are guided-wave devices have many advantages over bulk devices. There are two important, but distinctly different, groups of guided-wave optical amplifiers: fiber devices and semiconductor devices. Fiber devices can be further subdivided into two categories: those based on active rare-earth ion-doped fibers and those based on the nonlinear optical processes in fibers. Therefore, there are three types of established guided-wave optical amplifiers: (1) rare-earth ion-doped fiber amplifiers, (2) nonlinear Raman or Brillouin fiber amplifiers, and (3) semiconductor optical amplifiers. Each type can be made into lasers by arranging some proper optical feedback to the amplifiers. Optical amplifiers and lasers based on polymer waveguides are also of great interest, but they are not well established yet. Fiber amplifiers utilize the waveguiding effect of optical fibers, which are nonconductive dielectric glass materials but can be made very long. Semiconductor optical amplifiers use semiconductor waveguides, which are conductive but are of limited length. Fiber devices require optical pumping. Due to the fiber geometry, the only practical and efficient pumping arrangement is *longitudinal optical pumping*. The pump beam is launched into the fiber waveguide either through a fiber coupler or through the end of the fiber. As illustrated in Fig. 10.13, the longitudinal optical pumping arrangement can be unidirectional forward, with the pump and signal waves propagating codirectionally, as shown in Fig. 10.13(*d*), unidirectional backward, with the pump and signal waves propagating contradirectionally, as shown in Fig. 10.13(*e*), or bidirectional, as shown in Fig. 10.13(*f*). Semiconductor devices are normally pumped with electric current injection though they can be optically pumped as well. In this section, we consider only the rare-earth ion-doped fiber amplifiers. The basic principles of Raman and Brillouin amplifiers are discussed in Section 9.9. Semiconductor optical amplifiers are discussed in Section 13.8.

In comparison to bulk optical amplifiers, fiber amplifiers have several advantages. A few are unique to the fiber geometry, but most are common features of waveguide devices and are shared by semiconductor optical amplifiers as well. Some of the advantages are listed below.

1. **Low pump power.** With longitudinal optical pumping, the waveguiding nature of an optical fiber keeps the pump power confined and concentrated in the active

core region, allowing the pump power to be completely absorbed and utilized. In comparison, the pump beam for a longitudinally pumped bulk amplifier cannot be kept focused over a long distance because it is subject to the diffraction limit. In a fiber amplifier, the pump spot size is solely determined by the fiber core diameter, while the effective pumping length is determined only by the absorption coefficient of the fiber gain medium, which in turn is determined by the doping concentration of the rare-earth ions in the fiber core. Decoupling of the pump spot size from consideration of the effective pumping length makes the design of long fiber amplifiers with low rare-earth ion-doping concentrations and low pump powers possible, which are not possible for bulk optical amplifiers.

2. **Good overlap of pump and signal waves.** In a fiber amplifier, the pump and signal waves, though of different wavelengths, overlap over the entire length of the device due to the waveguiding effect of the fiber. This feature improves the efficiency and reduces the required pump power of a fiber amplifier.
3. **Easy control of transverse mode characteristics.** The transverse spatial characteristics of the output beam from a fiber amplifier can be easily and precisely controlled by choosing the fiber to have the desired mode property. To have a diffraction-limited single-mode output beam, even at a very high pumping level, it is only necessary to use a single-mode fiber for the signal wavelength because a single-mode waveguide is the most effective spatial filter that automatically produces a diffraction-limited beam.
4. **Reduced thermal effects.** The fiber geometry, with its small cross-sectional area and large length, naturally has a good efficiency for heat dissipation, thus eliminating the thermal lensing and stress problems often encountered in bulk devices at high pumping levels.
5. **Compatibility with fiber transmission systems.** This physical compatibility leads to efficient integration of fiber amplifiers in an optical fiber transmission system, greatly reducing coupling losses. It also allows large flexibility in the design and handling of the system, including the flexibility of using different pumping arrangements.

Rare-earth ion-doped fibers

Fiber laser amplifiers are based on rare-earth ion-doped fibers. The majority of rare-earth ion-doped fibers are low-loss silica or fluorozirconate glass fibers doped with active rare-earth ions, such as Pr^{3+}, Nd^{3+}, Sm^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, and Yb^{3+}. A rare-earth ion-doped fiber can be either a three-level or a four-level gain system. In some special instances, such as Er^{3+} pumped at 1.48 μm, a rare-earth ion can even operate as a quasi-two-level system. As examples, the energy levels of praseodymium, neodymium, and erbium ions are shown in Fig. 10.16. Some important optical transitions of these ions are summarized in Table 10.2.

Table 10.2 *Some optical transitions in three rare-earth ions*

	Pump		Laser			
Ion	Transition	Wavelength	Transition	Wavelength	System	τ_2
Pr^{3+}	$^3H_4 \rightarrow {}^1G_4$	1.017 μm	$^1G_4 \rightarrow {}^3H_5$	1.31 μm	4-level	100 μs
Nd^{3+}	$^4I_{9/2} \rightarrow {}^4F_{5/2}/^2H_{9/2}$	800 nm	$^4F_{3/2} \rightarrow {}^4I_{9/2}$	900 nm	3-level	500 μs
	$^4I_{9/2} \rightarrow {}^4F_{5/2}/^2H_{9/2}$	800 nm	$^4F_{3/2} \rightarrow {}^4I_{11/2}$	1.06 μm	4-level	500 μs
	$^4I_{9/2} \rightarrow {}^4F_{5/2}/^2H_{9/2}$	800 nm	$^4F_{3/2} \rightarrow {}^4I_{13/2}$	1.34 μm	4-level	500 μs
Er^{3+}	$^4I_{15/2} \rightarrow {}^4I_{9/2}$	800 nm	$^4I_{13/2} \rightarrow {}^4I_{15/2}$	1.53 μm	3-level	10 ms
	$^4I_{15/2} \rightarrow {}^4I_{11/2}$	980 nm	$^4I_{13/2} \rightarrow {}^4I_{15/2}$	1.53 μm	3-level	10 ms
	$^4I_{15/2} \rightarrow {}^4I_{13/2}$	1.48 μm	$^4I_{13/2} \rightarrow {}^4I_{15/2}$	1.53 μm	quasi-2-level	10 ms

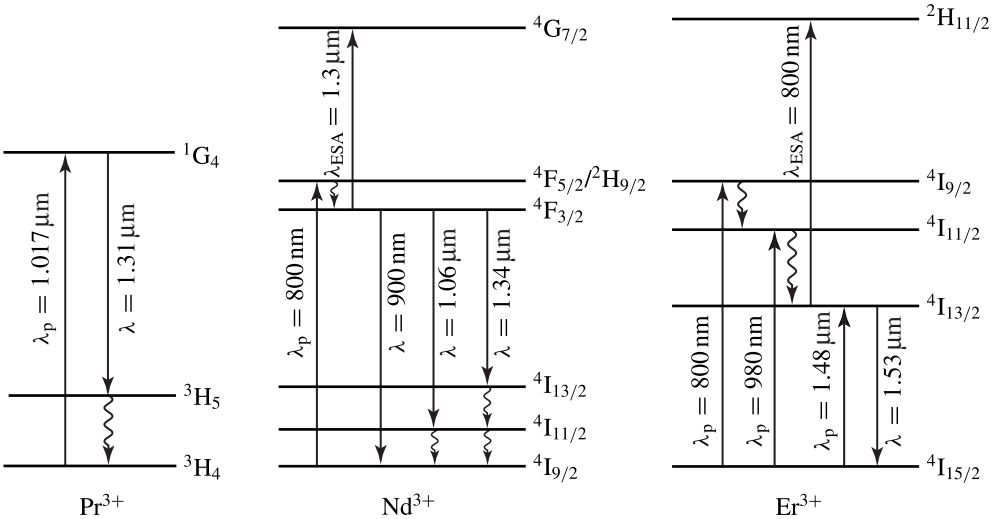


Figure 10.16 Energy levels of praseodymium, neodymium, and erbium ions.

One important property of rare-earth ion-doped fibers is that their transition characteristics, including the spectral broadening mechanism, the spectral shape and width, the spectral peak wavelength, and the fluorescence lifetime, are influenced by the molecular environment of the rare-earth ions. The most important factor is the structure and composition of the host material. Also important is the operating temperature. The absorption and emission spectral widths of a given ion doped in a glass fiber are generally very broad, much broader than those of the same ion doped in a crystalline material. The spectral characteristics can be significantly varied by using a completely different glass material for the fiber or by adjusting the composition of the glass. As an example, the absorption cross-section spectrum, $\sigma_a(\lambda)$, and the emission cross-section spectrum, $\sigma_e(\lambda)$, measured at room temperature in the spectral region around 1.53 μm for Er^{3+} doped in a silica fiber doped with Al_2O_3 and P_2O_5 are shown in Fig. 10.17(*a*), which can be compared with those shown in Fig. 10.17(*b*) for Er^{3+} doped in an Al_2O_3/GeO_2–silica

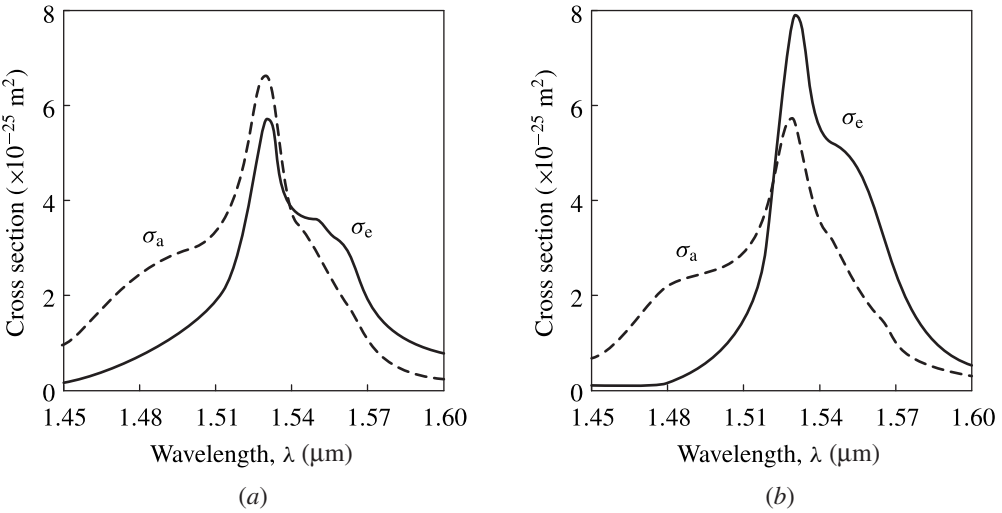


Figure 10.17 Absorption and emission cross-section spectra of Er^{3+} in (*a*) an Al_2O_3/P_2O_5–silica fiber and (*b*) an Al_2O_3/GeO_2–silica fiber. (Adapted from (*a*) Miniscalco, W. J., "Erbium-doped glasses for fiber amplifiers at 1500 nm," *Journal of Lightwave Technology* **9**(2): 234–250, Feb. 1991, and (*b*) Desurvire, E. and Simpson, J. R., "Evaluation of $^4I_{15/2}$ and $^4I_{13/2}$ Stark-level energies in erbium-doped aluminosilicate glass fibers," *Optics Letter* **15**(10): 547–549, May 1990.)

fiber. Other factors that affect the spectral characteristics and the fluorescence lifetime include the doping concentration of the active ion and the codopants. The laser emission wavelength corresponding to each laser transition shown in Table 10.2 can be varied and tuned within a rather broad range.

Because of their glass hosts, rare-earth ion-doped fibers have mixed homogeneous and inhomogeneous broadening characteristics. The relative significance between the two varies with the fiber host material, the dopant, the doping concentration, and temperature. For many rare-earth ion-doped fibers of interest, the homogeneous line broadening at room temperature is about the same as the inhomogeneous broadening. Experimental results on fiber amplifiers seem to be adequately explained by simple models that assume pure homogeneous broadening.

Fiber amplifiers

The development of rare-earth ion-doped fiber amplifiers was driven primarily by their applications in fiber-optic communication systems for amplifying weak optical signals. For this reason, the major effort has been on the development of erbium-doped fiber amplifiers (EDFAs) for amplifying optical signals in the spectral region around 1.55 μm, where silica fiber transmission lines have minimum attenuation loss. Also of interest are praseodymium-doped fiber amplifiers (PDFAs) and neodymium-doped fiber amplifiers (NDFAs) for the 1.3-μm spectral region, where many of the existing optical

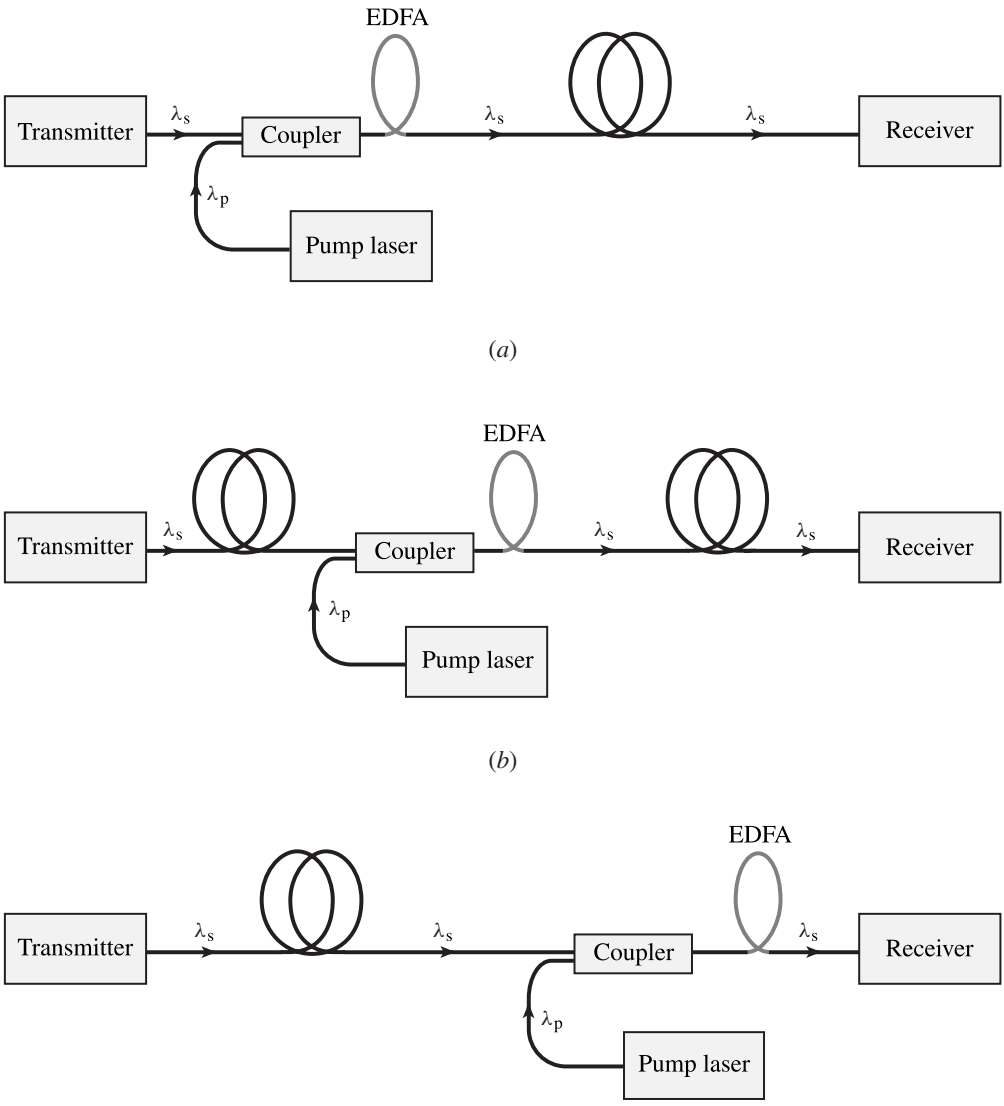


Figure 10.18 Use of a fiber amplifier as (*a*) a power amplifier, (*b*) an optical repeater, and (*c*) an optical preamplifier in a fiber-optic communication system.

communication systems operate because of the minimum dispersion and low attenuation loss of silica fibers in this spectral window. These fiber amplifiers can be used as power amplifiers (postamplifiers), optical repeaters (inline amplifiers), or optical preamplifiers, shown in Figs. 10.18(*a*), (*b*), and (*c*), respectively, in optical communication systems.

Fiber amplifiers have to be optically pumped. Practical considerations such as reliability, package size, cost, and power efficiency dictate that semiconductor lasers be used

to pump fiber amplifiers in most applications. Efficient semiconductor laser sources are available for the pump bands of Nd^{3+} and Er^{3+} listed in Table 10.2: AlGaAs/GaAs lasers for the 800-nm pump band, InGaAs/GaAs strained quantum-well lasers for the 980-nm pump band, and InGaAsP/InP lasers for the 1.48-μm pump band. The 1.017-μm pump band of Pr^{3+} is not so conveniently located for pumping with a semiconductor laser. However, by codoping with Yb^{3+}, a praseodymium-doped fiber can be pumped at 980 nm. An EDFA can be pumped at 800 or 980 nm, or 1.48 μm, but pumping in the 800-nm pump band is not very efficient due to a phenomenon known as excited-state absorption (ESA) at 800 nm. Therefore, practical EDFAs are pumped at either 980 nm or 1.48 μm. A silica-based NDFA operating in the 1.34-μm signal wavelength region also suffers from ESA, but due to absorption of signal photons rather than pump photons. This problem can be avoided in an NDFA that is based on a fluorozirconate glass fiber instead of a silica fiber.

It is important to recognize whether a particular fiber amplifier in a certain operating condition functions as a quasi-two-level system, a three-level system, or a four-level system because the characteristics of laser amplifiers vary significantly among the three systems. As indicated in Table 10.2, this depends on the combination of the active ion in the fiber, the pump transition used, and the laser wavelength of interest. For example, an EDFA for the signal laser wavelength of 1.53 μm functions as a three-level system when it is pumped at 800 or 980 nm and as a quasi-two-level system when pumped at 1.48 μm, but it is never a four-level system.

Rare-earth ion-doped fiber amplifiers have the general characteristics of laser amplifiers discussed in the preceding section. However, because of the waveguide structure of optical fibers, evaluation of certain parameters has to be modified when the general formulations in the preceding section are applied to fiber amplifiers. The required modification depends largely on the concentration profile of the active rare-earth ions in the fiber. In general, these ions are doped only in the core of the fiber, but they may distribute throughout the core area or reside only in a fraction of the core area. In the following discussions, we assume that the entire fiber core is uniformly doped with active ions at a concentration of N_t but the cladding contains no active ions. In this situation, only a fraction, quantified by the fiber mode confinement factor defined in (3.64), of the power of an optical beam guided in the fiber interacts with the active ions. Because the pump and the signal of a given fiber amplifier have different wavelengths, their confinement factors, Γ_p and Γ_s, respectively, have different values. With this understanding, we can find the following required modifications when applying the formulations in the preceding section to a fiber amplifier:

$$\alpha_p = \Gamma_p \sigma_a^p N_t, \tag{10.119}$$

$$P_p^{sat} = \pi w_p^2 \frac{h\nu_p}{\Gamma_p \eta_p \tau_2 \sigma_a^p}, \tag{10.120}$$

$$P_p^{tr} = \frac{\sigma_a}{\sigma_e - p\sigma_a} P_p^{sat}, \tag{10.121}$$

for the pump, where w_p is the effective mode radius of the pump beam, and

$$g_0(z) = \frac{\Gamma_s(\sigma_e - p\sigma_a)N_t}{1 + (1+p)P_p(z)/P_p^{sat}} \left[\frac{P_p(z)}{P_p^{sat}} - \frac{P_p^{tr}}{P_p^{sat}} \right]$$
$$= \frac{\Gamma_s(\sigma_e + \sigma_a)N_t}{1 + (1+p)P_p(z)/P_p^{sat}} \frac{P_p(z)}{P_p^{sat}} - \Gamma_s\sigma_a N_t, \tag{10.122}$$

$$P_{sat} = \pi w_s^2 \frac{h\nu_s}{\Gamma_s \tau_s \sigma_e}, \tag{10.123}$$

for the signal, where w_s is the effective mode radius of the signal beam. For a fiber amplifier of a length l, the integral of the unsaturated gain coefficient for both cases of $p \neq 0$ and $p = 0$ has the following closed-form solution:

$$\int_0^l g_0(z)\mathrm{d}z = \Gamma_s\sigma_e N_t l + \frac{\Gamma_s(\sigma_e + \sigma_a)N_t}{\alpha_p} \ln \frac{P_p^{out}}{P_p^{in}},$$
$$= \Gamma_s\sigma_e N_t l + \frac{\Gamma_s(\sigma_e + \sigma_a)N_t}{\alpha_p} \ln(1 - \zeta_p). \tag{10.124}$$

Similarly to the relation between (10.106) and (10.108), this expression can be transformed into

$$\int_0^l g_0(z)\mathrm{d}z = \begin{cases} \dfrac{\Gamma_s(\sigma_e + \sigma_a)N_t}{p\alpha_p} \ln \dfrac{P_p^{sat} + pP_p^{in}}{P_p^{sat} + pP_p^{out}} - \Gamma_s\sigma_a N_t l, & \text{for } p \neq 0, \\ \dfrac{\Gamma_s(\sigma_e + \sigma_a)N_t}{\alpha_p} \left(\dfrac{P_p^{in} - P_p^{out}}{P_p^{sat}} \right) - \Gamma_s\sigma_a N_t l, & \text{for } p = 0. \end{cases} \tag{10.125}$$

EXAMPLE 10.13 An EDFA uses a step-index Al_2O_3/GeO_2–silica fiber doped with an Er concentration of $N_t = 2.2 \times 10^{24}$ m^{-3} in its core of $a = 4.5$ μm radius. It is pumped in the forward direction at $\lambda_p = 1.48$ μm to amplify a signal at $\lambda_s = 1.53$ μm. At both wavelengths, the fiber is single moded supporting only the fundamental HE_{11} mode with effective mode radii of $w_p = 4.0$ μm and $w_s = 4.1$ μm and confinement factors of $\Gamma_p = 0.72$ and $\Gamma_s = 0.70$ for the pump and the signal, respectively. The fluorescence lifetime is $\tau_2 = 10$ ms. At the pump wavelength of 1.48 μm, $\sigma_a^p = 2.2 \times 10^{-25}$ m^2 and $\sigma_e^p = 1.2 \times 10^{-26}$ m^2. At the signal wavelength of 1.53 μm, $\sigma_e = 7.9 \times 10^{-25}$ m^2 and $\sigma_a = 5.75 \times 10^{-25}$ m^2. The background absorption of the fiber at both pump and signal wavelengths are negligible. The fiber length is chosen to be $l = 20$ m, and the input pump power is $P_p^{in} = 20$ mW. The pumping efficiency is $\eta_p = 1$. (a) Find P_p^{sat}, P_p^{tr}, and P_{sat}. (b) Find the unsaturated power gain G_0. (c) If the power of the input signal is $P_s^{in} = 1$ μW, what is the output signal power?

Solution (a) The pump photon energy is $h\nu_p = (1.2398/1.48)$ eV $= 0.838$ eV. The signal photon energy is $h\nu_s = (1.2398/1.53)$ eV $= 0.810$ eV. When pumped at $\lambda_p = 1.48$ μm, this EDFA is a quasi-two-level system with $p = \sigma_e^p/\sigma_a^p = 1.2/22 = 0.055$. We find, by using (10.120) and (10.121), respectively, that

$$P_p^{sat} = \pi w_p^2 \frac{h\nu_p}{\Gamma_p \eta_p \tau_2 \sigma_a^p} = \frac{\pi \times (4.0 \times 10^{-6})^2 \times 0.838 \times 1.6 \times 10^{-19}}{0.72 \times 1 \times 10 \times 10^{-3} \times 2.2 \times 10^{-25}} \text{ W} = 4.25 \text{ mW}$$

and

$$P_p^{tr} = \frac{\sigma_a}{\sigma_e - p\sigma_a} P_p^{sat} = \frac{5.75 \times 10^{-25}}{7.9 \times 10^{-25} - 0.055 \times 5.75 \times 10^{-25}} \times 4.25 \text{ mW} = 3.22 \text{ mW}.$$

Without pumping, $W_p\tau_2 = 0$, and

$$\tau_s = \tau_2 \left(1 + \frac{\sigma_a}{\sigma_e}\right) = \left(1 + \frac{5.75}{7.9}\right) \times 10 \text{ ms} = 17.3 \text{ ms}$$

from (10.76). Therefore, the intrinsic saturation power is

$$P_{sat} = \pi w_s^2 \frac{h\nu_s}{\Gamma_s \tau_s \sigma_e} = \frac{\pi \times (4.1 \times 10^{-6})^2 \times 0.810 \times 1.6 \times 10^{-19}}{0.70 \times 17.3 \times 10^{-3} \times 7.9 \times 10^{-25}} \text{ W} = 716 \text{ μW}.$$

For $P_p^{in} = 20$ mW, the pumping ratio is $s = P_p^{in}/P_p^{sat} = 20/4.25 = 4.71$. Therefore, $W_p\tau_2 = s = 4.71$, and

$$\tau_s = \tau_2 \frac{1 + \sigma_a/\sigma_e}{1 + (1+p)W_p\tau_2} = \frac{1 + 5.75/7.9}{1 + 1.055 \times 4.71} \times 10 \text{ ms} = 2.9 \text{ ms}$$

from (10.76). The saturation power for the signal at this pumping level is

$$P_{sat} = \pi w_s^2 \frac{h\nu_s}{\Gamma_s \tau_s \sigma_e} = \frac{\pi \times (4.1 \times 10^{-6})^2 \times 0.810 \times 1.6 \times 10^{-19}}{0.70 \times 2.9 \times 10^{-3} \times 7.9 \times 10^{-25}} \text{ W} = 4.27 \text{ mW}.$$

This is the signal saturation power at the input pump power of $P_p^{in} = 20$ mW. As the pump power decays along the fiber due to pump absorption, τ_s increases toward the value of 17.3 ms and, as a consequence, P_{sat} decreases toward its intrinsic value of 716 μW.

(b) To find G_0, we have to find the integral of $g_0(z)$ over the entire length of the fiber. This integral can be evaluated by using (10.124) or, equivalently, by using (10.125). For this fiber, we have

$$\alpha_p = \Gamma_p \sigma_a^p N_t = 0.72 \times 2.2 \times 10^{-25} \times 2.2 \times 10^{24} \text{ m}^{-1} = 0.348 \text{ m}^{-1}.$$

With $P_p^{sat} = 4.25$ mW found above and $P_p(0) = P_p^{in} = 20$ mW, we can solve (10.102) for $z = l = 20$ m to find that $P_p^{out} = P_p(l) = 984$ μW. All other parameters needed for

calculating the integral in (10.124) are known. We find from (10.124) that

$$\int_0^l g_0(z)\mathrm{d}z = 0.7 \times 7.9 \times 10^{-25} \times 2.2 \times 10^{24} \times 20 + \frac{0.7 \times (7.9 + 5.75) \times 10^{-25} \times 2.2 \times 10^{24}}{0.348} \times \ln \frac{0.984}{20} = 6.14.$$

Then, we have

$$G_0 = \mathrm{e}^{6.14} = 464,$$

which is 26.7 dB.

Although we have found G_0 by evaluating the integral of $g_0(z)$ through (10.124), it is instructive to find the distributions of the pump power $P_\mathrm{p}(z)$ and the unsaturated gain coefficient $g_0(z)$ as a function of distance along the EDFA. It can then be shown that the value of the integral of $g_0(z)$ obtained by directly integrating over its distribution is the same as that evaluated above using (10.124) to confirm the validity of (10.124). To find $P_\mathrm{p}(z)$ and $g_0(z)$ as a function of distance along the EDFA, we first find $P_\mathrm{p}(z)$ through (10.102). We then find $g_0(z)$ by using (10.122) and G_0 as a function of z by using (10.99). The results are plotted in Fig. 10.19. Also plotted in Fig. 10.19(*a*) is the function $\exp(-\alpha_\mathrm{p} z)$ for comparison with $P_\mathrm{p}(z)/P_\mathrm{p}^\mathrm{in}$ to show that the pump power decays

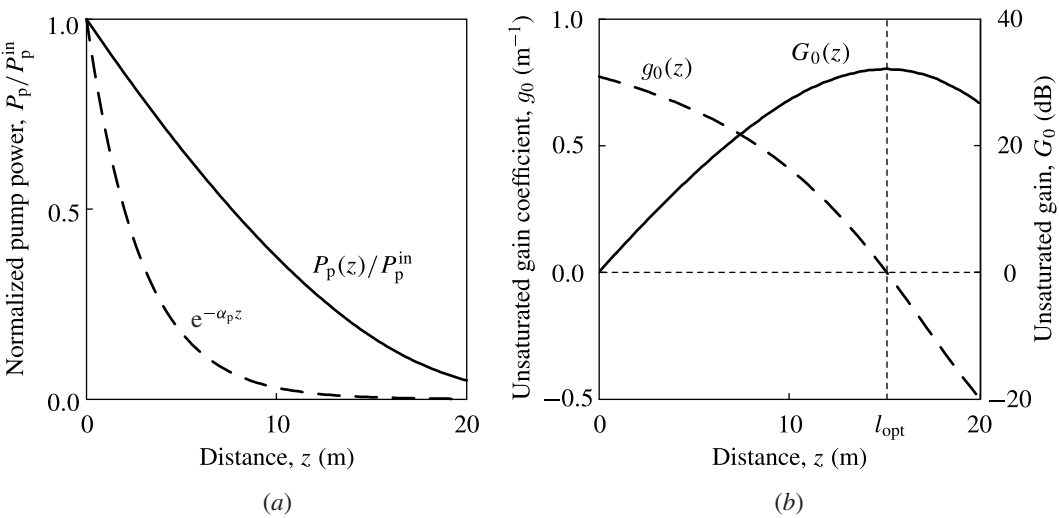


Figure 10.19 (*a*) Pump power evolution and (*b*) gain variation in an EDFA. Plotted as a function of distance z along the fiber from the input of the EDFA are (*a*) pump power normalized to the input pump power, $P_\mathrm{p}(z)/P_\mathrm{p}^\mathrm{in}$, and exponentially decaying function $\mathrm{e}^{-\alpha_\mathrm{p} z}$ for comparison with the pump power evolution and (*b*) unsaturated gain coefficient $g_0(z)$ per meter and unsaturated power gain $G_0(z)$ in decibels.

much more slowly than the exponential function because of the absorption saturation of the pump at the high pumping ratio of $s = 4.71$. We find from the curve in Fig. 10.19(b) for $G_0(z)$ that $G_0 = 26.7$ dB, or $G_0 = 464$ at $z = l = 20$ m, confirming the result obtained above.

(c) From (a), we know that the signal saturation power varies along the EDFA in the range of 716 μW $< P_{\text{sat}} <$ 4.27 mW. If we take $P_{\text{sat}} = 4.27$ mW, we find that $P_s^{\text{in}}/P_{\text{sat}} = 2.34 \times 10^{-4}$ for $P_s^{\text{in}} = 1$ μW. Then, with $G_0 = 464$ obtained above, we find from (10.98) that $G = 420$. If we take $P_{\text{sat}} = 716$ μW, we find $G = 308$. The actual gain is somewhere between these two limits, about $G = 364$. Therefore, the output signal power is about $P_s^{\text{out}} = 364$ μW. Here we have not considered the effect of the ASE. As the ASE can be significant, thus depleting a significant portion of the population inversion, the realistic gain available for the signal could be much smaller than that estimated here (see Problem 10.5.3). For an accurate solution, we need to solve (10.95) together with (10.101) numerically while including the effect of ASE in the process.

The length of a fiber amplifier can easily be made large. However, at a given pump power for a longitudinally pumped fiber amplifier, if the fiber gain medium is too long, part of the fiber gain medium will not be pumped because the pump power will be totally absorbed before any of it reaches the far end of the fiber. If the fiber gain medium is too short, part of the pump power will not be absorbed. Therefore, *when considering the optimum length of a fiber amplifier at a given pump power level, the difference between a four-level system and a three-level or a quasi-two-level system is significant.*

In a four-level system, such as a PDFA or an NDFA operating in the 1.3-μm region, unexcited active ions are transparent to the signal photons. Therefore, in order to obtain the maximum gain at a given pump power level for a longitudinally pumped four-level fiber amplifier, it is only necessary to make sure that the fiber gain medium is long enough so that all of the pump power is absorbed. Further increasing the length of the fiber amplifier does not have much effect on the overall gain of the amplifier, provided that the background attenuation coefficient of the host fiber is low. In a three-level or a quasi-two-level system, such as an EDFA operating in the 1.53-μm region, the situation is very different because unexcited active ions that remain in the ground state are in resonance with the signal frequency to absorb the signal photons. Therefore, at a given pump power level, there is an optimum length for a three-level or a quasi-two-level fiber amplifier to have the maximum overall gain. A three-level fiber amplifier that is shorter than the optimum length does not use the pump power efficiently, whereas one that is longer than the optimum length suffers from absorption of the signal by the gain medium in the unpumped section of the fiber.

The optimum length of a fiber amplifier can be found by considering the fact that the maximum value of the integral of $g_0(z)$ over the length of the fiber occurs when $g_0(l) = 0$ at the end of the fiber. Physically, this condition is equivalent to requiring

that $P_p(l) = P_p^{tr}$ so that G_0 is maximized, as can be seen by considering $g_0(l) = 0$ for (10.122). By applying this condition, we find from (10.102) that the optimum length of a quasi-two-level fiber amplifier, for which $p \neq 0$, is

$$l_{opt} = -\frac{1}{\alpha_p}\left(\ln\frac{P_p^{tr}}{P_p^{in}} + \frac{1}{p}\ln\frac{P_p^{sat} + pP_p^{tr}}{P_p^{sat} + pP_p^{in}}\right), \tag{10.126}$$

and from (10.103) that the optimum length of a three-level fiber amplifier, for which $p = 0$, is

$$l_{opt} = -\frac{1}{\alpha_p}\left(\ln\frac{P_p^{tr}}{P_p^{in}} + \frac{P_p^{tr} - P_p^{in}}{P_p^{sat}}\right). \tag{10.127}$$

The optimum length of a fiber amplifier clearly depends on the input pump power.

EXAMPLE 10.14 Find the optimum length of the EDFA described in Example 10.13 for an input pump power of $P_p^{in} = 20$ mW. What is the unsaturated power gain G_0 when the optimum length is used for such an EDFA?

Solution This EDFA is a quasi-two-level system with $p = 0.055$. From Example 10.13, we find that $\alpha_p = 0.348\ \text{m}^{-1}$, $P_p^{sat} = 4.25$ mW, and $P_p^{tr} = 3.22$ mW. Using (10.126), we find that the optimum length of this amplifier at an input power of $P_p^{in} = 20$ mW is

$$l_{opt} = -\frac{1}{0.348} \times \left(\ln\frac{3.22}{20} + \frac{1}{0.055} \times \ln\frac{4.25 + 0.055 \times 3.22}{4.25 + 0.055 \times 20}\right) \text{m} = 15.14\ \text{m}.$$

We find from Fig. 10.19(*b*) that $g_0(z) = 0$ and G_0 has a maximum value of 32 dB at $z = 15.14$ m, in agreement with what is found here. Therefore, the unsaturated power gain is $G_0 = 32$ dB for the EDFA with an optimum length of 15.14 m at $P_p^{in} = 20$ mW.

In comparison to other laser gain media, a rare-earth ion-doped fiber typically has a broad gain bandwidth, a relatively small emission cross section, and a long fluorescence lifetime, as can easily be seen from Table 10.2 and Fig. 10.17. These characteristics have very important implications for the practical applications of fiber amplifiers. The broad gain bandwidth allows tunability and tolerance on the wavelength of the input signal. The small emission cross section implies that the amplifier gain is not easily perturbed by variations of the signal power because variations in the population inversion caused by stimulated emission are small. Therefore, nonlinear effects such as distortion of the signal waveform and cross interference between different signal channels are minimized. Together with the broad gain bandwidth, this characteristic allows a single rare-earth ion-doped fiber amplifier to be used for amplifying multiple optical channels in a wavelength division multiplexing system. The long fluorescence lifetime

is no less important. In particular, the fluorescence lifetime of the $^4I_{13/2}$ level of an erbium-doped fiber is on the order of $\tau_2 = 10$ ms, though it varies somewhat among different host glasses. Indeed, the success of EDFAs is due in large part to this long fluorescence lifetime because it allows an EDFA to maintain a high gain at a modest pump power under constant CW pumping. In addition, a long fluorescence lifetime for the upper laser level means that population inversion and, therefore, gain do not respond to any pump fluctuation or noise that varies on a time scale less than the saturation lifetime τ_s. Though τ_s is smaller than τ_2 at a pumping level of $s = W_p\tau_2 = P_p^{in}/P_p^{sat} > 1$ required for an EDFA, it is still on the order of 1 ms at a high pumping level of $s = 10$. For this reason, an EDFA is not susceptible to noise or intermodulation distortions at frequencies higher than 1 kHz even at such a high pumping level.

Another significant characteristic of a rare-earth ion-doped glass fiber is that the orientation of the active rare-earth ions doped in the fiber is randomized because of the amorphous structure of the host glass. This effect results in polarization-independent absorption and emission cross sections. Consequently, the optical gain in a rare-earth ion-doped glass fiber amplifier is insensitive to the polarization state of the optical signal.

PROBLEMS

10.1.1 A He–Ne laser is a four-level system that has a spontaneous radiative lifetime of $\tau_{sp} = 300$ ns for emission at 632.8 nm wavelength. The upper and lower laser levels for this emission wavelength both relax radiatively, with $\tau_2 = \tau_2^{rad} = 30$ ns and $\tau_1 = \tau_1^{rad} = 10$ ns, respectively. Find the naturally broadened homogeneous linewidth and the lifetime-broadened linewidth. Compare them to the linewidth due to Doppler broadening found in Example 10.2. What is the expected lineshape?

10.1.2 The linewidth of an Ar-ion laser emitting at 488 nm is predominately inhomogeneously broadened due to Doppler broadening. The atomic mass number of Ar is 40, and the operating gas temperature of an Ar-ion laser is about 1200 °C. Find the emission linewidth at this wavelength.

10.1.3 A ruby laser is a three-level solid-state system. For its emission at 694.3 nm wavelength, the lower laser level is the ground state of the Cr^{3+} ions doped in a sapphire crystal. The upper laser level relaxes radiatively with $\tau_2 = \tau_2^{rad} = \tau_{sp} = 3$ ms. The measured linewidth at room temperature is $\Delta\nu = 330$ GHz.

a. Find the natural linewidth and the lifetime-broadened homogeneous linewidth.

b. If the measured linewidth at room temperature is entirely homogeneously broadened, what is the linewidth-broadening factor $\gamma_{21}^{dephase}$ due to dephasing through phonon collisions?

10.1.4 A copper vapor laser is a gas laser that operates at a gas temperature of about 1500 °C. The atomic mass number of Cu is 63. For the emission of a copper vapor laser at 510.5 nm, the upper laser level relaxes almost entirely radiatively only to the lower laser level with $\tau_2 = \tau_2^{\text{rad}} = \tau_{\text{sp}} = 500$ ns. The lower laser level is a long-living metastable state that relaxes nonradiatively with a lifetime on the order of $\tau_1 = 10$ μs.

a. Find the homogeneous linewidths due to natural broadening and lifetime broadening, respectively.
b. Find the inhomogeneous linewidth due to Doppler broadening.
c. Is the observed linewidth primarily homogeneously or inhomogeneously broadened? What is the expected lineshape?
d. Explain why the copper vapor laser is a three-level system though the lower laser level is not the ground state.
e. Give an explanation to account for the fact that a copper vapor laser cannot be operated in a CW mode but can only be pulsed.

10.1.5 The spectral distribution of blackbody radiation is normally measured with a spectrometer as a function of radiation wavelength instead of radiation frequency. Therefore, the experimentally measured spectral energy density of blackbody radiation is expressed as $u(\lambda)$ rather than as $u(\nu)$.

a. Use the relation $u(\lambda)\mathrm{d}\lambda = -u(\nu)\mathrm{d}\nu$ and Planck's formula for $u(\nu)$ given in (10.20) to show that

$$u(\lambda) = \frac{8\pi n^3 hc}{\lambda^5} \frac{1}{\mathrm{e}^{hc/\lambda k_{\text{B}} T} - 1}. \tag{10.128}$$

b. The blackbody radiation spectrum obtained in (a) as a function of radiation wavelength peaks at wavelength λ_{pk}. Show that λ_{pk} is a function of temperature only and that it varies inversely with temperature according to the following relation known as *Wien's displacement law*:

$$\lambda_{\text{pk}} T = 2.898 \times 10^{-3} \text{ m K}. \tag{10.129}$$

c. Find the peak wavelength of blackbody radiation at the human body temperature of 37 °C.
d. The sun has a surface temperature of 6000 K. Find the peak wavelength of solar radiation. Is the color of the sun seen at sea level determined by this peak wavelength?

10.1.6 Show that the peak value of the emission cross section of a laser gain medium at the center wavelength λ of the emission spectrum can be expressed as that in (10.39) for a homogeneously broadened medium with an ideal Lorentzian lineshape and as that in (10.40) for an inhomogeneously broadened medium with an ideal Gaussian lineshape.

10.1.7 Show that for a given peak emission wavelength λ and a given spontaneous radiative lifetime τ_{sp}, we have the following relation between the peak emission cross sections of a homogeneously broadened and an inhomogeneously broadened medium:

$$\frac{\sigma_e^h}{\sigma_e^{inh}} = \frac{1}{(\pi \ln 2)^{1/2}} \frac{\Delta\nu_{inh}}{\Delta\nu_h} = 0.678 \frac{\Delta\nu_{inh}}{\Delta\nu_h}. \tag{10.130}$$

How much larger is the peak emission cross section of an inhomogeneously broadened medium than that of a homogeneously broadened medium if the two media have the same emission wavelength, spontaneous radiative lifetime, and emission linewidth?

10.1.8 Use the data and information provided in Table 10.1 to calculate the peak emission cross sections for the following gain media: Ar ion, copper vapor, ruby, Nd : YAG, Ti : sapphire, and Cr : LiSAF. Compare the calculated results with the values of σ_e listed in the table for these media.

10.2.1 A Nd : YAG crystal doped with 1 at. % Nd, equivalent to 0.725 wt. % Nd, has a Nd concentration of 1.38×10^{26} m^{-3}. Use the data listed in Table 10.1 to find the percentage of active Nd ions that need to be pumped to the upper laser level $^4F_{3/2}$ in order to have a gain coefficient of 12.5 m^{-1} at 1.064 μm wavelength?

10.2.2 A Ti : sapphire crystal is doped with 0.1 wt. % Ti_2O_3 for a Ti concentration of 3.3×10^{25} m^{-3}. Use its spectra of $\sigma_a(\lambda)$ and $\sigma_e(\lambda)$ shown in Fig. 10.8 to answer the following questions.

a. What are the absorption coefficients for $\mathbf{E} \parallel c$ and $\mathbf{E} \perp c$ polarizations, respectively, at the absorption peak wavelength of 490 nm if the crystal is not pumped?

b. If 1% of the Ti ions are excited to the upper laser level, what are the gain coefficients for the two polarizations at the gain peak wavelength of 795 nm?

10.2.3 Explain why the peak value of the absorption cross section of Ti : sapphire is so much smaller than that of the emission cross section, as seen from Fig. 10.8.

10.2.4 The spontaneous emission spectrum of a laser medium is determined by the emission cross section spectrum of the medium, but the two spectra are not exactly the same. They have different shapes and peak at different frequencies.

a. Show by using (10.31) and (10.36) that

$$W_{sp}(\nu) = \frac{8\pi n^2 \nu^2}{c^2} \sigma_e(\nu). \tag{10.131}$$

b. By definition, the spontaneous lifetime τ_{sp} is the inverse of the total spontaneous relaxation rate:

$$\frac{1}{\tau_{sp}} = \int_0^\infty W_{sp}(\nu)\mathrm{d}\nu = \int_0^\infty W_{sp}(\lambda)\mathrm{d}\lambda. \tag{10.132}$$

Find $W_{sp}(\lambda)$ from this relation. Show that (10.45) and (10.46) can be obtained from (10.132).

c. Show that the spectral power density of spontaneous emission emitted by each atom in the upper laser level is

$$\hat{P}_{sp}(\nu) = \frac{8\pi n^2 h\nu^3}{c^2}\sigma_e(\nu) \tag{10.133}$$

per unit frequency interval and is

$$\hat{P}_{sp}(\lambda) = \frac{8\pi n^2 hc^2}{\lambda^5}\sigma_e(\lambda) \tag{10.134}$$

per unit wavelength interval.

d. Compared to the peak wavelength of the emission cross section spectrum $\sigma_e(\lambda)$, does the peak wavelength of the spontaneous emission spectrum $\hat{P}_{sp}(\lambda)$ shift to the longer or the shorter wavelength side?

e. Show that the total spontaneous emission power from a gain medium of a volume $\mathcal{V}$ that is pumped to have a population density N_2 in its upper laser level is

$$P_{sp} = \mathcal{V}N_2\frac{8\pi n^2 h}{c^2}\int_0^\infty \nu^3\sigma_e(\nu)\mathrm{d}\nu = \mathcal{V}N_2 8\pi n^2 hc^2\int_0^\infty \frac{\sigma_e(\lambda)}{\lambda^5}\mathrm{d}\lambda. \tag{10.135}$$

10.2.5 Modify the relations obtained in Problem 10.2.4(a) and (b) as needed for uniaxial crystals, such as ruby or Ti : sapphire, in which the emission cross section spectrum varies with the polarization of the emitted radiation with respect to the c axis of the crystal.

10.2.6 Show that the imaginary part of the resonant susceptibility given in (10.53) can be expressed in the following form when the g_1 states in level $|1\rangle$ and the g_2 states in level $|2\rangle$ are not split into sublevels:

$$\chi''_{res}(\omega) = \left(\frac{g_2}{g_1}N_1 - N_2\right)\frac{\pi^2 c^3}{n\omega^3\tau_{sp}}\hat{g}(\omega). \tag{10.136}$$

Once the lineshape function $\hat{g}(\omega)$ of a resonant transition is determined, the real part, $\chi'_{res}(\omega)$, of this resonant susceptibility can be found generally through the

Kramers–Kronig relations given in (1.177). For a homogeneously broadened transition, show that

$$\chi_{\text{res}}(\omega) = \chi'_{\text{res}}(\omega) + \mathrm{i}\chi''_{\text{res}}(\omega) \approx -\left(\frac{g_2}{g_1}N_1 - N_2\right)\frac{\pi c^3}{n\omega_{21}^3\tau_{\text{sp}}}\frac{1}{(\omega-\omega_{21})+\mathrm{i}\gamma_{21}}. \tag{10.137}$$

10.3.1 In this problem, we consider the optical gain in a medium.
   a. Show that the optical gain coefficient as a function of optical signal intensity for all three basic systems can be expressed in the form of (10.73) with the saturation intensity expressed in the form of (10.74).
   b. Show that the unsaturated gain coefficient g_0 and the saturation lifetime τ_{s} for the three systems are those given in (10.75)–(10.80).
   c. Show that the minimum pumping requirement obtained for $g_0 > 0$ is identical to that given in (10.69) in the case of a quasi-two-level system and that given in (10.71) in the case of a three-level system.

10.3.2 Show that the required pumping rate for a desired unsaturated gain coefficient of g_0 is that given in (10.81) for a quasi-two-level system, that given in (10.82) for a three-level system, and that given in (10.83) for a four-level system.

10.3.3 Use the data given in Table 10.1 to find the saturation intensities at transparency for the following gain media at their respective emission peak wavelengths: HeNe at 632.8 nm, Ar ion at 488 nm, Nd : YAG at 1.064 μm, Nd : glass at 1.054 μm, Er : fiber at 1.53 μm, Ti : sapphire at 800 nm, and Cr : LiSAF at 825 nm. For quasi-two-level systems such as Ti : sapphire and Cr : LiSAF, take $p = 0$ for simplicity.

10.3.4 Show that the unsaturated gain coefficient of any optically pumped system as a function of pump intensity can be expressed as the relation in (10.88), where $p \neq 0$ and $I_{\text{p}}^{\text{tr}} \neq 0$ for a quasi-two-level system, $p = 0$ but $I_{\text{p}}^{\text{tr}} \neq 0$ for a three-level system, and $p = 0$ and $I_{\text{p}}^{\text{tr}} = 0$ for a four-level system.

10.3.5 Nd : YAG is a four-level system for its 1.064 μm transition line. Being a four-level system, $\sigma_{\text{a}} = 0$. It has many pump bands, but a strong pump band exists at 808 nm wavelength with a peak absorption cross section of $\sigma_{\text{a}}^{\text{p}} = 5.6 \times 10^{-24}$ m^2. A Nd : YAG laser rod is doped with a Nd concentration of 1.38×10^{26} m^{-3}. From Table 10.1, $\tau_2 = 240$ μs and $\tau_{\text{sp}} = 515$ μs. Take an emission cross section of $\sigma_{\text{e}} = 9 \times 10^{-23}$ m^2. The pump quantum efficiency is $\eta_{\text{p}} = 47\%$.
   a. Find the pumping rate and the pump intensity required for an unsaturated gain coefficient of 5 m^{-1}. Compare the results obtained here for Nd : YAG to those obtained in Example 10.7 for ruby.
   b. What is the spontaneous emission power density if the Nd : YAG crystal is pumped to a gain coefficient of 5 m^{-1} for the 1.064-μm line? If a Nd : YAG crystal rod of 6 cm length and 4 mm cross-sectional diameter is uniformly

pumped, what is the spontaneous emission power in this situation? Compare the results with those obtained in Example 10.8 for ruby.

10.3.6 Ti : sapphire is a quasi-two-level system. A Ti : sapphire laser rod is doped with 0.024 wt. % Ti_2O_3 for a Ti concentration of 7.9×10^{24} m^{-3}. At the desired Ti : sapphire laser wavelength of $\lambda = 800$ nm for $\mathbf{E} \parallel c$ polarization, $\sigma_e = 3.4 \times 10^{-23}$ m^2 and $\sigma_a \approx 8 \times 10^{-26}$ m^2. At the pump wavelength of 532 nm, the absorption cross section is $\sigma_a^p = 7.4 \times 10^{-24}$ m^2, and the emission cross section is $\sigma_e^p \approx 3 \times 10^{-28}$ m^2. From Table 10.1, $\tau_2 = 3.2$ μs and $\tau_{sp} = 3.9$ μs. The pump quantum efficiency is $\eta_p = 80\%$.

a. Find the transparency pumping rate and the transparency pump intensity of the 532 nm pump required for Ti : sapphire to reach transparency for the 800 nm transition for $\mathbf{E} \parallel c$ polarization. Find the pumping rate and the pump intensity required for an unsaturated gain coefficient of 5 m^{-1} in the crystal under consideration. Compare the results obtained here for Ti : sapphire with those obtained in Example 10.7 for ruby.

b. Find the critical fluorescence power density corresponding to transparency for the 800 nm transition. What is the spontaneous emission power density if the Ti : sapphire crystal is pumped above transparency for a gain coefficient of 5 m^{-1} at 800 nm laser wavelength? Compare the results with those obtained in Example 10.8 for ruby.

10.3.7 Find the population densities, N_1 and N_2, in the lower and upper laser levels, respectively, as fractions of the total population density, N_t, for quasi-two-level, three-level, and four-level systems, respectively, when each medium is pumped at (a) its saturation pump intensity, P_p^{sat}, and (b) its transparency pump intensity, P_p^{tr}.

10.4.1 Show that the relation in (10.96) describes the spatial evolution of the amplified signal power in a laser amplifier that has a spatially varying unsaturated gain coefficient of $g_0(z)$. Show also that the power gain of an amplified signal can be found from the relation in (10.98).

10.4.2 Show that the integral of the unsaturated gain coefficient over the length of an amplifier has the form given in (10.106) irrespective of whether $p \neq 0$ or $p = 0$. Show also that it can be expressed in different forms for the cases of $p \neq 0$ and $p = 0$ given (10.108).

10.4.3 Show by using the relations in (10.87), (10.88), and (10.91) that when the effect of gain saturation is ignored, the spontaneous emission factor N_{sp} as defined in (10.114) can be expressed in the form of (10.115).

10.4.4 The Nd : YAG laser amplifier described in Example 10.9 is pumped with the same 808-nm pump source considered in Example 10.9 at the same pump power of $P_p = 2$ W. It is used to amplify a signal at 1.064 μm that has an input power of $P_s^{in} = 100$ mW. It has an optical bandwidth of 150 GHz, as found in Example 10.11.

a. What is the unsaturated power gain of the amplifier? What are the power gain and the output power of the signal? Compare the results with those obtained in Example 10.9.
b. What are the power conversion efficiency and the quantum efficiency? Compare the results with those obtained in Example 10.10.
c. Find the ASE noise power and the optical noise figure. Compare the results with those obtained in Example 10.12.

10.4.5 Answer the questions in Problem 10.4.4 for the Nd : YAG laser amplifier if it is pumped at $P_p = 3$ W for amplification of a signal that has an input power of $P_s^{in} = 100$ mW.

10.5.1 The spectra of the absorption and emission cross sections of an Er-doped Al_2O_3/P_2O_5–silica glass are shown in Fig. 10.17(*a*). The upper laser level is $^4I_{13/2}$ with a degeneracy of $g_2 = 14$, and the lower laser level is $^4I_{15/2}$ with $g_1 = 16$.
a. Find the spontaneous lifetime τ_{sp} through the relation in (10.46) by using the emission spectrum $\sigma_e(\lambda)$.
b. Find the spontaneous lifetime τ_{sp} through the relation in (10.46) by using the absorption spectrum $\sigma_a(\lambda)$. Compare the result with that obtained in (a).

10.5.2 Use the McCumber relation given in (10.48) to calculate the emission cross section spectrum $\sigma_e(\lambda)$ for Er-doped Al_2O_3/P_2O_5–silica glass from its absorption cross section spectrum $\sigma_a(\lambda)$ shown in Fig. 10.17(*a*). Compare the calculated $\sigma_e(\lambda)$ with the measured spectrum also shown in Fig. 10.17(*a*).

10.5.3 The EDFA described in Example 10.13 has a spontaneous linewidth of $\Delta\nu = 4.8$ THz. Consider its amplification of an input signal of 1 μW as described in Example 10.13(c).
a. Find the spontaneous emission factor N_{sp} for the EDFA at the input pump power level of 20 mW.
b. What is the ASE power if no optical filter is used? What is the ASE power if a bandpass optical filter having a bandwidth of $B_o = 400$ GHz centered at the signal wavelength is used? How do these ASE powers compare with the output signal power of the amplifier?
c. What is the noise figure of the amplifier?

10.5.4 The optimum length found in Example 10.14 is chosen for the EDFA pumped at 1.48 μm with an input pump power of 20 mW for amplification of an input signal of 1 μW at 1.53 μm. First find the output signal power. Then, answer the questions in Problem 10.5.3 for this EDFA of optimum length.

10.5.5 If the EDFA discussed in Example 10.13 is pumped at an input pump power of 10 mW, the gain would be much less than that found in Example 10.13.
a. Find the unsaturated power gain of the EDFA, which has a fiber length of 20 m.
b. Find the optimum length and the corresponding unsaturated power gain of the EDFA.

c. For a given input pump power, there is a maximum length of the fiber, beyond which $G_0 < 0$ dB and the EDFA becomes an attenuator of the signal. Find the maximum length at the input pump power level of 10 mW.

10.5.6 The EDFA described in Example 10.13 can be pumped at 980 nm instead. When it is pumped at this wavelength, it behaves as a three-level system with $\sigma_a^p = 2.58 \times 10^{-25}$ m^2. The fundamental mode of the fiber at this pump wavelength has an effective mode radius of $w_p = 3.3$ μm and a confinement factor of $\Gamma_p = 0.84$. The pumping efficiency is $\eta_p = 1$. Answer the questions in Example 10.13 for this EDFA pumped at 980 nm with an input pump power of 20 mW for the amplification of an input signal of 1 μW at 1.53 μm. Compare the results with those found in Example 10.13.

10.5.7 Answer the questions in Problem 10.5.3 for the EDFA under the operating conditions specified in Problem 10.5.6.

10.5.8 Find the optimum length of the EDFA considered in Problem 10.5.6, which is pumped at 980 nm with an input pump power of 20 mW for amplification of an input signal of 1 μW at 1.53 μm. Find the unsaturated power gain G_0 and the output signal power. Answer the questions in Problem 10.5.3 for this EDFA of optimum length. Compare the results with those found in Problem 10.5.4 for an EDFA of optimum length pumped at 1.48 μm.

SELECT BIBLIOGRAPHY

Becker, P. C., Olsson, N. A., and Simpson, J. R., *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology*. San Diego, CA: Academic Press, 1999.

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge University Press, 1996.

Desurvire, E., *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley, 1994.

Digonnet, M. J. F., ed., *Rare Earth Doped Fiber Lasers and Amplifiers*, 2nd edn. New York: Marcel Dekker, 2001.

France, P. W., ed., *Optical Fiber Lasers and Amplifiers*. London: Blackie, 1991.

Iizuka, K., *Elements of Photonics for Fiber and Integrated Optics*, Vol. II. New York: Wiley, 2002.

Kaminiskii, A. A., *Laser Crystals*, 2nd edn. Berlin: Springer-Verlag, 1990.

Lee, T. P., ed., *Current Trends in Optical Amplifiers and Their Applications*. Singapore: World Scientific, 1996.

Milonni, P. W. and Eberly, J. H., *Lasers*. New York: Wiley, 1988.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Siegman, A. E., *Lasers*. Mill Valley, CA: University Science Books, 1986.

Silfvest, W. T., *Laser Fundamentals*. Cambridge: Cambridge University Press 1996.

Verdeyen, J. T., *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Yariv, A., *Optical Electronics in Modern Communications*, 5th edn. Oxford: Oxford University Press, 1997.

Yeh, C., *Applied Photonics*. San Diego, CA: Academic Press, 1994.

ADVANCED READING LIST

Ainsile, J., "A review of the fabrication and properties of erbium-doped fibers for optical amplifiers," *Journal of Lightwave Technology* **9**(2): 220–227, Feb. 1991.

Bibeau, C., Payne, S. A., and Powell, H. T., "Direct measurement of the terminal laser level lifetime in neodymium-doped crystals and glasses," *Journal of the Optical Society of America B* **12**(10): 1981–1992, Oct. 1995.

Brignon, A., Feugnet, G., Huignard, J. P., and Pocholle, J. P., "Compact Nd : YAG and Nd : YVO_4 amplifiers end-pumped by a high-brightness stacked array," *IEEE Journal of Quantum Electronics* **34**(3): 577–585, Mar. 1998.

Giles, C. R., and Desurvire, E. "Modeling erbium doped fiber amplifiers," *Journal of Lightwave Technology* **9**(2): 271–283, Feb. 1991.

Haus, H. A., "Optimum noise performance of optical amplifiers," *IEEE Journal of Quantum Electronics* **37**(6): 813–823, June 2001.

Kück, S., "Laser-related spectroscopy of ion-doped crystals for tunable solid-state lasers," *Applied Physics B* **B72**(5): 515–562, Apr. 2001.

McCumber, D. E., "Theory of phonon-terminated optical masers," *Physical Review* **134**: A299–A306, 1964.

Miniscalco, W. J., "Erbium-doped glasses for fiber amplifiers at 1500 nm," *Journal of Lightwave Technology* **9**(2): 234–250, Feb. 1991.

Miniscalco, W. J., and Quimby, R. S., "General procedure for the analysis of Er^{3+} cross sections," *Optics Letters* **16**(4): 258–260, Feb. 1991.

Moulton, P. F., "Spectroscopic and laser characteristics of Ti : Al_2O_3," *Journal of the Optical Society of America B* **3**(1): 125–133, Jan. 1986.

Olsson, N. A., "Lightwave systems with optical amplifiers," *Journal of Lightwave Technology* **7**(7): 1071–1082, July 1989.

Payne, D. N., "Active fibres and optical amplifiers," *Fiber and Integrated Optics* **11**(3): 191–219, 1992.

Urquhart, P., "Review of rare earth doped fiber lasers and amplifiers," *IEE Proceedings, Part J Optoelectronics* **135**(6): 385–407, Dec. 1988.

Yamamoto Y. and Mukai, T., "Fundamentals of optical amplifiers," *Optical and Quantum Electronics* **21**: S1–S14, 1989.

11 Laser oscillators

There are a wide variety of lasers, covering a spectral range from the soft X-ray to the far infrared, delivering output powers from microwatts to terawatts, operating from continuous wave to femtosecond pulses, and having spectral linewidths from just a few hertz to many terahertz. The gain media utilized include plasma, free electrons, ions, atoms, molecules, gases, liquids, solids, and so on. The sizes range from microscopic, of the order of 10 μm^3, to gigantic, of an entire building, to stellar, of astronomical dimensions. An optical gain medium can amplify an optical field through stimulated emission. If the gain medium is sufficiently long, it is possible to generate laser light at one end of the medium through amplification of some initial optical field from spontaneous emission produced at the other end of the gain medium. Astrophysical laser action in space has been found to occur naturally, for example at the deep ultraviolet wavelength of 250 nm from the star Eta Carinae, at the near-infrared H_2 wavelength of 2.286 μm from the star NGC 7072, at the far-infrared wavelength of 169 μm in a disk of hydrogen gas surrounding the star MWC349 in the constellation Cygnus, and at the mid-infrared CO_2 wavelength of 10.6 μm in the Martian atmosphere. In a practical laser device, however, it is generally necessary to have certain positive optical feedback in addition to optical amplification provided by a gain medium. This requirement can be met by placing the gain medium in an optical resonator. The optical resonator provides selective feedback to the amplified optical field.

Lasers are indeed fascinating, but not all of them are of practical usefulness as photonic devices. In this chapter, we discuss the characteristics of laser oscillators in general. Optical fiber lasers are specifically discussed in Section 11.5. Semiconductor lasers are arguably the most important lasers in the photonics industry. They are covered in great detail in Chapter 13.

11.1 Resonant optical cavities

One major characteristic of laser light is that it is highly collimated and spatially and temporally coherent. This characteristic is a direct consequence of the fact that laser oscillation takes place only along a longitudinal axis of an optical resonator,

which can be either straight or folded. The gain medium emits spontaneous photons in all directions, but only the radiation that propagates along the longitudinal axis within a small divergence angle defined by the resonator obtains sufficient regenerative amplification to reach the threshold for oscillation. In order for the oscillating laser field in the longitudinal direction to be amplified most efficiently, any spontaneous photon emitted in a direction outside of that small angular range should not be allowed to compete for the gain. For this reason, a functional laser oscillator is necessarily an *open cavity* with optical feedback only along the longitudinal axis. Most of the randomly directed spontaneous photons escape from the cavity through the open sides very quickly. Only a very small fraction of them that happen to be emitted within the divergence angle of the laser field mix with the oscillating laser field to become the major incoherent noise source of the laser.

A laser cavity can take a variety of forms. Figure 11.1 shows the schematic structures of a few common laser cavities. Though a laser cavity is always an open cavity with a clearly defined longitudinal axis, the axis can lie on a straight line, as in Figs. 11.1(*a*) and (*e*), or it can be defined by a folded path, as in Figs. 11.1(*b*), (*c*), and (*d*). A linear cavity with two end mirrors, as in Fig. 11.1(*a*), is known as a *Fabry–Perot cavity* because it takes the form of a Fabry–Perot interferometer. A folded cavity can simply be a folded Fabry–Perot cavity with a standing oscillating field, as in Fig. 11.1(*b*). A folded cavity can also be a non-Fabry–Perot *ring cavity* that supports two independent oscillating fields traveling in opposite directions, as in Figs. 11.1(*c*) and (*d*). The optical feedback in a Fabry–Perot cavity is provided simply by the two end mirrors perpendicular to the longitudinal axis, as in Figs. 11.1(*a*) and (*b*). In a ring cavity, it is provided by the circulation of the laser field along the ring path defined by mirrors, as in Fig. 11.1(*c*), or by a fiber waveguide, as in Fig. 11.1(*d*). It can also be supplied by the distributed feedback of a distributed Bragg grating along the axis, as in Fig. 11.1(*e*). The cavity can also be constructed with an optical waveguide, as in the case of a semiconductor laser or a fiber laser. In the following discussions, we take the longitudinal axis to define the z coordinate, and the transverse coordinates perpendicular to the longitudinal axis to be the x and y coordinates. In a folded cavity, the z axis is thus also folded along with the longitudinal optical path.

In a ring cavity, an intracavity field completes one round trip by circulating inside the cavity in only one direction. The two contrapropagating fields that circulate in opposite directions in a ring cavity are independent of each other even when they have the same frequency. In a Fabry–Perot cavity, an intracavity field has to travel the length of the cavity twice in opposite directions to complete a round trip. The time it takes for an intracavity field to complete one round trip in the cavity is called the *round-trip time*, T:

$$T = \frac{\text{round-trip optical path length}}{c} = \frac{l_{\text{RT}}}{c}, \tag{11.1}$$

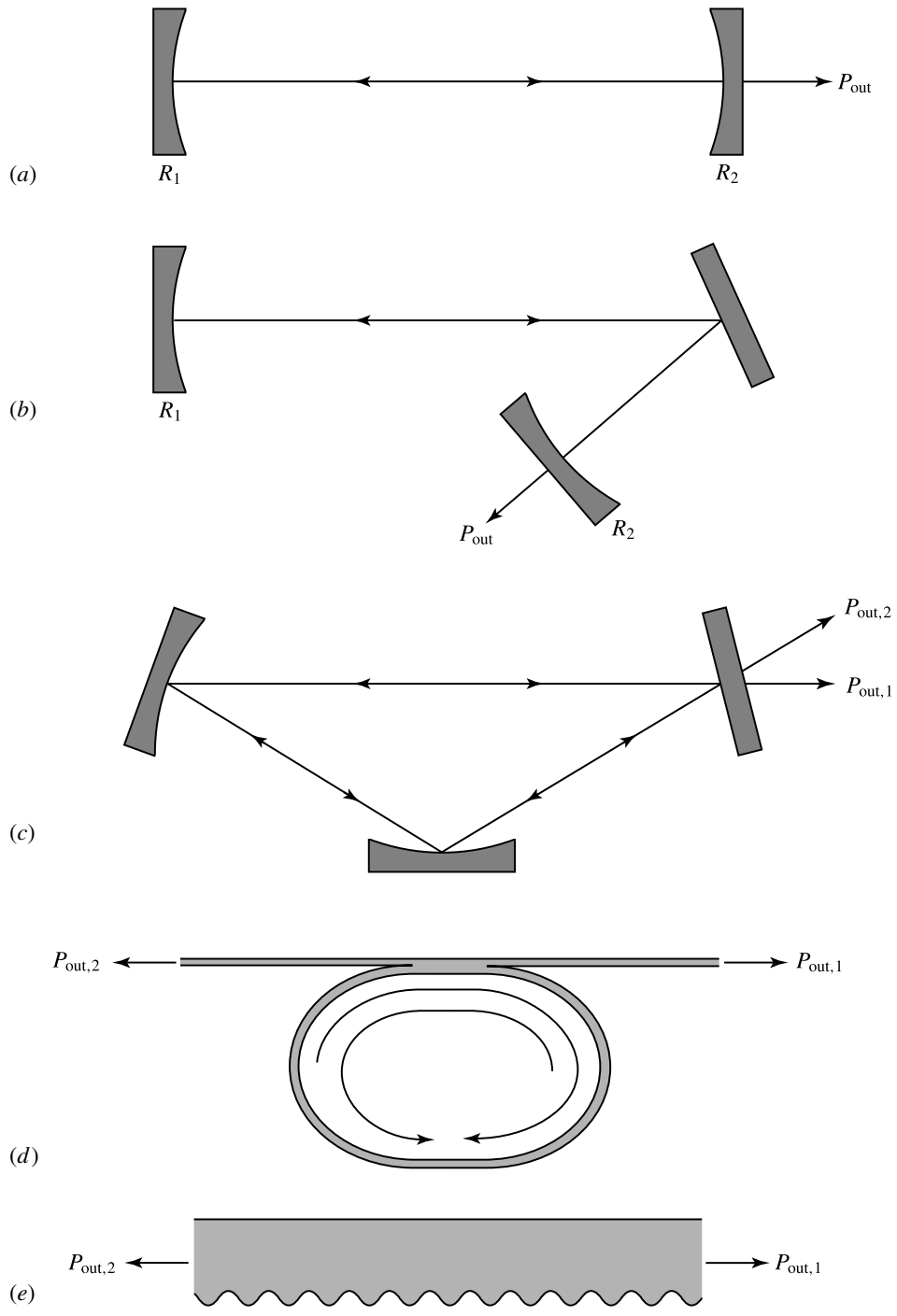


Figure 11.1 Schematics of a few common laser cavity structures: (*a*) linear cavity with end mirrors; (*b*) folded cavity with end mirrors; (*c*) three-mirror ring cavity with two independent, contrapropagating fields; (*d*) ring cavity with two independent, contrapropagating fields guided by an optical-fiber waveguide; and (*e*) cavity with a distributed Bragg grating.

where the *round-trip optical path length* l_{RT} takes into account the refractive index of the medium inside the cavity.

A laser consists of at least a gain medium in a resonant laser cavity. The gain medium may fill up the entire length of the cavity, or it may occupy a fraction of the cavity length. For a gain medium of a length l_{g} in a laser cavity of a length l, as shown in Fig. 11.2, we can define an *overlap factor* between the gain medium and the laser mode intensity distribution as the ratio

$$\Gamma = \frac{\iiint\limits_{\text{active}} |\mathbf{E}|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} |\mathbf{E}|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z} \approx \frac{\mathcal{V}_{\text{active}}}{\mathcal{V}_{\text{mode}}} \approx \frac{l_{\mathrm{g}}}{l}. \tag{11.2}$$

This ratio is commonly known as the *gain filling factor* for a gain medium that takes up only a fraction of the length of the laser cavity but is related to the mode confinement factor in a waveguide laser, such as the fiber laser or the semiconductor laser. When the gain medium fills up a laser cavity and covers the entire intracavity laser field distribution, $\Gamma = 1$; otherwise, $\Gamma < 1$. Take the refractive index of the gain medium to be n and that of the intracavity medium excluding the gain medium to be n_0, then the round-trip optical path length can be expressed as

$$l_{\mathrm{RT}} = \begin{cases} 2\Gamma n l + 2(1-\Gamma) n_0 l = 2\overline{n} l, & \text{for a linear cavity,} \\ \Gamma n l + (1-\Gamma) n_0 l = \overline{n} l, & \text{for a ring cavity,} \end{cases} \tag{11.3}$$

where

$$\overline{n} = \Gamma n + (1-\Gamma) n_0 \tag{11.4}$$

is the weighted average index of refraction throughout the laser cavity. When optical elements other than a gain medium exist in a laser cavity, $\overline{n}$ is still the weighted average index throughout the laser cavity with n_0 being the weighted average index of the background medium and such optical elements.

Consider an intracavity field, $\mathbf{E}_{\mathrm{c}}(z)$, at any point z on the longitudinal axis inside an optical cavity. When it completes a round trip back to position z, it is modified by a complex amplification or attenuation factor a to become $a\mathbf{E}_{\mathrm{c}}(z)$. The factor a can be expressed generally as

$$a = G \exp\left(\mathrm{i}\varphi_{\mathrm{RT}}\right), \tag{11.5}$$

where G is the *round-trip gain factor for the field amplitude*, equivalent to the *power gain in a single pass through a linear Fabry–Perot cavity*, and φ_{RT} is the *round-trip phase shift* for the intracavity field. Both G and φ_{RT} have real values, and $G \geq 0$. If $G > 1$, the intracavity field is amplified. If $G < 1$, the intracavity field is attenuated.

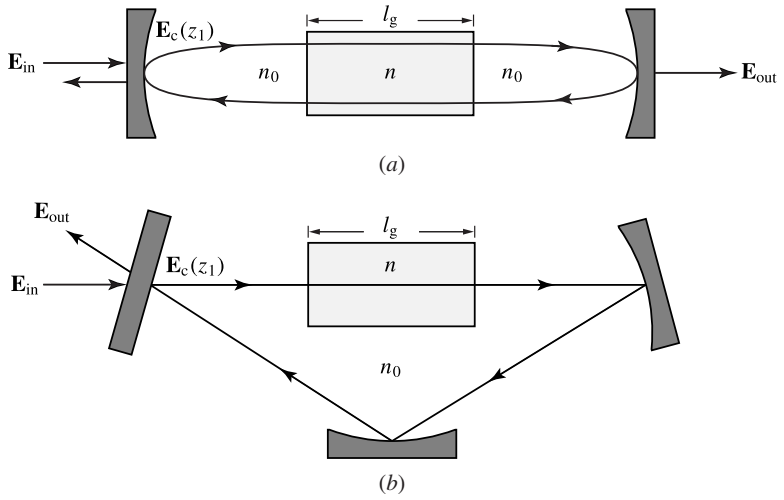


Figure 11.2 Passive laser cavities with a gain filling factor Γ under optical injection: (*a*) a Fabry–Perot cavity and (*b*) a ring cavity. The refractive index of the gain medium is n, while that of the background medium in the cavity is n_0. A laser cavity is passive when its gain medium is absent or is present but not pumped.

Longitudinal modes

We first consider the resonant characteristics of a passive optical cavity. A passive cavity cannot generate or amplify an optical field. In order to keep a resonant intracavity field in such a cavity, it is necessary to inject an input optical field, $\mathbf{E}_{in}$, to the cavity constantly. As shown in Fig. 11.2, the forward-traveling component of the intracavity field at location z_1 just inside the cavity next to the injection point is the sum of the transmitted input field and the fraction of the intracavity field returning after one round trip through the cavity:

$$\mathbf{E}_c(z_1) = t_{in}\mathbf{E}_{in} + a\mathbf{E}_c(z_1), \tag{11.6}$$

where t_{in} is the complex transmission coefficient for the input field. We find that

$$\mathbf{E}_c(z_1) = \frac{t_{in}}{1-a}\mathbf{E}_{in}. \tag{11.7}$$

The transmitted output field, $\mathbf{E}_{out}$, is proportional to the intracavity field: $\mathbf{E}_{out} \propto \mathbf{E}_c(z_1)$. Therefore, the output intensity is proportional to the input intensity through the following relationship:

$$I_{out} \propto \frac{I_{in}}{|1-a|^2} = \frac{I_{in}}{(1-G)^2 + 4G\sin^2(\varphi_{RT}/2)}. \tag{11.8}$$

The proportionality constant of this relationship depends on the transmission coefficient of the output port and the amount of intracavity absorption over the distance from point z_1 to the output point. The transmittance of the cavity is $T_c = I_{out}/I_{in}$, which is scaled by the value of this proportionality constant. For our discussions in the following,

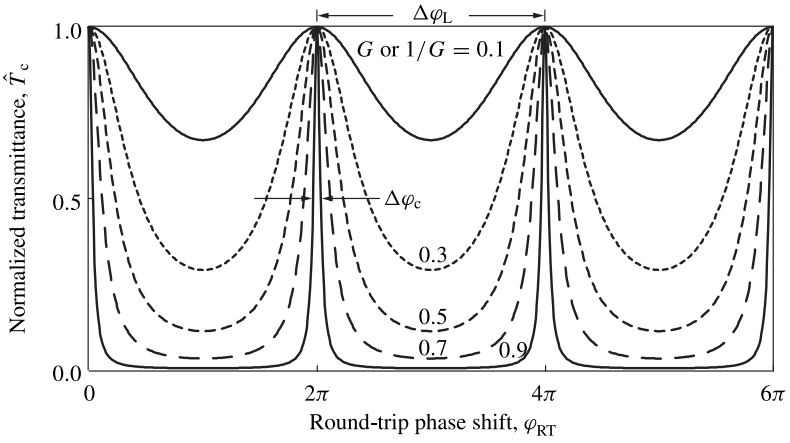


Figure 11.3 Normalized transmittance of a passive cavity as a function of the round-trip phase shift in the cavity. In a resonator with a fixed, frequency-independent optical path length, the round-trip phase shift is directly proportional to the optical frequency. The longitudinal mode frequencies are defined by the frequencies corresponding to the resonance peaks. The spectral shape for a gain of G is the same as that for a gain of $1/G$. Thus, the curve for $G = 10$ is the same as that for $G = 0.1$, that for $G = 2$ is the same as that for $G = 0.5$, and so on.

however, this proportionality constant is irrelevant. Therefore, we only have to consider the following normalized transmittance of the passive cavity:

$$\hat{T}_c = \frac{1}{1 + [4G/(1-G)^2]\sin^2(\varphi_{RT}/2)} = \frac{1}{1 + [(4/G)/(1-1/G)^2]\sin^2(\varphi_{RT}/2)}, \quad (11.9)$$

which is obtained by normalizing T_c to its peak value. In Fig. 11.3, $\hat{T}_c$ is plotted as a function of φ_{RT} for a few different values of G. We find that *the spectral shape for a gain of G is the same as that for a gain of* $1/G$.

At any given input field intensity, the intracavity field intensity of a passive cavity is proportional to $\hat{T}_c$ because the transmitted field is directly proportional to the intracavity field. Therefore, resonances of the cavity occur at the peaks of $\hat{T}_c$ where the intracavity intensity reaches a maximum with respect to a constant input field intensity. As can be seen from Fig. 11.3, the resonance condition of the cavity is that the round-trip phase shift is an integral multiple of 2π:

$$\varphi_{RT} = 2q\pi, \qquad q = 1, 2, \ldots \quad (11.10)$$

From (11.9) and Fig. 11.3, we find that the separation between two neighboring resonance peaks of $\hat{T}_c$ is

$$\Delta\varphi_L = 2\pi \quad (11.11)$$

and that the FWHM of each resonance peak of the cavity is

$$\Delta\varphi_c = 2\frac{1-G}{G^{1/2}}. \quad (11.12)$$

The *finesse*, F, of the cavity is the ratio of the separation to the FWHM of the peaks:

$$F = \frac{\Delta\varphi_{\mathrm{L}}}{\Delta\varphi_{\mathrm{c}}} = \frac{\pi G^{1/2}}{1-G}. \tag{11.13}$$

In the simplest situation where the optical field is a plane wave at a frequency ω, the round-trip phase shift can be generally expressed as

$$\varphi_{\mathrm{RT}} = \frac{\omega}{c} l_{\mathrm{RT}} + \varphi_{\mathrm{local}}, \tag{11.14}$$

where the first term on the right-hand side is the phase shift contributed by the propagation of the optical field over an optical path length of l_{RT}, and the second term, φ_{local}, is the sum of all the localized, and usually fixed, phase shifts such as those caused by reflection from the mirrors of a cavity. In the case when the frequency of the input field is fixed, the resonance condition given in (11.10) can be satisfied by varying the cavity path length l_{RT}, either by varying the physical length of the cavity or by varying the refractive index of the intracavity medium, or both. The optical cavity then functions as an *optical interferometer*, which is used to measure the frequency and the spectral width of an optical wave accurately.

When both the optical path length and the localized phase shifts are fixed, as is typically the case in a laser resonator, the resonance condition of $\varphi_{\mathrm{RT}} = 2q\pi$ is satisfied only if the optical frequency satisfies

$$\omega_q = \frac{c}{l_{\mathrm{RT}}}(2q\pi - \varphi_{\mathrm{local}}), \tag{11.15}$$

or

$$\nu_q = \frac{c}{l_{\mathrm{RT}}}\left(q - \frac{\varphi_{\mathrm{local}}}{2\pi}\right). \tag{11.16}$$

These discrete resonance frequencies are the *longitudinal mode* frequencies of the optical resonator because they are defined by the resonance condition of the round-trip phase shift along the longitudinal axis of the cavity. The frequency spacing, $\Delta\nu_{\mathrm{L}}$, between two neighboring longitudinal modes is known as the *free spectral range* of the optical resonator. The FWHM of a longitudinal mode spectral peak is $\Delta\nu_{\mathrm{c}}$. If the values of l_{RT} and φ_{local} are independent of frequency, then $\Delta\nu_{\mathrm{L}} \propto \Delta\varphi_{\mathrm{L}}$ and $\Delta\nu_{\mathrm{c}} \propto \Delta\varphi_{\mathrm{c}}$. Therefore, *the finesse of the resonator is the ratio of the free spectral range to the longitudinal mode width:*

$$F = \frac{\Delta\varphi_{\mathrm{L}}}{\Delta\varphi_{\mathrm{c}}} = \frac{\Delta\nu_{\mathrm{L}}}{\Delta\nu_{\mathrm{c}}}. \tag{11.17}$$

From (11.16), we find that

$$\Delta\nu_{\mathrm{L}} = \nu_{q+1} - \nu_q = \frac{c}{l_{\mathrm{RT}}} = \frac{1}{T}. \tag{11.18}$$

The longitudinal mode width can be expressed as

$$\Delta\nu_c = \frac{\Delta\nu_L}{F} = \frac{1-G}{\pi G^{1/2}}\Delta\nu_L. \tag{11.19}$$

Transverse modes

Any realistic optical cavity has a finite transverse cross-sectional area. Therefore, the resonant optical field inside an optical cavity cannot be a plane wave. Indeed, there exist certain normal modes for the transverse field distribution in a given optical cavity. Such transverse field patterns are known as the *transverse modes* of a cavity. *A transverse mode of an optical cavity is a stable transverse field pattern that reproduces itself after each round-trip pass in the cavity, except that it might be amplified or attenuated in magnitude and shifted in phase.*

The transverse modes of an optical cavity are defined by the transverse boundary conditions that are imposed by the transverse cross-sectional index profile of the cavity. For a cavity that utilizes an optical waveguide for lateral confinement of the optical field, the transverse modes are clearly the waveguide modes, such as the TE and TM modes of a slab waveguide or the TE, TM, HE, and EH modes of a cylindrical fiber waveguide. For a nonwaveguiding cavity, the transverse modes are TEM fields determined by the shapes and sizes of the end mirrors of the cavity, as well as by the properties of the medium and any other optical components inside the cavity. The *Gaussian modes* discussed in Section 1.7 are an important set of such unguided TEM modes.

In an optical cavity that supports multiple transverse modes, the round-trip phase shift is generally a function of the transverse mode indices m and n. Therefore, the resonance condition can be explicitly written as

$$\varphi_{mn}^{RT} = 2q\pi. \tag{11.20}$$

As a result, the resonance frequencies of the cavity are ω_{mnq}, or ν_{mnq}, which are functions of both longitudinal and transverse mode indices. For a given longitudinal mode index q, multiple resonance frequencies associated with different transverse modes can exist, as illustrated schematically in Fig. 11.4.

In a cavity that consists of an optical waveguide, the propagation constant $\beta_{mn}(\omega)$ is a function of the waveguide mode. If the physical length of the waveguide cavity is l, the effective round-trip optical path length of a waveguide mode is

$$l_{mn}^{RT} = \begin{cases} 2\dfrac{c}{\omega}\beta_{mn}(\omega)l, & \text{for a linear cavity,} \\ \dfrac{c}{\omega}\beta_{mn}(\omega)l, & \text{for a ring cavity.} \end{cases} \tag{11.21}$$

The round-trip optical path length, l_{mn}^{RT}, generally varies from one mode to another due to modal dispersion of the waveguide. In addition, the localized phase shift can also

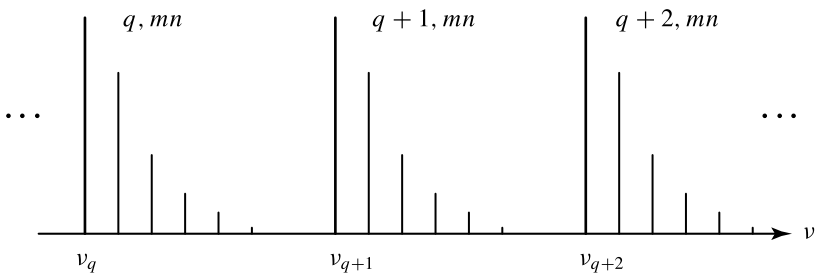


Figure 11.4 Cavity resonance frequencies associated with different longitudinal and transverse modes. For clarity, the heights of the transverse modes are made arbitrarily decreasing.

be mode dependent. Therefore, instead of ω_q given by (11.15) for a plane wave, the resonance frequencies ω_{mnq} for a waveguide cavity are the solutions of the following resonance condition:

$$\varphi_{mn}^{\mathrm{RT}} = \frac{\omega}{c} l_{mn}^{\mathrm{RT}} + \varphi_{mn}^{\mathrm{local}} = 2q\pi. \tag{11.22}$$

In a nonwaveguiding cavity, the propagation constant, k, is a property of the medium only and is not mode dependent. However, a mode-dependent on-axis phase variation $\zeta_{mn}(z)$ does exist, which is given in (1.140) for a Hermite–Gaussian mode as discussed in Section 1.7. The total on-axis phase variation for the TEM_{mn} Gaussian mode is $\varphi_{mn}(z) = kz + \zeta_{mn}(z)$, which includes the mode-independent phase shift kz and the mode-dependent phase shift $\zeta_{mn}(z)$. Consequently, the resonance condition for a Gaussian mode is a modification of that for a plane wave by adding the round-trip contribution of the mode-dependent phase shift:

$$\varphi_{mn}^{\mathrm{RT}} = \frac{\omega}{c} l_{\mathrm{RT}} + \zeta_{mn}^{\mathrm{RT}} + \varphi_{mn}^{\mathrm{local}} = 2q\pi, \tag{11.23}$$

where the localized phase shift can, in general, be mode dependent.

Cavity lifetime and quality factor

Here we consider some important parameters of a passive optical cavity with no optical gain. Such a passive optical cavity with $\chi_{\mathrm{res}} = 0$, thus $g = 0$, is also known as a *cold cavity*. To be specific, we identify the round-trip gain factor for the field amplitude in a cold cavity as G_{c}, or as G_{mn}^{c} for the transverse mode mn.

Because there is no optical gain in a cold cavity, $G_{\mathrm{c}} < 1$. Any optical field that initially exists in the cavity gradually decays as it circulates inside the cavity. Because the field amplitude is attenuated by a factor of G_{c} per round trip, the intensity and thus the number of intracavity photons are attenuated by a factor of G_{c}^2 per round trip. We can define a *photon lifetime*, also called *cavity lifetime*, τ_{c}, and a *cavity decay rate*, γ_{c}, for a cold cavity through the following relation:

$$G_{\mathrm{c}}^2 = \mathrm{e}^{-T/\tau_{\mathrm{c}}} = \mathrm{e}^{-\gamma_{\mathrm{c}} T}. \tag{11.24}$$

Therefore,

$$\tau_c = -\frac{T}{2\ln G_c}. \tag{11.25}$$

The cavity decay rate is the decay rate of the optical energy stored in a cavity and is given by

$$\gamma_c = \frac{1}{\tau_c} = -\frac{2}{T}\ln G_c. \tag{11.26}$$

In general, the value of G_c for a given cavity is mode dependent. Usually, the fundamental transverse mode has the lowest loss because its field distribution is most transversely concentrated toward the center of the cavity defined by the longitudinal axis. As the order of a mode increases, its loss in the cavity increases due to the increased diffraction loss associated with transverse spreading of its field distribution. Consequently, both τ_c and γ_c are also mode dependent: τ^c_{mnq} and γ^c_{mnq}. Unless a specific mode-discriminating mechanism is introduced in a cavity, either intentionally or unintentionally, the fundamental mode generally has the largest value of τ_c and the lowest value of γ_c.

The quality factor, Q, of a resonator is generally defined as the ratio of the resonance frequency and the energy damping rate of the resonator:

$$Q = \omega_{res}\left(\frac{\text{energy stored in the resonator}}{\text{average power dissipation}}\right) = \frac{\omega_{res}}{\gamma}, \tag{11.27}$$

where ω_{res} is the resonance frequency of the resonator and γ is the energy decay rate of the resonator. Therefore, the quality factor of a cold cavity is

$$Q = \frac{\omega_q}{\gamma_c} = \omega_q \tau_c, \tag{11.28}$$

where ω_q is the longitudinal mode frequency. For a low-loss, high-Q cavity, G_c is not much less than unity, and it can be easily shown by using (11.19) and (11.24) that

$$\Delta\nu_c \approx \frac{1}{2\pi\tau_c} = \frac{\gamma_c}{2\pi} \tag{11.29}$$

and

$$Q \approx \frac{\nu_q}{\Delta\nu_c}. \tag{11.30}$$

Note that though it is not explicitly spelled out in (11.28) and (11.30), the quality factor is a function of not only the longitudinal-mode index q but also the transverse-mode indices m and n: $Q = Q_{mnq}$. To be precise, (11.28) should be written as

$$Q_{mnq} = \frac{\omega_{mnq}}{\gamma^c_{mnq}} = \omega_{mnq}\tau^c_{mnq}. \tag{11.31}$$

For an optical cavity, the dependence of Q_{mnq} on the longitudinal-mode index q is generally negligible because q is a very large value except in the case of a very short

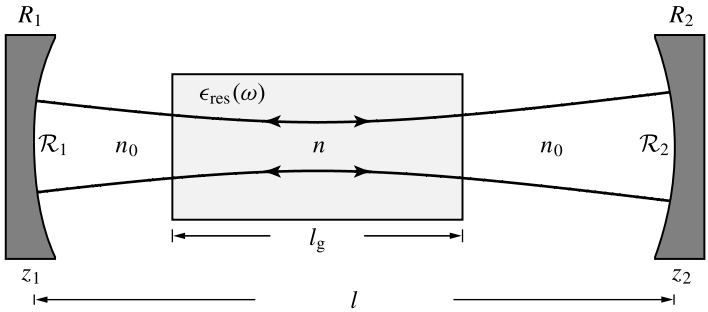


Figure 11.5 Fabry–Perot cavity containing an optical gain medium with a filling factor Γ. Changes of Gaussian beam divergence at the boundaries of the gain medium are ignored in this plot.

microcavity. However, the dependence of Q_{mnq} on the transverse-mode indices m and n cannot be ignored. Indeed, Q_{00q} for the fundamental transverse mode is generally larger than Q_{mnq} for any high-order transverse mode because the fundamental transverse mode generally has the lowest loss.

Fabry–Perot cavity

The most common laser cavity is a Fabry–Perot cavity consisting of two end mirrors and an optical gain medium, shown in Fig. 11.5. The radii of curvature of the left and right mirrors are $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. The sign of the radius of curvature is taken to be positive for a concave mirror and negative for a convex mirror. For the cavity shown in Fig. 11.5, which is formed with two concave mirrors, $\mathcal{R}_1 > 0$ and $\mathcal{R}_2 > 0$.

Most of the important features of a nonwaveguiding Fabry–Perot laser cavity can be obtained by applying the following simple concept: for the cavity to be a stable cavity in which a Gaussian mode can establish, the radii of curvature of both end mirrors have to match the wavefront curvatures of the Gaussian mode at the surfaces of the mirrors: $\mathcal{R}(z_1) = -\mathcal{R}_1$ and $\mathcal{R}(z_2) = \mathcal{R}_2$, where z_1 and z_2 are, respectively, the coordinates of the left and right mirrors measured from the location of the Gaussian beam waist. Based on this concept, we have from (1.136) the following two relations:

$$z_1 + \frac{z_{\rm R}^2}{z_1} = -\mathcal{R}_1 \quad \text{and} \quad z_2 + \frac{z_{\rm R}^2}{z_2} = \mathcal{R}_2. \tag{11.32}$$

From these relations, we find that (see Problem 11.1.1)

$$z_{\rm R}^2 = \frac{l(\mathcal{R}_1 - l)(\mathcal{R}_2 - l)(\mathcal{R}_1 + \mathcal{R}_2 - l)}{(\mathcal{R}_1 + \mathcal{R}_2 - 2l)^2}, \tag{11.33}$$

where $l = z_2 - z_1$ is the length of the cavity defined by the separation between the mirrors. Given the values of $\mathcal{R}_1$, $\mathcal{R}_2$, and l, stable Gaussian modes exist for the cavity if both relations in (11.32) can be satisfied with a positive, real parameter $z_{\rm R} > 0$ for a finite, positive spot size w_0 according to (1.134). Then the cavity is stable. If the relations in (11.32) cannot be simultaneously satisfied with a positive, real value for

z_R, then the cavity is unstable because no stable Gaussian mode can be established in the cavity. Application of this concept yields the following *stability criterion* for a Fabry–Perot cavity (see Problem 11.1.2):

$$0 \le \left(1 - \frac{l}{\mathcal{R}_1}\right)\left(1 - \frac{l}{\mathcal{R}_2}\right) \le 1. \tag{11.34}$$

In a stable resonator cavity, the mode-dependent on-axis phase shift in a single pass through the cavity from the left to the right mirror for the TEM_{mn} Hermite–Gaussian mode is simply $\zeta_{mn}(z_2) - \zeta_{mn}(z_1)$. Therefore, the round-trip mode-dependent on-axis phase shift is

$$\zeta_{mn}^{\text{RT}} = 2[\zeta_{mn}(z_2) - \zeta_{mn}(z_1)]. \tag{11.35}$$

With some modifications, the same concept can be used to find the characteristics and stability criterion of a cavity with multiple mirrors, such as a folded Fabry–Perot cavity or a ring cavity.

We consider a cavity that contains an isotropic gain medium with a filling factor Γ. The surfaces of the gain medium are antireflection coated so that there is no reflection inside the cavity other than the reflection at the two end mirrors. If the gain medium fills up the entire cavity, we simply make $\Gamma = 1$ in the results obtained below. The Fabry–Perot cavity has a physical length l between the two end mirrors. The field amplitude reflection coefficients are r_1 and r_2 for the left and right mirrors, respectively. They are generally complex to account for the phase changes on reflection, φ_1 and φ_2, respectively, and can be written as

$$r_1 = R_1^{1/2} e^{i\varphi_1}, \qquad r_2 = R_2^{1/2} e^{i\varphi_2}, \tag{11.36}$$

where R_1 and R_2 are the intensity reflectivities of the left and right mirrors, respectively.

The dielectric property of the intracavity gain medium contains the permittivity of the background material plus a contribution from the resonant susceptibility, $\chi_{\text{res}}(\omega)$, that characterizes the laser transition. To identify the effect of each contribution clearly, it is instructive to express the permittivity of the gain medium explicitly, including the contribution of the resonant laser transition, as

$$\epsilon_{\text{res}}(\omega) = \epsilon(\omega) + \epsilon_0 \chi_{\text{res}}(\omega), \tag{11.37}$$

where $\epsilon(\omega) = n^2$ is the background permittivity of the gain medium excluding the contribution of the resonant laser transition. In a cold cavity, $\chi_{\text{res}} = 0$. Therefore, the weighted average of the propagation constant for the intracavity field in a cold cavity is

$$\overline{k} = \frac{\overline{n}\omega}{c} = \Gamma k + (1 - \Gamma)k_0, \tag{11.38}$$

where $k = n\omega/c$ is the propagation constant in the gain medium and $k_0 = n_0\omega/c$ is that in the surrounding background medium. The round-trip optical path length in this cavity is $l_{\text{RT}} = 2\overline{n}l$.

Usually there is an intracavity background loss contributed by a variety of different mechanisms, such as scattering or absorption, that are irrelevant to the laser transition. In addition, mode-dependent diffraction losses exist for the intracavity optical field due to the finite sizes of the end mirrors. The combined effect of these losses can be accounted for by taking a spatially averaged, mode-dependent loss coefficient, $\overline{\alpha}_{mn}$, so that the effective propagation constant is complex with a mode-dependent imaginary part: $\overline{k} + \mathrm{i}\overline{\alpha}_{mn}/2$. This loss is known as the *distributed loss* of the laser cavity mode. In general, $\overline{\alpha} \ll \overline{k}$ for any practical gain medium.

By following a mode field over one round trip in the cavity, we find that

$$a = r_1 r_2 \exp\left(\mathrm{i}2\overline{k}l - \overline{\alpha}_{mn}l + \mathrm{i}\zeta_{mn}^{\mathrm{RT}}\right) \tag{11.39}$$

for the TEM_{mn} Hermite–Gaussian mode. Therefore, both the round-trip gain factor and the round-trip phase shift are mode dependent:

$$G_{mn}^{\mathrm{c}} = R_1^{1/2} R_2^{1/2} \exp\left(-\overline{\alpha}_{mn}l\right) \tag{11.40}$$

and

$$\varphi_{mn}^{\mathrm{RT}} = 2\overline{k}l + \zeta_{mn}^{\mathrm{RT}} + \varphi_1 + \varphi_2. \tag{11.41}$$

Using (11.41) for the resonance condition in (11.20), we find the following resonance frequencies of the cold Fabry–Perot cavity:

$$\omega_{mnq}^{\mathrm{c}} = \frac{c}{2\overline{n}l}(2q\pi - \zeta_{mn}^{\mathrm{RT}} - \varphi_1 - \varphi_2) \tag{11.42}$$

and $\nu_{mnq}^{\mathrm{c}} = \omega_{mnq}^{\mathrm{c}}/2\pi$, where the superscript c indicates the fact that the frequencies are associated with a *cold* cavity with $\chi_{\mathrm{res}} = 0$. These frequencies are clearly functions of the transverse-mode indices because of the mode-dependent phase shift ζ_{mn}^{RT}. However, because ζ_{mn}^{RT} is not a function of the longitudinal-mode index q, the frequency separation between different longitudinal modes of the same transverse mode group is a constant:

$$\Delta\nu_{\mathrm{L}} = \nu_{mn,q+1}^{\mathrm{c}} - \nu_{mnq}^{\mathrm{c}} = \frac{c}{2\overline{n}l} = \frac{1}{T}. \tag{11.43}$$

Here we assume that the background optical property of the medium is not very dispersive so that the background refractive index $\overline{n}$ can be considered a constant independent of optical frequency in the narrow range between neighboring modes of interest. Using (11.13) and (11.40), the finesse of the lossy cavity is

$$F = \frac{\pi R_1^{1/4} R_2^{1/4} \mathrm{e}^{-\overline{\alpha}_{mn}l/2}}{1 - R_1^{1/2} R_2^{1/2} \mathrm{e}^{-\overline{\alpha}_{mn}l}}, \tag{11.44}$$

which is mode dependent due to the mode-dependent loss $\overline{\alpha}_{mn}$. The longitudinal mode width, $\Delta\nu_{\mathrm{c}} = \Delta\nu_{\mathrm{L}}/F$, is also mode dependent for the same reason. For a cavity with a negligible loss, we can take $\overline{\alpha}_{mn} = 0$; then, the expression in (11.44) reduces to the

familiar formula for the finesse of a lossless Fabry–Perot interferometer:

$$F = \frac{\pi R_1^{1/4} R_2^{1/4}}{1 - R_1^{1/2} R_2^{1/2}}. \tag{11.45}$$

Therefore, for a nondispersive, lossless Fabry–Perot cavity, $\Delta\nu_L$, F, and $\Delta\nu_c$ are also independent of longitudinal and transverse mode indices though the mode frequency ν_{mnq} is a function of all three mode indices.

Using (11.25) and (11.40), the mode-dependent photon lifetime of the Fabry–Perot cavity can be expressed as

$$\tau_{mnq}^c = \frac{\overline{n}l}{c(\overline{\alpha}_{mn} l - \ln\sqrt{R_1 R_2})}, \tag{11.46}$$

and the mode-dependent cavity decay rate as

$$\gamma_{mnq}^c = \frac{c}{\overline{n}}\left(\overline{\alpha}_{mn} - \frac{1}{l}\ln\sqrt{R_1 R_2}\right). \tag{11.47}$$

Clearly, both τ_{mnq}^c and γ_{mnq}^c are also mode dependent due to the mode-dependent distributed loss $\overline{\alpha}_{mn}$. However, they are independent of the longitudinal mode index q under the assumption that the background refractive index $\overline{n}$ and the loss $\overline{\alpha}_{mn}$, as well as the mirror reflectivities R_1 and R_2, are not sensitive to the frequency differences among different longitudinal modes. If any of these parameters vary significantly within the range of longitudinal modes of interest, then the dependence of τ_{mnq}^c and γ_{mnq}^c on the index q cannot be ignored.

The Fabry–Perot cavity for a typical laser is a high-Q cavity. Even in a high-gain laser with low mirror reflectivities, Q is still very large. For example, consider a high-gain InGaAsP/InP semiconductor laser emitting at 1.3 μm wavelength with $n = 3.5$, $l = 300$ μm, and $R_1 = R_2 = 0.3$. Assuming a negligibly small $\overline{\alpha}$ for simplicity, we find that $\tau_c = 2.9$ ps, $T = 7$ ps, $\Delta\nu_L = 142.86$ GHz, $F = 2.46$, and $\Delta\nu_c = 58$ GHz. Using (11.28), we obtain $Q = 4.2 \times 10^3$, while the approximate relation (11.30) yields a slightly smaller value of $Q = 4.0 \times 10^3$. A Q value on the order of 10^3 is relatively low for a laser cavity. Even so, the difference between (11.30) and (11.28) is only about 5%. For a low-loss cavity, Q can easily be as high as 10^8, and the result obtained from (11.30) is essentially the same as that from (11.28). See Example 11.1 below for a very different laser cavity.

EXAMPLE 11.1 A Nd : YAG microchip laser, shown in Fig. 11.6, is made of a Nd : YAG crystal of the same properties as that of the Nd : YAG laser amplifier described in Example 10.9, except that it is thinner and its surfaces are coated differently to form a laser cavity. It consists of parallel Nd : YAG plates of 500 μm thickness. The surfaces of the plate are coated for $R_1 = 100\%$ and $R_2 = 99.7\%$ at the 1.064 μm laser wavelength to form the laser cavity but for $R_1 = R_2 = 0$ at the 808 nm pump wavelength to allow only a single pass of the pump beam. The refractive index of Nd : YAG is $n = 1.82$.

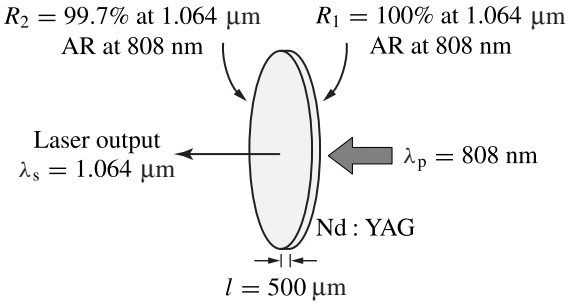


Figure 11.6 Schematics of a fiber-coupled, end-pumped Nd : YAG microchip laser. AR means antireflection.

The distributed loss of the laser cavity is found to be $\overline{\alpha} = 0.5\ \text{m}^{-1}$. (a) Find the round-trip optical path length, the round-trip time, and the longitudinal mode spacing of this cavity. (b) Find the finesse of this cavity. (c) What are the cavity decay rate and the photon lifetime? (d) What are the longitudinal mode width and the Q factor of the cold cavity?

Solution (a) This is a Fabry–Perot laser cavity with a filling factor of $\Gamma = 1$. Therefore, $\overline{n} = n = 1.82$. Then the round-trip optical path length is

$$l_{\text{RT}} = 2nl = 2 \times 1.82 \times 500\ \mu\text{m} = 1.82\ \text{mm}.$$

The round-trip time is

$$T = \frac{l_{\text{RT}}}{c} = \frac{1.82 \times 10^{-3}}{3 \times 10^8}\ \text{s} = 6.07\ \text{ps}.$$

The longitudinal mode spacing is

$$\Delta\nu_{\text{L}} = \frac{1}{T} = \frac{1}{6.07 \times 10^{-12}}\ \text{Hz} = 164.8\ \text{GHz}.$$

(b) The finesse of this cavity has to be found by using (11.44), not (11.45), because there is a distributed loss of $\overline{\alpha} = 0.5\ \text{m}^{-1}$. For $l = 500\ \mu\text{m}$, $\overline{\alpha}l = 2.5 \times 10^{-4}$. With $R_1 = 100\%$ and $R_2 = 99.7\%$, we find that the finesse is

$$F = \frac{\pi R_1^{1/4} R_2^{1/4} \text{e}^{-\overline{\alpha}l/2}}{1 - R_1^{1/2} R_2^{1/2} \text{e}^{-\overline{\alpha}l}} = \frac{\pi \times (0.997)^{1/4} \times \exp(-2.5 \times 10^{-4}/2)}{1 - (0.997)^{1/2} \times \exp(-2.5 \times 10^{-4})} = 1793.$$

(c) The cavity decay rate can be calculated using (11.47):

$$\gamma_{\text{c}} = \frac{c}{\overline{n}} \left(\overline{\alpha} - \frac{1}{l} \ln \sqrt{R_1 R_2} \right) = \frac{3 \times 10^8}{1.82} \times \left(0.5 - \frac{\ln \sqrt{0.997}}{500 \times 10^{-6}} \right)\ \text{s}^{-1} = 5.78 \times 10^8\ \text{s}^{-1}.$$

Then, the photon lifetime is

$$\tau_{\text{c}} = \frac{1}{\gamma_{\text{c}}} = \frac{1}{5.78 \times 10^8}\ \text{s} = 1.73\ \text{ns}.$$

(d) Using the definition of the finesse in (11.17), or (11.19), we find the following longitudinal mode width for the cold cavity:

$$\Delta\nu_c = \frac{\Delta\nu_L}{F} = \frac{164.8 \times 10^9}{1793}\ \text{Hz} = 91.91\ \text{MHz}.$$

Using the approximate relation given in (11.29), we find that $\Delta\nu_c \approx \gamma_c/2\pi = 91.99$ MHz, which is almost the same as the accurate value obtained above. Using the definition given in (11.28), the Q factor of the cold cavity is found to be

$$Q = \frac{\omega}{\gamma_c} = \frac{2\pi c}{\lambda\gamma_c} = \frac{2\pi \times 3 \times 10^8}{1.064 \times 10^{-6} \times 5.78 \times 10^8} = 3.065 \times 10^6.$$

Though the cavity is very short, it is a high-Q cavity because of the very high reflectivities of both of its end mirrors. If we use the approximate relation in (11.30), we find $Q = 3.068 \times 10^6$, which is only slightly different from that found by using (11.28). For a high-Q cavity, there is little difference between (11.30) and (11.28) for the Q value and between (11.29) and (11.17) for the mode width.

11.2 Laser oscillation

In the preceding section, it is mentioned that a practical laser device can be constructed by placing an optical gain medium inside an optical resonator. The gain medium provides amplification to the intracavity optical field while the resonator provides optical feedback. A laser is basically a coherent optical oscillator, and the basic function of an oscillator is to generate a coherent signal through resonant oscillation without an input signal. Therefore, no external optical field is injected into the optical cavity for laser oscillation. The intracavity optical field has to grow from the field generated by spontaneous emission from the intracavity gain medium. When steady-state oscillation is reached, the coherent laser field at any given location inside the cavity should become a constant with time in both phase and magnitude. In the model shown in Fig. 11.2, the situation of steady-state laser oscillation requires that $\mathbf{E}_{in} = 0$ and $\mathbf{E}_c = \text{constant} \neq 0$. Therefore, from (11.6), the *condition for steady-state laser oscillation* is

$$a = G\exp\left(i\varphi_{RT}\right) = 1. \tag{11.48}$$

To illustrate the implications of this condition, we consider in the following the simple Fabry–Perot laser shown in Fig. 11.5 that contains an isotropic gain medium with a filling factor Γ.

The total permittivity of the gain medium, including the contribution of the resonant laser transition, is $\epsilon_{res} = \epsilon + \epsilon_0\chi_{res}$ given in (11.37). Therefore, the total complex propagation constant of the gain medium including the contribution of the resonant

transition is

$$k_t = \omega \mu_0^{1/2} (\epsilon + \epsilon_0 \chi_{res})^{1/2}$$
$$= k + \Delta k_{res} - i\frac{g}{2}, \tag{11.49}$$

where

$$\Delta k_{res} \approx k \frac{\chi'_{res}}{2n^2} = \frac{\omega}{2nc} \chi'_{res}, \tag{11.50}$$

$$g \approx -k \frac{\chi''_{res}}{n^2} = -\frac{\omega}{nc} \chi''_{res}. \tag{11.51}$$

Here g is the gain coefficient of the laser medium associated with the laser transition identified in (10.55), and Δk_{res} is the corresponding change in the optical wavenumber in the medium caused by the change in the refractive index associated with population changes in the resonant laser levels. As discussed in Section 10.2, when population inversion is achieved, $\chi''_{res} < 0$ so that the gain coefficient g has a positive value.

By replacing k for a cold medium with k_t for a pumped gain medium, we find that $\overline{k}$ in (11.39) has to be replaced with $\overline{k} + \Gamma \Delta k_{res} - i\Gamma g/2$ when an actively pumped laser cavity is considered. We then find the mode-dependent round-trip gain factor

$$G_{mn} = R_1^{1/2} R_2^{1/2} \exp\left[(\Gamma g - \overline{\alpha}_{mn})\, l\right] \tag{11.52}$$

and mode-dependent round-trip phase shift

$$\varphi_{mn}^{RT} = 2(\overline{k} + \Gamma \Delta k_{res}) l + \zeta_{mn}^{RT} + \varphi_1 + \varphi_2. \tag{11.53}$$

Because both G_{mn} and φ_{mn}^{RT} are real parameters, the condition in (11.48) can be satisfied for a given laser mode to oscillate only if the *gain condition*

$$G_{mn} = 1 \tag{11.54}$$

and the *phase condition*

$$\varphi_{mn}^{RT} = 2q\pi, \qquad q = 1, 2, \ldots, \tag{11.55}$$

are simultaneously fulfilled. Note that both G_{mn} and φ_{mn}^{RT} are frequency dependent.

Laser threshold

The condition in (11.54) implies that there are a threshold gain and a corresponding threshold pumping level for laser oscillation. For the Fabry–Perot laser shown in Fig. 11.5, which has a length l and contains a gain medium of a length l_g for a filling factor of $\Gamma = l_g/l$, the *threshold gain coefficient*, g_{mn}^{th}, for the TEM_{mn} mode can be

found from

$$\Gamma g_{mn}^{\text{th}} = \overline{\alpha}_{mn} - \frac{1}{l} \ln \sqrt{R_1 R_2}, \tag{11.56}$$

or

$$g_{mn}^{\text{th}} l_{\text{g}} = \overline{\alpha}_{mn} l - \ln \sqrt{R_1 R_2}. \tag{11.57}$$

Because the distributed loss $\overline{\alpha}_{mn}$ is mode dependent, the threshold gain coefficient g_{mn}^{th} varies from one transverse mode to another. In addition, the effective gain coefficient can be different for different transverse modes because different transverse modes have different field distribution patterns and thus overlap with the gain volume differently. The transverse mode that has the lowest loss and the largest effective gain at any given pumping level reaches threshold first and starts oscillating at the lowest pumping level. In a typical laser, the transverse mode that reaches threshold first is normally the fundamental mode.

Unless a frequency-selecting mechanism is placed in a laser to create a frequency-dependent loss that varies from one longitudinal mode to another, the threshold gain coefficient g_{mn}^{th} does not vary much among the mnq longitudinal modes that share the common mn transverse mode pattern. It is possible, however, to introduce a frequency-selecting device to a laser cavity so that $\overline{\alpha}_{mn}$ and, consequently, g_{mn}^{th} become highly frequency dependent for the purpose of frequency selection or frequency tuning of the laser output.

The power required to pump a laser to reach its threshold is called the *threshold pump power*, P_{p}^{th}. Because the threshold gain coefficient is mode dependent and frequency dependent, the threshold pump power is also mode dependent and frequency dependent. The threshold pump power of a laser mode can be found by calculating the power required for the gain medium to have an unsaturated gain coefficient equal to the threshold gain coefficient of the mode: $g_0 = g_{mn}^{\text{th}}$, assuming uniform pumping throughout the medium. For a quasi-two-level or three-level laser, there is also a *transparency pump power*, P_{p}^{tr}, corresponding to $g_0 = 0$, assuming uniform pumping throughout the gain medium. In the situation of nonuniform pumping, these conditions for reaching threshold and transparency have to be modified, as discussed below. Clearly, $P_{\text{p}}^{\text{tr}} < P_{\text{p}}^{\text{th}}$ by definition.

In a nonwaveguiding laser, the transverse cross section of the gain medium is normally larger than the cross-sectional area of a laser mode. In this situation, it is not necessary to pump the entire gain medium, but only the volume of the gain medium seen by the laser mode. Calculation of P_{p}^{th} depends on the specifics of the pump source and the pumping geometry. Nevertheless, if we consider a *saturation pump power*, $P_{\text{p}}^{\text{sat}}$, as the pump power required for the pumping rate to be $W_{\text{p}} = 1/\tau_2$ following the same concept of $I_{\text{p}}^{\text{sat}}$ as defined in (10.85), we can find P_{p}^{th} in terms of $P_{\text{p}}^{\text{sat}}$. Because of absorption of the pump power by the gain medium, the pump power distribution in the pumped volume of a gain medium is often spatially nonuniform. The distribution of the pump

power in a laser medium cannot be easily generalized because it is a function of many parameters specific to a particular pump source, a particular pumping geometry, and a given gain medium.

A case of common interest for solid-state lasers, however, is the longitudinal optical pumping considered in Section 10.4 for laser amplifiers. In this situation, the laser threshold is reached when

$$\int_0^{l_g} g_0(z)\mathrm{d}z = g_{th} l_g. \tag{11.58}$$

For single-pass optical pumping, as considered in Section 10.4, if transverse divergence of the pump beam is negligible, the integral of the unsaturated gain coefficient over a gain medium of a length l has the closed-formed solutions given in (10.106) and (10.108). By taking (10.108) with $l = l_g$ and using the condition in (11.58) for the laser threshold, we find that the threshold pump power of the laser can be expressed in terms of the pump power utilization factor ζ_p, which is defined in (10.107), as

$$P_p^{th} = \begin{cases} \dfrac{1}{p} \dfrac{\exp\left[p \dfrac{\sigma_a N_t + g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right] - 1}{1 - (1 - \zeta_p^{th}) \exp\left[p \dfrac{\sigma_a N_t + g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right]} P_p^{sat}, & \text{for } p \neq 0, \\ \dfrac{\sigma_a N_t + g_{th}}{(\sigma_e + \sigma_a)N_t} \dfrac{\alpha_p l_g}{\zeta_p^{th}} P_p^{sat}, & \text{for } p = 0, \end{cases} \tag{11.59}$$

where ζ_p^{th} is the pump power utilization factor at the laser threshold. It can be found by applying (10.106) with $l = l_g$ to the condition in (11.58) that

$$\zeta_p^{th} = 1 - \exp\left[-\frac{\sigma_e N_t - g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right]. \tag{11.60}$$

By plugging this relation for ζ_p^{th} into (11.59), the threshold pump power can be explicitly expressed in terms of the laser parameters as (see Problem 11.2.1)

$$P_p^{th} = \begin{cases} \dfrac{1}{p} \dfrac{\exp\left[p \dfrac{\sigma_a N_t + g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right] - 1}{1 - \exp\left[-\dfrac{(\sigma_e - p\sigma_a)N_t - (1 + p)g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right]} P_p^{sat}, & \text{for } p \neq 0, \\ \dfrac{\sigma_a N_t + g_{th}}{(\sigma_e + \sigma_a)N_t} \dfrac{\alpha_p l_g}{1 - \exp\left[-\dfrac{\sigma_e N_t - g_{th}}{(\sigma_e + \sigma_a)N_t} \alpha_p l_g\right]} P_p^{sat}, & \text{for } p = 0. \end{cases} \tag{11.61}$$

The transparency condition is such that the integral of the unsaturated gain coefficient over the length of the gain medium is zero. Therefore, by replacing g_{th} with 0 in (11.61),

the transparency pump power of a laser gain medium can be found:

$$P_{\mathrm{p}}^{\mathrm{tr}} = \begin{cases} \dfrac{1}{p}\dfrac{\exp\{[p\sigma_{\mathrm{a}}/(\sigma_{\mathrm{e}}+\sigma_{\mathrm{a}})]\alpha_{\mathrm{p}}l_{\mathrm{g}}\}-1}{1-\exp\{-[(\sigma_{\mathrm{e}}-p\sigma_{\mathrm{a}})/(\sigma_{\mathrm{e}}+\sigma_{\mathrm{a}})]\alpha_{\mathrm{p}}l_{\mathrm{g}}\}}P_{\mathrm{p}}^{\mathrm{sat}}, & \text{for } p \neq 0, \\ \dfrac{\sigma_{\mathrm{a}}}{(\sigma_{\mathrm{e}}+\sigma_{\mathrm{a}})}\dfrac{\alpha_{\mathrm{p}}l_{\mathrm{g}}}{1-\exp\{-[\sigma_{\mathrm{e}}/(\sigma_{\mathrm{e}}+\sigma_{\mathrm{a}})]\alpha_{\mathrm{p}}l_{\mathrm{g}}\}}P_{\mathrm{p}}^{\mathrm{sat}}, & \text{for } p = 0. \end{cases} \tag{11.62}$$

The relation obtained in (11.61) for the threshold pump power $P_{\mathrm{p}}^{\mathrm{th}}$ and that obtained in (11.62) for the transparency pump power $P_{\mathrm{p}}^{\mathrm{tr}}$ are generally valid at all pumping levels for single-pass longitudinal optical pumping. They are applicable no matter whether there is significant absorption saturation of the pump power in the gain medium or not. They are not valid for multiple-pass longitudinal optical pumping, however. The absorption saturation of the pump power is negligible under the condition that $s_{\mathrm{th}} = P_{\mathrm{p}}^{\mathrm{th}}/P_{\mathrm{p}}^{\mathrm{sat}} \ll 1$. In this situation, the pump power decays exponentially along the longitudinal pumping axis in each pass through the gain medium. Then, a closed-form solution of $P_{\mathrm{p}}^{\mathrm{th}}$ that has a common form for both single-pass and multiple-pass longitudinal pumping arrangements can be found in terms of the pump power utilization factor $\zeta_{\mathrm{p}}^{\mathrm{th}}$. In a single-pass arrangement under the condition that $s_{\mathrm{th}} \ll 1$, $\zeta_{\mathrm{p}}^{\mathrm{th}} \approx 1 - \mathrm{e}^{-\alpha_{\mathrm{p}}l_{\mathrm{g}}}$, assuming no reflection of the pump beam at the pump input surface of the gain medium. In a multiple-pass situation, however, $\zeta_{\mathrm{p}}^{\mathrm{th}}$ has to be properly evaluated to account for the total pump power absorbed by the gain medium in all passes (see Problem 11.2.2).

EXAMPLE 11.2 The Nd : YAG microchip laser considered in Example 11.1 is pumped through a multimode fiber of 200 μm core diameter in the same manner as that for the amplifier described in Example 10.9. Both the pump and the laser spots have circular TEM_{00} mode profiles of 100 μm radius. Relevant parameters, from Example 10.9, are $N_{\mathrm{t}} = 1.52 \times 10^{26}\ \mathrm{m}^{-3}$ for the Nd concentration, $\sigma_{\mathrm{a}}^{\mathrm{p}} = 3.0 \times 10^{-24}\ \mathrm{m}^2$ with a pump quantum efficiency of $\eta_{\mathrm{p}} = 80\%$ for the pump at 808 nm, $\sigma_{\mathrm{e}} = 3.1 \times 10^{-23}\ \mathrm{m}^2$ with $\tau_2 = 240\ \mu\mathrm{s}$ for the laser transition at 1.064 μm. (a) Find the threshold gain coefficient for the laser. (b) Find the threshold pump power of the laser.

Solution (a) This laser has a filling factor of $\Gamma = 1$. It also has $\overline{\alpha} = 0.5\ \mathrm{m}^{-1}$, $l = 500\ \mu\mathrm{m}$, $R_1 = 100\%$, and $R_2 = 99.7\%$, as given in Example 11.1. Therefore, according to (11.56), the threshold gain coefficient is

$$g_{\mathrm{th}} = \frac{1}{\Gamma}\left(\overline{\alpha} - \frac{1}{l}\ln\sqrt{R_1R_2}\right) = \frac{1}{1}\left(0.5 - \frac{1}{500\times10^{-6}}\ln\sqrt{0.997}\right)\ \mathrm{m}^{-1} = 3.5\ \mathrm{m}^{-1}.$$

(b) The laser is a four-level system with $\sigma_{\mathrm{a}} = 0$ and $p = 0$. The threshold pump power can be found directly by using (11.61) for $p = 0$. However, it is instructive to find $\zeta_{\mathrm{p}}^{\mathrm{th}}$ through (11.60) first and then use (11.59) to find $P_{\mathrm{p}}^{\mathrm{th}}$. Because $\alpha_{\mathrm{p}} = \sigma_{\mathrm{a}}^{\mathrm{p}}N_{\mathrm{t}} = 456\ \mathrm{m}^{-1}$

and $l_g = 500$ μm, we find that $\alpha_p l_g = 0.228$. Therefore, for this single-pass pumping arrangement, we have

$$\zeta_p^{th} = 1 - \exp\left(-\frac{3.1 \times 10^{-23} \times 1.52 \times 10^{26} - 3.5}{3.1 \times 10^{-23} \times 1.52 \times 10^{26}} \times 0.228\right) = 0.2037.$$

Because the pump spot size and all other pump parameters are the same as those used for the amplifier in Example 10.9, we find from Example 10.9 that $P_p^{sat} = 13.4$ W. Then, using (11.59) for $p = 0$, we find the following threshold pump power:

$$P_p^{th} = \frac{3.5}{3.1 \times 10^{-23} \times 1.52 \times 10^{26}} \times \frac{0.228}{0.2037} \times 13.4 \text{ W} = 11.1 \text{ mW}.$$

By comparing $\zeta_p^{th} = 0.2037$ found above to $1 - e^{-\alpha_p l_g} = 1 - e^{-0.228} = 0.2039$, we find that the pump power decays along the longitudinal axis of the gain medium almost exponentially. This characteristic indicates that there is almost no absorption saturation of the pump power, which can be understood from the fact that $g_{th}/\sigma_e N_t = 7.4 \times 10^{-4} \ll 1$ and $P_p^{th}/P_p^{sat} = 8.3 \times 10^{-4} \ll 1$. The pump power is not fully utilized in this single-pass pumping arrangement because only 20.4% of the input pump power is absorbed by the gain medium. The laser threshold can be lowered by taking a multiple-pass arrangement (see Problem 11.2.3) or by properly increasing the length of the gain medium (see Problem 11.2.4) to increase the utilization fraction of the pump power.

Mode pulling

Comparing (11.53) for an active Fabry–Perot laser with (11.41) for its cold cavity, we find that the round-trip phase shift for a field in a laser cavity is a function of χ'_{res} through its dependence on Δk_{res}. Consequently, the longitudinal mode frequencies ω_{mnq} at which a laser oscillates are not exactly the same as the longitudinal mode frequencies ω^c_{mnq} given in (11.42) for the cold Fabry–Perot cavity.

Using (11.53) and (11.55), we find that the longitudinal mode frequencies of a Fabry–Perot laser are related to those of its cold cavity by

$$\omega_{mnq} = \omega^c_{mnq}\left(1 + \frac{\chi'_{res}}{2n\bar{n}}\right)^{-1} \approx \omega^c_{mnq}\left(1 - \frac{\chi'_{res}}{2n\bar{n}}\right). \tag{11.63}$$

Clearly, the mode frequencies ω_{mnq} at which a laser oscillates differ from the cold cavity mode frequencies because they vary with the resonant susceptibility, which depends on the level of population inversion in the gain medium. This dependence of the oscillating laser mode frequencies on the population inversion in the gain medium is caused by the fact that the refractive index and the gain of the medium are intimately connected to each other, as is dictated by the Kramers–Kronig relation. This effect causes a frequency shift of

$$\delta\omega_{mnq} = \omega_{mnq} - \omega^c_{mnq} \approx -\frac{\chi'_{res}}{2n\bar{n}}\omega^c_{mnq} \tag{11.64}$$

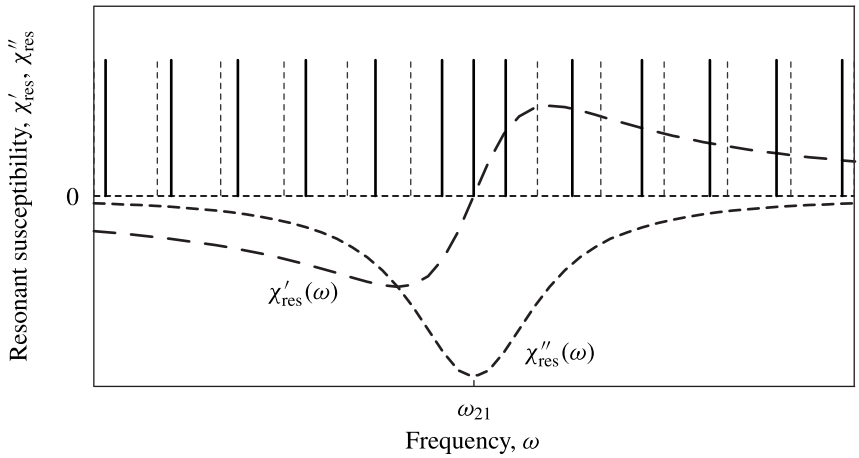


Figure 11.7 Frequency-pulling effect for laser modes. Compared to the resonance frequencies of the cold cavity shown in dotted lines, the mode frequencies of an active laser shown in solid lines are pulled toward the resonant transition frequency of the gain medium in the situation of population inversion. The real and imaginary parts of the gain susceptibility as a function of optical frequency are shown.

for the oscillation frequency of mode mnq. Because of the frequency dependence of χ'_{res}, the dependence of this frequency shift on χ'_{res} results in the *mode-pulling* effect demonstrated in Fig. 11.7. Near the resonant transition frequency, ω_{21}, of the gain medium, χ_{res} is highly dispersive.

In the presence of population inversion, $\chi''_{res}(\omega) < 0$ for either $\omega < \omega_{21}$ or $\omega > \omega_{21}$, but $\chi'_{res}(\omega) < 0$ for $\omega < \omega_{21}$ and $\chi'_{res}(\omega) > 0$ for $\omega > \omega_{21}$. As a result, $\omega_{mnq} > \omega^c_{mnq}$ for $\omega^c_{mnq} < \omega_{21}$ and $\omega_{mnq} < \omega^c_{mnq}$ for $\omega^c_{mnq} > \omega_{21}$. Therefore, in comparison to the resonance frequencies of the cold cavity, the oscillating mode frequencies of a laser are pulled toward the transition frequency of the gain medium. In addition, the longitudinal modes belonging to a common transverse mode are no longer equally spaced in frequency. In a laser of relatively high gain and large dispersion, such as in a semiconductor laser, this can result in a large variation in the frequency spacing among the oscillating modes.

Because of the frequency dependence of the gain coefficient g due to the frequency dependence of χ''_{res}, different longitudinal modes not only experience different values of refractive index but also see different values of gain coefficient, as also illustrated in Fig. 11.7. A longitudinal mode whose frequency is close to the gain peak at the transition resonance frequency has a higher gain than one whose frequency is far away from the gain peak.

Oscillating laser modes

Because of the frequency dependence of the gain coefficient, the net gain, $g - g^{th}_{mn}$, of a laser is always frequency dependent and varies among different transverse modes and

among different longitudinal modes no matter whether the threshold gain coefficient g_{mn}^{th} for any given transverse mode is frequency dependent or not. At a low pumping level before the laser starts oscillating, the net gain is negative for all laser modes. As the pumping level increases, the mode that reaches its threshold first will start oscillating. Once a laser starts oscillating in one mode, whether any other longitudinal or transverse modes have the opportunity to oscillate through further increase of the pumping level is a complicated issue of mode interaction and competition that depends on a variety of parameters, including the properties of the gain medium, the structure of the laser, the pumping geometry, the optical nonlinearity in the system, and the operating condition of the laser. Here we only discuss some basic concepts in the situation of steady-state oscillation of a CW laser. Interaction and competition among laser modes are more complicated when a laser is pulsed than when it is in CW operation. Therefore, some of the conclusions obtained here may not be valid for a pulsed laser.

The gain condition in (11.54) implies that once a given laser mode is oscillating in steady state, the gain that is available to this mode does not increase with increased pumping above the threshold pumping level because G_{mn} for a laser mode has to be kept at unity for steady-state oscillation. Thus the effective gain coefficient for an oscillating mode is "clamped" at the threshold level of the mode so long as the pumping level is kept at or above threshold. The mechanism for holding down the gain coefficient at the threshold level is the effect of gain saturation discussed in Section 10.3. An increase in the pumping level above threshold only increases the field intensity for the oscillating mode in the cavity, but the gain coefficient is saturated at the threshold value by the high intensity of the intracavity laser field. The fact that the gain of a laser mode oscillating in the steady state is saturated at the threshold value has a significant effect on the mode characteristics of a CW laser.

When the gain medium of a laser is homogeneously broadened, all modes that occupy the same spatial gain region compete for the gain from the population inversion in the same group of active atoms. When the mode that first reaches threshold starts oscillating, the entire gain curve supported by this group of atoms saturates. Because this oscillating mode is normally the one that has a longitudinal mode frequency closest to the gain peak and a transverse mode pattern with the lowest loss, the gain curve is saturated in such a manner that its value at this longitudinal mode frequency is clamped at the threshold value of the transverse mode that has the lowest threshold gain coefficient among all transverse modes. If the gain peak does not happen to coincide with this mode frequency, it still lies above the threshold when the gain curve is saturated, as shown in Fig. 11.8. Nevertheless, all other longitudinal modes belonging to this transverse mode have frequencies away from the gain peak. Therefore, even with increased pumping, they do not have sufficient gain to reach threshold because the entire gain curve shared by these modes is saturated, as illustrated in Fig. 11.8. Other transverse modes that are supported solely by this group of saturated, homogeneously broadened atoms do not have the opportunity to oscillate, either because the gain curve is saturated below

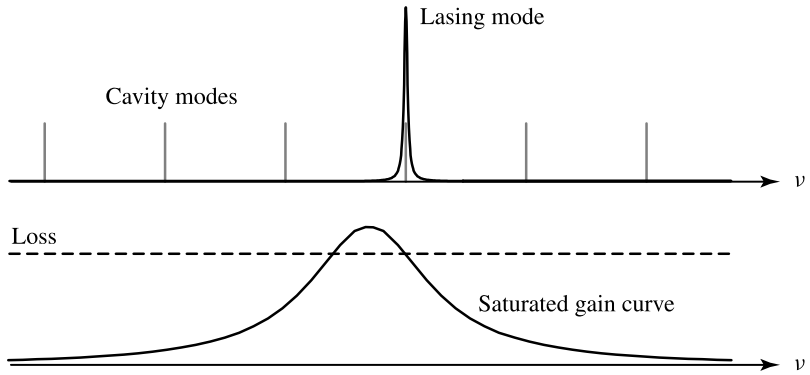


Figure 11.8 Gain saturation in a laser in the case of homogeneous broadening. Only one longitudinal mode whose frequency is closest to the gain peak oscillates. The entire gain curve is saturated such that the gain at the single lasing frequency remains at the loss level.

their threshold levels. Nevertheless, as different transverse modes have different spatial field distributions, a high-order transverse mode may draw its gain from a gain region outside of the region saturated by a low-order mode. Therefore, when the pumping level is increased, a high-order transverse mode may still reach its relatively high threshold for oscillation after a low-order transverse mode of a low threshold already oscillates. Consequently, for a homogeneously broadened CW laser in steady-state oscillation, only one among all of the longitudinal modes belonging to the same transverse mode will oscillate, but it is possible for more than one transverse mode to oscillate simultaneously at a high pumping level. Note that this conclusion does not hold true for a pulsed laser. It is possible for multiple longitudinal modes all belonging to the same transverse mode to oscillate simultaneously in a pulsed laser even when the gain medium is homogeneously broadened.

In a laser containing an inhomogeneously broadened gain medium, there are different groups of active atoms in the same spatial region. Each group saturates independently. Two modes occupying the same spatial gain region do not compete for the same group of atoms if the separation of their frequencies is larger than the homogeneous linewidth of each group of atoms. When one longitudinal mode reaches threshold and oscillates, only the gain coefficient around its frequency is saturated, the gain coefficient at other frequencies continues to increase with increased pumping. As the pumping level increases, other longitudinal modes will reach threshold and oscillate successively. As a result, at a sufficiently high pumping level, multiple longitudinal modes belonging to the same transverse mode can oscillate simultaneously. The saturation of the gain coefficient around each of the frequencies of these oscillating modes, but not across the entire gain curve, creates the effect of *spectral hole burning* in the gain curve of an inhomogeneously broadened laser medium, as illustrated in Fig. 11.9. Different transverse modes also saturate independently in an inhomogeneously broadened

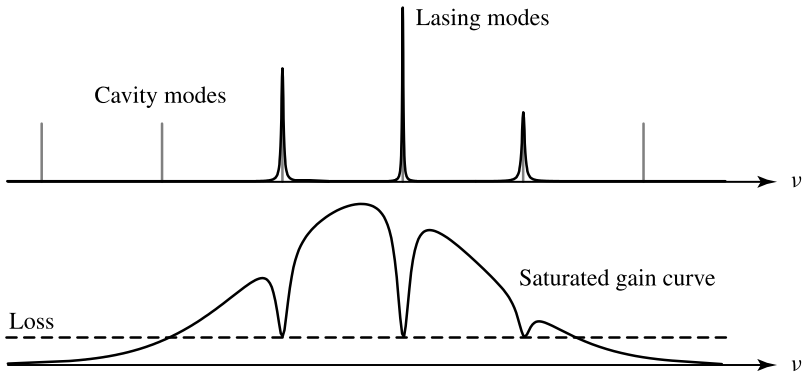


Figure 11.9 Spectral hole burning effect in the gain saturation of a laser in the case of inhomogeneous broadening. Multiple longitudinal modes oscillate simultaneously at a sufficiently high pumping level. The gain at each lasing frequency is saturated at the loss level. Mode pulling effect is ignored in this illustration.

medium if their longitudinal mode frequencies are sufficiently separated. Therefore, an inhomogeneously broadened laser can also oscillate in multiple transverse modes.

The linewidth of an oscillating laser mode is still described by (11.19). From this relation, we see that in practice the round-trip field gain factor G of a laser in steady-state oscillation cannot be exactly equal to unity because the laser linewidth cannot be zero, due to the existence of spontaneous emission. In reality, in steady-state oscillation the value of G is slightly less than unity, with the small difference made up by spontaneous emission. Clearly, the linewidth of an oscillating laser mode is determined by the amount of spontaneous emission that is channeled into the laser mode. Therefore, (11.19) is not very useful for calculating the linewidth of a laser mode in steady-state oscillation without knowing the exact value of G in the presence of spontaneous emission. Instead, a detailed analysis taking into account spontaneous emission yields the following *Shawlow–Townes relation* for the linewidth of a laser mode in terms of the laser parameters:

$$\Delta\nu_{\mathrm{ST}} = \frac{2\pi h\nu(\Delta\nu_{\mathrm{c}})^2}{P_{\mathrm{out}}} N_{\mathrm{sp}} = \frac{h\nu}{2\pi\tau_{\mathrm{c}}^2 P_{\mathrm{out}}} N_{\mathrm{sp}}, \tag{11.65}$$

where P_{out} is the output power of the laser mode being considered and N_{sp} is the spontaneous emission factor defined in (10.114). The effect of spontaneous emission on the linewidth of an oscillating laser mode enters the relation in (11.65) through the population densities of the upper and the lower laser levels in the form of the spontaneous emission factor. Because $N_{\mathrm{sp}} \geq 1$, the ultimate lower limit of the laser linewidth, which is known as the *Shawlow–Townes limit*, is that given in (11.65) with $N_{\mathrm{sp}} = 1$.

EXAMPLE 11.3 The Nd : YAG crystal used for the microchip laser described in Examples 11.1 and 11.2 has a spontaneous linewidth of 150 GHz at the laser wavelength

of 1.064 μm. (a) How many longitudinal modes will oscillate when the laser operates at room temperature? (b) What is the linewidth of an oscillating laser mode when the laser has an output power of 1 mW?

Solution (a) Only one longitudinal mode will oscillate above the laser threshold, for two reasons. First, Nd : YAG is predominantly homogeneously broadened at room temperature. Second, according to Example 11.1, the longitudinal mode spacing for this microchip laser is $\Delta\nu_L = 164.8$ GHz, which is larger than the entire Nd : YAG linewidth of 150 GHz.

(b) The laser photon energy is $h\nu = (1.2398/1.064)$ eV $= 1.165$ eV. From Example 11.1, $\Delta\nu_c = 91.91$ MHz. Because this laser is a four-level system, $N_{sp} = 1$. From (11.65), we find the following Shawlow–Townes linewidth for the oscillating laser mode:

$$\Delta\nu_{ST} = \frac{2\pi \times 1.165 \times 1.6 \times 10^{-19} \times (91.91 \times 10^6)^2}{1 \times 10^{-3}} \times 1 \text{ Hz} = 9.9 \text{ Hz}.$$

Compared to the longitudinal mode width of 91.91 MHz for the cold cavity, this oscillating mode width is nearly seven orders of magnitude smaller. This linewidth-narrowing effect is caused by the coherent nature of the stimulated emission and is a fundamental feature of lasers. Note, however, that the Shawlow–Townes linewidth is only the theoretical lower limit of an oscillating laser mode. In practice, the linewidth of a laser is often broadened far above this limit by other mechanisms, such as fluctuations in the pump power and temperature, mechanical vibrations, and electronic noise from the circuit supporting the operation of the laser. The Shawlow–Townes limit can be approached only by making every effort to eliminate all external effects that broaden the laser linewidth.

11.3 Laser power

In this section, we consider the output power of a laser. Because the situation of a multimode laser can be quite complicated due to mode competition, we consider for simplicity only a homogeneously broadened, CW laser oscillating in a single longitudinal and transverse mode. Therefore, the parameters mentioned in this section are not labeled with mode indices because all of these parameters are clearly associated with the only oscillating mode being considered. The simple case of a Fabry–Perot cavity that contains an isotropic gain medium with a filling factor Γ as shown in Fig. 11.5 is considered. To illustrate the general concepts, we first consider the situation when the gain medium is uniformly pumped so that the entire gain medium has a spatially independent gain coefficient g. We then consider at the end of this section the case of optically pumped lasers, as also considered for the laser threshold in the preceding section, taking into account the longitudinal spatial dependence of the gain coefficient.

For the single oscillating mode of the Fabry–Perot laser considered here, the round-trip gain factor G is that given by (11.52), and the cavity decay rate γ_c defined by (11.24) is that given by (11.47). Therefore,

$$G^2 = \exp(2\Gamma g l - \gamma_c T). \tag{11.66}$$

Because G^2 is the net amplification factor of the intracavity field energy, or photon number, in a round-trip time T of the laser cavity, we can define an *intracavity energy growth rate*, or *intracavity photon growth rate*, Γg, for the oscillating laser mode through the following relation:

$$G^2 = \exp[(\Gamma g - \gamma_c)T)], \tag{11.67}$$

for a laser containing a gain medium with a filling factor Γ. We find, by comparing (11.67) with (11.66), that

$$g = \frac{2gl}{T} = \frac{cg}{\overline{n}} \tag{11.68}$$

and, by comparing (11.47) with (11.56), that

$$\gamma_c = \Gamma \frac{2g_{th}l}{T} = \Gamma \frac{cg_{th}}{\overline{n}}. \tag{11.69}$$

Note that while g and g_{th} are measured per meter, g and γ_c are measured per second.

The relation in (11.68) translates the gain coefficient that characterizes space-dependent amplification of a laser field propagating through the intracavity gain medium into an intracavity energy growth rate that characterizes time-dependent amplification of the energy in a laser mode by the gain medium. The relation in (11.69) clearly indicates that the threshold intracavity energy growth rate for laser oscillation is the cavity decay rate:

$$\Gamma g_{th} = \gamma_c. \tag{11.70}$$

This relation can also be obtained by applying the threshold gain condition of $G = 1$ given in (11.54) to the relation in (11.67). It is easy to understand because for a laser mode to oscillate, the growth of intracavity photons in that mode through amplification by the gain medium has to at least match the decay of photons caused by all of the loss mechanisms combined. Therefore, we shall call the energy growth rate Γg and the cavity decay rate γ_c, both of which are specific to a laser mode, the *gain parameter* and the *loss parameter*, respectively, of the laser mode.

By using temporal growth and decay rates instead of spatial gain and loss coefficients to describe the characteristics of a laser, we are in effect moving from a spatially distributed description of the laser to a lumped-device description. In the lumped-device description, a laser mode is considered an integral entity with its spatial characteristics effectively integrated into the parameters Γg and γ_c. The detailed spatial characteristics of the mode are irrelevant and are lost in this description. Therefore, instead of the

intensity of the oscillating laser field, we have to consider the *intracavity photon density*, S, of the oscillating laser mode. For a Fabry–Perot laser containing a gain medium with a filling factor Γ so that the average refractive index inside the cavity is $\overline{n} = \Gamma n + (1 - \Gamma)n_0$ as defined in (11.4), the average photon density in the cavity is

$$S = \frac{\overline{n}I}{ch\nu}, \tag{11.71}$$

where I is the spatially averaged intensity inside the laser cavity and $h\nu$ is the photon energy of the oscillating laser mode. Because the gain parameter g is directly proportional to the gain coefficient g of the gain medium, the relation between the unsaturated, small-signal gain parameter g_0 and the saturated gain parameter g for a laser mode in the lumped-device description can be obtained by converting the relation between g_0 and g discussed in Section 10.3 through the relation in (11.68). Therefore, for the gain parameter of a laser mode, we have

$$g = \frac{g_0}{1 + S/S_{\text{sat}}}, \tag{11.72}$$

where

$$g_0 = \frac{cg_0}{\overline{n}} \tag{11.73}$$

is the *unsaturated gain parameter* and

$$S_{\text{sat}} = \frac{\overline{n}I_{\text{sat}}}{ch\nu} = \frac{\overline{n}}{c\tau_s\sigma_e} \tag{11.74}$$

is the *saturation photon density*.

When a CW laser oscillates in the steady state, the value of Γg for the oscillating mode is clamped at its threshold value of γ_c, just as the value of g is clamped at g_{th}. Therefore, by setting Γg to equal γ_c with g given in (11.72), we find that the photon density for a CW laser mode in steady-state oscillation is

$$S = \left(\frac{\Gamma g_0}{\gamma_c} - 1\right) S_{\text{sat}} = (r - 1)S_{\text{sat}}, \qquad \text{for } r \geq 1. \tag{11.75}$$

The *dimensionless pumping ratio*, r, represents that a laser is pumped at r times its threshold. It is defined as

$$r = \frac{\Gamma g_0}{\gamma_c} = \frac{g_0}{g_{\text{th}}}. \tag{11.76}$$

Note that (11.75) is valid only for $r \geq 1$ when the laser oscillates because only then is the laser gain saturated. For $r < 1$, the laser does not reach threshold. The laser cavity is then filled with spontaneous photons at a density that is small in comparison to the high density of coherent photons when the laser reaches threshold and oscillates.

From the photon density of the oscillating laser mode, we can easily find the following total intracavity energy contained in this mode:

$$U_{\text{mode}} = \mathcal{V}_{\text{mode}} S h\nu, \tag{11.77}$$

where $\mathcal{V}_{\text{mode}}$ is the volume of the oscillating mode. The *mode volume* can be found by integrating the normalized intensity distribution of the mode over the three-dimensional space defined by the laser cavity. It is usually a fraction of the volume of the cavity. The output power of the laser is simply the coherent optical energy emitted from the laser per second. Therefore, it is simply the product of the mode energy and the *output-coupling rate*, γ_{out}, of the cavity:

$$P_{\text{out}} = U_{\text{mode}}\gamma_{\text{out}} = \mathcal{V}_{\text{mode}}Sh\nu\gamma_{\text{out}} = (r-1)\mathcal{V}_{\text{mode}}S_{\text{sat}}h\nu\gamma_{\text{out}}. \tag{11.78}$$

The output-coupling rate is also called the *output-coupling loss parameter* because it contributes to the total loss of a laser cavity and is a fraction of the total loss parameter γ_{c}. One can indeed write $\gamma_{\text{c}} = \gamma_{\text{i}} + \gamma_{\text{out}}$, where γ_{i} is the internal loss of the laser, which does not contribute to output coupling of the laser power. As an example, for the Fabry–Perot laser with its γ_{c} given by (11.47), we have the internal loss given by $\gamma_{\text{i}} = c\overline{\alpha}_{mn}/\overline{n}$ and the output-coupling loss given by

$$\gamma_{\text{out}} = -\frac{c}{\overline{n}l}\ln\sqrt{R_1R_2}. \tag{11.79}$$

In this case, γ_{out} is the total output-coupling loss through both mirrors. Therefore, P_{out} given in (11.78) is the total output power emitted through both mirrors. For the power output through each mirror, we find that

$$\gamma_{\text{out},1} = -\frac{c}{\overline{n}l}\ln\sqrt{R_1} \quad \text{and} \quad \gamma_{\text{out},2} = -\frac{c}{\overline{n}l}\ln\sqrt{R_2} \tag{11.80}$$

and that

$$P_{\text{out},1} = U_{\text{mode}}\gamma_{\text{out},1} = P_{\text{out}}\frac{\gamma_{\text{out},1}}{\gamma_{\text{out}}} \quad \text{and} \quad P_{\text{out},2} = U_{\text{mode}}\gamma_{\text{out},2} = P_{\text{out}}\frac{\gamma_{\text{out},2}}{\gamma_{\text{out}}}. \tag{11.81}$$

It is convenient to define the *saturation output power* as

$$P_{\text{out}}^{\text{sat}} = \mathcal{V}_{\text{mode}}S_{\text{sat}}h\nu\gamma_{\text{out}}. \tag{11.82}$$

Using the definition of S_{sat} in (11.74), it can be shown that

$$P_{\text{out}}^{\text{sat}} = -P_{\text{sat}}\ln\sqrt{R_1R_2}, \tag{11.83}$$

where P_{sat} is the saturation power of the gain medium found by integrating I_{sat} over the cross-sectional area of the gain medium. Combining (11.78) with (11.82), we can express the output laser power in terms of $P_{\text{out}}^{\text{sat}}$ as

$$P_{\text{out}} = (r-1)P_{\text{out}}^{\text{sat}}. \tag{11.84}$$

Note that $P_{\text{out}}^{\text{sat}}$ *is not the level at which the output power of a laser saturates*. Its physical meaning can be easily seen from (11.83) and (11.84). From (11.83), we find that the output power of a laser is $P_{\text{out}}^{\text{sat}}$ when the intracavity laser power is P_{sat} of the gain medium. From (11.84), we find that $P_{\text{out}} = P_{\text{out}}^{\text{sat}}$ when $r = 2$; in other words, a laser has an output power of $P_{\text{out}}^{\text{sat}}$ when it is pumped at twice its threshold level.

In order to express the output laser power explicitly as a function of the pump power, it is necessary to specify the pumping mechanism and the pumping geometry. For this purpose, we consider longitudinal optical pumping with negligible transverse pump beam divergence but with a spatially varying gain coefficient $g(z)$ along the longitudinal axis of the gain medium, as is the case considered in the preceding section for the threshold pump power obtained in (11.61). In this situation, all of the results obtained so far in this section are still applicable if we make the following substitution for the gain coefficient:

$$g = \frac{1}{l_g}\int_0^{l_g} g(z)\mathrm{d}z = \frac{1}{\Gamma l}\int_0^{l_g} g(z)\mathrm{d}z. \tag{11.85}$$

Then, in the case when $p = 0$ or $p \ll 1$, the pumping ratio at an input pump power of P_p can be expressed as (see Problem 11.3.1)

$$r = \frac{\Gamma g_0}{\gamma_c} = \frac{\int_0^{l_g} g_0(z)\mathrm{d}z}{g_{th} l_g} \approx \frac{\zeta_p P_p - \zeta_p^{tr} P_p^{tr}}{\zeta_p^{th} P_p^{th} - \zeta_p^{tr} P_p^{tr}}, \tag{11.86}$$

where P_p^{th} and P_p^{tr} are the threshold pump power and the transparency pump power of the laser found in (11.61) and (11.62), respectively, and ζ_p, ζ_p^{th}, and ζ_p^{tr} are the pump power utilization factors at the pumping levels of P_p, P_p^{th}, and P_p^{tr}, respectively. In the case of $P_p \ll P_p^{sat}$ when the absorption saturation of the pump is negligible, $\zeta_p \approx \zeta_p^{th} \approx \zeta_p^{tr}$. In the presence of significant absorption saturation of the pump, however, we find that $\zeta_p < \zeta_p^{th} < \zeta_p^{tr}$ because $P_p > P_p^{th} > P_p^{tr}$.

By substituting (11.86) for r in (11.84), we find the following relation between the output power of a laser and the power launched to pump the laser (see Problem 11.3.1):

$$P_{out} = \frac{\zeta_p P_p - \zeta_p^{th} P_p^{th}}{\zeta_p^{th} P_p^{th} - \zeta_p^{tr} P_p^{tr}} P_{out}^{sat}. \tag{11.87}$$

This relation is obtained by using (11.86) under the following assumptions: (1) $p = 0$ or $p \ll 1$, (2) P_{out}^{sat} is a constant throughout the gain medium, and (3) the intracavity laser photon density is relatively uniformly distributed. It works best in the situation where (1) the gain medium is a four-level or three-level system so that $p = 0$, (2) $W_p \tau_2 \ll 1$ so that τ_s and P_{out}^{sat} are not spatially varying, and (3) $R_1 R_2$ approaches unity so that the intracavity photon density distribution is quite uniform. A laser that satisfies these conditions is considered in Example 11.4.

Alternatively, by consideration of energy conservation, the output laser power can be found through the following relation:

$$P_{out} = \eta_p \frac{\gamma_{out}}{\gamma_c}\frac{h\nu}{h\nu_p}(\zeta_p P_p - \zeta_p^{th} P_p^{th}) = \eta_p \frac{\gamma_{out}}{\gamma_c}\frac{\lambda_p}{\lambda}(\zeta_p P_p - \zeta_p^{th} P_p^{th}), \tag{11.88}$$

where η_p is the pump quantum efficiency defined in (10.84), and λ and λ_p are the laser and pump wavelengths, respectively. This relation is quite general. It is not subject to the conditions that limit the applicability of (11.87). Under the conditions when (11.87) is valid, it can be shown that (11.87) yields exactly the same result as (11.88). When the conditions for (11.87) are not fully satisfied so that (11.87) no longer yield a reliable result, (11.88) can still be used to find the output laser power.

The relations in (11.87) and (11.88) state that the output power of a laser grows linearly with the pump power above threshold. It also indicates that the laser has zero output power before it reaches threshold. Upon reaching the threshold, the optical output of the device also shows dramatic spectral narrowing that accompanies the start of laser oscillation. According to (11.65), the linewidth of an oscillating laser mode continues to narrow with increasing laser power as the laser is pumped higher and higher above threshold. These are the unique characteristics that distinguish a laser from other types of light sources such as fluorescent light emitters and luminescent light sources. However, a real laser does not have exactly such ideal characteristics, mainly because of the presence of spontaneous emission and nonlinearities in the gain medium.

Figure 11.10 shows typical characteristics of the output power of a single-mode laser as a function of pump power. On the one hand, the linear relation in (11.87) between P_{out} and P_p is a consequence of applying the linear relation between g_0 and P_p to (11.76). As discussed in Section 10.3, the linear relation between g_0 and P_p derived from (10.88) is itself an approximation near the transparency point of a gain medium. As the pump power increases to a sufficiently high level, the unsaturated gain coefficient of a medium cannot continue to increase linearly with pump power

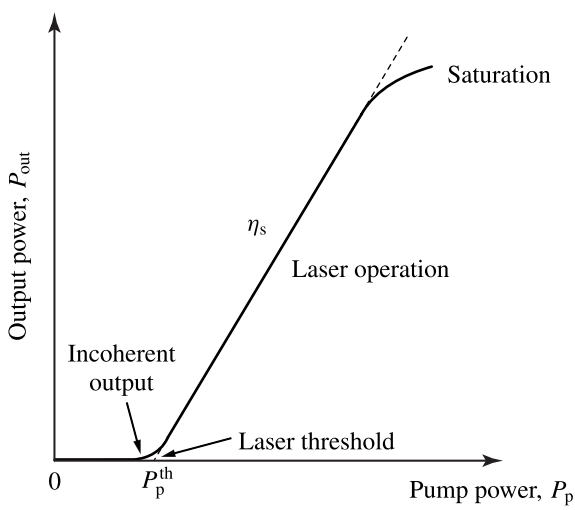


Figure 11.10 Typical characteristics of the output power of a single-mode laser as a function of pump power.

because of depletion of the ground-level population. Therefore, we should expect the output power of a laser not to continue its linear dependence on pump power but to increase less than linearly with pump power at high pumping levels. On the other hand, once the gain medium of a laser is pumped so that its upper laser level begins to be populated, it emits spontaneous photons regardless of whether the laser is oscillating or not. Clearly, the output power of a laser pumped below threshold is not exactly zero because spontaneous power is already emitted from the laser before the laser reaches threshold. Though this spontaneous power is incoherent and is generally small in a practical laser, it is significant for a laser below and right at threshold. Above threshold, it is the major source of incoherent noise for the coherent field of the laser output.

EXAMPLE 11.4 Find the pump power required for the Nd : YAG microchip laser described in Examples 11.1–11.3 to have an output power of 1 mW.

Solution This laser satisfies the conditions required for the application of (11.87). We should first find P_p through (11.84) and (11.86) in the spirit of (11.87). Then we show that the same result is obtained from (11.88).

The required pump power for a desired output power can be found using (11.84) to obtain the value of the pumping ratio r for a given value of P_{out}. To use (11.84) for this purpose, we have to find P_{out}^{sat} through the values of $\mathcal{V}_{mode}$, S_{sat}, and γ_{out}. Because $l = 500\ \mu m$ and $w = 100\ \mu m$,

$$\mathcal{V}_{mode} = \pi w^2 l = \pi \times (100 \times 10^{-6})^2 \times 500 \times 10^{-6}\ m^3 = 1.57 \times 10^{-11}\ m^3.$$

Because $s_{th} = P_p^{th}/P_p^{sat} = 8.3 \times 10^{-4} \ll 1$ from Example 11.2, we expect $s = W_p\tau_2 = P_p/P_p^{sat} \ll 1$ so that $\tau_s = \tau_2 = 240\ \mu s$ for the operating pump power range of this laser. With $\overline{n} = n = 1.82$ for this laser, we then have

$$S_{sat} = \frac{\overline{n}}{c\tau_s\sigma_e} = \frac{1.82}{3 \times 10^8 \times 240 \times 10^{-6} \times 3.1 \times 10^{-23}}\ m^{-3} = 8.15 \times 10^{17}\ m^{-3}.$$

Because $R_1 = 100\%$ and $R_2 = 99.7\%$, we have

$$\gamma_{out} = \gamma_{out,2} = -\frac{c}{\overline{n}l}\ln\sqrt{R_2} = -\frac{3 \times 10^8}{1.82 \times 500 \times 10^{-6}} \times \ln\sqrt{0.997}\ s^{-1} = 4.95 \times 10^8\ s^{-1}.$$

We then find that

$$\begin{aligned} P_{out}^{sat} &= \mathcal{V}_{mode} S_{sat} h\nu\gamma_{out} \\ &= 1.57 \times 10^{-11} \times 8.15 \times 10^{17} \times 1.165 \times 1.6 \times 10^{-19} \times 4.95 \times 10^8\ W \\ &= 1.18\ mW \end{aligned}$$

We can then use (11.84) to find that

$$P_{out} = 1.18(r - 1)\ mW.$$

For an output of $P_{\text{out}} = 1$ mW, we find that the required pumping ratio is

$$r = 1 + \frac{1}{1.18} = 1.85.$$

Because this laser is a four-level laser, we have $P_{\text{p}}^{\text{tr}} = 0$. We already found from Example 11.2 that $P_{\text{p}}^{\text{th}} = 11.1$ mW and that the absorption saturation of the pump is negligible. Therefore, $\zeta_{\text{p}} \approx \zeta_{\text{p}}^{\text{th}} \approx 0.204$. Using these parameters, the required pump power is found from (11.86) to be

$$P_{\text{p}} = \frac{\zeta_{\text{p}}^{\text{th}}}{\zeta_{\text{p}}} r P_{\text{p}}^{\text{th}} \approx r P_{\text{p}}^{\text{th}} = 1.85 \times 11.1 \text{ mW} = 20.5 \text{ mW}.$$

Alternatively, we can find P_{p} using (11.88). For this purpose, we have, from Example 11.2, the following parameters: $\eta_{\text{p}} = 0.8$, $\lambda = 1.064$ μm, $\lambda_{\text{p}} = 808$ nm, and $g_{\text{th}} = 3.5 \text{ m}^{-1}$. We also find that

$$\gamma_{\text{c}} = \Gamma \frac{c g_{\text{th}}}{\overline{n}} = 1 \times \frac{3 \times 10^8 \times 3.5}{1.82} \text{ s}^{-1} = 5.77 \times 10^8 \text{ s}^{-1}.$$

For $P_{\text{out}} = 1$ mW and $P_{\text{p}}^{\text{th}} = 11.1$ mW, we find from (11.88) that

$$\begin{aligned} P_{\text{p}} &= \frac{1}{\zeta_{\text{p}} \eta_{\text{p}}} \frac{\gamma_{\text{c}}}{\gamma_{\text{out}}} \frac{\lambda_{\text{p}}}{\lambda} P_{\text{out}} + \frac{\zeta_{\text{p}}^{\text{th}}}{\zeta_{\text{p}}} P_{\text{p}}^{\text{th}} \\ &= \frac{1}{0.204 \times 0.8} \times \frac{5.77 \times 10^8}{4.95 \times 10^8} \times \frac{808 \times 10^{-9}}{1.064 \times 10^{-6}} \times 1 \text{ mW} + \frac{0.204}{0.204} \times 11.1 \text{ mW} \\ &= 20.5 \text{ mW}. \end{aligned}$$

We see that the same result is obtained.

The overall efficiency of a laser, known as the *power conversion efficiency*, is

$$\eta_{\text{c}} = \frac{P_{\text{out}}}{P_{\text{p}}}. \tag{11.89}$$

The linear dependence of the laser output power on the pump power indicated by (11.87) leads to the concept of the *differential power conversion efficiency*, also known as the *slope efficiency*, of a laser, defined as

$$\eta_{\text{s}} = \frac{\text{d}P_{\text{out}}}{\text{d}P_{\text{p}}} = \frac{\zeta_{\text{p}} P_{\text{out}}^{\text{sat}}}{\zeta_{\text{p}}^{\text{th}} P_{\text{p}}^{\text{th}} - \zeta_{\text{p}}^{\text{tr}} P_{\text{p}}^{\text{tr}}} = \frac{\zeta_{\text{p}} P_{\text{out}}}{\zeta_{\text{p}} P_{\text{p}} - \zeta_{\text{p}}^{\text{th}} P_{\text{p}}^{\text{th}}} = \eta_{\text{p}} \zeta_{\text{p}} \frac{\gamma_{\text{out}}}{\gamma_{\text{c}}} \frac{\lambda_{\text{p}}}{\lambda}. \tag{11.90}$$

Referring to the power characteristics of the laser shown in Fig. 11.10, the threshold of a given laser can usually be lowered by increasing the finesse, thus lowering the values of γ_{c} and γ_{out}, of the laser cavity, but only at the expense of reducing the differential power conversion efficiency of the laser. Clearly, in the linear region of the laser power characteristics where the relation given by (11.87) is valid, η_{s} is a constant that is independent of the operating point of the laser. In contrast, η_{c} increases with pump

power, but η_c is always smaller than η_s in the linear region. At high pumping levels where the laser output power does not increase linearly with pump power, η_s is no longer independent of the operating point. It can even become smaller than η_c in some unfavorable situations.

The quantum efficiency of a laser oscillator is defined differently from that of a laser amplifier. The *external quantum efficiency*, η_e, also known as the *differential quantum efficiency*, measures the efficiency of converting pump photons or pump electrons *above threshold* into laser photons at the laser output. Furthermore, an *internal quantum efficiency*, η_i, can be defined to measure the efficiency of converting the pump photons or the pump electrons above threshold into *intracavity* laser photons. They are defined through the following relation:

$$\eta_e = \frac{\Phi_{out}}{\zeta_p \Phi_p - \zeta_p^{th} \Phi_p^{th}} \quad \text{and} \quad \eta_i = \frac{\gamma_c}{\gamma_{out}} \eta_e, \tag{11.91}$$

where $\Phi_{out} = P_{out}/h\nu$ is the output photon flux of the laser, Φ_p is the pump photon flux, in the case of optical pumping, or the pump electron flux, in the case of electrical pumping, Φ_p^{th} is the threshold pump photon or electron flux, and ζ_p is the pump power utilization factor. In the case of optical pumping, $\Phi_p = P_p/h\nu_p$. Then, the external quantum efficiency is directly related to the slope efficiency and the pump quantum efficiency as

$$\eta_e = \frac{\lambda}{\lambda_p} \frac{\eta_s}{\zeta_p} = \eta_p \frac{\gamma_{out}}{\gamma_c}. \tag{11.92}$$

From (11.91) and (11.92), we find that $\eta_i = \eta_p$. Because $\gamma_{out} < \gamma_c$, the external quantum efficiency η_e is smaller than the internal quantum efficiency η_i for a typical laser. This reflects the fact that, because of the presence of losses in the laser cavity other than the output coupling loss, not all photons generated inside a laser cavity contribute to the output of the laser. Furthermore, the internal quantum efficiency η_i is equal to the pump quantum efficiency η_p, reflecting the fact that after a laser is pumped above threshold, every additional atom excited to the upper laser level results in the contribution of one photon in the oscillating laser mode through stimulated emission.

EXAMPLE 11.5 Find the power conversion efficiency, the slope efficiency, and the external and internal quantum efficiencies of the Nd : YAG microchip laser described in Example 11.4 operating at an output power of 1 mW.

Solution From Example 11.4, we find that $P_p = 20.5$ mW for $P_{out} = 1$ mW. The power conversion efficiency in this operating condition is

$$\eta_c = \frac{P_{out}}{P_p} = \frac{1}{20.5} = 4.9\%.$$

This laser has a threshold pump power of $P_p^{th} = 11.1$ mW found in Example 11.2. Also from Example 11.2, we know that $\zeta_p \approx \zeta_p^{th} \approx 0.204$ because of negligible absorption saturation of the pump. The slope efficiency can then be found from (11.90) to be

$$\eta_s = \frac{\zeta_p P_{out}}{\zeta_p P_p - \zeta_p^{th} P_p^{th}} \approx \frac{P_{out}}{P_p - P_p^{th}} = \frac{1}{20.5 - 11.1} = 10.6\%.$$

With $\lambda = 1.064$ μm and $\lambda_p = 808$ nm, the external quantum efficiency is thus found from (11.92) to be

$$\eta_e = \frac{\lambda}{\lambda_p}\frac{\eta_s}{\zeta_p} = \frac{1.064 \times 10^{-6}}{808 \times 10^{-9}} \times \frac{10.6\%}{0.204} = 68.4\%.$$

For this laser, we have $\gamma_c = 5.78 \times 10^8$ s^{-1} from Example 11.1 and $\gamma_{out} = 4.95 \times 10^8$ s^{-1} from Example 11.4. The internal quantum efficiency can be found by using (11.91) to be

$$\eta_i = \frac{\gamma_c}{\gamma_{out}}\eta_e = \frac{5.78 \times 10^8}{4.95 \times 10^8} \times 68.4\% = 80\%.$$

We see that, as expected, η_i is the same as η_p, which is 80% as given in Example 11.2.

This laser has a power conversion efficiency of 4.9% compared to a slope efficiency of 10.6%. Compared to the high quantum efficiencies of $\eta_e = 68.4\%$ and $\eta_i = 80\%$, these power efficiencies are relatively low. The power conversion efficiency can be increased by operating the laser at a higher pumping level, but it cannot exceed the slope efficiency, which is a constant before nonlinearity saturates the laser output at a significantly high pumping level. The reason for η_c to be always smaller than η_s before saturation is that a laser has to overcome its threshold before it starts to oscillate. The reason for the low power efficiencies in this example is that only 20.4% of the pump power is absorbed by the gain medium because the pump beam passes through the gain medium in only one single pass. Therefore, close to 80% of the input pump power simply passes through the gain medium without being utilized. Both η_c and η_s for this laser can be increased by taking a multiple-pass arrangement (see Problem 11.3.3) or by properly increasing the length of the gain medium (see Problem 11.3.5) to increase the utilization factor ζ_p of the pump power. However, quantum efficiencies are not increased by such steps. Indeed, if the cavity parameters are changed, the external quantum efficiency η_e might even be reduced while the slope efficiency η_s is increased.

11.4 Pulsed lasers

In the CW operation of a laser, the laser is pumped continuously at a constant pumping level, and the loss of the laser cavity is also kept constant so that the laser has a constant output power when it reaches steady state. A laser can also be pulsed to deliver short optical pulses at its output. In pulsed operation, the net gain seen by the laser

field is not kept constant but is temporally varied by pulse pumping the gain medium and/or by modulating the cavity loss. Depending on the laser material, the cavity design, and the technique employed for the pulsed operation, laser pulses of temporal pulsewidths ranging from the order of microseconds to the order of femtoseconds with large ranges of pulse repetition rates, from single shots to gigahertz, and pulse energies, from femtojoules per pulse to joules per pulse, can be generated. Many effective techniques have been developed for the generation of laser pulses. The simplest approach is *gain switching*, while the most important and most commonly employed techniques are *Q switching* and *mode locking*. Here we only discuss the basic principles of these techniques.

Gain switching

Gain switching is a technique that is used to generate very short laser pulses through the control of oscillator transients. The concept of gain switching is straightforward: the gain parameter Γg of a laser is switched on rapidly above the laser threshold, which is defined by the loss parameter γ_c, by fast, pulsed pumping so that a very short laser pulse is generated through the transient effects of the laser oscillator.

In a gain-switched laser, the gain medium is pumped so fast that the population inversion builds up more rapidly than the photons in the cavity. The gain is thus raised considerably above the threshold before the laser field starts to build up from the initial noise level in the cavity. The transient effects that follow under the excessively high-gain condition result in the generation of a short, powerful laser pulse. Because the intracavity laser photon density grows exponentially in time with the net gain, for a gain-switched laser pulse to have a short risetime it is only necessary to create a large excess gain above the threshold before stimulated emission starts to reduce it by depleting the population inversion. This condition requires hard and fast pumping. In addition, a long lifetime τ_2 for the upper laser level in comparison to the pump-pulse duration also helps in building up the excess population inversion. To make sure of a short falltime for the gain-switched pulse, first the gain has to be terminated when the photon density builds up to its peak value, and then the intracavity photons have to be depleted quickly. The first of these two conditions requires short pumping duration, while the second requires a short photon lifetime, corresponding to a large photon decay rate γ_c. In addition, if gain saturation occurs, the laser gain can be terminated even more rapidly, thus reducing the pulse falltime substantially. Figure 11.11 illustrates the basic concept of gain switching.

From the above discussion, one can see that the conditions for the generation of very short laser pulses by gain switching are (1) a large excess population inversion at the onset of laser oscillation, (2) a short photon lifetime τ_c, and (3) sufficient gain saturation after pulse buildup. The technical aspect of gain switching is in the choice of the pump and laser parameters best to satisfy these conditions. Successful gain switching of a

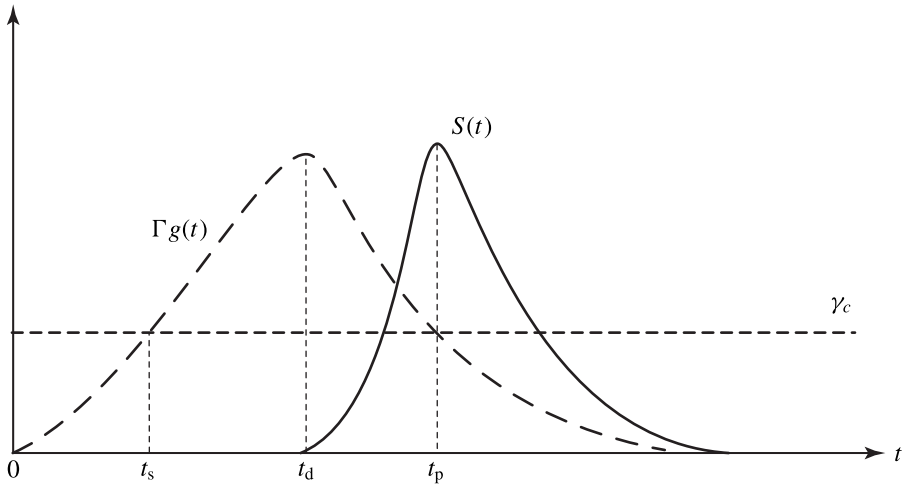


Figure 11.11 Temporal evolutions of gain parameter and intracavity photon density in a gain-switched laser.

laser can be accomplished by choosing (1) a very short and strong pump, (2) a short laser cavity, and (3) a laser medium that has a low saturation intensity and a long fluorescence lifetime for the upper laser level in comparison to the pump-pulse duration.

Physical constraints sometimes make it impossible to fulfill all these requirements. However, it is not necessary to satisfy all of these conditions fully before a short laser pulse can be generated by gain switching. For example, by pumping the gain medium hard enough, a very short pulse can be generated even when the pump pulse is longer than the fluorescence lifetime of the upper laser level. Certainly, if the fluorescence lifetime is long, the gain medium can be pumped less hard for a desired pulsewidth, or a shorter pulse can be generated with the same pump. Therefore, a typical gain-switched laser that generates ultrashort pulses is a laser with a very short cavity pumped by a strong pump pulse that has a temporal duration on the order of or shorter than the fluorescence lifetime of the gain medium. Short laser pulses are ideal sources for optically pumping a secondary gain-switched laser. This approach has been demonstrated for gain switching in solid-state lasers, dye lasers, and semiconductor lasers. In gain-switched semiconductor lasers that are electrically pumped, the pump current pulses are typically one order of magnitude shorter than the carrier lifetime of the semiconductor gain medium.

The dynamics of the laser oscillator are solely responsible for the behavior of a gain-switched laser pulse. Therefore, if not well-controlled, many transient phenomena, such as spiking and relaxation oscillation, can take place. The output then consists of a series of spikes if the relaxation oscillation is not damped. However, by choosing a cavity of a proper photon lifetime and by controlling the level and duration of pumping, a single clean pulse without relaxation oscillation spikes can be generated. It is then only necessary to pump the laser medium as fast as possible in a cavity as short as possible

to generate a very short pulse. No special optical elements are required in the laser cavity. By repetitively gain switching a laser with a periodic train of pump pulses, a train of regularly spaced, gain-switched pulses can be generated. The gain switching technique has been used to generate single pulses of temporal widths ranging from 1 ps in an optically pumped short-cavity GaAs laser to a few hundred nanoseconds in CO_2 lasers.

To generate an ultrashort laser pulse by gain switching, it is important to have an extremely small cavity lifetime τ_c because the shortest cavity photon decay time is limited by τ_c. Sometimes τ_c can be smaller than the cavity round-trip time T because of high intracavity loss or high output-coupling loss. Unless the small τ_c is caused by a high distributed loss, however, the shortest pulse that can be generated by gain switching from a laser with $\tau_c < T$ is limited by T rather than τ_c. This is because it takes at least one round trip to deplete all the intracavity photons by output coupling through the mirrors. This limitation applies also to Q switching, which is also a transient technique. It does not apply to mode locking, which does not rely on transient phenomena to generate ultrashort laser pulses. Therefore, the cavity length usually imposes a direct physical limitation on the laser dynamics so that the shortest pulsewidth generated by the transient technique of gain switching or Q switching can only be as short as the cavity round-trip time.

Q switching

Q switching is the most widely used technique for the generation of high-intensity giant laser pulses of short duration. Similar to gain switching, Q switching relies on the transient dynamics of a laser to generate very short pulses. However, it does not require extremely fast pumping, as does the technique of gain switching. In fact, it is possible to pump the gain medium continuously while switching the cavity Q factor repetitively to generate a periodic train of Q-switched pulses.

The principle of Q switching is based on delaying the onset of laser oscillation relative to the start of pumping to accumulate a large population inversion. This task is accomplished by reducing the laser cavity Q factor in the early stage of pumping to prohibit the depletion of population inversion caused by premature laser oscillation. Upon reaching a large population inversion, the Q factor is rapidly increased, resulting in a large excess gain above threshold and a burst of high-intensity short pulse driven by the transient dynamics of the laser. Because $Q = \omega/\gamma_c$ according to (11.28), modulating the cavity Q factor is equivalent to modulating the cavity loss rate γ_c. The basic principle of Q switching is illustrated in Fig. 11.12. In contrast to gain switching where the cavity loss rate is kept constant, $\gamma_c(t)$ for a Q-switched laser is a time-varying function.

In the *pumping phase* of a Q-switched laser, population inversion builds up without being depleted by stimulated emission. Clearly, if the pump-pulse duration is much shorter than the fluorescence lifetime of the gain medium, gain switching can be very

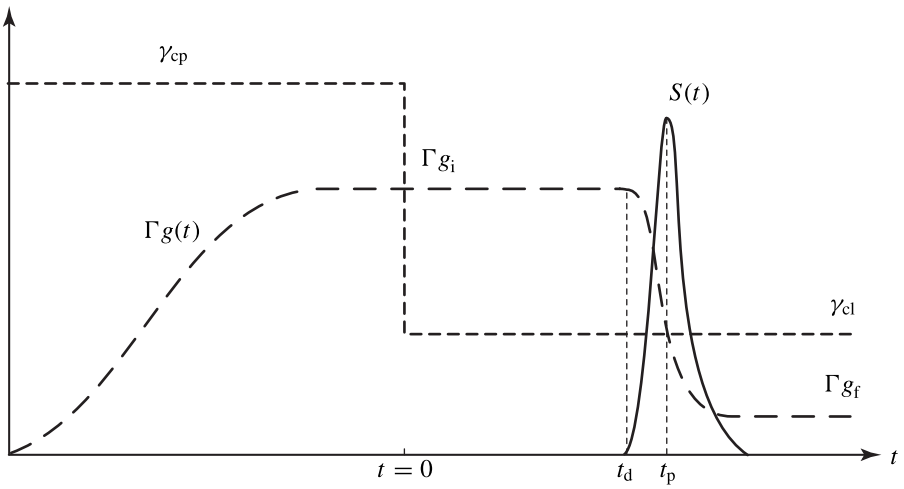


Figure 11.12 Temporal evolutions of cavity loss rate, gain parameter, and intracavity photon density in a Q-switched laser. This illustration is based on an initial pumping ratio of $r = 2.2$.

effective, and there is no need for Q switching. This is the condition discussed above for gain switching. When the pump pulse is long, the gain grows slowly. However, given sufficient time, the gain can still accumulate to a substantial value if τ_2 is sufficiently long. This is the situation when Q switching can be effectively implemented to generate a short pulse. Therefore, a large τ_2 is even more desirable for Q switching than for gain switching. For most efficient utilization of the pump energy, the pump duration should not be too much longer than τ_2 although it does not have to be short. Because of spontaneous relaxation, population inversion, if not depleted by laser oscillation, cannot continue to build up much longer than a period of τ_2. For a repetitively Q-switched laser under continuous pumping, this fact means that the overall efficiency of the laser drops when the repetition rate is below $1/\tau_2$.

The major difference between Q switching and gain switching is in the pumping phase when $\gamma_c(t)$ is kept high in the case of Q switching. In the lasing phase, gain-switched and Q-switched lasers are driven by the same transient laser dynamics initiated by the initial excess population inversion. This can be seen by comparing Fig. 11.12 to Fig. 11.11.

It is clear that the conditions for the generation of very short laser pulses by gain switching discussed earlier apply equally well to Q switching. Q switching differs from gain switching only in the technical aspect of how the large initial excess population inversion is achieved. Gain switching relies on a very fast and strong pump pulse to achieve a high peak gain before stimulated emission starts. This condition is not required for Q switching as the high cavity loss of a Q-switched laser in the pumping phase prohibits the laser from oscillating. Thus the requirement on the pump for Q switching is less demanding than that for gain switching. In comparison to a gain-switched laser, the technical demand in a Q-switched laser is shifted from the pump to the Q switch. In

order to generate a very short pulse by Q switching, it is therefore desirable to have (1) an effective Q switch that switches the cavity Q very rapidly from a very low value to a high value at the moment the gain reaches a desired high level, (2) a short laser cavity, and (3) a laser medium that has a low saturation intensity and a long fluorescence lifetime. Similarly to the case of gain switching, these requirements do not have to be fully satisfied, but a Q-switched laser pulse cannot be shorter than the cavity round-trip time and a laser with a high saturation intensity has to be operated at a high power level.

In the ideal situation, the value of $\gamma_c(t)$ is switched abruptly from a high level, γ_{cp}, for the pumping phase to a low level, γ_{cl}, for the *lasing phase*. In practice, because it takes a time delay, t_d shown in Fig. 11.12, after the time of switching into the lasing phase for the Q-switched pulse to build up to a significant level, the condition of ideal Q switching can be approximated by *fast Q switching* where the transition from the pumping phase to the lasing phase is completed within a time duration less than the time delay of the pulse buildup. In this ideal, or nearly ideal, situation of fast Q switching, the characteristics of the Q-switched pulse are completely determined by the initial pumping ratio at the onset of the lasing phase:

$$r = \frac{\Gamma g_i}{\gamma_{cl}} \tag{11.93}$$

where g_i is the *initial gain parameter* at the onset of the lasing phase, as illustrated in Fig. 11.12. In a Q-switched laser, the initial gain parameter g_i for the lasing phase actually shoots over the threshold level defined by γ_{cl} because of the Q-switching action. Therefore, the parameter r defined in (11.93) has a somewhat different meaning from that defined in (11.76). Clearly, we always have $r > 1$ for Q-switching operation. The *peak output power* of a Q-switched pulse is approximately given by

$$P_{pk} \approx \frac{\tau_2}{\tau_{cl}}(r - \ln r - 1)P_{out}^{sat}, \tag{11.94}$$

where $\tau_{cl} = 1/\gamma_{cl}$ is the photon lifetime in the lasing phase and $P_{out}^{sat} = \mathcal{V}_{mode} S_{sat} h\nu \gamma_{out}$ as defined in (11.82). For $1.2 < r < 5$, the FWHM pulsewidth, Δt_{ps}, can be quite accurately approximated by the following formula:

$$\Delta t_{ps} = \frac{2.5}{(r - \ln r - 1)^{1/2}}\tau_{cl}, \tag{11.95}$$

which is obtained by approximate analytical fitting of Q-switched pulses. The energy of a Q-switched pulse can be approximated by

$$U \approx P_{pk}\Delta t_{ps}. \tag{11.96}$$

Clearly, the larger the value of r, the more dramatic the Q-switching behavior is, resulting in a more powerful Q-switched pulse with a higher peak power and a smaller pulsewidth.

Depending on the technique used to modulate the cavity Q factor, the type of Q switching can be generally categorized as active or passive. In active Q switching, the Q switch that modulates the cavity Q is controlled by an externally applied signal. Various techniques have been developed for active Q switching, including mechanical modulation, electro-optic modulation, acousto-optic modulation, and magneto-optic modulation. Today, most of the actively Q-switched lasers use electro-optic or acousto-optic modulators, which modulate the cavity Q by modulating the loss in the cavity. Both of these two types of modulators are controlled by external electronic signals, which have the advantages of stability and flexibility of Q modulation and ease of synchronization with measurement apparatus. Typical electro-optic modulators are based on the Pockels effect. They have fast switching times in the nanosecond range with a large Q modulation and can be controlled with precise timing, but they often require a large switching voltage and are difficult to operate at a high repetition rate. The acousto-optic modulators are Bragg diffractors driven by an RF signal. They have a slower switching speed and a smaller Q modulation than electro-optic modulators, thus producing longer pulses. They can be easily operated at a high and variable repetition rate and are primarily used in continuously pumped, repetitively Q-switched lasers at a repetition rate in the kilohertz range. High-frequency intracavity electro-optic modulation based on the electroabsorption effect can be applied to a semiconductor laser for the generation of actively Q-switched picosecond pulses at a high repetition rate in the gigahertz range.

In a passively Q-switched laser, the Q switch is typically a nonlinear optical element that changes the cavity loss by responding directly to the intracavity laser intensity. The most commonly used passive Q switch is a saturable absorber, the optical properties of which are discussed in Section 9.7. With a proper arrangement, any all-optical switch, such as a Kerr lens, also discussed in Section 9.7, can function as a passive Q switch. Passive Q switching has the advantage of being simple and inexpensive, but the pulses generated by passive Q switching are often subject to larger intensity fluctuations and timing jitter than those generated by active Q switching.

Solid-state lasers, such as Nd : YAG, ruby, and Ti : sapphire, are primary candidates for Q switching because they normally have a long fluorescence lifetime. The pulsewidth of a pulse generated by a Q-switched solid-state laser is typically in the range of a few nanoseconds to hundreds of nanoseconds. Many useful laser materials, such as laser dyes and semiconductors, have a very small τ_2 on the order of nanoseconds or less. They are not easy to Q switch unless a very efficient pump source and a very fast Q switch are used. As a consequence, the Q-switched pulses generated by these lasers are typically on the order of tens or hundreds of picoseconds although pulses as short as a few picoseconds at a repetition rate as high as a few tens of gigahertz have been generated by passive Q switching of semiconductor lasers. It is also possible to combine Q switching with mode locking in a Q-switched mode-locked laser to generate a train of very short mode-locked pulses under a long Q-switched envelope.

EXAMPLE 11.6 The characteristics of the Nd : YAG microchip laser described in Examples 11.1–11.5 in ideal Q-switching or gain-switching operation are considered in this example. For the gain-switching operation, all of the laser parameters, including those of the cavity and the gain medium, remain the same as those described in Examples 11.1–11.5 except that the pump is an optical pulse at the pump wavelength of 808 nm. For the Q-switching operation, a Q switch introduces an additional high loss to the laser in the pumping phase, but the laser parameters in the lasing phase are the same as those for the gain-switching operation. A possible Q-switching mechanism is passive Q switching by codoping the Nd : YAG with Cr^{4+} ions as the saturable absorber. For direct comparison with CW operation, we take the pumping ratio to be $r = 1.85$, as found in Example 11.4 for a CW output power of 1 mW. (a) What are the required conditions for the laser to be nearly ideally Q-switched? (b) Find the peak power, pulsewidth, and pulse energy of the ideally Q-switched pulse. Compare the peak power of the Q-switched pulse to that of the CW power of 1 mW. (c) What are the required conditions for the laser to be nearly ideally gain switched? (d) What are the characteristics of the ideally gain-switched pulse?

Solution (a) Because the laser parameters in the lasing phase are the same as those of the CW laser described in Examples 11.1–11.5, we have $\gamma_{cl} = \gamma_c = 5.78 \times 10^8\ s^{-1}$ and $\tau_{cl} = \tau_c = 1.73$ ns from Example 11.1. Two of the three conditions for ideal Q switching, namely, a short laser cavity and a gain medium with a long fluorescence lifetime and a low saturation intensity, are already met by this laser. Therefore, the only requirement that has to be considered is an effective Q switch that switches the cavity from a high loss of γ_{cp} to a low loss of γ_{cl}. First, γ_{cp} has to be larger than Γg_i, which for a pumping ratio of $r = 1.85$ is $\Gamma g_i = r\gamma_{cl} = 1.85\gamma_{cl}$. Thus, an effective Q switch has to keep $\gamma_{cp} > 1.85\gamma_{cl} = 1.07 \times 10^9\ s^{-1}$. Next, the Q switch has to switch fast enough. Quantitatively, the Q switch has to switch the cavity loss from the high value of γ_{cp} to the low value of γ_{cl} within a time interval Δt_{QS} that is shorter than the pulse delay time t_d, as shown in Fig. 11.12, for the process to qualify as fast Q switching.

The pulse delay time can be estimated by considering the fact that the pulse grows from a seed of spontaneous emission to the saturation photon density exponentially with a rate of $\Gamma g_i - \gamma_{cl}$. The saturation photon density is S_{sat}, which has a value of $S_{sat} = 8.15 \times 10^{17}\ m^{-3}$ found in Example 11.4 for this laser. The seed of spontaneous emission is one photon per mode, which translates into a spontaneous photon density of $1/\mathcal{V}_{mode}$, with $\mathcal{V}_{mode} = 1.57 \times 10^{-11}\ m^3$, also found in Example 11.4 for this laser. If we take t_d to be the time it takes the photon density of the oscillating laser mode to grow exponentially with a rate of $\Gamma g_i - \gamma_{cl}$ from $1/\mathcal{V}_{mode}$ to S_{sat}, then t_d can be found as

$$t_d = \frac{1}{\Gamma g_i - \gamma_{cl}} \ln \frac{S_{sat}}{1/\mathcal{V}_{mode}} = \frac{\tau_{cl}}{r-1} \ln(S_{sat}\mathcal{V}_{mode})$$

$$= \frac{1.73}{1.85-1} \ln(8.15 \times 10^{17} \times 1.57 \times 10^{-11})\ \text{ns} = 33.3\ \text{ns}.$$

Therefore, the requirements for ideal Q switching of this laser at the given pumping ratio of $r = 1.85$ are $\gamma_{cp} > 1.07 \times 10^9\ s^{-1}$ and $\Delta t_{QS} \ll 33.3$ ns.

(b) From Example 11.4, we find that $P_{out}^{sat} = 1.18$ mW and $\tau_2 = 240\ \mu s$ for this laser. Therefore, from (11.94), the peak power of the ideally Q-switched pulse is

$$P_{pk} \approx \frac{\tau_2}{\tau_{cl}}(r - \ln r - 1)P_{out}^{sat}$$
$$= \frac{240 \times 10^{-6}}{1.73 \times 10^{-9}} \times (1.85 - \ln 1.85 - 1) \times 1.18 \times 10^{-3}\ \text{W} = 38.4\ \text{W}.$$

Compared to the 1 mW output power of the laser in CW operation, the peak power of this Q-switched pulse is 3.84×10^4 times higher primarily because of the fact that τ_2 is five orders of magnitude larger than τ_{cl}. This demonstrates that a gain medium that has a large τ_2 makes a good Q-switched laser.

The pulsewidth is found from (11.95) to be

$$\Delta t_{ps} = \frac{2.5}{(r - \ln r - 1)^{1/2}}\tau_{cl} = \frac{2.5}{(1.85 - \ln 1.85 - 1)^{1/2}} \times 1.73\ \text{ns} = 8.93\ \text{ns}.$$

Compared to the cavity round-trip time of $T = 6.07$ ps found in Example 11.1, which sets the ultimate lower limit for the pulsewidth of a Q-switched pulse, this pulsewidth is quite long. It can be shortened by pumping the laser higher to increase the pumping ratio r and by using an output-coupling mirror of a lower reflectivity to reduce τ_{cl}.

The pulse energy is simply

$$U \approx P_{pk}\Delta t_{ps} = 38.4 \times 8.93 \times 10^{-9}\ \text{J} = 343\ \text{nJ}.$$

This pulse energy is not very high because of the small amount of energy that can be stored in the small gain volume of the microchip laser. To increase the Q-switched pulse energy, one must increase the volume of the gain medium as well as the mode volume of the oscillating laser field.

(c) Two of the three conditions for ideal gain switching are the same as those for ideal Q switching, which are already met by this laser. The only condition remaining to be considered is a very short and strong pump for gain switching. Whether a pump pulse is short or not is relative to the fluorescence lifetime τ_2 of the gain medium. Because τ_2 is the relaxation time constant of the excited population in the upper laser level, the pump energy can be efficiently stored in the population inversion of the gain medium if the pump-pulse duration is much smaller than τ_2. If τ_2 is much smaller than the pump-pulse duration, then the pump energy cannot be efficiently stored in the population inversion of the gain medium because population relaxation during the pumping process is significant. From this discussion, we understand that for ideal gain switching, it is necessary that the pump pulse be much shorter than τ_2 of the gain medium. It does not have to be extremely short, however. A pump pulse that has a duration of $\tau_2/10$ is short enough, while one that has a duration of $\tau_2/100$ is close to an ideal delta pulse pump. For ideal gain switching of this Nd : YAG laser with $\tau_2 = 240\ \mu s$, we need a short pump

pulse of a few microseconds or less in duration that has a sufficiently high energy to pump the laser to the desired pumping ratio of $r = 1.85$ in such a short duration.

(d) The characteristics of an ideally gain-switched pulse are the same as those of an ideally Q-switched pulse found in (b). The only difference is in the pump pulse. Ideal Q switching can be accomplished with a relatively long pump pulse so long as the Q switch satisfies the conditions discussed in (a).

Mode locking

Mode locking is the most important technique for the generation of repetitive, ultrashort laser pulses. The principle of mode locking is very different from those of gain switching and Q switching in that it is not based on the transient dynamics of a laser. Instead, a mode-locked laser operates in a dynamic steady state.

A pulsed laser can oscillate in multiple longitudinal modes regardless of whether the gain medium is homogeneously or inhomogeneously broadened. Mode locking refers to the situation when all of the oscillating longitudinal modes of a laser are locked in phase. When this phase locking is accomplished, constructive interference of all of the oscillating modes results in a short pulse circulating inside the cavity, which is regeneratively amplified by the gain medium after periodically delivering an output pulse through an output-coupling mirror in each round trip. The mode-locking operation is accomplished by a nonlinear optical element known as the *mode locker* that is placed inside the laser cavity, typically near one end of the cavity if the laser has the configuration of a linear cavity. Viewed in the frequency domain, mode locking is a process that generates a train of short laser pulses by locking multiple longitudinal laser modes in phase. The function of the mode locker in the frequency domain is thus to lock the phases of the oscillating modes together through nonlinear interactions among the mode fields. In the time domain, the mode-locking process can be understood as a regenerative pulse-generating process by which a short pulse circulating inside the laser cavity is formed when the laser reaches steady state. The action of the mode locker in the time domain resembles that of a pulse-shaping optical shutter that opens periodically in synchronism with the arrival at the mode locker of the laser pulse circulating in the cavity. Consequently, the output of a mode-locked laser is a train of regularly spaced pulses of identical pulse envelope.

The simplest case of multimode oscillation is when there are only two oscillating longitudinal modes of frequencies ω_1 and ω_2. Then, the total laser field at a fixed location is

$$E(t) = \mathcal{E}_1 e^{i\varphi_1(t)} e^{-i\omega_1 t} + \mathcal{E}_2 e^{i\varphi_2(t)} e^{-i\omega_2 t}, \tag{11.97}$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are the magnitudes of the field amplitudes and φ_1 and φ_2 are the phases. With all the phase information included in φ_1 and φ_2, $\mathcal{E}_1$ and $\mathcal{E}_2$ are positive,

real quantities. The intensity of this laser is given by

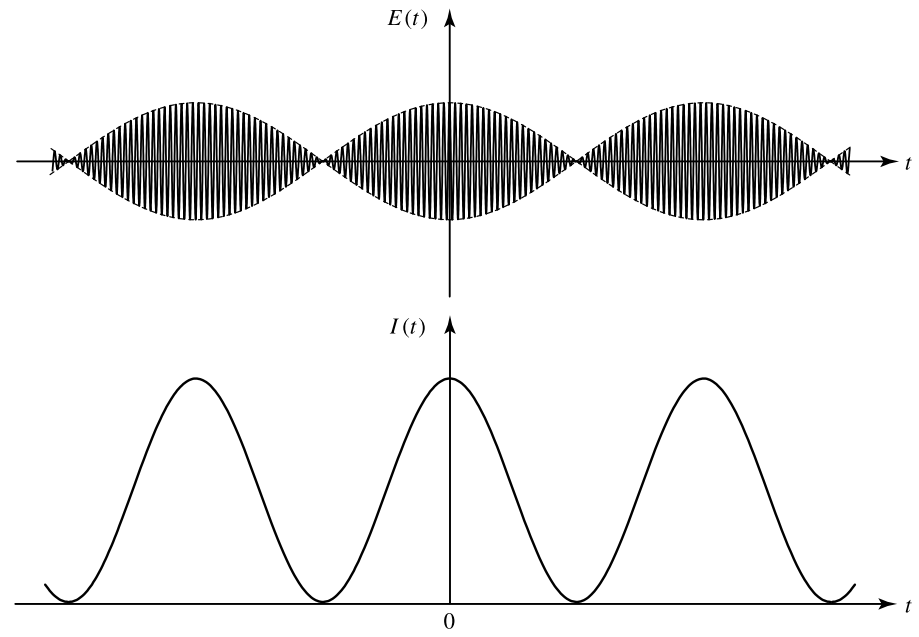
$$I(t) = 2c\epsilon_0 n|E(t)|^2$$
$$= 2c\epsilon_0 n\left\{\mathcal{E}_1^2 + \mathcal{E}_2^2 + 2\mathcal{E}_1\mathcal{E}_2\cos[(\omega_1 - \omega_2)t - \varphi_1(t) + \varphi_2(t)]\right\}. \tag{11.98}$$

In general, the phases can vary with time. If $\varphi_1(t)$ and $\varphi_2(t)$ vary randomly with time on a characteristic time scale that is shorter than $2\pi/(\omega_1 - \omega_2)$, the beat note of the two frequencies cannot be observed even with a very fast detector. In this situation, the output of the laser has a constant intensity that is the incoherent sum of the intensities of the individual modes. This situation simply represents the ordinary multimode oscillation of a CW laser. If φ_1 and φ_2 are time independent, the laser intensity given in (11.98) becomes periodically modulated with a period of $2\pi/(\omega_1 - \omega_2)$ defined by the beat frequency, as shown in Fig. 11.13(a). The modulation depth of this intensity profile depends on the ratio between $\mathcal{E}_1$ and $\mathcal{E}_2$. When $\mathcal{E}_1 = \mathcal{E}_2$, the modulation depth is 100% with $I_{\min} = 0$. In this situation, $I(t)$ resembles a train of periodic "pulses" that have a duty cycle of 50% and a peak intensity of twice the average intensity. This is simply *coherent mode beating* and is the best one can do with two oscillating modes.

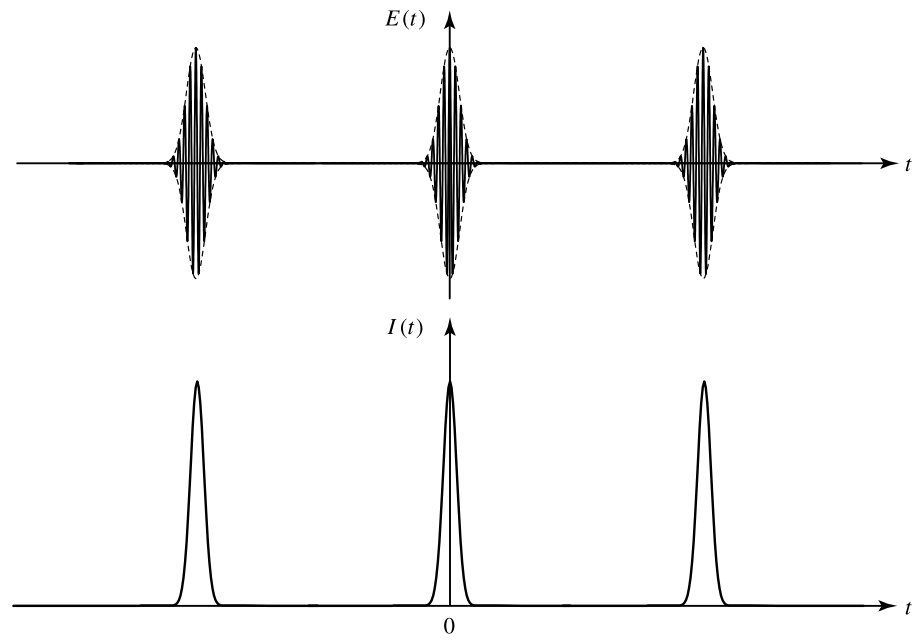
The periodic intensity profile created by two-mode beating, which is shown in Fig. 11.13(a), is certainly far from what we normally expect from a train of mode-locked pulses. As the number of modes that are locked in phase increases, the characteristics of periodic pulses become increasingly apparent in the output of the laser, as demonstrated in Fig. 11.13(b). At a given pulse repetition rate, we can reduce the pulse duty cycle, thus shortening the pulsewidth and simultaneously increasing the peak pulse intensity, by increasing the number of modes. In general, a practically useful mode-locked laser oscillates in a large number of modes. For a laser of many oscillating modes, we have the following laser field:

$$E(t) = \sum_q \mathcal{E}_q e^{i\varphi_q} e^{-i\omega_q t}, \tag{11.99}$$

where again $\mathcal{E}_q$ are taken to be positive, real quantities representing the magnitudes of the field amplitudes, and the summation is taken over all of the oscillating modes. As discussed above, if the phases φ_q vary randomly with time, (11.99) describes the field of a CW multimode laser, which is of no interest here. For mode locking, we consider the situation when φ_q are time independent. In the case of only two oscillating modes with φ_1 and φ_2 being time-independent constants, the phase difference $\varphi_1 - \varphi_2$ merely shifts the mode beating pattern with respect to the origin of the time axis and is of no physical significance. With more than two oscillating modes, however, only one phase can be arbitrary because the relative phases among different modes are significant. Consequently, the temporal characteristics of the combined laser field described in (11.99) depend on the phase relationships among the oscillating modes, as well as on the distribution of the field magnitudes $\mathcal{E}_q$ and the frequency spacing between neighboring modes.



(*a*)



(*b*)

Figure 11.13 (*a*) Field and intensity variations of a laser caused by beating between two longitudinal modes of constant phases. (*b*) Field and intensity variations of a laser with multiple longitudinal modes locked in phase.

We consider the situation when the oscillating laser modes are equally spaced with a longitudinal mode spacing of $\Delta\omega_{\mathrm{L}}$. The magnitudes and phases of the mode fields are functions of the mode frequencies, but not all of the phases vary with time. Their spectral distribution can be described by a complex spectral envelope function $\mathcal{E}(\omega)$ through

$$\mathcal{E}_q \mathrm{e}^{\mathrm{i}\varphi_q} = \frac{\Delta\omega_{\mathrm{L}}}{2\pi}\mathcal{E}(\omega_q - \omega_0). \tag{11.100}$$

For simplicity, we have chosen ω_0 to be a longitudinal mode frequency near the center of the spectrum. Thus we have $\omega_q = \omega_0 + n\Delta\omega_{\mathrm{L}}$, and the total field in (11.99) can then be transformed as follows:

$$\begin{aligned}
E(t) &= \frac{\Delta\omega_{\mathrm{L}}}{2\pi}\sum_{q=-\infty}^{\infty}\mathcal{E}(\omega_q - \omega_0)\mathrm{e}^{-\mathrm{i}\omega_q t} \\
&= \frac{\Delta\omega_{\mathrm{L}}}{2\pi}\mathrm{e}^{-\mathrm{i}\omega_0 t}\sum_{n=-\infty}^{\infty}\mathcal{E}(n\Delta\omega_{\mathrm{L}})\mathrm{e}^{-\mathrm{i}n\Delta\omega_{\mathrm{L}} t} \\
&= \frac{\Delta\omega_{\mathrm{L}}}{2\pi}\mathrm{e}^{-\mathrm{i}\omega_0 t}\mathcal{F}^{-1}\mathcal{F}\left\{\sum_{n=-\infty}^{\infty}\mathcal{E}(n\Delta\omega_{\mathrm{L}})\mathrm{e}^{-\mathrm{i}n\Delta\omega_{\mathrm{L}} t}\right\} \\
&= \Delta\omega_{\mathrm{L}}\mathrm{e}^{-\mathrm{i}\omega_0 t}\mathcal{F}^{-1}\left\{\sum_{n=-\infty}^{\infty}\mathcal{E}(n\Delta\omega_{\mathrm{L}})\delta(\omega - n\Delta\omega_{\mathrm{L}})\right\} \\
&= \mathrm{e}^{-\mathrm{i}\omega_0 t}\mathcal{F}^{-1}\left\{\mathcal{E}(\omega)\cdot\Delta\omega_{\mathrm{L}}\sum_{n=-\infty}^{\infty}\delta(\omega - n\Delta\omega_{\mathrm{L}})\right\} \\
&= \mathrm{e}^{-\mathrm{i}\omega_0 t}\mathcal{E}(t) * \sum_{m=-\infty}^{\infty}\delta(t - mT) \\
&= \mathrm{e}^{-\mathrm{i}\omega_0 t}\sum_{m=-\infty}^{\infty}\mathcal{E}(t - mT),
\end{aligned} \tag{11.101}$$

where $T = 2\pi/\Delta\omega_{\mathrm{L}}$, and

$$\mathcal{E}(t) = \mathcal{F}^{-1}\{\mathcal{E}(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{E}(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}\omega. \tag{11.102}$$

In (11.101) and (11.102), $\mathcal{F}\{\cdot\}$ means taking the Fourier transform from the time domain to the frequency domain, and $\mathcal{F}^{-1}\{\cdot\}$ means taking the inverse Fourier transform from the frequency domain back to the time domain.

The result in (11.101) is obtained under the assumption that the phases φ_q do not vary with time. It shows that when phases φ_q do not vary with time, the total field $E(t)$ is a periodic function of time with a period T determined by the mode spacing and a temporal profile $\mathcal{E}(t)$ determined by the spectral envelope. Figure 11.14 shows the spectral and temporal characteristics of the field and intensity profiles of a completely

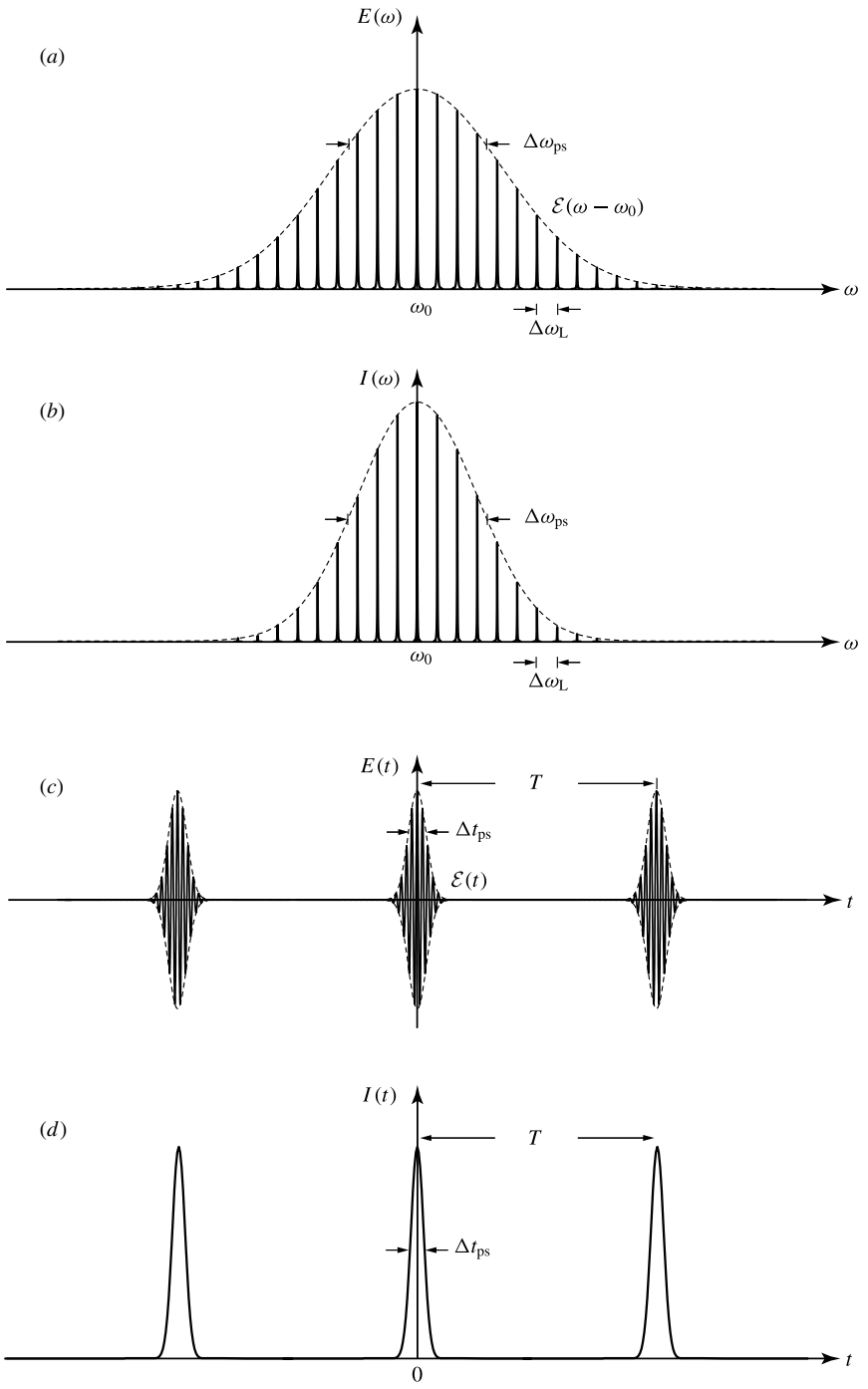


Figure 11.14 (*a*) Spectral field distribution, (*b*) spectral intensity distribution, (*c*) temporal field variation, and (*d*) temporal intensity variation of a completely mode-locked laser. For simplicity, it is assumed that the center of the spectral distribution coincides with the frequency of a longitudinal mode. Note that $\Delta\omega_{ps}$ and Δt_{ps} are defined as the FWHMs of $I(\omega)$ and $I(t)$, respectively.

mode-locked laser, in which all of the longitudinal modes are locked to the same phase. The spectral width, $\Delta\omega_{\rm ps}$, of a laser pulse is defined as the FWHM of the spectral intensity distribution, $I(\omega)$, as shown in Fig. 11.14(*b*). Correspondingly, the temporal pulsewidth, $\Delta t_{\rm ps}$, is defined as the FWHM of the temporal intensity profile of an individual pulse, as illustrated in Fig. 11.14(*d*). Because of the Fourier-transform relationship, given in (11.102), between the temporal field profile, $\mathcal{E}(t)$, and the spectral field profile, $\mathcal{E}(\omega)$, the temporal and spectral widths of a pulse are subject to the following relation:

$$\Delta\nu_{\rm ps}\Delta t_{\rm ps} \geq K, \tag{11.103}$$

where $\Delta\nu_{\rm ps} = \Delta\omega_{\rm ps}/2\pi$ and K is a constant of the order of unity that depends on the pulse shape. For any pulse with a given pulse shape, the best one can hope for is $\Delta\nu_{\rm ps}\Delta t_{\rm ps} = K$. When this is accomplished, the pulse is said to be *Fourier-transform limited*, or simply *transform limited*. A transform-limited pulse is one that has the smallest pulsewidth $\Delta t_{\rm ps} = K/\Delta\nu_{\rm ps}$ for a given pulse spectral width $\Delta\nu_{\rm ps}$.

Two pulse shapes are of most interest for mode-locked lasers. One is the Gaussian pulse, and the other is the sech^2 pulse. For the Gaussian pulse, both $\mathcal{E}(\omega)$ and $\mathcal{E}(t)$ are Gaussian functions because the Fourier transform of a Gaussian function is another Gaussian function, and both its temporal intensity profile and spectral intensity profile are also Gaussian. For a sech^2 pulse, both $\mathcal{E}(\omega)$ and $\mathcal{E}(t)$ are its sech functions because the Fourier transform of a sech function is another sech function, and both its temporal intensity profile and its spectral intensity profile are sech^2 functions. The transform-limit constants are $K = 2\ln 2/\pi = 0.4413$ for a Gaussian pulse and $K = 4\ln^2(1+\sqrt{2})/\pi^2 = 0.3148$ for a sech^2 pulse. Actively mode-locked pulses tend to have Gaussian shapes, whereas passively mode-locked pulses often have sech^2 shapes.

When all of the modes of a laser are locked to a common phase, we can set $\varphi_q = \varphi_0 = 0$ because a constant common phase has no physical significance. This is the ideal situation of *complete mode locking*. From (11.100), we find that the spectral envelope is a real function when $\varphi_q = 0$. This implies that $\mathcal{E}(\omega) = \mathcal{E}^*(\omega)$ and

$$\begin{aligned}\mathcal{E}(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{E}(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{E}^*(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \\ &= \mathcal{E}^*(-t).\end{aligned} \tag{11.104}$$

Therefore, $I(t) = I(-t)$ if the laser pulse is completely mode locked.

From the above discussions, it can be concluded that *a completely mode-locked laser pulse has a symmetric temporal intensity profile and is transform limited. It does not necessarily have a symmetric spectral intensity profile, but an asymmetric temporal*

pulse shape or a deviation from the transform limit signifies incomplete mode locking. The reverse is not true, however, because *a transform-limited pulse is not necessarily completely mode locked.* When the longitudinal laser modes are not locked in phase, pulses can still be formed, but with less than ideal characteristics.

A completely mode-locked pulse, being transform limited, satisfies the condition $\Delta t_{\text{ps}} \Delta \nu_{\text{ps}} = K$. Because the number of oscillating modes can be estimated with

$$N \approx \frac{\Delta \nu_{\text{ps}}}{\Delta \nu_{\text{L}}}, \tag{11.105}$$

the temporal width of a mode-locked pulse is inversely proportional to the number of oscillating modes:

$$\Delta t_{\text{ps}} \approx \frac{K}{N \Delta \nu_{\text{L}}} = K \frac{T}{N}. \tag{11.106}$$

At a fixed longitudinal mode spacing $\Delta \nu_{\text{L}}$, hence a fixed pulse repetition rate $f_{\text{ps}} = 1/T$, the pulsewidth can be shortened by increasing the number of oscillating modes. It is common to expect a pulsewidth that is two to five orders of magnitude smaller than the pulse spacing in a train of mode-locked pulses. The relation in (11.106) indicates that this requires locking of hundreds to hundreds of thousands of oscillating modes.

An inhomogeneously broadened laser naturally oscillates in multiple longitudinal modes. In such a laser, the mode locker only has to lock these modes in phase to produce a train of mode-locked pulses. However, many mode-locked lasers that produce ultrashort pulses are homogeneously broadened. In the free-running steady state of a homogeneously broadened laser, only one longitudinal mode will oscillate because of homogeneous saturation across the gain spectrum. Even though it is possible to force multimode oscillation in a homogeneously broadened laser when it is pulsed, the homogeneously broadened gain medium has a natural tendency to narrow the spectral bandwidth of the oscillating laser field. Therefore, besides locking the phases of the oscillating laser modes together, the mode locker has the function of expanding the spectral width of the laser pulse to counteract the spectral narrowing effect of the gain medium.

For the pulses generated by a given mode-locked laser, the pulse spectral bandwidth $\Delta \nu_{\text{ps}}$ is ultimately limited by the spontaneous linewidth $\Delta \nu$ of the gain medium because $\Delta \nu$ sets the limit for the gain bandwidth of the laser. Therefore, the mode-locked pulses that can be generated from a given laser, regardless of whether it is homogeneously or inhomogeneously broadened, are subject to the following absolute limitation:

$$\Delta t_{\text{ps}} \geq \frac{K}{\Delta \nu_{\text{ps}}} \geq \frac{K}{\Delta \nu}, \tag{11.107}$$

where $\Delta \nu$ has the values listed in Table 10.1 for many representative laser gain media. For most mode-locked lasers, only a fraction of the laser gain bandwidth is utilized so that $\Delta \nu_{\text{ps}}$ is only a fraction of $\Delta \nu$. This fraction of bandwidth utilization depends on a number of operating parameters, including the modulation strength and the modulation

frequency of the mode locker, as well as the type of mode locker used. Increasing this fraction is the key to reducing the temporal pulsewidth of mode-locked pulses.

A continuously mode-locked laser delivers a steady train of short pulses at a constant average output power $\overline{P}$, while each pulse has a high peak power P_{pk}. Effectively, the energy of laser output in each pulse repetition period T is concentrated within the duration of the pulsewidth Δt_{ps}. Therefore, the peak power of the pulses is enhanced over the average laser power by a factor of $T/\Delta t_{\text{ps}}$ in accordance with

$$P_{\text{pk}} = K' \frac{T}{\Delta t_{\text{ps}}} \overline{P} = K' \frac{\overline{P}}{f_{\text{ps}} \Delta t_{\text{ps}}} = \frac{K'}{K} N \overline{P}, \tag{11.108}$$

where K' is a constant of the order of unity that depends on the pulse shape. For a Gaussian pulse $K' = 2\sqrt{\ln 2}/\sqrt{\pi} = 0.9394$. For a sech^2 pulse, $K' = \ln(1 + \sqrt{2}) = 0.8814$. From (11.108), we see that the enhancement of the pulse peak power over the average power is proportional to the number of locked modes.

EXAMPLE 11.7 By properly incorporating a suitable mode locker in the laser cavity, a CW Nd : YAG laser can often be mode locked with little additional loss, thus maintaining average power while delivering a regular train of ultrashort laser pulses. A mode-locked Nd : YAG laser consists of a Nd : YAG gain medium that has a spontaneous linewidth of $\Delta\nu = 150$ GHz in a Fabry–Perot cavity that has a round-trip optical path length of $l_{\text{RT}} = 2$ m. The laser is continuously pumped to have an average output power of $\overline{P} = 2$ W. The mode locker used in this laser generates pulses of Gaussian temporal and spectral shapes. (a) What is the repetition rate of the mode-locked pulses? Does it vary with pulsewidth or laser output power? (b) If transform-limited pulses of $\Delta t_{\text{ps}} = 100$ ps are generated, how much of the bandwidth of the gain medium is utilized? How many longitudinal modes should oscillate and be locked to generate such pulses? (c) What is the peak power of the pulses? (d) What is the pulsewidth of the shortest pulses that can possibly be generated from this laser? Under what conditions can such pulses be generated?

Solution (a) The pulse repetition rate is determined by the cavity round-trip time T, which in turn is determined by the round-trip optical path length l_{RT}. Therefore,

$$f_{\text{ps}} = \frac{1}{T} = \frac{c}{l_{\text{RT}}} = \frac{3 \times 10^8}{2} \text{ Hz} = 150 \text{ MHz}.$$

It does not vary with either pulsewidth or laser output power. We also find from this result that $T = 6.7$ ns.

(b) For transform-limited Gaussian pulses of $\Delta t_{\text{ps}} = 100$ ps, we have

$$\Delta\nu_{\text{ps}} = \frac{K}{\Delta t_{\text{ps}}} = \frac{0.4413}{100 \times 10^{-12}} \text{ Hz} = 4.413 \text{ GHz}.$$

Because $\Delta\nu = 150$ GHz, we have $\Delta\nu_{ps}/\Delta\nu = 4.413/150 = 2.94\%$. Therefore, only 2.94% of the bandwidth of the gain medium is used. The longitudinal mode spacing is simply the same as the pulse repetition rate: $\Delta\nu_L = f_{ps} = 150$ MHz. The number of oscillating modes that are locked to generate these pulses can be found from (11.105):

$$N = \frac{\Delta\nu_{ps}}{\Delta\nu_L} = \frac{4.413 \times 10^9}{150 \times 10^6} = 30.$$

Only 30 oscillating modes are required because the pulsewidth of 100 ps is relatively long for mode-locked pulses in a cavity that has a round-trip time of $T = 6.7$ ns.

(c) The peak power of these Gaussian pulses can be found by using (11.108) with $K' = 0.9394$:

$$P_{pk} = K'\frac{T}{\Delta t_{ps}}\overline{P} = 0.9394 \times \frac{6.67 \times 10^{-9}}{100 \times 10^{-12}} \times 2 \text{ W} = 125 \text{ W}.$$

This peak power is only about 63 times the average power because of the moderate value of the $T/\Delta t_{ps}$ ratio and the correspondingly small number of oscillating modes.

(d) The pulsewidth is ultimately limited by the condition given in (11.107). For Gaussian pulses, $K = 0.4413$. Thus, the shortest pulses that can be generated from this laser have the following pulsewidth:

$$\Delta t_{ps}^{min} = \frac{K}{\Delta\nu} = \frac{0.4413}{150 \times 10^9} \text{ s} = 2.94 \text{ ps}.$$

Such pulses are generated under the following conditions: (1) the entire bandwidth of the laser gain medium is utilized so that $\Delta\nu_{ps} = \Delta\nu$, and (2) the pulses are transform limited so that $\Delta t_{ps}\Delta\nu_{ps} = K$. To utilize the entire bandwidth of the laser gain medium is not a simple matter. Aside from pumping the laser sufficiently to realize its entire potential gain bandwidth, it requires that all optical elements in the laser cavity, including the mirrors and the mode locker, have bandwidths larger than $\Delta\nu$. It also requires that the mode-locking mechanism be strong enough to force all modes across the entire bandwidth to oscillate and lock in phase. The generation of transform-limited pulses is not a trivial matter, either. It requires elimination or compensation of all possible sources of dispersion in the laser while using an effective mode-locking scheme to lock all oscillating modes perfectly in phase.

In comparison to the transient techniques of gain switching and Q switching, the requirements and optimum conditions for operation of a mode-locked laser that is regeneratively pulsed are very different. Because a *regeneratively pulsed laser* does not function by control of the laser transient, it does not depend on rapid depletion of intracavity photons to generate a short pulse, as do *transiently pulsed lasers*. Consequently, it does not require a very short photon lifetime and a correspondingly short laser cavity.

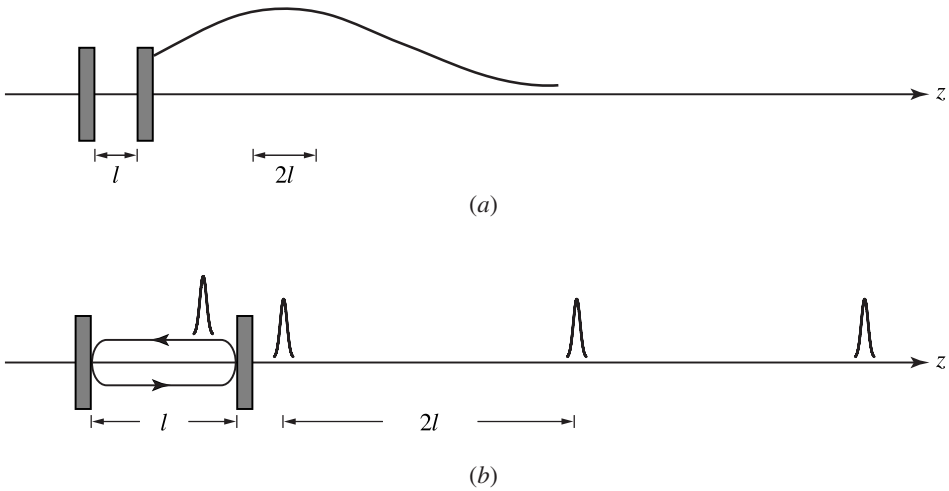


Figure 11.15 Comparison between (*a*) a transiently pulsed laser such as a gain-switched or Q-switched laser and (*b*) a regeneratively pulsed laser such as a mode-locked laser.

On the contrary, it requires a sufficiently long cavity for the pulse to fit in and circulate around. In general, a pulse generated by the transient technique of gain switching or Q switching cannot be shorter than the cavity round-trip time. Therefore, it has a spatial span longer than the cavity length. In contrast, a pulse generated by a regenerative approach such as mode locking always has a spatial span much shorter than the cavity length and can have a pulsewidth significantly shorter than the cavity round-trip time. This comparison is illustrated in Fig. 11.15. In a transiently pulsed laser, photon energy is distributed throughout the laser cavity, and a pulse is generated through fast temporal evolution of this distributed energy. As a result, the laser can be modeled as a lumped device. In a regeneratively pulsed laser, however, the photon energy is localized and circulates in the cavity. Therefore, a mode-locked laser cannot be modeled as a lumped device.

Another difference between a transiently pulsed laser and a regeneratively pulsed laser is the characteristic requirements of the gain medium. A transiently pulsed laser requires a long fluorescence lifetime and prefers to have it as long as possible. The fluorescence lifetime τ_2 varies among different types of regeneratively pulsed lasers, but in general a particularly large τ_2 is not required. A *synchronously pumped laser* can successfully operate on a gain medium that has a very small τ_2 or even one that has no energy-storage mechanism such as in the case of a synchronously pumped OPO. In certain mode-locked systems, τ_2 is preferred to be sufficiently large but not so large as to cause competition between transient oscillation and the buildup of mode-locked pulses. Therefore, the τ_2 requirement of a mode-locked laser is a sophisticated issue that depends on the specific type and mechanism of the mode-locking operation.

Although repetitive pulses can also be generated by the repetitive operation of transiently pulsed lasers, a regeneratively pulsed laser offers certain advantages. The pulses

generated from a regeneratively pulsed laser do not have to build up from noise once the laser has reached steady state. In the steady state, the pulses bear no relation to the initial noise from which they have developed. Thus these pulses tend to have much better characteristics than those generated by transient techniques. The pulses are usually very smooth and maintain a very high degree of coherence from one to another over a long period of time, making them very useful for many time-resolved spectroscopic applications. They are not affected by the transient effects, such as relaxation oscillations, of transiently pulsed lasers. In a mode-locked laser, the regeneratively generated pulses can be transform limited if all of the oscillating modes are completely locked in phase. Finally, a regeneratively pulsed laser can generate much shorter pulses at a higher repetition rate than a transiently pulsed laser of the same gain medium can.

Similarly to Q switching, mode locking can also be either active or passive, depending on the type of mode locker used. In an actively mode-locked laser, operation of the mode locker is controlled by an externally applied signal. In a passively mode-locked laser, the mode locker functions directly in response to the optical field in the laser cavity through a nonlinear optical mechanism. Mode locking can take the form of either periodic loss modulation or periodic gain modulation. The most important mode-locking techniques are illustrated in Fig. 11.16.

For active mode locking with loss modulation, the most commonly employed technique is acousto-optic modulation with an externally applied RF signal, as shown in Fig. 11.16(*a*). An acousto-optic modulator used for mode locking is different from one used for Q switching: an acousto-optic mode locker is a standing-wave Bragg diffractor that is turned on continuously, but an acousto-optic Q switch is a traveling-wave Bragg diffractor that is turned on and off to switch the cavity Q between different values. A very important technique, known as *synchronous pumping* and illustrated in Fig. 11.16(*b*), for generating ultrashort laser pulses can be considered as active mode locking by gain modulation. For synchronous pumping, the gain medium in the laser cavity is localized and placed near one end of the cavity and is pumped periodically, either optically or electrically, with a train of very short pulses at the same repetition rate as that of the periodic arrival at the gain medium of the pulse circulating inside the laser cavity.

Passive mode locking can be accomplished by using a saturable absorber localized and placed near one end of a linear laser cavity, as shown in Fig. 11.16(*c*). Sometimes, a saturable absorber used for passive Q switching can also be used for passive mode locking, but in general the requirements for passive mode locking are very different from those for passive Q switching. Passive mode locking with a saturable absorber can be arranged in a ring configuration shown in Fig. 11.16(*d*) for *colliding-pulse mode locking*. In this mode-locking scheme, there are two intracavity laser pulses that circulate in opposite directions and collide at the saturable absorber to enhance the pulse-shortening function of the saturable absorber. Passive mode locking can also be accomplished without the use of a saturable absorber by employing nonlinear refractive

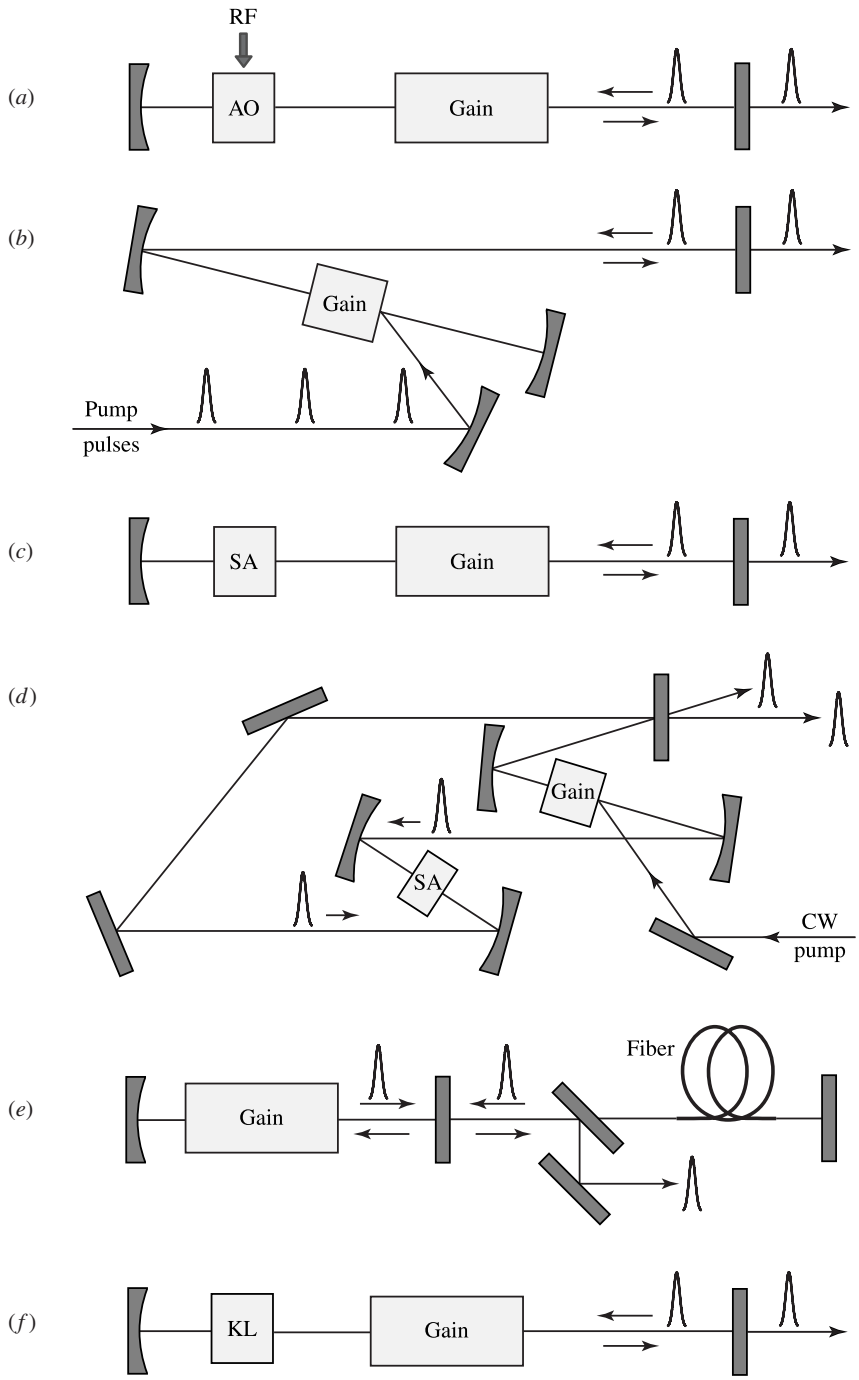


Figure 11.16 Representative mode-locking techniques: (*a*) active mode locking with an acousto-optic modulator, (*b*) synchronous pumping, (*c*) passive mode locking with a saturable absorber, (*d*) colliding-pulse mode locking, (*e*) additive-pulse mode locking, and (*f*) Kerr-lens mode locking. AO represents an acousto-optic modulator. SA represents a saturable absorber. KL represents a Kerr lens.

index changes through the real part of $\chi^{(3)}$, or even $\chi^{(2)}$, of a nonlinear optical element. Two very important concepts belonging to this category are *additive-pulse mode locking* and *Kerr-lens mode locking*, which are illustrated in Figs. 11.16(e) and (f), respectively.

Mode locking has been applied to a wide variety of laser materials to generate laser pulses with pulsewidths ranging from the order of 10 fs to the order of 1 ns. For a given laser material, passive mode locking typically generates shorter pulses than active mode locking, but active mode locking often produces pulses with less fluctuation and jitter. Some systems combine active mode locking with passive mode locking in a form of hybrid mode locking to realize the advantages of both. It is also possible to combine mode locking with a transient pulsing technique. In this situation, the laser does not reach a complete steady state. An important example of this possibility is the operation of *Q-switched mode-locked lasers* by combining Q switching and mode locking. Because the transiently Q-switched pulse has a duration longer than the cavity round-trip time, the result is a finite train of equally spaced mode-locked pulses with unequal amplitudes under a Q-switched envelope.

EXAMPLE 11.8 Nd : YAG lasers can undertake all modes of laser operation, including CW, gain-switching, Q-switching, and mode-locking operations, as Examples 11.1–11.7 illustrate. With the exception of synchronous pumping, almost all other mode-locking techniques can be successfully employed to mode lock Nd : YAG lasers, either in a pure form of CW mode locking or in a hybrid form that combines Q switching with mode locking, or otherwise. These being the facts, however, the microchip Nd : YAG laser with its cavity parameters described in Examples 11.1–11.6 cannot be mode locked by any means. Give quantitative reasons for this problem.

Solution First, consider the fact that the longitudinal mode spacing of this microchip laser is $\Delta\nu_{\mathrm{L}} = 164.8$ GHz, as found in Example 11.3, while the entire linewidth of the Nd : YAG plate used for this laser is only $\Delta\nu = 150$ GHz. Although a homogeneously broadened laser can oscillate in multiple longitudinal modes when the laser is mode locked, as discussed in the text above, this microchip laser can only oscillate in a single longitudinal mode regardless of how it is being operated because $\Delta\nu_{\mathrm{L}} > \Delta\nu$, not because it is homogeneously broadened. Clearly, there is no possibility of mode locking if a laser can only oscillate in one longitudinal mode.

We can see the problem from another angle in the time domain. According to the illustration in Fig. 11.15(b) and the discussions in the text, a mode-locked pulse must have a spatial span that is much shorter than the cavity length to allow it to circulate inside the cavity as a regenerative pulse. For a laser with a linear Fabry–Perot cavity such as the microchip laser under consideration, the mode-locked pulse has to fit loosely into the length of the cavity to allow it to circulate inside without wrapping itself up, thus having a pulsewidth that is much shorter than one-half of the cavity round-trip time: $\Delta t_{\mathrm{ps}} \ll T/2$.

From Example 11.1, we find that $T = 6.07$ ps for this laser. Therefore, any mode-locked pulse that can possibly be generated from this laser has a pulsewidth subject to the limitation $\Delta t_{ps} \ll 3.03$ ps. However, the pulsewidth of a mode-locked pulse is also subject to the limitation given in (11.107). With $\Delta\nu = 150$ GHz, we have $\Delta t_{ps} > 2.94$ ps according to the calculation in Example 11.7 if the pulse has a Gaussian shape. These two conflicting limitations cannot be satisfied simultaneously, thus excluding any possibility of mode locking this laser.

11.5 Optical fiber lasers

A fiber laser can be constructed by simply creating some form of optical feedback to a fiber amplifier. Nevertheless, while most interest in fiber amplifiers has concentrated on the 1.3- and 1.55-μm spectral regions for optical communication systems, the development of fiber lasers has covered a broad spectral range, from a holmium-doped fiber laser at 550 nm and a praseodymium-doped fiber laser at 610 nm in the visible spectral region to an erbium-doped fiber laser at 2.7 μm and a holmium-doped fiber laser at 2.9 μm in the infrared region. Besides, the active ions used for fiber lasers include almost all rare-earth ions doped in either silica or fluoride glass fibers.

The operation of a fiber laser follows the general laser principles discussed in earlier sections. Besides CW oscillation, fiber lasers can also be Q switched or mode locked to deliver very short and intense laser pulses. The geometry and the waveguiding nature of the fiber gain medium, however, lead to many unique configurations, along with some special characteristics, for fiber lasers.

Several different cavity configurations for fiber lasers are shown in Fig. 11.17. The most straightforward configuration, shown in Fig. 11.17(a), is a Fabry–Perot cavity created by attaching a dichroic mirror to each end of a fiber that contains a section of rare-earth ion-doped fiber. The dichroic mirrors are selected to have high reflectivities at the laser wavelength but have nearly 100% transmittance at the pump wavelength. The pump beam is launched through one end of the fiber, while the laser output exits from the other end. An alternative configuration, which has two different arrangements shown in Figs. 11.17(b) and (c), is a transversely coupled fiber Fabry–Perot cavity in which a fiber directional coupler is used to couple the pump power into, and part of the resonating laser power out of, the cavity. In this configuration, both the pump and the laser beams never leave the fiber, avoiding the coupling of these beams in and out of the fiber through mirrors and lenses. An all-fiber Fabry–Perot laser, shown in Fig. 11.17(d), can be constructed using two fiber loop reflectors in place of mirrors. Similarly to the dichroic mirrors used in the Fabry–Perot cavity of Fig. 11.17(a), the fiber loop reflectors are chosen to have 100% transmittance for the pump wave and high reflectance for the resonating laser wave. Another all-fiber configuration is the fiber ring cavities shown in Figs. 11.17(e) and (f). Note the significant difference between a fiber

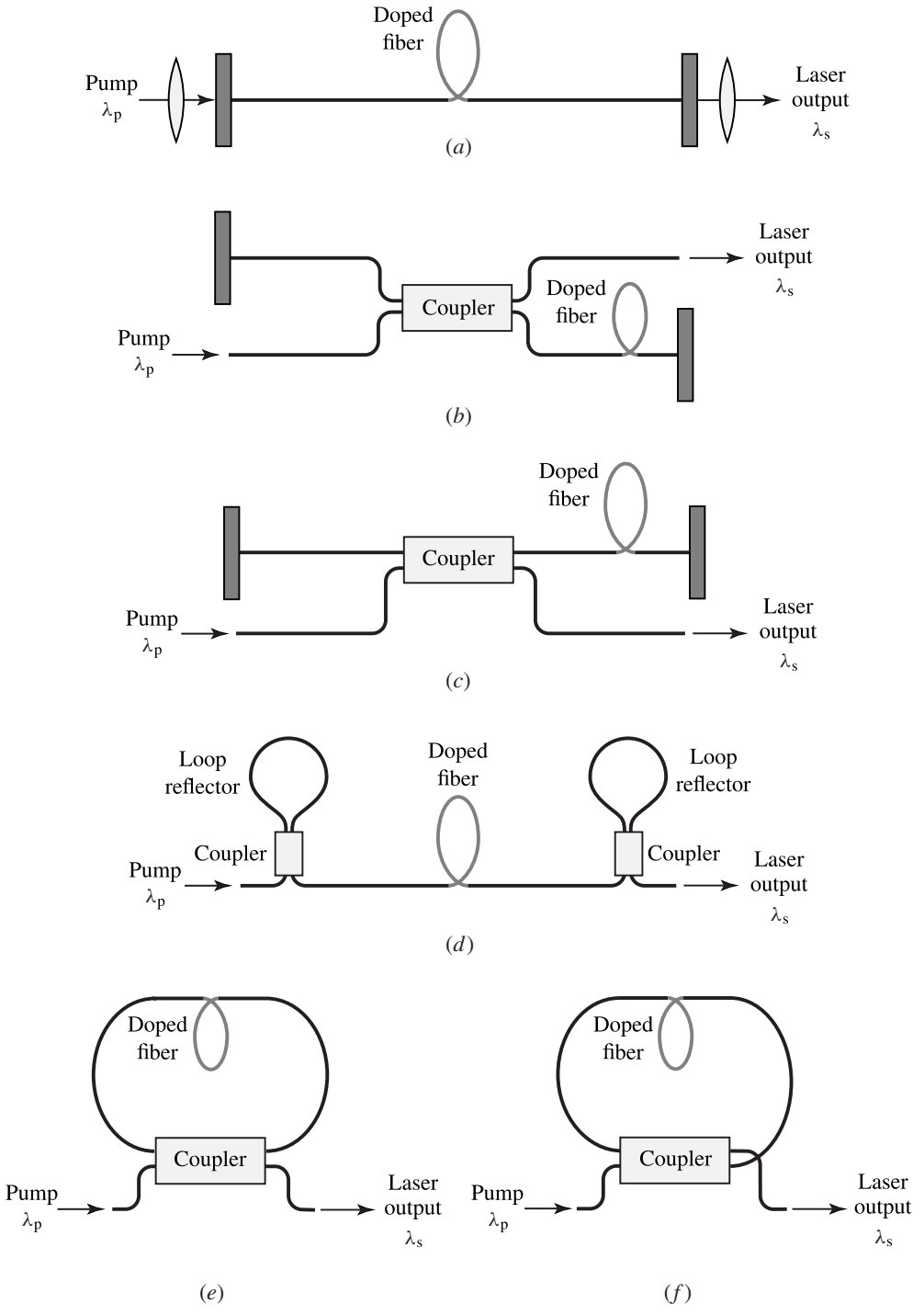


Figure 11.17 Fiber laser cavity configurations: (*a*) a Fabry–Perot cavity with end mirrors, (*b*) and (*c*) two arrangements of a transversely coupled fiber Fabry–Perot cavity, (*d*) an all-fiber Fabry–Perot cavity with fiber loop reflectors, (*e*) and (*f*) two arrangements of an all-fiber ring cavity.

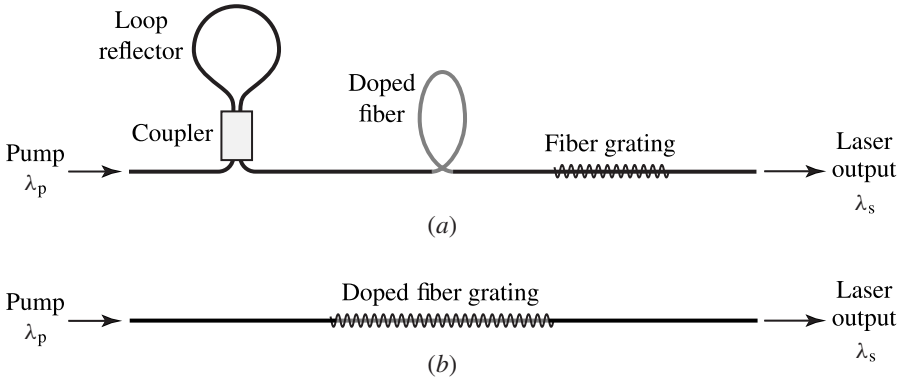


Figure 11.18 (*a*) Fiber DBR laser and (*b*) fiber DFB laser for single-longitudinal-mode laser oscillation.

ring cavity and a fiber loop reflector. A ring cavity is an optical resonator, which stores energy, but a loop reflector is a nonresonant optical interferometer, which does not store energy.

Because the host materials are glasses, most rare-earth ion-doped fibers are at least partially inhomogeneously broadened at room temperature. This property, coupled with the broad gain bandwidth and the usually long cavity length of a fiber laser, leads to the fact that a fiber laser normally oscillates in multiple longitudinal modes. There are a few approaches to forcing a fiber laser to oscillate in a single longitudinal mode, thus delivering a narrow-linewidth, single-frequency laser output. An all-fiber approach is to use frequency-selective fiber Bragg gratings. The frequency of the single-frequency laser output can be tuned if tunable fiber gratings are used. One possible arrangement, shown in Fig. 11.18(*a*), is a kind of *fiber DBR laser*, in which a fiber grating is used as a frequency-selective distributed Bragg reflector to replace the output-coupling loop reflector of Fig. 11.17(*d*). Another possibility, shown in Fig. 11.18(*b*), is a *fiber DFB laser*, in which no localized reflector is used but laser oscillation is accomplished by frequency-selective distributed feedback of a fiber grating throughout the entire section of the rare-earth ion-doped fiber gain medium. Because of the waveguiding nature of a fiber, the transverse-mode characteristics of a fiber laser are not a function of the cavity configuration but are solely determined by the mode characteristics of the fiber waveguide. By using a single-mode fiber, single-transverse-mode oscillation of a fiber laser can be guaranteed, irrespective of other parameters of the fiber laser resonator.

Because a fiber laser generally has a longitudinal optical pumping arrangement, the general characteristics of solid-state lasers with longitudinal optical pumping discussed in the earlier sections apply equally well. As in the case of the fiber amplifiers discussed in Section 10.5, some quantitative modifications on the formulations of certain relations are needed to account for the waveguiding nature of the fiber. Specifically, it is necessary to incorporate the confinement factors Γ_p for the pump beam and Γ_s for the signal beam

into the formulation. Except for the relation for the transparency pump power, P_p^{tr}, given in (10.121), all of the formulations listed in (10.119)–(10.125) for a fiber amplifier are valid for a fiber laser. The transparency pump power of a laser is defined differently from that of an amplifier. The transparency pump power of a fiber laser is still that given in (11.62) without modification. The only additional formulations that have to be modified for a fiber laser are those for the pump power utilization factor at threshold and the threshold pump power:

$$\zeta_p^{th} = 1 - \exp\left[-\frac{\Gamma_s\sigma_e N_t - g_{th}}{\Gamma_s(\sigma_e + \sigma_a)N_t}\alpha_p l_g\right] \tag{11.109}$$

and

$$P_p^{th} = \begin{cases} \dfrac{1}{p}\dfrac{\exp\left[p\dfrac{\Gamma_s\sigma_a N_t + g_{th}}{\Gamma_s(\sigma_e + \sigma_a)N_t}\alpha_p l_g\right] - 1}{1 - \exp\left[-\dfrac{\Gamma_s(\sigma_e - p\sigma_a)N_t - (1+p)g_{th}}{\Gamma_s(\sigma_e + \sigma_a)N_t}\alpha_p l_g\right]} P_p^{sat}, & \text{for } p \neq 0, \\ \dfrac{\Gamma_s\sigma_a N_t + g_{th}}{\Gamma_s(\sigma_e + \sigma_a)N_t}\dfrac{\alpha_p l_g}{1 - \exp\left[-\dfrac{\Gamma_s\sigma_e N_t - g_{th}}{\Gamma_s(\sigma_e + \sigma_a)N_t}\alpha_p l_g\right]} P_p^{sat}, & \text{for } p = 0. \end{cases} \tag{11.110}$$

By setting g_{th} to zero in (11.109) and (11.110), ζ_p^{tr} and P_p^{tr} can be found. With these modifications, all of the relations found in Sections 11.1–11.4 are applicable to fiber lasers. However, because at least one or two of the conditions for the applicability of (11.87) are often violated in a fiber laser, the output power of a fiber laser should be found by using (11.88).

EXAMPLE 11.9 Because an erbium-doped fiber is a high-gain medium that can have a long length, an EDFA can be made into an erbium-doped fiber laser with a relatively small amount of optical feedback. The simplest approach is to cleave the two ends of the EDFA and leave them uncoated for a 4% reflectivity each, thus forming a Fabry–Perot cavity of $R_1 = R_2 = 4\%$ and a gain-medium length of $l_g = l$ with a unity filling factor of $\Gamma = 1$. The EDFA of $l = 20$ m pumped at $\lambda_p = 1.48$ μm with a gain peak at $\lambda = 1.53$ μm described in Example 10.13 is made into a fiber laser in this manner. Aside from the absorption associated with the laser transition levels, this fiber has a background distributed loss of $\overline{\alpha} = 2$ dB km^{-1}. (a) Find the threshold pump power P_p^{th} of this fiber laser. What is the pump power utilization factor ζ_p^{th} at the laser threshold? (b) Find the transparency pump power P_p^{tr} of this fiber laser. What is the pump power utilization factor ζ_p^{tr} at transparency? (c) What is the output power of the laser if it is pumped with an input pump power of $P_p = 20$ mW?

Solution We find from Example 10.13 the following parameters for this fiber: $\eta_p = 1$, $p = 0.055$, $\alpha_p = 0.3485$ m^{-1}, and $P_p^{sat} = 4.25$ mW at the pump wavelength of

$\lambda_p = 1.48\ \mu m$; $\sigma_a = 5.75 \times 10^{-25}\ m^2$, $\sigma_e = 7.9 \times 10^{-25}\ m^2$, and $\Gamma_s = 0.70$ at the signal wavelength of $\lambda = 1.53\ \mu m$; $N_t = 2.2 \times 10^{24}\ m^{-3}$. Therefore, $\alpha_p l_g = 6.97$, $\sigma_a N_t = 1.265$, $\sigma_e N_t = 1.738$, $(\sigma_e + \sigma_a)N_t = 3.003$, and $(\sigma_e - p\sigma_a)N_t = 1.668$. The distributed loss is $\overline{\alpha} = 2\ dB\ km^{-1} = 0.46\ km^{-1} = 4.6 \times 10^{-4}\ m^{-1}$. Because $l_g = l = 20$ m and $R_1 = R_2 = 0.04$, the threshold gain coefficient of the laser is

$$g_{th} = \overline{\alpha} - \frac{\ln \sqrt{R_1 R_2}}{l_g} = \left(4.6 \times 10^{-4} - \frac{\ln 0.04}{2}\right)\ m^{-1} = 0.1614\ m^{-1}.$$

(a) The threshold pump power can be found from (11.110) for $p \neq 0$:

$$P_p^{th} = \frac{1}{0.055} \frac{\exp\left(0.055 \times \dfrac{0.7 \times 1.265 + 0.1614}{0.7 \times 3.003} \times 6.97\right) - 1}{1 - \exp\left(-\dfrac{0.7 \times 1.668 - 1.055 \times 0.1614}{0.7 \times 3.003} \times 6.97\right)} \times 4.25\ mW$$
$$= 16.87\ mW.$$

The pump power utilization factor at threshold can be found from (11.109):

$$\zeta_p^{th} = 1 - \exp\left(-\frac{0.7 \times 1.738 - 0.1614}{0.7 \times 3.003} \times 6.97\right) = 0.970.$$

(b) The transparency pump power can be found from (11.110) for $p \neq 0$ by setting g_{th} to zero:

$$P_p^{tr} = \frac{1}{0.055} \frac{\exp\left(0.055 \times \dfrac{0.7 \times 1.265}{0.7 \times 3.003} \times 6.97\right) - 1}{1 - \exp\left(-\dfrac{0.7 \times 1.668}{0.7 \times 3.003} \times 6.97\right)} \times 4.25\ mW$$
$$= 13.83\ mW.$$

The pump power utilization factor at transparency can be found from (11.109) by setting g_{th} to zero:

$$\zeta_p^{tr} = 1 - \exp\left(-\frac{0.7 \times 1.738}{0.7 \times 3.003} \times 6.97\right) = 0.982.$$

(c) The output power of the laser can be found using (11.88). We find from Example 10.13 that $P_p^{out} = 0.984$ mW when $P_p^{in} = 20$ mW. Thus, $\zeta_p = (20 - 0.984)/20 = 0.951$. We also find that

$$\frac{\gamma_{out}}{\gamma_c} = \frac{-\ln \sqrt{R_1 R_2}}{g_{th} l} = \frac{-\ln 0.04}{0.1614 \times 20} = 0.997.$$

We can then find the output laser power from (11.88):

$$P_{out} = \eta_p \frac{\gamma_{out}}{\gamma_c} \frac{\lambda_p}{\lambda} (\zeta_p P_p - \zeta_p^{th} P_p^{th})$$
$$= 1 \times 0.997 \times \frac{1.48}{1.53} \times (0.951 \times 20 - 0.970 \times 16.87)\ mW$$
$$= 2.56\ mW.$$

We see from this example that $P_p > P_p^{th} > P_p^{tr}$ and $\zeta_p < \zeta_p^{th} < \zeta_p^{tr}$, as expected from the discussion following (11.86) in Section 11.3. We also see that this laser has a relatively high transparency pump power compared to its threshold pump power, with P_p^{th} only 23% above P_p^{tr}, even though this laser has end mirrors of very low reflectivities of $R_1 = R_2 = 4\%$. Because P_p^{th} has to be always larger than P_p^{tr}, this situation indicates that the laser threshold cannot be significantly reduced by using coated end mirrors of high reflectivities to reduce the output coupling loss of the laser. It is possible to minimize the threshold pump power by choosing an optimum fiber length (see Problem 11.5.1). It is also possible to maximize the output power at a given pumping level by properly choosing the fiber length (see Problem 11.5.2). These two lengths are different because the former is independent of the input pump power while the latter varies with the pump power.

One unique feature of a fiber laser is that the fiber gain medium can be made long to reduce the laser threshold. However, for the same reason as discussed in Section 10.5 for fiber amplifiers, the effect of increasing the length of the fiber gain medium on the threshold of a fiber laser depends on the nature of the particular fiber gain medium used. For a four-level fiber laser, the threshold pump power decreases inversely with the length of the rare-earth ion-doped fiber gain medium, assuming that the background attenuation coefficient of the host glass is small. If the gain medium of a fiber laser functions as a three-level or a quasi-two-level system, however, there is an optimum length for the minimum pump power threshold. Increasing the length of the fiber gain medium beyond the optimum length results in an increase in the threshold.

A rare-earth ion-doped fiber pumped by a properly chosen semiconductor laser is a high-gain device because of the high optical intensity in the fiber waveguide, the long fluorescence lifetime of the rare-earth ions, and the efficient use of the narrow-band pump power matching the absorption band of the rare-earth ions. This high gain and the fact that a fiber has a large length and a small cross-sectional area make it possible for laser action in a rare-earth ion-doped fiber without a resonant cavity to take place through amplified spontaneous emission (ASE). The condition for such laser action to occur is when the spontaneous emission originating from one end of the fiber is amplified through the fiber to an intensity that saturates the gain at the other end:

$$h\nu \Delta\nu G_0 = P_{sat}, \tag{11.111}$$

where $h\nu$ is the energy of the spontaneous photon, $\Delta\nu$ is the bandwidth of the spontaneous emission, G_0 is the integrated unsaturated power gain over the length of the fiber, and P_{sat} is the saturation power as defined in (10.95). Such a device is known as an *ASE fiber laser* or a *mirrorless fiber laser* but is often also called a *superfluorescent fiber laser*.[1] An ASE fiber laser has spatial coherence and, if a single-mode fiber is used, a

[1] Superfluorescence and ASE are fundamentally different phenomena. Calling an ASE laser a superfluorescent laser, though common in the literature, is technically inaccurate. For details, see Siegman, A. E., *Lasers*. Mill Valley, CA: University Science Books, 1986, p. 551.

single transverse mode pattern. However, it does not have a longitudinal mode structure. Besides, its output has a broad spectrum and very little temporal coherence. It serves as a high-power, broadband light source, which is very useful in many applications where temporal coherence is not needed or is avoided. Unlike an ordinary resonant laser, an ASE laser has no distinctive threshold. Its gain is never clamped at any particular level.

PROBLEMS

11.1.1 Show by using (11.32) that for a linear Fabry–Perot cavity of a length l, the locations of the end mirrors measured from the beam waist are

$$z_1 = -\frac{l(\mathcal{R}_2 - l)}{\mathcal{R}_1 + \mathcal{R}_2 - 2l} \quad \text{and} \quad z_2 = \frac{l(\mathcal{R}_1 - l)}{\mathcal{R}_1 + \mathcal{R}_2 - 2l}, \tag{11.112}$$

for the left mirror with a radius of curvature of $\mathcal{R}_1$ and the right mirror with a radius of curvature of $\mathcal{R}_2$, respectively. Show by using this result that the Rayleigh range of the Gaussian beam defined by this cavity is that given by (11.33).

11.1.2 A stable Fabry–Perot cavity must have a positive, real value for its Rayleigh range z_R. By applying this concept to the relation in (11.33), show that the following statements are true.

a. It is not possible for a Fabry–Perot cavity to be stable if both mirrors are convex, but it is possible if both mirrors are concave or if one is concave and the other is convex.

b. The stability criterion for a Fabry–Perot cavity of any combination of mirrors is that given by (11.34).

11.1.3 A Fabry–Perot optical cavity of a length l consists of one concave mirror with $\mathcal{R}_1 = 1$ m and one planar mirror with $\mathcal{R}_2 = \infty$ in free space with $n_0 = 1$. The cavity length can be varied.

a. For a stable cavity, what is the range of values that can be chosen for the cavity length l?

b. Where is the waist of the Gaussian beam defined by this optical cavity as the cavity length is varied within the range found in (a)?

c. Within the stability range, there is a cavity length for which the waist spot size of the Gaussian beam is maximized. Find this cavity length and the corresponding maximum waist spot size for the optical wavelength at $\lambda = 1$ μm.

d. What should the cavity length be if a waist spot size of $w_0 = 300$ μm for $\lambda = 1$ μm is desired?

11.1.4 An empty Fabry–Perot cavity in free space has a cavity length of $l = 0.5$ m and mirror reflectivities of $R_1 = 100\%$ and $R_2 = 90\%$.

a. What are the round-trip time and the longitudinal mode spacing of this cavity?

b. Find the finesse and the longitudinal mode width of this cavity.
c. What are the cavity decay rate, the photon lifetime, and the Q factor for $\lambda = 1$ μm?
d. If the cavity length remains unchanged but the reflectivities of the mirrors are changed to $R_1 = R_2 = 90\%$, which of the parameters found in (a)–(c) will change in value? How do they change?

11.1.5 A ring cavity consists of three mirrors of $R_1 = 99\%$, $R_2 = 95\%$, and $R_3 = 90\%$ in free space. To form the ring cavity, the mirrors are arranged with the following inter-mirror spacings: $l_{12} = 0.5$ m, $l_{23} = 0.4$ m, and $l_{31} = 0.3$ m. The only losses of this cavity are those from the transmission of the mirrors.
a. What are the round-trip time and the longitudinal mode spacing of this cavity?
b. Find the finesse and the longitudinal mode width of this cavity.
c. What are the cavity decay rate, the photon lifetime, and the Q factor for $\lambda = 1$ μm?

11.1.6 The cavity round-trip time T and the photon lifetime τ_c are two characteristic time constants of a resonant optical cavity. If both time constants are known, all of the other parameters of the cavity can be found. It is most convenient to express other cavity parameters of interest in terms of these two parameters because both of them can be measured experimentally.
a. Express the finesse, the longitudinal mode spacing, the longitudinal mode width, the cavity decay rate, and the Q factor of a cavity in terms of T and τ_c.
b. Find these parameters for a cavity that has $T = 10$ ps and $\tau_c = 1$ ns. The optical wavelength of interest is 1 μm.

11.1.7 The relation between the two characteristic time constants, the cavity round-trip time T and the photon lifetime τ_c, of a resonant optical cavity determines whether the cavity has a high Q or low Q for a given cavity length. A high-Q cavity can support a low-gain laser, whereas a low-Q cavity requires a high-gain laser medium to make a laser feasible. The condition that $T = \tau_c$ sets the two types of cavity apart. Express this condition in terms of the reflectivities of the cavity mirrors assuming no distributed loss in the cavity. Express it in terms of the finesse of the cavity. The results found apply to any types of cavity, including Fabry–Perot and ring cavities.

11.2.1 Show by following the procedure through (11.59) and (11.60) that the threshold pump power of an optically pumped laser in a single-pass longitudinal pumping arrangement is that given in (11.61).

11.2.2 In this problem, we consider an optically pumped laser with either single-pass or multiple-pass longitudinal pumping with negligible transverse pump beam divergence. Under the condition that $s_{th} = P_p^{th}/P_p^{sat} \ll 1$, the pump power decays exponentially along the longitudinal pumping axis. The gain medium can be any system so that the parameter p can be either zero or nonzero.

a. Show that the threshold pump power of a laser with single-pass longitudinal optical pumping is

$$P_{\mathrm{p}}^{\mathrm{th}} = \frac{\sigma_{\mathrm{a}} N_{\mathrm{t}} + g_{\mathrm{th}}}{(\sigma_{\mathrm{e}} + \sigma_{\mathrm{a}}) N_{\mathrm{t}}} \frac{\alpha_{\mathrm{p}} l_{\mathrm{g}}}{\zeta_{\mathrm{p}}^{\mathrm{th}}} P_{\mathrm{p}}^{\mathrm{sat}}, \quad \text{for either } p \neq 0 \text{ or } p = 0, \tag{11.113}$$

where $\zeta_{\mathrm{p}}^{\mathrm{th}} \approx 1 - \mathrm{e}^{-\alpha_{\mathrm{p}} l_{\mathrm{g}}}$ assuming no reflection of the pump beam at the pump input surface of the gain medium. What is the transparency pump power?

b. Show that under the same condition that $s_{\mathrm{th}} \ll 1$ so that the power of the pump beam decays exponentially along the longitudinal axis in each pass through the gain medium, the threshold pump power of a laser that is longitudinally pumped in a multiple-pass arrangement is also given by (11.113) with $\zeta_{\mathrm{p}}^{\mathrm{th}}$ accounting for the total power absorbed in all passes given by

$$\zeta_{\mathrm{p}}^{\mathrm{th}} = \frac{(1 - R_{1\mathrm{p}})(1 + R_{2\mathrm{p}} \mathrm{e}^{-\alpha_{\mathrm{p}} l_{\mathrm{g}}})}{1 - R_{1\mathrm{p}} R_{2\mathrm{p}} \mathrm{e}^{-2\alpha_{\mathrm{p}} l_{\mathrm{g}}}} (1 - \mathrm{e}^{-\alpha_{\mathrm{p}} l_{\mathrm{g}}}), \tag{11.114}$$

where $R_{1\mathrm{p}}$ is the reflectivity of the pump beam at the pump input facet of the gain medium and $R_{2\mathrm{p}}$ is that at the other facet of the gain medium. What is the transparency pump power?

11.2.3 The threshold pump power of the Nd : YAG microchip laser described in Example 11.2 can be lowered by allowing the pump beam to make multiple passes through the gain medium to utilize the pump power better. For the pump beam to make two passes through the gain medium, the pump input surface of the Nd : YAG plate is coated for 100% transmission for a pump wavelength at 808 nm while the other surface is coated for 100% reflection at 808 nm wavelength. All other parameters of the laser, including $R_1 = 100\%$ and $R_2 = 99.7\%$ for the laser wavelength at 1.064 μm and the length of the gain medium $l_{\mathrm{g}} = 500$ μm, remain unchanged.

a. Find the value of $g_{\mathrm{th}} l_{\mathrm{g}}$ and the threshold gain coefficient g_{th}. Compare them with those found in Example 11.2.

b. Find the threshold pump power for the laser in this double-pass pumping arrangement. Compare it with that found in Example 11.2 for the single-pass arrangement.

c. What is the linewidth of an oscillating laser mode when the laser has an output power of 1 mW? How does it compare with that found in Example 11.3?

11.2.4 The threshold pump power of the Nd : YAG microchip laser described in Example 11.2 can be lowered by properly increasing the length of the gain medium to utilize the pump power better. To compare this approach with the double-pass approach described in Problem 11.2.3, we double the length to $l_{\mathrm{g}} = 1$ mm while allowing the pump beam to make only one single pass through the gain medium so that the fraction of the pump power absorbed

by the gain medium is the same in these two cases. All other parameters of the laser, including $R_1 = 100\%$ and $R_2 = 99.7\%$ for the laser wavelength at 1.064 μm, remain the same as those described in Example 11.2.

a. Find the value of $g_{th}l_g$ and the threshold gain coefficient g_{th}. Compare them with those found in Example 11.2 and those found in Problem 11.2.3(a).
b. Find the threshold pump power for this laser. Compare it with that found in Example 11.2 and that found in Problem 11.2.3(b).
c. What is the linewidth of an oscillating laser mode when the laser has an output power of 1 mW? How is it compared to that found in Example 11.3 and that found in Problem 11.2.3(c)?

11.2.5 A CW Ti : sapphire laser emitting at $\lambda = 800$ nm is constructed by placing a Ti : sapphire laser rod of length $l_g = 2$ cm in a resonant cavity of length $l = 25$ cm. The Ti : sapphire rod is doped with 0.024 wt. % Ti_2O_3 for a Ti concentration of 7.9×10^{24} m^{-3} and is longitudinally pumped with the second harmonic of a Nd : YAG laser beam at $\lambda_p = 532$ nm. Ti : sapphire is a quasi-two-level system. At the desired Ti : sapphire laser wavelength of $\lambda = 800$ nm for $\mathbf{E} \parallel c$ polarization, $\sigma_e = 3.4 \times 10^{-23}$ m^2 and $\sigma_a = 8 \times 10^{-26}$ m^2. At the pump wavelength of 532 nm, $\sigma_e = 3 \times 10^{-28}$ m^2 and $\sigma_a = 7.4 \times 10^{-24}$ m^2. The refractive index of Ti : sapphire is 1.76. The fluorescence lifetime is $\tau_2 = 3.2$ μs. The pump quantum efficiency is $\eta_p = 80\%$. The ends of the rod are cut at the Brewster angle so that there is negligible reflection for both the pump and the laser beams. The pump beam is focused to an average spot size of $2w = 100$ μm in diameter across the length of the laser rod to match the spot size of the laser beam defined by the resonant cavity. The laser cavity is formed by mirrors of 100% reflectivity at 800 nm wavelength, except for the output coupling mirror of a reflectivity $R = 95\%$ that allows the laser beam to be transmitted. The only internal loss of the laser is attributable to absorption of the laser rod at the 800 nm laser wavelength.

a. At what angle are the ends of the laser rod cut?
b. Find the values of $g_{th}l_g$ and g_{th} for the threshold of this laser.
c. What percentage of input pump power is absorbed?
d. Find the transparency pump power of this quasi-two-level laser.
e. Find the threshold pump power of the laser.

11.2.6 The threshold of a laser depends on the reflectivities of the mirrors that form the laser cavity. For the Ti : sapphire laser described in Problem 11.2.5, a threshold pump power of $P_p^{th} = 1.5$ W is desired by properly choosing the reflectivity R of the output coupling mirror while keeping all of the other parameters of the laser rod, the optical cavity, and the pump unchanged. What reflectivity of the output coupling mirror should be chosen for this purpose?

11.2.7 What are the major consequences of the effect of mode pulling on the oscillating mode characteristics of a laser?

11.2.8 Single-frequency CW lasers are very useful in many applications. Consider both homogeneously and inhomogeneously broadened lasers. Discuss how a CW laser in steady-state oscillation can be made to oscillate in only one frequency.

11.3.1 Show that when $p = 0$ or $p \ll 1$ for an optically pumped laser that has a spatially varying gain coefficient due to longitudinal pumping, the pumping ratio r can be expressed in the form of (11.86), and the output laser power as a function of pump power takes the form of (11.87).

11.3.2 Find the pump power required for the Nd : YAG microchip laser with double-pass pumping described in Problem 11.2.3 to have an output power of 1 mW. Compare it with the 20.5 mW required for the laser with single-pass pumping to have 1 mW output power found in Example 11.4. What is the output power of the laser with double-pass pumping if it is pumped with a pump power of 20.5 mW?

11.3.3 Find the power conversion efficiency, the slope efficiency, the external quantum efficiency, and the internal quantum efficiency of the Nd : YAG microchip laser with double-pass pumping described in Problems 11.2.3 and 11.3.2. Compare the results with those of the same laser with single-pass pumping found in Example 11.5.

11.3.4 Find the pump power required for the Nd : YAG microchip laser described in Problem 11.2.4 to have an output power of 1 mW. This laser has a Nd : YAG plate of doubled length $l_g = 1$ mm, but is pumped in a single pass. What is the output power of this laser if it is pumped with a pump power of 20.5 mW? Compare the results with those found in Problem 11.3.2 for the laser that has a short Nd : YAG plate but is pumped in double passes.

11.3.5 Find the power conversion efficiency, the slope efficiency, the external quantum efficiency, and the internal quantum efficiency of the Nd : YAG microchip laser with a doubled Nd : YAG length of $l_g = 1$ mm described in Problems 11.2.4 and 11.3.4. Compare the results with those found in Example 11.5 and Problem 11.3.3 for the other two cases.

11.3.6 Find the pump power required for the Ti : sapphire laser described in Problem 11.2.5 to have an output power of 1 W. Find its power conversion efficiency, slope efficiency, external quantum efficiency, and internal quantum efficiency.

11.3.7 Find the pump power required for the Ti : sapphire laser described in Problem 11.2.6 to have an output power of 1 W. Find its power conversion efficiency, slope efficiency, external quantum efficiency, and internal quantum efficiency. Compare the results found for this laser with those found in Problem 11.3.6 for the Ti : sapphire that has an output coupling mirror of different reflectivity.

11.4.1 Answer the following questions regarding the principles of gain switching and Q switching.

a. What are the two fundamental parameters of a laser material that determine the characteristics of pulsed lasers using such a material as the gain medium?
b. What are the most favorable conditions for gain switching?
c. What are the most favorable conditions for Q switching?
d. In what situation can the same pulse generated by Q switching a laser be generated by gain switching the same laser so that the functioning of the Q switch becomes redundant?

11.4.2 Answer the following questions regarding gain-switched and Q-switched pulses.
a. What is the absolutely shortest pulse one can generate by gain switching a given laser? What has to be done to approach that limit?
b. Answer the questions in (a) for Q switching instead of gain switching.
c. With given cavity mirrors and a given gain medium for a gain-switched or Q-switched laser, name two different approaches that can easily be taken to shorten the output laser pulsewidth.

11.4.3 The pulsewidth of the Q-switched pulse generated by the Q-switched Nd : YAG microchip laser considered in Example 11.6 can be reduced by increasing the pumping level while keeping the laser parameters unchanged. By so doing, the pumping ratio r is increased while the photon lifetime τ_{cl} in the lasing phase remains unchanged. When the pulsewidth changes, the peak power and pulse energy all change accordingly. Meanwhile, the requirements on the Q switch to act as an ideal Q switch are changed. A Q-switched pulse of $\Delta t_{ps} = 3$ ns is desired.
a. What is the required pumping ratio for the Q-switched pulse to have $\Delta t_{ps} = 3$ ns?
b. What are the peak power and energy of this Q-switched pulse?
c. What are the requirements on the Q switch for ideal Q switching of this laser?

11.4.4 The Nd : YAG laser described in Example 11.7 can also be gain switched or Q switched. Without additional information on the parameters of the laser or its cavity, find the upper and lower limits of the pulsewidth of a gain-switched or Q-switched pulse that can be generated from this laser. What is the most likely range for the pulsewidth?

11.4.5 Consider the following four different modes of laser operation: CW, gain switching, Q switching, and mode locking. How many longitudinal modes will oscillate in each situation if the laser is homogeneously broadened? How many longitudinal modes will oscillate in each situation if the laser is inhomogeneously broadened?

11.4.6 Answer the following questions regarding mode locking.
a. What can be said about the temporal pulse characteristics of a completely mode-locked pulse no matter how it is generated?

b. What can be said about a pulse generated by a mode-locked laser if the temporal pulse shape is asymmetric?
c. What are the expected shapes of the temporal and spectral envelopes of optical pulses generated by active and passive mode locking, respectively?
d. For most modes of laser operation, it is desirable that the fluorescence lifetime of the gain medium be as long as possible, but there are exceptions. Give two examples of laser operation in which a gain medium with a very long fluorescence lifetime is not desirable. What is the desirable fluorescence lifetime in each of those two situations?

11.4.7 Show that a laser pulse that has a Gaussian temporal intensity profile also has a Gaussian spectral intensity profile. Show that for such Gaussian pulses the transform-limit constant K defined in (11.103) has the value of $K = 2\ln 2/\pi = 0.4413$. Show also that for such Gaussian pulses the constant K' defined in (11.108) has the value of $K' = 2\sqrt{\ln 2}/\sqrt{\pi} = 0.9394$. Note that

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x = \sqrt{\pi}. \tag{11.115}$$

11.4.8 Show that a laser pulse that has a sech^2 temporal intensity profile also has a sech^2 spectral intensity profile. Show that for such sech^2 pulses the transform-limit constant K defined in (11.103) has the value of $K = 4\ln^2(1+\sqrt{2})/\pi^2 = 0.3148$. Show also that for such sech^2 pulses the constant K' defined in (11.108) has the value of $K' = \ln(1+\sqrt{2}) = 0.8814$. Note that

$$\int_{-\infty}^{\infty} \mathrm{sech}^2 x\,\mathrm{d}x = 2. \tag{11.116}$$

11.4.9 Find the pulsewidths of the shortest possible laser pulses that can be *directly* generated by mode-locked lasers of the various gain media listed in Table 10.1. Assume that the pulses have Gaussian shape. Can even shorter pulses be generated by gain switching or Q switching? After a pulse is emitted from a laser, it can often be made shorter than its limit found here by indirect means such as pulse compression or pulse truncation. Why is this possible?

11.4.10 The solid-state laser material Nd : glass has a gain bandwidth of $\Delta\lambda_g = 22$ nm peaked at the wavelength $\lambda = 1.054$ μm. In comparison, the laser material Nd : YLF has a gain bandwidth of $\Delta\lambda_g = 1.35$ nm peaked at the wavelength $\lambda = 1.053$ μm. They are normally excited by optical pumping.
a. If the entire gain bandwidth of the laser material is utilized in the generation of an optical pulse that has a Gaussian temporal pulse shape, what are the shortest pulsewidths possible using Nd : glass and Nd : YLF, respectively, as the gain medium?

b. It is practically very difficult, if not entirely impossible, to generate such pulses as mentioned in (a) by gain switching Nd : glass and Nd : YLF lasers. Discuss quantitatively the physical limitations that create such a difficulty.
c. If such pulses were to be generated by active Q switching, what practical difficulties would be encountered?

11.4.11 Several types of lasers are very versatile in terms of their mode of operation. For example, by simply turning the active mode locker on or off, a Nd : YAG laser can be switched between CW operation and continuous mode-locking operation without much change in its average output power. As another example, a semiconductor laser can be biased at a constant DC level while being switched between CW or repetitive gain-switching operations, the latter of which delivers a regular train of gain-switched pulses. The relation between the peak power and the average power of the pulses of a constant pulsewidth Δt_{ps} in a pulse train of a pulse repetition rate f_{ps} is given in (11.108). The laser beam at the fundamental frequency ω is sent through a second-harmonic crystal to generate its second harmonic at the second-harmonic frequency 2ω with an efficiency that is proportional to the instantaneous power of the laser such that $P_{2\omega}(t) = a P_{\omega}^2(t)$, with a being a constant. The average powers $\overline{P}_{\omega}$ and $\overline{P}_{2\omega}$ of the fundamental and the second harmonic, respectively, are monitored for both CW and pulsed operations of the laser. Clearly, $(\overline{P}_{2\omega})_{\text{pulsed}}$ is higher than $(\overline{P}_{2\omega})_{\text{CW}}$ if $(\overline{P}_{\omega})_{\text{pulsed}}$ is comparable to $(\overline{P}_{\omega})_{\text{CW}}$. Show that the pulsewidth can be found from

$$\Delta t_{\text{ps}} = A \frac{(\overline{P}_{2\omega}/\overline{P}_{\omega}^2)_{\text{CW}}}{f_{\text{ps}}(\overline{P}_{2\omega}/\overline{P}_{\omega}^2)_{\text{pulsed}}}, \tag{11.117}$$

where A is a constant that depends on the pulse shape. Show also that $A = (2 \ln 2/\pi)^{1/2} = 0.6643$ for Gaussian pulses and that $A = (2/3)\ln(1 + \sqrt{2}) = 0.5876$ for sech^2 pulses. This problem describes a convenient way of measuring the pulsewidth of repetitive pulses if the pulse shape and the pulse repetition rate are both known. Note that

$$\int_{-\infty}^{\infty} \text{sech}^4 x \, \text{d}x = \frac{4}{3}. \tag{11.118}$$

(See Chen, Y. C. and Liu, J. M., "Measurement of picosecond semiconductor laser pulse duration with internally generated second harmonic emission," *Applied Physics Letters* **47**(7): 662–664, Oct. 1985.)

11.5.1 Find the optimum fiber length that minimizes the threshold pump power of the erbium-doped fiber laser described in Example 11.9. What is this minimum threshold pump power? What is the transparency pump power of the laser if this optimum fiber length is used? What is the output laser power if the laser is pumped at $P_{\text{p}} = 20$ mW? Find the values of $\zeta_{\text{p}}^{\text{th}}$, $\zeta_{\text{p}}^{\text{tr}}$, and ζ_{p}.

11.5.2 Find the fiber length that maximizes the output power of the erbium-doped fiber laser described in Example 11.9 at the pumping level of $P_p = 20$ mW. What is this maximized output laser power? What are the threshold pump power and the transparency pump power of the laser if this fiber length is used? Find the values of ζ_p^{th}, ζ_p^{tr}, and ζ_p. Compare these results with those found in Problem 11.5.1 to explain why the maximum output power at the given pumping level does not take place for a fiber length that minimizes the threshold pump power.

11.5.3 The erbium-doped fiber laser described in Example 11.9 can be pumped at 980 nm instead. When it is pumped at this wavelength, it behaves as a three-level system with $\sigma_a^p = 2.58 \times 10^{-25}$ m^2. The fundamental mode of the fiber at this pump wavelength has an effective mode radius of $w_p = 3.3$ μm and a confinement factor of $\Gamma_p = 0.84$. The pumping efficiency is $\eta_p = 1$. Answer the questions in Example 11.9 for this erbium-doped fiber laser pumped at 980 nm with an input pump power of 20 mW for a laser output at 1.53 μm. Compare the results with those found in Example 11.9.

11.5.4 Answer the questions in Problem 11.5.1 for the erbium-doped fiber laser under the operating conditions specified in Problem 11.5.3.

11.5.5 Answer the questions in Problem 11.5.2 for the erbium-doped fiber laser under the operating conditions specified in Problem 11.5.3.

SELECT BIBLIOGRAPHY

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge University Press, 1996.

Digonnet, M. J. F., ed., *Rare Earth Doped Fiber Lasers and Amplifiers*, 2nd edn. New York: Marcel Dekker, 2001.

France, P. W., ed., *Optical Fiber Lasers and Amplifiers*. London: Blackie, 1991.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Iizuka, K., *Elements of Photonics in Free Space and Special Media*, Vol. I. New York: Wiley, 2002.

Inguscio, M. and Wallenstein, R., eds., *Solid State Lasers: New Development and Applications*. New York: Plenum Press, 1993.

Koechner, W., *Solid-State Laser Engineering*, 2nd edn. Berlin: Springer-Verlag, 1988.

Milonni, P. W. and Eberly, J. H., *Lasers*. New York: Wiley, 1988.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Siegman, A. E., *Lasers*. Mill Valley, CA: University Science Books, 1986.

Silfvest, W. T., *Laser Fundamentals*. Cambridge: Cambridge University Press, 1996.

Verdeyen, J. T., *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Yariv, A., *Optical Electronics in Modern Communications*, 5th edn. Oxford: Oxford University Press, 1997.

Yeh, C., *Applied Photonics*. San Diego, CA: Academic Press, 1994.

ADVANCED READING LIST

Byer, R. L., "Nonlinear optics and solid-state lasers: 2000," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 911–930, Nov.–Dec. 2000.

Chesler, R. B., Karr, M. A., and Geusic, J. E., "An experimental and theoretical study of high repetition rate *Q*-switched Nd : YAlG lasers," *Proceedings of the IEEE* **58**(12): 1899–1914, Dec. 1970.

Fermann, M. E., Galvanauskas, A., Sucha, G., and Harter, D., "Fiber-lasers for ultrafast optics," *Applied Physics B* **65**(2): 259–275, 1997.

French, P. M. W., "The generation of ultrashort laser pulses," *Report on Progress in Physics* **58**: 169–267, 1995.

Kranzelbinder, G. and Leising, G., "Organic solid-state lasers," *Report on Progress in Physics* **63**(5): 729–762, May 2000.

Krupke, W. F., "Ytterbium solid-state lasers: the first decade," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1287–1296, Nov.–Dec. 2000.

Malcolm, G. P. A. and Ferguson, A. I., "Diode-pumped solid-state lasers," *Contemporary Physics* **32**(5): 305–319, Sep.–Oct. 1991.

Sorokin, E., Sorokina, I. T., and Wintner, E., "Diode-pumped ultra-short-pulse solid-state lasers," *Applied Physics B* **72**(1): 3–14, Jan. 2001.

Urquhart, P., "Review of rare earth doped fiber lasers and amplifiers," *IEE Proceedings*, *Part J Optoelectronics* **135**(6): 385–407, Dec. 1988.

Vanherzeele, H., "Optimization of a CW mode-locked frequency-doubled Nd : $LiYF_4$ laser," *Applied Optics* **27**(17): 3608–3615, Sep. 1988.

Part V

Semiconductor optoelectronics

12 Semiconductor basics

Semiconductors are important materials. Because of their unique electronic properties, they are the materials of choice for modern electronic devices. Silicon, in particular, has become the most important material for the electronics industry. Besides their unique properties for electronics applications, semiconductors also have many other important properties that are very useful for photonic device applications. In earlier chapters, we have already seen that III–V semiconductors are useful materials for optical waveguides and electro-optic devices. Many semiconductors are also used for acousto-optic devices and nonlinear optical devices. In such applications, which are based solely on the dielectric properties of semiconductors, semiconductors are nothing but another group of dielectric optical materials. Nevertheless, semiconductors do have many optoelectronic properties that are not shared by other dielectric materials. These optoelectronic properties make semiconductors once again, beyond their unique position in the electronics industry, the key materials for many important optoelectronic devices, such as light-emitting diodes, semiconductor lasers, and photodetectors. These devices are covered in the following two chapters. In this chapter, we review the basic properties of semiconductors that are relevant to their optoelectronic device applications.

12.1 Semiconductors

In Chapter 10, optical transitions between discrete atomic or molecular energy levels are considered, though the atoms or molecules may be embedded in a host solid-state material as dopants. In a semiconductor, however, the allowed states of the electrons of its constituent atoms form continuous *energy bands* rather than discrete levels. The optical processes associated with such electrons are a strong function of the characteristics of the energy bands.

For a solid material in thermal equilibrium at a temperature T, the probability of any electronic state at an energy E being occupied by an electron is given by the Fermi–Dirac distribution function:

$$f(E) = \frac{1}{e^{(E-E_F)/k_B T} + 1}, \tag{12.1}$$

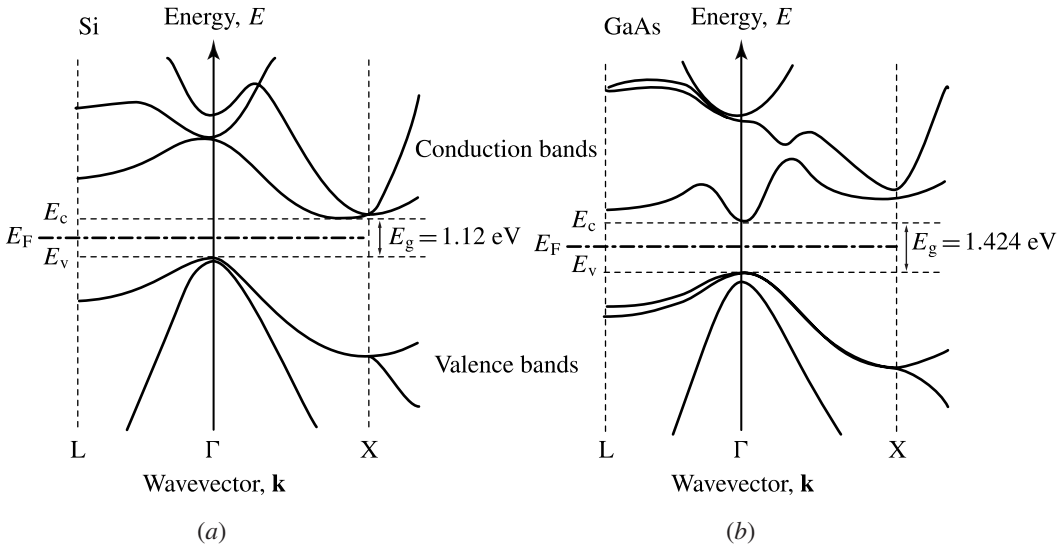


Figure 12.1 Energy band structures of (*a*) Si, which is an indirect-gap semiconductor, and (*b*) GaAs, which is a direct-gap semiconductor. In each semiconductor, the Fermi level lies in the bandgap that separates the conduction and valence bands. (Based on data from assorted sources.)

where E_F is the *Fermi level* of the material and k_B is the Boltzmann constant. At a very low temperature approaching 0 K, all of the states below the Fermi level are occupied while those above it are empty. If the Fermi level lies within an energy band, that band is partially filled. A solid with one or more partially filled bands is a metal because electrons in a partially filled band can move under an electric field to conduct an electric current. If the Fermi level lies between two separate energy bands, as illustrated in Fig. 12.1, all of the energy bands are either completely filled or completely empty at $T \rightarrow 0$ K. The filled bands are known as the *valence bands*, whereas the empty bands are known as the *conduction bands*. The lowest conduction band and the highest valence band are separated by an energy gap. The energy separation between the bottom of the lowest conduction band and the top of the highest valence band is called the *bandgap*, E_g. A material is an insulator if all of its bands are either completely filled or completely empty because, due to the Pauli exclusion principle, electrons in a completely filled energy band cannot move under an electric field. At a nonzero temperature, however, electrons in high valence bands have a probability of being thermally excited to low conduction bands. The probability of this thermal excitation is a function of E_g/k_BT; it increases with rising temperature but decreases with increasing bandgap. The electrons that are excited to conduction bands become *conduction electrons*. The removal of electrons from valence bands results in electron deficiencies, known as *holes*, in the form of unoccupied electron states among occupied states. An electron in a conduction band is a *carrier* of negative charge, whereas a hole in a valence band behaves like a carrier of positive charge. Both contribute to the electrical conductivity of a semiconductor. A semiconductor is an insulator at $T \rightarrow 0$ K, but has appreciable conductivity as the temperature rises.

The energy of an electron is a function of its quantum-mechanical wavevector, **k**. In a semiconductor, this dependence of electron energy on its wavevector forms the *band structure* of the semiconductor. The illustration in Fig. 12.1 shows the band structures of Si and GaAs. In the case of Si, the minimum of conduction bands and the maximum of valence bands do not occur at the same **k** value. A semiconductor that has such a band characteristic is called an *indirect-gap semiconductor*, and its bandgap is referred to as an *indirect bandgap*. In contrast, a semiconductor like GaAs is a *direct-gap semiconductor* because its band structure is characterized by a *direct bandgap*, with the minimum of the conduction bands and the maximum of the valence bands occurring at the same value of **k**, which in this particular instance is $\mathbf{k} = 0$.

The minimum of the conduction bands is called the *conduction-band edge*, E_c, and the maximum of the valence bands is called the *valence-band edge*, E_v. The bandgap, E_g, is the energy difference between E_c and E_v:

$$E_g = E_c - E_v. \tag{12.2}$$

The bandgap of a semiconductor is typically less than 4 eV. With the exception of some IV–VI compound semiconductors, such as lead salts, the bandgap of a semiconductor normally decreases with increasing temperature. The bandgaps and other properties of many important semiconductors are listed in Table 12.1. In this table, λ_g is the free-space optical wavelength of a photon that has an energy equal to the bandgap of a given material: $\lambda_g = hc/E_g$.

A semiconductor can be an elemental material or a compound material. The group IV elements Si and Ge are elemental semiconductors. Crystalline C can take the form either of diamond, which is more an insulator than a semiconductor because of its large bandgap of 5.47 eV at room temperature, or graphite, which is a semimetal. Though C is not a semiconductor, Si and C can form the IV–IV compound semiconductor SiC, which has many different structural forms with different bandgaps. Si and Ge can be mixed to form the IV–IV alloy semiconductor Si_xGe_{1-x}. These group IV crystals and IV–IV compounds are indirect-gap materials.

The most important semiconductors for photonic devices, however, are the III–V compound semiconductors, which are formed by combining group III elements, such as Al, Ga, and In, with group V elements, such as N, P, As, and Sb. A *binary* compound consists of two elements. There are more than ten binary III–V semiconductors, such as GaAs, InP, AlAs, and InSb. Different binary III–V compounds can be alloyed with varying compositions to form *mixed crystals* of *ternary* compound alloys and *quaternary* compound alloys. A ternary III–V compound consists of three elements, two group III elements and one group V element, such as $Al_xGa_{1-x}As$, or one group III element and two group V elements, such as $GaAs_{1-x}P_x$. A quaternary III–V compound consists of two group III elements and two group V elements, such as $In_{1-x}Ga_xAs_{1-y}P_y$. A III–V compound can be either a direct-gap or an indirect-gap material. A III–V compound with a small bandgap tends to be a direct-gap material, whereas one with a large bandgap tends to be an indirect-gap material.

Table 12.1 *Properties of some important semiconductors*[a,b]

Semiconductor		Type[c]	Bandgap, E_g (eV)		λ_g (nm)	Refractive index		Lattice constant (Å)
			At 0 K	At 300 K	At 300 K	At λ_g	At 1 μm	At 300 K
IV	C[d]	I	5.48	5.47	227	2.71	2.39	3.5668
	Si	I	1.17	1.12	1110	3.58	3.61	5.4310
	Ge	I	0.74	0.66	1880	4.12	4.38	5.6579
IV–IV	SiC	I	2.39–3.33	2.36–3.30	380–530	–	–	–
	Si_xGe_{1-x}	I	0.74–1.17	0.66–1.12	1110–1880	–	–	–
III–V[e]	AlN	D	6.29	6.20	200	2.80	2.17	$a = 3.112$ $c = 4.980$
	AlP	I	2.49	2.41	515	2.96	2.77	5.4635
	AlAs	I	2.23	2.17	572	3.19	2.95	5.6605
	AlSb	I	1.69	1.62	768	3.50	3.46	6.1355
	GaN	D	3.50	3.44	360	2.70	2.34	$a = 3.189$ $c = 5.185$
	GaP	I	2.34	2.26	549	3.43	3.17	5.4505
	GaAs	D	1.52	1.42	871	3.63	3.51	5.6533
	GaSb	D	0.81	0.73	1700	3.75	4.10	6.0959
	InN[f]	D	1.92	1.90	653	–	–	$a = 3.540$ $c = 5.800$
	InP	D	1.42	1.35	919	3.40	3.33	5.8687
	InAs	D	0.43	0.35	3540	3.52	3.63	6.0583
	InSb	D	0.24	0.17	7290	4.00	4.26	6.4794

[a] Madelung, O., ed., *Semiconductors: Basic Data*, 2nd edn. Berlin: Springer, 1996.

[b] Seraphin, B. O. and Bennett, H. E., "Optical constants", in eds. R. K. Willardson and A. C. Beer, *Semiconductors and Semimetals*, Vol. 3. New York: Academic Press, 1967, Chapter 12.

[c] D, direct gap; I, indirect gap.

[d] The diamond form of carbon.

[e] The data for nitrides, AlN, GaN, and InN, are those of hexagonal wurtzite structure. Other III–V compounds are of cubic zinc blende structure.

[f] The bandgap of InN in the old literature is in the range of 1.9–2 eV, but some recent reports point to the possibility of a lower bandgap at about 0.7 eV. This controversy is not fully resolved yet. See Bhuiyan, A. G., Hashimoto, A., and Yamamoto, A., "Indium nitride (InN): a review on growth, characterization, and physics," *Journal of Applied Physics* **94**(5): 2779–2808, Sep. 2003.

Among the III–V compounds, the nitrides are quite unique. The binary nitride semiconductors AlN, GaN, and InN, as well as their ternary alloys such as InGaN, are all direct-gap semiconductors. These direct-gap semiconductors form a complete series of materials that have bandgap energies ranging from 1.9 eV for InN to 6.2 eV for AlN, corresponding to the spectral range from 650 to 200 nm. Therefore, the nitride compounds and their alloys cover almost the entire visible spectrum and extend to the ultraviolet region. They are particularly important for the development of semiconductor lasers, light-emitting diodes, and semiconductor photodetectors in the blue, violet, and ultraviolet spectral regions. Another unique property of nitride semiconductors is

that they crystallize preferentially in hexagonal wurtzite structure, which has uniaxial optical properties. However, nitride semiconductors can also crystallize in cubic zinc blend structure, which is the common structure of all III–V compounds.

Group II elements, such as Zn, Cd, and Hg, can also be combined with group VI elements, such as S, Se, and Te, to form binary II–VI semiconductors. Among such compounds, the Zn and Cd compounds, ZnS, ZnSe, ZnTe, CdS, CdSe, and CdTe, are direct-gap semiconductors with large bandgaps ranging from 1.5 eV for CdTe to 3.78 eV for ZnS, whereas the Hg compounds HgSe and HgTe are semimetals with negative bandgaps and HgS has two forms, α-HgS being a large-gap semiconductor and β-HgS being a semimetal. The II–VI compounds can be further mixed to form mixed II–VI compound alloys, such as the ternary alloys $Hg_xCd_{1-x}Te$ and $Hg_xCd_{1-x}Se$. The ternary II–VI alloys that include Hg can have a wide range of bandgaps covering the visible to the mid infrared spectral regions.

In addition to III–V and II–VI compounds, the IV–VI lead-salt compound semiconductors, PbS, PbSe, and PbTe, as well as their alloys like $Pb_xSn_{1-x}Te$ and PbS_xSe_{1-x}, are also useful. These lead-salt compounds are direct-gap semiconductors with small bandgaps in the range of 0.145–0.41 eV. These lead-salt compounds have the unusual property that their bandgaps increase with increasing temperature, whereas the bandgaps of other semiconductors decrease with increasing temperature.

By examining the data listed in Table 12.1, an important trend regarding the bandgaps and the refractive indices of semiconductors can be observed: *as the atomic weight of a component in a semiconductor increases by moving down a particular column of the periodic table, the bandgap decreases while the refractive index at a given optical wavelength corresponding to a photon energy below the bandgap increases.* These characteristics can be seen in the group IV elemental semiconductors as the bandgap decreases, while the refractive index increases, from C through Si to Ge. For the III–V compounds, these characteristics can be seen by comparing those compounds of the same group III element but different group V elements or those of the same group V element but different group III elements. For example, the bandgap decreases but the refractive index increases from AlP through AlAs to AlSb as the atomic weight increases from P to As to Sb among the group V elements. As another example, the bandgap decreases but the refractive index increases from AlAs through GaAs to InAs as the atomic weight increases from Al through Ga to In among the group III elements. Similar characteristics exist for II–VI compounds. Therefore, one can expect that among CdS, CdSe, and CdTe, for example, CdS has the largest bandgap while CdTe has the smallest bandgap. As another example, ZnSe has a larger bandgap than CdSe while HgSe has a negative bandgap.

Lattice-matched compounds

Two crystals that have the same lattice structure and the same lattice constant are *latticed matched*. Figure 12.2 shows the lattice constant as a function of bandgap

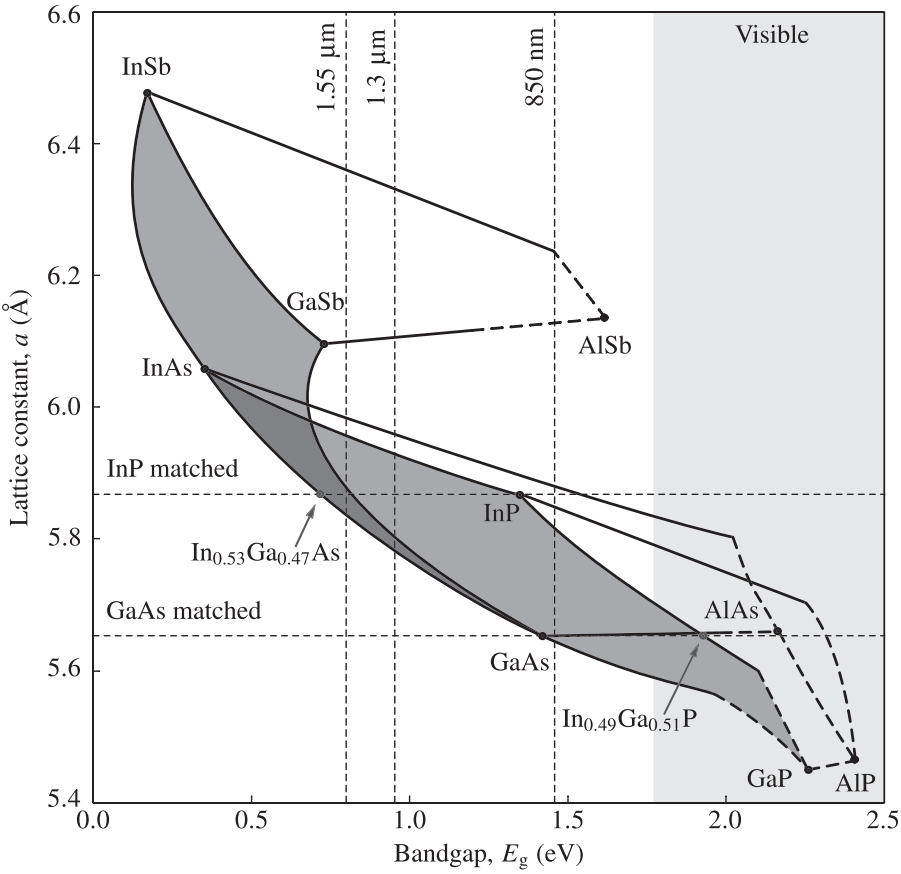


Figure 12.2 Lattice constant versus bandgap for III–V compound semiconductors. In this plot, a ternary compound lies on the curve connecting the two binary compounds that form the ternary alloy. Solid curves represent direct-gap semiconductors, and dashed curves represent indirect-gap semiconductors. A quaternary compound lies in the area defined by the four binary compounds that form the quaternary alloy. All compositions that are lattice matched to a given binary compound lie on the horizontal line passing through the binary compound. (Based on data from assorted sources.)

for III–V compounds at 300 K. In this plot, a ternary compound lies on the curve connecting the two binary compounds that form the ternary alloy. For example, the lattice constant and the bandgap of $Al_xGa_{1-x}As$ for $0 \leq x \leq 1$ can be found on the curve connecting AlAs and GaAs. A quaternary compound lies in the area defined by the four binary compounds that form the quaternary alloy. For example, the parameters of $In_{1-x}Ga_xAs_yP_{1-y}$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$ are found within the area bounded by the curves connecting InAs, GaAs, InP, and GaP. In Fig. 12.2, all compositions that are lattice matched to a given binary compound lie on the horizontal line passing through the binary compound. Because the lattice constants of different compounds

vary with temperature at different rates, two compositions that are lattice matched at a particular temperature are normally not matched at other temperatures.

As can be seen from Table 12.1 and Fig. 12.2, the lattice constants of AlAs and GaAs have a very small mismatch of only 0.13% at 300 K. They are the most closely lattice matched among all pairs of III–V binary compounds. In fact, they are perfectly lattice matched at 900 °C. Consequently, the lattice constant of the ternary compound $Al_xGa_{1-x}As$ varies very little with x, and $Al_xGa_{1-x}As$ is closely lattice matched to AlAs and GaAs over the entire composition range. Similar characteristics exist for the $AlP-Al_xGa_{1-x}P-GaP$ system and the $AlSb-Al_xGa_{1-x}Sb-GaSb$ system as well, but the lattice constants of $Al_xGa_{1-x}P$ and $Al_xGa_{1-x}Sb$ have larger variations with x than that of $Al_xGa_{1-x}As$. Other than these systems, a ternary compound can be lattice matched to a binary compound or to another ternary compound only at a particular composition but not over the entire composition range. Such lattice-matched compositions can be found using Fig. 12.2.

The quaternary compound alloys normally have a large flexibility for *lattice matching* to other compounds because each quaternary system has two variable composition parameters and, consequently, occupies an area, rather than a curve as for a ternary system, in Fig. 12.2. For example, the $In_{1-x}Ga_xAs_yP_{1-y}$ system, which lies in the lower left shaded area of Fig. 12.2, can be lattice matched to InP over a composition range of $0 \leq y \leq 1$ and $x \approx 0.47y$, and it can be lattice matched to GaAs over another composition range of $0 \leq y \leq 1$ and $1 - x \approx 0.49(1 - y)$.

As can be seen in Fig. 12.2, the bandgap of a ternary or a quaternary compound varies with the composition of the compound when the lattice constant of the compound is, for the purpose of lattice matching, kept at a fixed value as the composition varies. As the bandgap varies, the refractive index of the compound also varies and, as discussed above, they vary in opposite directions. A semiconductor optical waveguide, as discussed in Chapter 2, or a semiconductor heterostructure device, as discussed in Section 13.5, generally consists of layers of compound semiconductors of different compositions that are all lattice matched to a binary compound substrate, such as GaAs, InP, InAs, or GaSb, on which the structure is fabricated. The bandgaps and refractive indices of lattice-matched compounds are very important factors to be considered in the design and fabrication of semiconductor optical waveguides and heterostructure devices. In the following, the properties of two important systems, namely, $Al_xGa_{1-x}As$ lattice matched to a GaAs substrate and $In_{1-x}Ga_xAs_yP_{1-y}$ lattice matched to an InP substrate, are summarized.

$Al_xGa_{1-x}As/GaAs$

Over the entire composition range of $0 \leq x \leq 1$, $Al_xGa_{1-x}As$ is closely, though not perfectly, lattice matched to GaAs. Because this ternary compound is an alloy of indirect-gap AlAs and direct-gap GaAs, it is a direct-gap semiconductor for small values of x

in the range $0 \leq x < 0.45$ but becomes an indirect-gap semiconductor for large values of x in the range $0.45 < x \leq 1$. Its bandgap in electron volts at 300 K as a function of the composition parameter x can be described by

$$E_g(x) = 1.424 + 1.247x, \qquad \text{direct gap for } 0 \leq x < 0.45, \tag{12.3}$$

$$E_g(x) = 1.900 + 0.125x + 0.143x^2, \qquad \text{indirect gap for } 0.45 < x \leq 1. \tag{12.4}$$

The direct bandgap ranges from 1.424 to 1.985 eV, corresponding to an optical wavelength λ_g in the range between 870 and 625 nm. The indirect bandgap covers a range from 1.985 to 2.168 eV, corresponding to λ_g in the range between 625 and 572 nm. The bandgap of GaAs decreases with increasing temperature. It has a value of 1.5216 eV at 0 K. The temperature dependence of the bandgap of GaAs is

$$E_g = 1.5216 - \frac{5.405 \times 10^{-4}\, T^2}{T + 204}\ (\text{eV}). \tag{12.5}$$

The refractive index of $Al_xGa_{1-x}As$ is a function of x as well as of the optical wavelength because of dispersion. At an optical wavelength of $\lambda = 900$ nm, corresponding to a photon energy of 1.38 eV, which is below the bandgap of $Al_xGa_{1-x}As$ over the entire composition range, the refractive index as a function of x can be approximated by

$$n(x) = 3.593 - 0.710x + 0.091x^2, \qquad \text{at } \lambda = 900 \text{ nm for } 0 \leq x \leq 1. \tag{12.6}$$

We see, by examining the variations of $n(x)$ and $E_g(x)$ with the parameter x, that as the value of x increases, the bandgap of $Al_xGa_{1-x}As$ increases but its refractive index at the fixed wavelength of 900 nm decreases.

The refractive index of GaAs at 300 K as a function of optical wavelength in the spectral range of $\lambda \geq 870$ nm for photon energies below the GaAs bandgap is given by the following Sellmeier equation:

$$n^2 = 8.950 + \frac{2.054\lambda^2}{\lambda^2 - 0.390}, \tag{12.7}$$

where λ is in micrometers. The refractive index of GaAs varies with temperature approximately as

$$\frac{1}{n}\frac{\mathrm{d}n}{\mathrm{d}T} = 4.5 \times 10^{-5}\ \text{K}^{-1}. \tag{12.8}$$

$In_{1-x}Ga_xAs_yP_{1-y}$/InP

The $In_{1-x}Ga_xAs_yP_{1-y}$ quaternary compounds that are lattice matched to InP are direct-gap semiconductors over the entire lattice-matched composition range of $0 \leq y \leq 1$ and $x = 0.47y$. At 300 K, the bandgap in electron volts as a function of the composition

parameter y is given by

$$E_g(y) = 1.350 - 0.72y + 0.12y^2, \quad \text{direct gap for } 0 \le y \le 1 \text{ and } x = 0.47y. \quad (12.9)$$

Therefore, the direct bandgap of $In_{1-x}Ga_xAs_yP_{1-y}$ that is lattice matched to InP covers the range from 0.75 to 1.35 eV, corresponding to λ_g in the range between 919 nm and 1.65 μm. The bandgap of InP decreases with increasing temperature. It has a value of 1.4206 eV at 0 K. The temperature dependence of the bandgap of InP is

$$E_g = 1.4206 - \frac{4.906 \times 10^{-4} T^2}{T + 327} \qquad \text{(eV)}. \quad (12.10)$$

The refractive index is a function of the composition parameter y and optical wavelength. Two optical wavelengths, 1.3 and 1.55 μm, are of particular interest for lasers and LEDs based on the $In_{1-x}Ga_xAs_yP_{1-y}$/InP system because they lie in the windows of minimum dispersion and minimum loss, respectively, in silica fibers. The 0.954 and 0.8 eV photon energies of 1.3 and 1.55 μm wavelengths are below the bandgap of $In_{1-x}Ga_xAs_yP_{1-y}$ for $0 \le y \le 0.6$ and $0 \le y \le 0.9$, respectively, and $x = 0.47y$ for lattice matching to InP. At these wavelengths for the respective composition ranges, the refractive index of $In_{1-x}Ga_xAs_yP_{1-y}$ that is lattice matched to InP can be approximated by

$$n(y) = 3.205 + 0.34y + 0.21y^2, \qquad \text{at } \lambda = 1.3 \text{ μm for } 0 \le y \le 0.6, \quad (12.11)$$

$$n(y) = 3.166 + 0.26y + 0.09y^2, \qquad \text{at } \lambda = 1.55 \text{ μm for } 0 \le y \le 0.9. \quad (12.12)$$

From the relations in (12.9), (12.11), and (12.12), we see that as the value of the composition parameter y increases, the bandgap of $In_{1-x}Ga_xAs_yP_{1-y}$ lattice matched to InP decreases, but its refractive index at a fixed wavelength of 1.3 or 1.55 μm increases.

The refractive index of InP at 300 K as a function of optical wavelength in the spectral range of $\lambda \ge 920$ nm for photon energies below the InP bandgap is given by the following Sellmeier equation:

$$n^2 = 7.255 + \frac{2.316\lambda^2}{\lambda^2 - 0.3922}, \quad (12.13)$$

where λ is in micrometers. The refractive index of InP varies with temperature approximately as

$$\frac{1}{n}\frac{dn}{dT} = 2.7 \times 10^{-5} \text{ K}^{-1}. \quad (12.14)$$

EXAMPLE 12.1 An InGaAsP quaternary compound that is lattice matched to InP at 300 K has a bandgap optical wavelength of $\lambda_g = 1.223$ μm. Find the energy of its bandgap. Find its refractive indices at 1.3 and 1.55 μm wavelengths, respectively. What is the composition of this quaternary compound?

Solution For $\lambda_g = 1.223$ μm, the bandgap

$$E_g = \frac{hc}{\lambda_g} = \frac{1.2398}{1.223} \text{ eV} = 1.014 \text{ eV}.$$

According to (12.9), the composition parameter y can be found by solving

$$0.12y^2 - 0.72y + 1.35 = 1.014,$$

which yields $y = 0.51$. Using (12.11) and (12.12), we then find that the refractive indices are

$$n = 3.21 + 0.34 \times 0.51 + 0.21 \times 0.51^2 = 3.438$$

at $\lambda = 1.3$ μm and

$$n = 3.17 + 0.26 \times 0.51 + 0.09 \times 0.51^2 = 3.326$$

at $\lambda = 1.55$ μm.

Because $x = 0.47y$ for an InGaAsP compound that is lattice matched to InP, we find that $x = 0.47 \times 0.51 = 0.24$ for $y = 0.51$. Therefore, the composition of this quaternary compound is $In_{0.76}Ga_{0.24}As_{0.51}P_{0.49}$.

12.2 Electron and hole concentrations

The *electron concentration* in a semiconductor is the number of conduction electrons in the conduction bands per unit volume of the semiconductor, and the *hole concentration* is the number of holes in the valence bands per unit volume of the semiconductor. The concentrations of electrons and holes in a semiconductor are determined by many factors, including the bandgap and band structure of the semiconductor, the types and concentrations of the impurities doped in the semiconductor, temperature, and any external disturbances to the semiconductor.

Density of states

Because electrons are subject to the Pauli exclusion principle, which requires that no more than one electron can occupy the same quantum-mechanical state, the number of electrons in a particular energy band is determined by both the number of available states in that band and the probability of occupancy for each state. In a bulk semiconductor, the number of electron states in a given energy band is linearly proportional to the volume of the semiconductor. Therefore, a very useful concept is the *density of states*, which in a three-dimensional system like a bulk semiconductor is the number of states per unit material volume.

In a bulk semiconductor, the density of electron states within the energy range between E and $E + \mathrm{d}E$ for $E \geq E_\mathrm{c}$ near the conduction-band edge is

$$\rho_\mathrm{c}(E)\mathrm{d}E = \sum_\mathrm{c} \frac{4\pi(2m_\mathrm{c})^{3/2}}{h^3}(E - E_\mathrm{c})^{1/2}\mathrm{d}E = \frac{4\pi(2m_\mathrm{e}^*)^{3/2}}{h^3}(E - E_\mathrm{c})^{1/2}\mathrm{d}E, \quad (12.15)$$

where m_c is the *density of states effective mass* of an electron in a conduction band c and m_e^* is the density of states effective mass for electrons in all equivalent conduction bands. For example, there are six equivalent conduction-band minima in Si and four in Ge, but only one in a direct-gap semiconductor like GaAs. Therefore, $m_\mathrm{e}^* = 6^{2/3}m_\mathrm{c} = 1.08m_0$ for Si, $m_\mathrm{e}^* = 4^{2/3}m_\mathrm{c} = 0.55m_0$ for Ge, and $m_\mathrm{e}^* = m_\mathrm{c} = 0.067m_0$ for GaAs, where m_0 is the free electron mass. Similarly, the density of states within the energy range between E and $E + \mathrm{d}E$ for $E \leq E_\mathrm{v}$ near the valence-band edge is

$$\rho_\mathrm{v}(E)\mathrm{d}E = \sum_\mathrm{v} \frac{4\pi(2m_\mathrm{v})^{3/2}}{h^3}(E_\mathrm{v} - E)^{1/2}\mathrm{d}E = \frac{4\pi(2m_\mathrm{h}^*)^{3/2}}{h^3}(E_\mathrm{v} - E)^{1/2}\mathrm{d}E, \quad (12.16)$$

where m_v is the density of states effective mass of a hole in a valence band v and m_h^* is the density of states effective mass of holes in all valence bands that are degenerate at the valence-band edge. In both direct-gap and indirect-gap semiconductors including GaAs, Si, and Ge, there are normally two hole bands, known as the *heavy-hole band* and the *light-hole band*, of different effective masses that are degenerate at the valence-band edge. Therefore,

$$m_\mathrm{h}^* = (m_\mathrm{hh}^{3/2} + m_\mathrm{lh}^{3/2})^{2/3}, \quad (12.17)$$

where m_hh and m_lh are the effective masses in the heavy-hole and light-hole bands, respectively. We have $m_\mathrm{h}^* = 0.56m_0$ for Si, $m_\mathrm{h}^* = 0.31m_0$ for Ge, and $m_\mathrm{h}^* = 0.52m_0$ for GaAs.

Carriers in equilibrium

In thermal equilibrium, the probability of occupancy for a given electron state at an energy E is described by the Fermi–Dirac function $f(E)$ given in (12.1). Therefore, the concentration of conduction electrons (negatively charged carriers) whose energies fall between E and $E + \mathrm{d}E$ is

$$n_0(E)\mathrm{d}E = f(E)\rho_\mathrm{c}(E)\mathrm{d}E \qquad (\mathrm{m}^{-3}), \quad (12.18)$$

and the concentration of holes (positively charged carriers) whose energies fall between E and $E + \mathrm{d}E$ is

$$p_0(E)\mathrm{d}E = [1 - f(E)]\rho_\mathrm{v}(E)\mathrm{d}E \qquad (\mathrm{m}^{-3}). \quad (12.19)$$

Note that the probability of finding a hole at an energy E is $1 - f(E)$ because a hole is an unoccupied electron state. The total concentrations of electrons and holes in thermal equilibrium are, respectively,

$$n_0 = \int_{E_c}^{\infty} n_0(E)\mathrm{d}E = \int_{E_c}^{\infty} f(E)\rho_c(E)\mathrm{d}E = \int_{E_c}^{\infty} \frac{\rho_c(E)\mathrm{d}E}{\mathrm{e}^{(E-E_F)/k_BT} + 1} \tag{12.20}$$

and

$$p_0 = \int_{-\infty}^{E_v} p_0(E)\mathrm{d}E = \int_{-\infty}^{E_v} [1 - f(E)]\rho_v(E)\mathrm{d}E = \int_{-\infty}^{E_v} \frac{\rho_v(E)\mathrm{d}E}{\mathrm{e}^{(E_F-E)/k_BT} + 1}. \tag{12.21}$$

Using (12.15) and (12.16), the electron and hole concentrations given by (12.20) and (12.21) can be expressed as

$$n_0 = N_c(T)F_{1/2}\left(\frac{E_F - E_c}{k_BT}\right), \tag{12.22}$$

$$p_0 = N_v(T)F_{1/2}\left(\frac{E_v - E_F}{k_BT}\right), \tag{12.23}$$

respectively, where N_c and N_v are the *effective densities of states* for conduction and valence bands, respectively, defined as

$$N_c(T) = 2\left(\frac{2\pi m_e^* k_B T}{h^2}\right)^{3/2}, \qquad N_v(T) = 2\left(\frac{2\pi m_h^* k_B T}{h^2}\right)^{3/2}, \tag{12.24}$$

and $F_{1/2}(\xi)$ is the Fermi–Dirac integral of order $1/2$ defined as

$$F_{1/2}(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{x^{1/2}\mathrm{d}x}{\mathrm{e}^{(x-\xi)} + 1} = \mathrm{e}^{\xi} \frac{2}{\sqrt{\pi}} \int_0^{\infty} \frac{x^{1/2}\mathrm{d}x}{\mathrm{e}^{x} + \mathrm{e}^{\xi}}. \tag{12.25}$$

The relations given in (12.22) and (12.23) for the electron and hole concentrations are very general. They are valid for a semiconductor in thermal equilibrium with its environment no matter whether the semiconductor is doped with impurity or not. They are also valid for any value of E_F with respect to E_c and E_v. It can be seen clearly from these relations that the values of n_0 and p_0 are strongly dependent on the value of the Fermi energy E_F. The impurities in a semiconductor affect the electron and hole concentrations through changing the value of the Fermi energy E_F.

From (12.22) and (12.23), it can be clearly seen that both n_0 and p_0 are determined by the Fermi–Dirac integral $F_{1/2}(\xi)$. For this reason, the characteristics of this integral as a function of its variable ξ are plotted in Fig. 12.3. For $\xi = 0$, we find that $F_{1/2}(0) = 0.765\,15$. We see that $F_{1/2}(\xi) \approx \mathrm{e}^{\xi}$ for large negative values of ξ. This approximation has an error of less than 1% for $\xi \leq -3.6$. Therefore, when the Fermi level is far away from both band edges so that $(E_c - E_F)/k_BT \geq 3.6$ and $(E_F - E_v)/k_BT \geq 3.6$, the

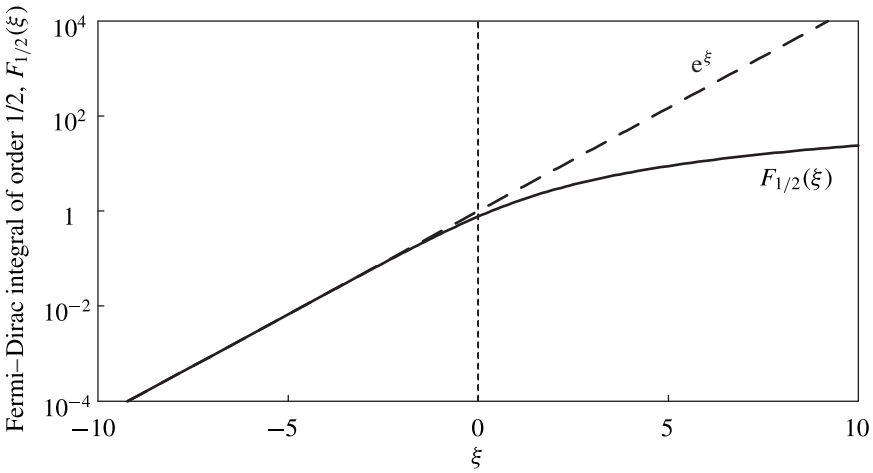


Figure 12.3 Fermi–Dirac integral of order 1/2, $F_{1/2}(\xi)$, as a function of the variable ξ. Also plotted is the behavior of the exponential function e^{ξ} to show that $F_{1/2}(\xi) \approx e^{\xi}$ for $\xi \ll -1$.

electron and hole concentrations given in (12.22) and (12.23) can be approximated to an accuracy of better than 99% by the following expressions:

$$n_0 = N_c(T)e^{-(E_c - E_F)/k_B T}, \tag{12.26}$$

$$p_0 = N_v(T)e^{-(E_F - E_v)/k_B T}. \tag{12.27}$$

In an *intrinsic semiconductor*, impurities contribute negligibly to the electron and hole concentrations. Practically all of the conduction electrons in an intrinsic semiconductor come from thermal excitation from the valence bands. Consequently, there are as many holes as electrons so that $n_0 = p_0$. For most semiconductors of interest, we have $E_g \gg k_B T$ at a temperature below the melting temperature of the semiconductor. Under these conditions, we can apply the relation $n_0 = p_0$ to (12.26) and (12.27) to find that the Fermi level of an intrinsic semiconductor is given by

$$E_{Fi} = \frac{E_c + E_v}{2} + \frac{k_B T}{2} \ln \frac{N_v}{N_c} = \frac{E_c + E_v}{2} + \frac{3k_B T}{4} \ln \frac{m_h^*}{m_e^*}, \tag{12.28}$$

which lies very close to the middle of the bandgap. Therefore, we have

$$n_0 = p_0 = n_i(T) \approx \sqrt{N_c N_v} e^{-E_g/2k_B T} = 2\left(\frac{2\pi k_B T}{h^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} e^{-E_g/2k_B T} \tag{12.29}$$

for an intrinsic semiconductor, where the *intrinsic carrier concentration*, n_i, is a function of temperature.

EXAMPLE 12.2 Calculate the values of the effective densities of states N_c and N_v for GaAs at 300 K. Use the results to find the electron and hole concentrations and the Fermi level for intrinsic GaAs at 300 K.

Solution For GaAs, $m_e^* = 0.067m_0$ and $m_h^* = 0.52m_0$. Using the constants $m_0 = 9.11 \times 10^{-31}$ kg, $k_B = 1.38 \times 10^{-23}$ J K^{-1}, and $h = 6.626 \times 10^{-34}$ J s, the effective densities of states at $T = 300$ K can be calculated from (12.24) to be

$$N_c = 2 \times \left[\frac{2\pi \times 0.067 \times 9.11 \times 10^{-31} \times 1.38 \times 10^{-23} \times 300}{(6.626 \times 10^{-34})^2}\right]^{3/2} \text{m}^{-3}$$
$$= 4.35 \times 10^{23} \text{ m}^{-3}$$

and

$$N_v = 2 \times \left[\frac{2\pi \times 0.52 \times 9.11 \times 10^{-31} \times 1.38 \times 10^{-23} \times 300}{(6.626 \times 10^{-34})^2}\right]^{3/2} \text{m}^{-3}$$
$$= 9.41 \times 10^{24} \text{ m}^{-3}.$$

At $T = 300$ K, the bandgap of GaAs is $E_g = 1.424$ eV, and $k_B T = 25.9$ meV. Therefore,

$$\frac{E_g}{k_B T} = \frac{1.424}{25.9 \times 10^{-3}} = 54.98 \text{ and } \frac{E_g}{2k_B T} = 27.49.$$

Because the Fermi level of an intrinsic semiconductor lies close to the center of the bandgap, $(E_c - E_{Fi})/k_B T \approx (E_{Fi} - E_v)/k_B T \approx E_g/(2k_B T) = 27.49 \gg 3.6$. Therefore, the approximations given in (12.26) and (12.27) and, consequently, those given in (12.28) and (12.29) are all valid in this situation. For this intrinsic GaAs, we then find from (12.29) that

$$n_0 = p_0 = n_i = \sqrt{4.35 \times 10^{23} \times 9.41 \times 10^{24}} \times e^{-27.49} \text{ m}^{-3}$$
$$= 2.33 \times 10^{12} \text{ m}^{-3},$$

and from (12.28) that

$$E_{Fi} - \frac{E_c + E_v}{2} = \frac{25.9}{2} \times \ln \frac{9.41 \times 10^{24}}{4.35 \times 10^{23}} \text{ meV} = 39.8 \text{ meV}.$$

Because the center of the bandgap is at $(E_c + E_v)/2$, this intrinsic Fermi level is 39.8 meV above the center of the bandgap. The reason for this shift of E_F above the bandgap center is that $m_e^* < m_h^*$ for GaAs. Compared with the bandgap of 1.424 eV, this shift away from the bandgap center is small, verifying the statement that the intrinsic Fermi level lies very close to the center of the bandgap.

In an *extrinsic semiconductor*, however, n_0 and p_0 are different from n_i because of the contribution of carriers from the impurities in a semiconductor. An impurity atom that can be positively ionized to contribute a conduction electron is a *donor*, and one that can be negatively ionized to contribute a hole to the valence bands is an *acceptor*. In general, the requirement for charge neutrality in a semiconductor leads to the

following relation:

$$n_0 + N_a^- = p_0 + N_d^+, \tag{12.30}$$

where N_a^- is the concentration of the immobile *negatively ionized acceptors* and N_d^+ is that of the immobile *positively ionized donors*. When $N_d^+ > N_a^-$, the semiconductor is an *n-type semiconductor* with $n_0 > p_0$. In an n-type semiconductor, electrons are the *majority carriers*, and holes are the *minority carriers*. When $N_a^- > N_d^+$, the semiconductor is a *p-type semiconductor* with $p_0 > n_0$. In a *p-type semiconductor*, holes are the majority carriers, and electrons are the minority carriers.

The Fermi level of an intrinsic semiconductor lies very close to the middle of the bandgap and is only a weak function of temperature. In contrast, the Fermi level of an extrinsic semiconductor is a function of the types and concentrations of the impurities. *In an n-type semiconductor, it moves towards the conduction-band edge; in a p-type semiconductor, it moves towards the valence-band edge.* Up to a moderate doping concentration, the Fermi level remains in the bandgap. Such a semiconductor is called a *nondegenerate semiconductor*.

For a nondegenerate semiconductor, no matter whether it is intrinsic or extrinsic, (12.26) and (12.27) are valid for n_0 and p_0, respectively. Therefore, the carrier concentrations of a nondegenerate semiconductor in thermal equilibrium satisfy the following *law of mass action*:

$$n_0 p_0 = n_i^2(T). \tag{12.31}$$

The values of n_0 and p_0 in a nondegenerate semiconductor can be found by solving (12.30) and (12.31) simultaneously. Then, from (12.26) and (12.27), the Fermi level of a nondegenerate semiconductor can be found:

$$E_F = E_c - k_B T \ln \frac{N_c}{n_0} = E_v + k_B T \ln \frac{N_v}{p_0}. \tag{12.32}$$

This relation is valid for both intrinsic and extrinsic situations so long as the semiconductor is nondegenerate. In the intrinsic case, this relation is equivalent to (12.28), as $E_F = E_{Fi}$ for an intrinsic semiconductor. In the extrinsic case, it is valid for both n-type and p-type semiconductors.

In the case when $N_c \gg N_d^+ - N_a^- \gg n_i$, the semiconductor is nondegenerate and has $n_0 \approx N_d^+ - N_a^- \gg p_0 \approx n_i^2/n_0$. The Fermi level for such a nondegenerate n-type semiconductor shifts from E_{Fi} toward the conduction-band edge; it can be approximated as

$$E_F \approx E_c - k_B T \ln \frac{N_c}{N_d^+ - N_a^-}. \tag{12.33}$$

In the case when $N_v \gg N_a^- - N_d^+ \gg n_i$, the semiconductor is nondegenerate and has $p_0 \approx N_a^- - N_d^+ \gg n_0 \approx n_i^2/p_0$. The Fermi level for such a nondegenerate p-type

semiconductor shifts from E_{Fi} toward the valence-band edge; it can be approximated as

$$E_F \approx E_v + k_B T \ln \frac{N_v}{N_a^- - N_d^+}. \tag{12.34}$$

These approximations fail when the net impurity concentration is sufficiently high to make the semiconductor degenerate or when it is too low so that n_0 and p_0 remain close to n_i.

EXAMPLE 12.3 A piece of n-type GaAs is doped with a net impurity concentration of $N_d^+ - N_a^- = 5 \times 10^{18}$ m^{-3}. Is it degenerate or nondegenerate? Find its electron and hole concentrations and its Fermi level at $T = 300$ K. How much is the shift of the Fermi level, measured from the intrinsic Fermi level, caused by the doping of the impurity?

Solution From Example 12.2, we know that $N_c = 4.35 \times 10^{23}$ m^{-3} at $T = 300$ K. We also find from Example 12.2 that $n_i = 2.33 \times 10^{12}$ m^{-3}. This n-type GaAs is nondegenerate because $N_d^+ - N_a^- \ll N_c$ for the given impurity concentration. The general procedure for finding n_0 and p_0 is to solve the simultaneous equations of $n_0 - p_0 = N_d^+ - N_a^-$, from (12.30), and $n_0 p_0 = n_i^2$, from (12.31), with the known values of $N_d^+ - N_a^-$ and n_i. However, because $N_d^+ - N_a^- \gg n_i$ for the given problem, we find that

$$n_0 \approx N_d^+ - N_a^- = 5 \times 10^{18} \text{ m}^{-3}$$

and

$$p_0 = \frac{n_i^2}{n_0} = 1.1 \times 10^6 \text{ m}^{-3}.$$

The Fermi level for this nondegenerate n-type GaAs can be found by using (12.32):

$$E_F = E_c - 25.9 \times \ln \frac{4.35 \times 10^{23}}{5 \times 10^{18}} \text{ meV} = E_c - 294.6 \text{ meV}.$$

Compared with the intrinsic Fermi level, E_{Fi}, found in Example 12.2, we find that

$$\begin{aligned} E_F - E_{Fi} &= \frac{E_c - E_v}{2} - 294.6 \text{ meV} - 39.8 \text{ meV} \\ &= \frac{E_g}{2} - 334.4 \text{ meV} \\ &= \frac{1424}{2} \text{ meV} - 334.4 \text{ meV} \\ &= 377.6 \text{ meV}. \end{aligned}$$

Therefore, the Fermi level of this n-type GaAs is shifted away from the intrinsic Fermi level by 377.6 meV toward the conduction-band edge.

In a heavily p-doped semiconductor, the Fermi level can move into the valence band. Similarly, the Fermi level can move into the conduction band in a heavily n-doped semiconductor. When the Fermi level lies within a valence band or within a conduction band, we have a *degenerate semiconductor*. The electron and hole concentrations are still given by (12.22) and (12.23), respectively. Nevertheless, the law of mass action expressed in (12.31) and the position of the Fermi level given by (12.32) are not valid for a degenerate semiconductor because either (12.26) or (12.27) can have a significant error when the Fermi level lies above the conduction-band edge or below the valence-band edge.

EXAMPLE 12.4 What is the impurity doping concentration required for n-type GaAs to become degenerate at 300 K?

Solution An n-type semiconductor becomes degenerate when its Fermi level lies at or above its conduction-band edge: $E_F \geq E_c$. From (12.22), we find that this condition requires that

$$n_0 \geq N_c(T)F_{1/2}(0) = 0.765\,15N_c(T)$$

at a given temperature T. For GaAs at $T = 300$ K, we find that $n_0 \geq 3.33 \times 10^{23}$ m^{-3} from this relation because $N_c = 4.35 \times 10^{23}$ m^{-3}, as found in Example 12.2. Because $n_0 \gg n_i \gg p_0$ in this situation, we find from (12.30) by neglecting p_0 in comparison to n_0 that the required impurity concentration for n-type GaAs to become degenerate at 300 K is simply

$$N_d^+ - N_a^- = n_0 \geq 3.33 \times 10^{23}\ \text{m}^{-3}.$$

Carriers in quasi-equilibrium

Electrons and holes in excess of their respective thermal equilibrium concentrations can be generated in a semiconductor by current injection or optical excitation. When this situation occurs, the carriers will relax toward thermal equilibrium through both intraband and interband processes. Intra-conduction-band relaxation allows electrons to reach thermal equilibrium among themselves through electron–electron collisions and electron–phonon interactions, while intra-valence-band relaxation allows holes to also reach thermal equilibrium among themselves through similar processes. The time constants of such intraband relaxation processes are generally in the range of 10 fs to 1 ps, depending on the concentration of the excess carriers. Thermal equilibrium between electrons and holes is reached through electron–hole recombination processes, the time constants of which typically vary from the order of 100 ps to the order of 1 ms, depending on the properties of the specific semiconductor and the carrier concentration. Consequently, thermal equilibrium in the conduction bands and that in the valence

bands can be separately reached in less than 1 ps, but complete thermal equilibrium for the entire system would not usually be reached for at least a few hundred picoseconds. If the external excitation persists, the semiconductor can reach a *quasi-equilibrium* state in which electrons and holes are not characterized by a common Fermi level but are characterized by two separate *quasi-Fermi levels*. In such a quasi-equilibrium state, instead of a single Fermi–Dirac distribution function given in (12.1) for both conduction and valence bands, the probability of occupancy in the conduction bands and that in the valence bands are described by two separate Fermi–Dirac distribution functions:

$$f_{\mathrm{c}}(E) = \frac{1}{\mathrm{e}^{(E-E_{\mathrm{Fc}})/k_{\mathrm{B}}T} + 1}, \tag{12.35}$$

for the conduction bands, and

$$f_{\mathrm{v}}(E) = \frac{1}{\mathrm{e}^{(E-E_{\mathrm{Fv}})/k_{\mathrm{B}}T} + 1}, \tag{12.36}$$

for the valence bands, where E_{Fc} and E_{Fv} are quasi-Fermi levels for the conduction and valence bands, respectively.

As an intrinsic property of the band structure, the densities of states, $\rho_{\mathrm{c}}(E)$ and $\rho_{\mathrm{v}}(E)$, given in (12.15) and (12.16) for the conduction and valence bands, respectively, are independent of equilibrium or nonequilibrium of the carriers. Therefore, in quasi-equilibrium, the electron concentration as a function of energy is

$$n(E)\mathrm{d}E = f_{\mathrm{c}}(E)\rho_{\mathrm{c}}(E)\mathrm{d}E \qquad (\mathrm{m}^{-3}), \tag{12.37}$$

and the hole concentration as a function of energy is

$$p(E)\mathrm{d}E = [1 - f_{\mathrm{v}}(E)]\rho_{\mathrm{v}}(E)\mathrm{d}E \qquad (\mathrm{m}^{-3}). \tag{12.38}$$

The total concentrations of electrons and holes in quasi-equilibrium are, respectively,

$$n = \int_{E_{\mathrm{c}}}^{\infty} n(E)\mathrm{d}E = \int_{E_{\mathrm{c}}}^{\infty} f_{\mathrm{c}}(E)\rho_{\mathrm{c}}(E)\mathrm{d}E = \int_{E_{\mathrm{c}}}^{\infty} \frac{\rho_{\mathrm{c}}(E)\mathrm{d}E}{\mathrm{e}^{(E-E_{\mathrm{Fc}})/k_{\mathrm{B}}T} + 1} \tag{12.39}$$

and

$$p = \int_{-\infty}^{E_{\mathrm{v}}} p(E)\mathrm{d}E = \int_{-\infty}^{E_{\mathrm{v}}} [1 - f_{\mathrm{v}}(E)]\rho_{\mathrm{v}}(E)\mathrm{d}E = \int_{-\infty}^{E_{\mathrm{v}}} \frac{\rho_{\mathrm{v}}(E)\mathrm{d}E}{\mathrm{e}^{(E_{\mathrm{Fv}}-E)/k_{\mathrm{B}}T} + 1}. \tag{12.40}$$

Using (12.15) and (12.16), the electron and hole concentrations for a semiconductor in quasi-equilibrium can be expressed in a form similar to that of (12.22) and (12.23):

$$n = N_{\mathrm{c}}(T)F_{1/2}\left(\frac{E_{\mathrm{Fc}} - E_{\mathrm{c}}}{k_{\mathrm{B}}T}\right), \tag{12.41}$$

$$p = N_{\mathrm{v}}(T)F_{1/2}\left(\frac{E_{\mathrm{v}} - E_{\mathrm{Fv}}}{k_{\mathrm{B}}T}\right). \tag{12.42}$$

We find from (12.41) and (12.42) that *the electron concentration, n, and the hole concentration, p, in a quasi-equilibrium state are completely quantified by the quasi-Fermi levels,* E_{Fc} *and* E_{Fv}*, respectively.* We then find

$$np = N_c(T)N_v(T)F_{1/2}\left(\frac{E_c - E_{Fc}}{k_B T}\right)F_{1/2}\left(\frac{E_{Fv} - E_v}{k_B T}\right). \tag{12.43}$$

In the situation when $(E_c - E_{Fc})/k_B T \geq 3.6$ and $(E_{Fv} - E_v)/k_B T \geq 3.6$, the quasi-equilibrium electron and hole concentrations given in (12.41) and (12.42) can be approximated to an accuracy of better than 99% by

$$n = N_c(T)e^{-(E_c - E_{Fc})/k_B T}, \tag{12.44}$$

$$p = N_v(T)e^{-(E_{Fv} - E_v)/k_B T}. \tag{12.45}$$

Then,

$$np = N_c(T)N_v(T)e^{-(E_g - \Delta E_F)/k_B T} = n_i^2(T)e^{\Delta E_F/k_B T} = n_0 p_0 e^{\Delta E_F/k_B T}, \tag{12.46}$$

where

$$\Delta E_F = E_{Fc} - E_{Fv} \tag{12.47}$$

is the separation between the quasi-Fermi levels. Because of the splitting of the quasi-Fermi levels in a quasi-equilibrium state, the law of mass action given in (12.31) is no longer valid but is replaced by (12.46). Note that these approximations are not valid if the quasi-equilibrium electron and hole concentrations are high enough to push any one of the quasi-Fermi levels to the vicinity of any band edge or beyond. Such a situation can happen even in an intrinsic semiconductor under high electrical or optical excitation of carriers.

In a quasi-equilibrium state, high concentrations of electrons and holes that are in excess of equilibrium concentrations can be generated by electrical or optical excitation. Compared with the equilibrium electron and hole concentrations, n_0 and p_0 given in (12.22) and (12.23), respectively, we find that $n > n_0$ if $E_{Fc} > E_F$ and $p > p_0$ if $E_F > E_{Fv}$. Therefore, the existence of quasi-equilibrium electron and hole concentrations that are higher than the equilibrium concentrations is characterized by the splitting of quasi-Fermi levels with $\Delta E_F > 0$. Quasi-equilibrium in a semiconductor is maintained when the carrier generation rate is equal to the carrier recombination rate.

EXAMPLE 12.5 An equal number of excess electrons and holes of a concentration of $\Delta n = \Delta p = 5 \times 10^{18}$ m^{-3} is generated at $T = 300$ K in an intrinsic GaAs sample by optical excitation. Find the quasi-Fermi levels, E_{Fc} and E_{Fv}. What is the separation between these quasi-Fermi levels?

Solution Because $n_i = 2.33 \times 10^{12}$ m$^{-3} \ll \Delta n = \Delta p$, we have $n = p = n_i + \Delta n \approx \Delta n = 5 \times 10^{18}$ m^{-3} in this situation. From the values of $N_c = 4.35 \times 10^{23}$ m^{-3}

and $N_v = 9.41 \times 10^{24}$ m^{-3} found in Example 12.2 for GaAs at 300 K, we find that $n \ll N_c$ and $p \ll N_v$. Therefore, both E_{Fc} and E_{Fv} are still sufficiently far away from the band edges so that (12.44) and (12.45) are valid. We then find from (12.44) that

$$\begin{aligned} E_{Fc} &= E_c - k_B T \ln \frac{N_c}{n} \\ &= E_c - 25.9 \times \ln \frac{4.35 \times 10^{23}}{5 \times 10^{18}} \text{ meV} \\ &= E_c - 294.6 \text{ meV}, \end{aligned}$$

and from (12.45) that

$$\begin{aligned} E_{Fv} &= E_v + k_B T \ln \frac{N_v}{p} \\ &= E_v + 25.9 \times \ln \frac{9.41 \times 10^{24}}{5 \times 10^{18}} \text{ meV} \\ &= E_v + 374.2 \text{ meV}. \end{aligned}$$

Therefore, E_{Fc} lies at 294.6 meV below the conduction-band edge, and E_{Fv} lies at 374.2 meV above the valence-band edge. Because the bandgap of GaAs at 300 K is $E_g = 1.424$ eV, the separation of these two quasi-Fermi levels is

$$\begin{aligned} \Delta E_F = E_{Fc} - E_{Fv} &= E_c - E_v - 294.6 \text{ meV} - 374.2 \text{ meV} \\ &= E_g - 668.8 \text{ meV} \\ &= 755.2 \text{ meV}. \end{aligned}$$

12.3 Carrier recombination

In a semiconductor, electrons in the conduction bands and holes in the valence bands can be generated through many mechanisms, including thermal excitation, current injection, and optical excitation. Meanwhile, an electron in a conduction band and a hole in a valence band can be eliminated together through a recombination process. In an equilibrium state, electron–hole generation is exactly balanced by electron–hole recombination.

Recombination processes

There are many different electron–hole recombination processes. Based on the mechanisms responsible for these processes, they are classified into three general categories: (1) the *Shockley–Read recombination* processes, (2) the *bimolecular recombination* processes, and (3) the *Auger recombination* processes. These basic mechanisms are schematically illustrated in Fig. 12.4. A Shockley–Read process involves one carrier at a time; a bimolecular process takes place with an electron and a hole simultaneously;

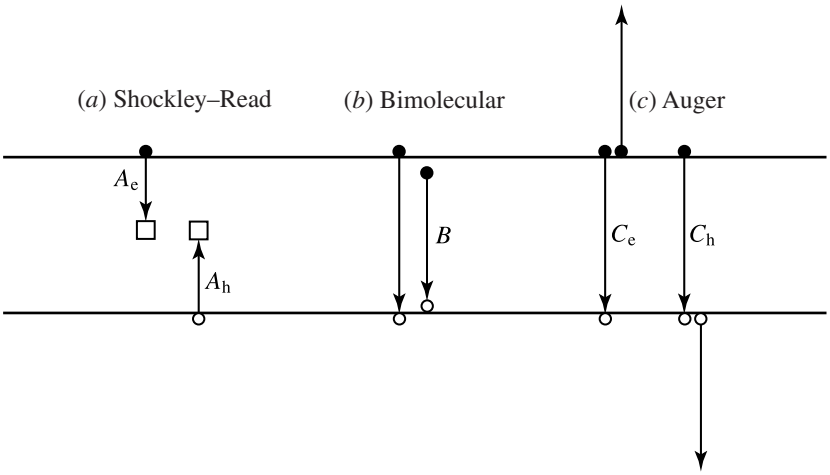


Figure 12.4 Carrier recombination processes in a semiconductor. (*a*) Shockley–Read recombination processes: (i) A_e, electron capture and (ii) A_h, hole capture. (*b*) Bimolecular recombination processes: (i) band-to-band recombination and (ii) exciton recombination. (*c*) Auger recombination processes: (i) C_e, two electrons and one hole and (ii) C_h, two holes and one electron.

an Auger process is a three-body process with three participating carriers at the same time. Notwithstanding these differences, any of the three types of recombination processes ends with the annihilation of one electron with one hole. Depending on whether or not electron–hole recombination in a particular process results in the emission of electromagnetic radiation, a process can also be classified as a *radiative recombination* process or as a *nonradiative recombination* process.

A Shockley–Read process is a recombination process that takes place through the capture of one carrier, either an electron or a hole, at a time by *recombination centers*, as illustrated in Fig. 12.4(*a*). A recombination center is created by an imperfection, such as a defect or an impurity, in a semiconductor. Its energy level lies somewhere in the bandgap of the semiconductor. The density of the recombination centers in a given piece of semiconductor is determined by the density of the defects and impurities in the semiconductor and is independent of the electron or hole concentration. There are four processes associated with the recombination centers: electron capture, hole capture, electron emission, and hole emission. A recombination center is available for the capture of an electron only if it is not already occupied by an electron, whereas it is available for the capture of a hole only if it is occupied by an electron.

Two basic time constants can be defined: $\tau_{e0}^{-1} = v_{th}\sigma_e N_r$ for electrons and $\tau_{h0}^{-1} = v_{th}\sigma_h N_r$ for holes, where v_{th} is the thermal velocity of an electron, σ_e and σ_h are the capture cross sections[1] of the recombination center for an electron and a hole, respectively, and N_r is the concentration of the recombination centers. The physical meaning

[1] Note that σ_e here is not to be confused with the emission cross section of a laser transition.

of τ_{e0} is that n/τ_{e0} is the electron capture rate if all recombination centers are empty and are thus able to trap electrons, whereas that of τ_{h0} is that p/τ_{h0} is the hole capture rate if all recombination centers are occupied and are thus able to trap holes. In many situations, however, the recombination centers are neither all empty nor all occupied. In such situations, the Shockley–Read recombination lifetimes of electrons and holes are different from τ_{e0} and τ_{h0}, respectively.

After a recombination center first captures a carrier, either an electron or a hole, a Shockley–Read electron–hole recombination process is completed only if the recombination center subsequently captures another carrier of opposite charge to result in the annihilation of an electron–hole pair. A carrier can also be captured and then reemitted without completing the recombination process. The net Shockley–Read recombination rate for electrons is the difference of the electron capture rate and the electron emission rate by recombination centers, whereas that for holes is the difference of the hole capture rate and the hole emission rate. Because only one carrier is captured at a time, the rates of electron capture and hole capture in the Shockley–Read recombination processes are linearly proportional to the total electron and hole concentrations, respectively. We therefore express the net electron recombination rate in the form of $A_e(n - n_0)$ and the net hole recombination rate in the form of $A_h(p - p_0)$. Because a completed recombination process always results in the annihilation of an electron–hole pair, the net electron and hole recombination rates are necessarily equal: $A_e(n - n_0) = A_h(p - p_0)$ though A_e and A_h can be different if the excess electron concentration $\Delta n = n - n_0$ is different from the excess hole concentration $\Delta p = p - p_0$. The net Shockley–Read recombination rate is found from balancing the capture and emission of electrons and holes in steady state to be

$$R_{SR} = A_e(n - n_0) = A_h(p - p_0) = \frac{np - n_0 p_0}{\tau_{h0}(n + n_1) + \tau_{e0}(p + p_1)}, \tag{12.48}$$

where n_1 and p_1 characterize the emission rates of the recombination centers for electrons and holes, respectively, and $n_1 p_1 = n_0 p_0 = n_i^2(T)$. We see from this relation that the two coefficients A_e and A_h both vary with temperature, the doping condition of the semiconductor, and the excess carrier concentrations. *A Shockley–Read recombination process can be either radiative or nonradiative, depending on the type of recombination centers involved in the process.*

A bimolecular recombination process always involves an electron and a hole at the same time. As illustrated in Fig. 12.4(*b*), there are primarily two types of bimolecular electron–hole recombination processes: *band-to-band recombination*, which takes place between an electron in a conduction band and a hole in a valence band, and *exciton recombination*, which is the recombination of an electron–hole pair that forms a free or bound exciton. An electron–hole pair in a semiconductor can be held together by their Coulomb attraction to form an exciton like an electron–proton pair forming a hydrogen atom. A *free exciton* is free to wander around in the semiconductor; its energy

is reduced by the energy needed to hold the electron and hole together and is slightly less than the bandgap of the semiconductor. A *bound exciton* is localized and bound to an impurity center in the semiconductor; its energy depends on the properties of the impurity and is generally lower than that of a free exciton in the same semiconductor. The bimolecular recombination processes have the same contribution to the rate of electron recombination and the rate of hole recombination. This rate is proportional to the product of the electron and hole concentrations and can be expressed as Bnp, where B is the *bimolecular recombination coefficient*. *Bimolecular recombination processes are radiative processes.*

An Auger recombination process is a three-body process that requires the participation on each occasion of two electrons and one hole, with a rate of $C_e n^2 p$, or one electron and two holes, with a rate of $C_h n p^2$, where C_e and C_h are the *Auger recombination coefficients*. The total Auger recombination rate for both electrons and holes is $C_e n^2 p + C_h n p^2$. As illustrated in Fig. 12.4(c), in an Auger recombination process, the energy released by band-to-band recombination of an electron and a hole is picked up by a third carrier, either another electron or another hole, as kinetic energy of the third carrier. This energy is eventually converted to the thermal energy of the semiconductor lattice as the excited third carrier relaxes toward the band edge. Consequently, *an Auger process is nonradiative.*

Each recombination process has a corresponding inverse process for generating free electrons and holes. Irrespective of the absence or presence of an external excitation such as current injection or optical excitation, free electrons and holes are continuously generated by thermal excitation through these inverse processes. The inverse processes of electron capture and hole capture in the Shockley–Read processes are electron emission and hole emission, respectively. An electron captured by a recombination center can be thermally reemitted back to the conduction band, and a hole captured by a recombination center can be thermally reemitted back to the valence band. Similarly to the difference in electron and hole capture rates by the recombination centers, the processes of electron and hole emission by the recombination centers generally have different thermal generation rates for electrons and holes. The thermal generation rates of electrons and holes in the Shockley–Read processes are already accounted for in the net Shockley–Read recombination rate given in (12.48). The inverse process of band-to-band bimolecular recombination is the generation of an electron–hole pair by thermal excitation of a valence electron to the conduction band. This process has the same generation rate of Bn_0p_0 for electrons and holes as they are generated in pairs. The inverse of an Auger process can also generate additional electrons and holes. This process also generates electrons and holes in pairs and has the same generation rate of $C_e n_0^2 p_0 + C_h n_0 p_0^2$ for electrons and holes. To summarize, the total *thermal generation rates* for electrons, G_e^0, and for holes, G_h^0, are generally different in the presence of recombination centers, but are the same if the density of the recombination centers is very small compared to the equilibrium electron and hole concentrations.

These thermal generation rates are a function of temperature and the properties of the semiconductor and its impurities.

From the above discussions, the net *recombination rate* for electrons and that for holes can be respectively expressed as

$$R_{\mathrm{e}} = A_{\mathrm{e}}n + Bnp + C_{\mathrm{e}}n^2p + C_{\mathrm{h}}np^2 - G_{\mathrm{e}}^0 \tag{12.49}$$

and

$$R_{\mathrm{h}} = A_{\mathrm{h}}p + Bnp + C_{\mathrm{e}}n^2p + C_{\mathrm{h}}np^2 - G_{\mathrm{h}}^0, \tag{12.50}$$

where $G_{\mathrm{e}}^0 = A_{\mathrm{e}}n_0 + Bn_0p_0 + C_{\mathrm{e}}n_0^2p_0 + C_{\mathrm{h}}n_0p_0^2$ and $G_{\mathrm{h}}^0 = A_{\mathrm{h}}p_0 + Bn_0p_0 + C_{\mathrm{e}}n_0^2p_0 + C_{\mathrm{h}}n_0p_0^2$. Because electrons and holes always recombine in pairs, they must have the same net recombination rate:

$$R = R_{\mathrm{e}} = R_{\mathrm{h}}. \tag{12.51}$$

The net bimolecular recombination rate is clearly the same for electrons and holes, and so is the net Auger recombination rate. For the Shockley–Read processes, $A_{\mathrm{e}}n$ and $A_{\mathrm{h}}p$ might be different, but the difference is exactly balanced by the difference between $A_{\mathrm{e}}n_0$ and $A_{\mathrm{h}}p_0$. Therefore, as indicated in (12.48), the net Shockley–Read recombination rate is also the same for electrons and holes.

When a semiconductor is in thermal equilibrium with its environment, recombination of the carriers has to be exactly balanced by thermal generation of the carriers so that the electron and hole concentrations are maintained at their respective equilibrium values of n_0 and p_0. Therefore, the net recombination rate is zero, $R = R_{\mathrm{e}} = R_{\mathrm{h}} = 0$, when a semiconductor is in thermal equilibrium.

In practice, the A, B, and C coefficients that characterize the three basic recombination processes in (12.49) and (12.50) are not completely independent of the carrier concentrations. Despite this fact, we can still see significant differences between the functional dependencies of the recombination rates on the carrier concentrations for the three different processes. Besides, the A, B, and C coefficients are generally different by many orders of magnitude, with A often being the largest and C being the smallest. For these reasons, the significance of each of the three different recombination processes varies strongly with the concentrations of the carriers. Only the Shockley–Read process is important at low carrier concentrations, whereas the Auger process can be significant only at very high carrier concentrations. Between the two limits, the bimolecular recombination process can be the dominant recombination process. The specific quantitative carrier concentrations for each process to be significant vary from one kind of semiconductor to another and from one specific sample to another, depending on many factors such as band structure, bandgap, type of impurity, doping concentration, defect density, and temperature. In general, however, the B coefficients of direct-gap semiconductors such as GaAs and InP are much larger, by orders of magnitude, than those of indirect-gap semiconductors such as Si and Ge. Therefore, carrier

recombination in most indirect-gap semiconductors is predominantly nonradiative and is often completely characterized by the Shockley–Read process with a net recombination rate of $R = R_{SR}$ given in (12.48) for carrier concentrations up to a pretty high level.

Carrier lifetime

When the electron and hole concentrations in a semiconductor are higher than their respective equilibrium concentrations, due to current injection or optical excitation for example, the excess carriers will relax toward their respective thermal equilibrium concentrations through recombination processes. The relaxation time constant for excess electrons is the *electron lifetime*, defined as

$$\tau_e = \frac{n - n_0}{R}, \tag{12.52}$$

and that of excess holes is the *hole lifetime*, defined as

$$\tau_h = \frac{p - p_0}{R}. \tag{12.53}$$

From these relations, we find that

$$\frac{\tau_e}{\tau_h} = \frac{\Delta n}{\Delta p}, \tag{12.54}$$

where $\Delta n = n - n_0$ and $\Delta p = p - p_0$ are the excess electron and hole concentrations, respectively.

The lifetime of the minority carriers in a semiconductor is called the *minority carrier lifetime*, and that of the majority carriers is called the *majority carrier lifetime*. In an n-type semiconductor, τ_e is the majority carrier lifetime, and τ_h is the minority carrier lifetime. In a p-type semiconductor, τ_e becomes the minority carrier lifetime, and τ_h is the majority carrier lifetime. When the density of the recombination centers is not small compared to the thermal-equilibrium carrier concentrations, the excess minority carrier density can be less than the excess majority carrier density. In this situation, $n - n_0 \neq p - p_0$ and, consequently, $\tau_e \neq \tau_h$. When the concentrations of free electrons and holes are much less than the density of the recombination centers, the majority carrier lifetime can be much greater than the minority carrier lifetime.

A sufficient condition for electrons and holes to have the same lifetime is that the electron and hole concentrations are both very large compared to the density of the recombination centers. When this condition is satisfied, $n - n_0 = p - p_0$; then, we can define the excess carrier density as

$$N = n - n_0 = p - p_0, \tag{12.55}$$

which is also the density of excess free electron–hole pairs. Then, the free electrons and the free holes have the same lifetime: $\tau_e = \tau_h = \tau_s$, which is the *spontaneous carrier*

recombination lifetime of the excess electron–hole pairs given by

$$\tau_s = \frac{N}{R}. \tag{12.56}$$

This relation is valid in practical operating conditions of semiconductor lasers and light-emitting diodes. Under the condition of (12.55), we also find from (12.48) that $A_e = A_h = A$. By applying (12.55) to (12.49), (12.50), and (12.51) and then using (12.56), we find that

$$\frac{1}{\tau_s} = A + B(N + n_0 + p_0) + C_e[N^2 + (2n_0 + p_0)N + (n_0^2 + 2n_0p_0)] + C_h[N^2 + (2p_0 + n_0)N + (p_0^2 + 2n_0p_0)]. \tag{12.57}$$

This spontaneous carrier recombination lifetime is the saturation lifetime of a semiconductor because it has the effect of the saturation lifetime τ_s defined in (10.74) in defining the saturation intensity of a semiconductor, as further discussed following (13.118).

EXAMPLE 12.6 The n-type GaAs considered in Example 12.3 under optical excitation with $n - n_0 = p - p_0 = N$ is found to have the following recombination coefficients: $A = 5.0 \times 10^5\ \text{s}^{-1}$, $B = 8.0 \times 10^{-11}\ \text{cm}^3\ \text{s}^{-1} = 8.0 \times 10^{-17}\ \text{m}^3\ \text{s}^{-1}$, and $C = C_e + C_h = 5.0 \times 10^{-30}\ \text{cm}^6\ \text{s}^{-1} = 5.0 \times 10^{-42}\ \text{m}^6\ \text{s}^{-1}$. (In the literature, the coefficients B and C are commonly quoted in the units of $\text{cm}^3\ \text{s}^{-1}$ and $\text{cm}^6\ \text{s}^{-1}$, respectively. Here we convert them to $\text{m}^3\ \text{s}^{-1}$ and $\text{m}^6\ \text{s}^{-1}$, respectively, for the convenience of computation in SI units.) Assume that $C_e = C_h = C/2$ for simplicity, as the ratio between C_e and C_h is not found. Find the ranges of the excess carrier concentration N where each of the three different recombination processes dominates. Plot the spontaneous carrier lifetime τ_s as a function of the excess carrier concentration N for N in the range between 10^{18} and $10^{26}\ \text{m}^{-3}$.

Solution For $n - n_0 = p - p_0 = N$ and $C_e = C_h = C/2$ considered in this problem, we have, from (12.57),

$$\frac{1}{\tau_s} = A + B(N + n_0 + p_0) + C\left[N^2 + \frac{3}{2}(n_0 + p_0)N + \frac{1}{2}(n_0^2 + p_0^2) + 2n_0p_0\right].$$

Because the values of A, B, and C are different by many orders of magnitude, each term dominates over a certain range of N values. To find these ranges, we only have to compare two neighboring terms at a time. The Shockley–Read recombination process dominates when $A > B(N + n_0 + p_0)$, thus

$$N < \frac{A}{B} - n_0 - p_0 = 6.25 \times 10^{21}\ \text{m}^{-3}.$$

The Auger recombination process becomes important when $C[N^2 + \frac{3}{2}(n_0 + p_0)N + \frac{1}{2}(n_0^2 + p_0^2) + 2n_0p_0] > B(N + n_0 + p_0)$. Because $N \gg n_0, p_0$ when this condition is

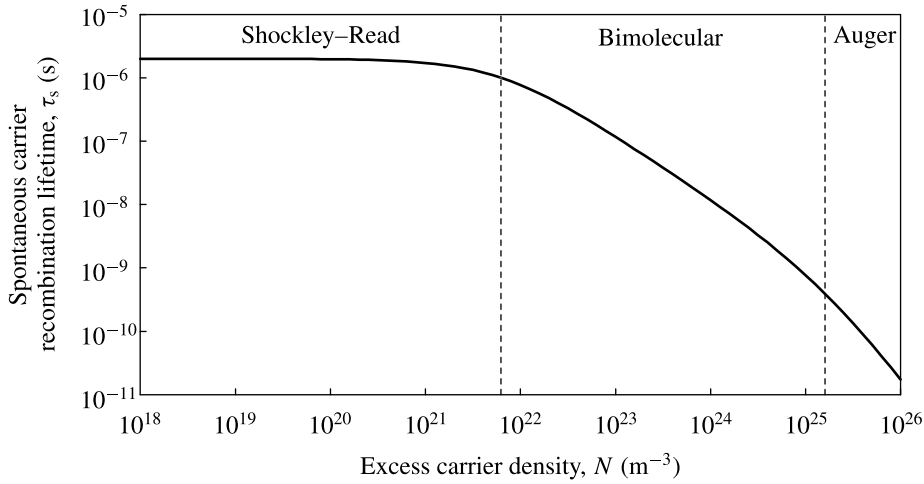


Figure 12.5 Spontaneous carrier recombination lifetime as a function of excess carrier density.

satisfied, we find that the Auger recombination process is important for

$$N > \frac{B}{C} = 1.6 \times 10^{25}\ \text{m}^{-3}.$$

In the middle range, for $6.25 \times 10^{21}\ \text{m}^{-3} < N < 1.6 \times 10^{25}\ \text{m}^{-3}$, bimolecular recombination is the dominant process.

The carrier lifetime as a function of excess carrier concentration is plotted in Fig. 12.5. It can be seen from the change of slope in the curve that the Shockley–Read recombination process dominates for $N < 6.25 \times 10^{21}\ \text{m}^{-3}$, where τ_s is almost constant; the bimolecular recombination process dominates for $6.25 \times 10^{21}\ \text{m}^{-3} < N < 1.6 \times 10^{25}\ \text{m}^{-3}$, where τ_s decreases approximately linearly with increasing N; and finally, the Auger process becomes significant for $N > 1.6 \times 10^{25}\ \text{m}^{-3}$, where τ_s decreases with increasing N more than linearly.

12.4 Current density

An electric current in a semiconductor results from the flow of electrons and holes. The *current density* flowing in a semiconductor is the current flowing through a unit cross-sectional area of the semiconductor; its unit is amperes per square meter.

There are two mechanisms that can cause the flow of electrons and holes: *drift*, in the presence of an electric field, and *diffusion*, in the presence of a spatial gradient in the carrier concentration. The electron current density, $\boldsymbol{J}_e$, and the hole current density, $\boldsymbol{J}_h$, can be expressed as

$$\boldsymbol{J}_e = e\mu_e n\boldsymbol{E}_e + eD_e\boldsymbol{\nabla} n, \tag{12.58}$$

$$\boldsymbol{J}_h = e\mu_h p\boldsymbol{E}_h - eD_h\boldsymbol{\nabla} p, \tag{12.59}$$

respectively, and the total current density is the sum of the two:

$$\boldsymbol{J} = \boldsymbol{J}_\mathrm{e} + \boldsymbol{J}_\mathrm{h}. \tag{12.60}$$

In (12.58) and (12.59), e is the electronic charge; μ_e and μ_h are the electron and hole *mobilities*, respectively; $\boldsymbol{E}_\mathrm{e}$ and $\boldsymbol{E}_\mathrm{h}$ are the electric fields seen by electrons and holes, respectively; and D_e and D_h are the *diffusion coefficients* of electrons and holes, respectively. The electron and hole mobilities strongly depend on temperature, the type of semiconductor, and the impurities and defects in the semiconductor. They generally decrease with increasing concentration of impurities and defects. For most semiconductors of interest, the electron mobility is larger than the hole mobility. At 300 K, $\mu_\mathrm{e} = 1350\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ and $\mu_\mathrm{h} = 480\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ for intrinsic Si, $\mu_\mathrm{e} = 3900\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ and $\mu_\mathrm{h} = 1900\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ for intrinsic Ge, and $\mu_\mathrm{e} = 8500\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ and $\mu_\mathrm{h} = 400\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ for intrinsic GaAs. For nondegenerate semiconductors, the diffusion coefficients are related to the mobilities by the following *Einstein relations*:

$$D_\mathrm{e} = \frac{k_\mathrm{B}T}{e}\mu_\mathrm{e} \quad \text{and} \quad D_\mathrm{h} = \frac{k_\mathrm{B}T}{e}\mu_\mathrm{h}. \tag{12.61}$$

From these relations, we have, at 300 K, $D_\mathrm{e} = 35\ \mathrm{cm^2\ s^{-1}}$ and $D_\mathrm{h} = 12.5\ \mathrm{cm^2\ s^{-1}}$ for intrinsic Si, $D_\mathrm{e} = 100\,\mathrm{cm^2\ s^{-1}}$ and $D_\mathrm{h} = 50\,\mathrm{cm^2\ s^{-1}}$ for intrinsic Ge, $D_\mathrm{e} = 220\,\mathrm{cm^2\ s^{-1}}$ and $D_\mathrm{h} = 10\,\mathrm{cm^2\ s^{-1}}$ for intrinsic GaAs. For degenerate semiconductors, (12.61) is not valid, and the Einstein relations between the diffusion coefficients and the mobilities must be generalized by taking into account the Fermi integrals (see Problem 12.4.1). The mobilities and diffusion coefficients, as well as the effective masses, of electrons and holes for intrinsic Si, Ge, and GaAs are summarized in Table 12.2.

The electric fields seen by electrons and holes can be expressed in terms of the gradients in the conduction- and valence-band edges, respectively:

$$\boldsymbol{E}_\mathrm{e} = \frac{\boldsymbol{\nabla} E_\mathrm{c}}{e} \quad \text{and} \quad \boldsymbol{E}_\mathrm{h} = \frac{\boldsymbol{\nabla} E_\mathrm{v}}{e}. \tag{12.62}$$

In a homogeneous semiconductor where the conduction- and valence-band edges are parallel to each other, $\boldsymbol{E}_\mathrm{e}$ and $\boldsymbol{E}_\mathrm{h}$ are generally the same. However, in an inhomogeneous semiconductor, such as in a graded-gap superlattice where the conduction-band edge

Table 12.2 *Electronic properties of some intrinsic semiconductors at* 300 K

	Effective mass		Mobility ($\mathrm{cm^2\ V^{-1}\ s^{-1}}$)		Diffusion coefficient ($\mathrm{cm^2\ s^{-1}}$)	
	m_e^*/m_0	m_h^*/m_0	μ_e	μ_h	D_e	D_h
Si	1.08	0.56	1350	480	35	12.5
Ge	0.55	0.31	3900	1900	100	50
GaAs	0.067	0.52	8500	400	220	10

is not parallel to the valence-band edge, $\boldsymbol{E}_e$ and $\boldsymbol{E}_h$ can be quite different. Using (12.62), the drift components of the electron and hole current densities can be expressed, respectively, as

$$\boldsymbol{J}_e^{\text{drift}} = e\mu_e n\boldsymbol{E}_e = \mu_e n\nabla E_c, \tag{12.63}$$

$$\boldsymbol{J}_h^{\text{drift}} = e\mu_h p\boldsymbol{E}_h = \mu_h p\nabla E_v. \tag{12.64}$$

Using (12.44) and (12.45) for the electron and hole concentrations and the relations between the diffusion coefficients and the mobilities of the carriers, the diffusion components of the electron and hole current densities can be expressed, respectively, as

$$\boldsymbol{J}_e^{\text{diffusion}} = eD_e\nabla n = \mu_e n\nabla E_{Fc} - \mu_e n\nabla E_c, \tag{12.65}$$

$$\boldsymbol{J}_h^{\text{diffusion}} = -eD_h\nabla p = \mu_h p\nabla E_{Fv} - \mu_h p\nabla E_v. \tag{12.66}$$

By combining the drift and diffusion components, we find that the total electron and hole current densities can be simply expressed in terms of the gradients in the quasi-Fermi levels:

$$\boldsymbol{J}_e = \mu_e n\nabla E_{Fc}, \tag{12.67}$$

$$\boldsymbol{J}_h = \mu_h p\nabla E_{Fv}. \tag{12.68}$$

Consequently, the total current density can be expressed as

$$\boldsymbol{J} = \boldsymbol{J}_e + \boldsymbol{J}_h = \mu_e n\nabla E_{Fc} + \mu_h p\nabla E_{Fv}. \tag{12.69}$$

The relations in (12.67)–(12.69) are quite general. *They are valid for both nondegenerate and degenerate semiconductors, which can be either homogeneous or inhomogeneous* (see Problem 12.4.1). Note, however, that (12.61) is valid only for nondegenerate semiconductors.

Some very important conclusions can be drawn from the relation in (12.69). A semiconductor in thermal equilibrium carries no net electric current, meaning that $\boldsymbol{J} = 0$ in thermal equilibrium. We also know that when a semiconductor is in thermal equilibrium, the electrons and holes in it are characterized by a common Fermi level: $E_F = E_{Fc} = E_{Fv}$. From (12.69), we find that these two facts indicate that $\nabla E_F = 0$ when a semiconductor is in thermal equilibrium. Consequently, a semiconductor in thermal equilibrium is characterized by a single, constant Fermi level throughout its entire volume no matter whether the semiconductor is homogeneous or inhomogeneous and regardless of the detailed structures in the semiconductor. On the other hand, when a semiconductor carries an electric current, it must be in a quasi-equilibrium state with separate quasi-Fermi levels: $E_{Fc} \neq E_{Fv}$. Furthermore, these quasi-Fermi levels must not be constant in space but must have nonvanishing spatial gradients in order to support an electric current.

Conductivity

The *electric conductivity*, σ, of a material is the proportionality constant between the current density and the electric field. In a semiconductor, the conductivity is contributed by both electrons and holes. Only the drift current has to be considered because the diffusion current is not generated by an electric field. For a homogeneous semiconductor, $\boldsymbol{E}_e = \boldsymbol{E}_h = \boldsymbol{E}$. Taking $\nabla n = \nabla p = 0$ to eliminate the diffusion current, we find from (12.58)–(12.60) that $\boldsymbol{J} = e(\mu_e n + \mu_h p)\boldsymbol{E} = \sigma \boldsymbol{E}$ for a homogeneous semiconductor. We thus find the following relation for the conductivity of a semiconductor:

$$\sigma = e(\mu_e n + \mu_h p), \tag{12.70}$$

which is measured per ohm per meter but is usually quoted per ohm per centimeter. The conductivity of a semiconductor in thermal equilibrium is $\sigma_0 = e(\mu_e n_0 + \mu_h p_0)$, often known as the *dark conductivity*; that of an intrinsic semiconductor is $\sigma_i = e(\mu_e + \mu_h)n_i$, known as the *intrinsic conductivity*. The resistivity of a semiconductor is simply the inverse of its conductivity: $\rho = 1/\sigma$, in ohm-meters but usually also given in ohm-centimeters.

As can be seen from (12.70), the conductivity of a semiconductor increases with increasing carrier concentrations. For semiconductors with low impurity concentrations, μ_e and μ_h vary little with the impurity concentration; therefore, the conductivity increases with doping density. The conductivity does not continue to increase linearly with doping density at high impurity concentrations because the mobilities decrease at high impurity concentrations. Because $\mu_e > \mu_h$ for most semiconductors of interest, an n-type semiconductor generally has a higher conductivity than a p-type one of the same impurity concentration. The conductivity of a given semiconductor is a strong function of temperature because carrier concentrations and carrier mobilities are both sensitive to temperature.

EXAMPLE 12.7 Find the intrinsic conductivity and the intrinsic resistivity of GaAs at 300 K.

Solution We find from Example 12.2 that $n_i = 2.33 \times 10^{12}$ m^{-3} for GaAs at 300 K. From Table 12.2, $\mu_e = 8500$ cm^2 V^{-1} s^{-1} $= 0.85$ m^2 V^{-1} s^{-1} and $\mu_h = 400$ cm^2 V^{-1} s^{-1} $= 0.04$ m^2 V^{-1} s^{-1}. Therefore, the intrinsic conductivity is

$$\begin{aligned}\sigma_i &= 1.6 \times 10^{-19} \times (0.85 + 0.04) \times 2.33 \times 10^{12}\ \Omega^{-1}\ \text{m}^{-1} \\ &= 3.32 \times 10^{-7}\ \Omega^{-1}\ \text{m}^{-1} \\ &= 3.32 \times 10^{-9}\ \Omega^{-1}\ \text{cm}^{-1}.\end{aligned}$$

The intrinsic resistivity is

$$\rho_i = \frac{1}{\sigma_i} = 3.01 \times 10^6\ \Omega\ \text{m} = 3.01 \times 10^8\ \Omega\ \text{cm}.$$

Though GaAs is a semiconductor, intrinsic GaAs is often considered to be semi-insulating because of its high resistivity.

12.5 Semiconductor junctions

Useful semiconductor devices are made of inhomogeneous semiconductors that have either spatially nonuniform doping distribution or spatially varying bandgaps or both. There are two categories of semiconductor junctions: *homojunctions* and *heterojunctions*. A homojunction is formed by different doping in the same semiconductor, whereas a heterojunction is formed between two different semiconductors. In addition, a metal–semiconductor junction can be formed between a metal and a semiconductor. The immense possibilities of such structures are witnessed by the existence of the great variety of semiconductor devices and by the constant invention of new devices. In this section, we review the basic properties of the semiconductor homojunctions and heterojunctions.

The most important homojunctions are the *p–n junctions*. A p–n junction is formed between a p-type region and an n-type region with different doping in the same semiconductor. A homojunction can also be a *p–i junction*, which is formed between a p-type region and an undoped intrinsic region of the same semiconductor, or an *i–n junction*, which is formed between an undoped intrinsic region and an n-type region of the same semiconductor.

A heterojunction is normally formed between two lattice-matched semiconductors of different bandgaps. To name a heterojunction, the conductivity type of the small-gap semiconductor is represented by a lowercase letter, n, p, or i, and the conductivity type of the large-gap semiconductor is represented by an uppercase letter, N, P, or I. Because the two semiconductors that form a heterojunction have different bandgaps, they can be either of different conductivity types or of the same conductivity type. Junctions formed between dissimilar semiconductors of the same conductivity type, such as p–P and n–N junctions, are *isotype heterojunctions*; those formed between dissimilar semiconductors of different conductivity types, such as p–N and P–n junctions, are *anisotype heterojunctions*.

A semiconductor junction can be either an *abrupt junction*, which has a sudden change of doping and/or bandgap from one region to the other region, or a *graded junction*, where the change of doping and/or bandgap is gradual. The basic principles of abrupt and graded junctions are the same though there are quantitative differences in the properties of these two different types.

In this section, we consider only abrupt p–n, p–N, and P–n junctions. For simplicity, we assume that the p region is doped with a concentration N_a of fully ionized acceptors, and the n region is doped with a concentration N_d of fully ionized donors. In the vicinity of a junction between the p and n regions, there exists a *depletion layer*, where majority

carriers, holes on the p side and electrons on the n side, are depleted. A junction reaches thermal equilibrium when it is not subject to an external excitation. In thermal equilibrium, the p and n regions outside the depletion layer are *homogeneous regions* because the characteristics of the semiconductor in these regions approach those of homogeneous semiconductors in thermal equilibrium. The homogeneous p region has a majority hole concentration of $p_{p0} = N_a^- = N_a$ and a minority electron concentration of n_{p0}, whereas the homogeneous n region has a majority electron concentration of $n_{n0} = N_d^+ = N_d$ and a minority hole concentration of p_{n0}. The equilibrium state can be perturbed with a bias voltage. For a junction under bias, *diffusion regions* exist between the depletion layer and the homogeneous regions on both p and n sides. In the diffusion regions, the characteristics of the semiconductor are dominated by diffusion of the minority carriers, which are electrons on the p side and holes on the n side.

The major differences among p–n homojunctions and p–N and P–n heterojunctions are their energy band structures. The energy bands and the built-in potential for each type of junction are considered in this section to illustrate the differences among these junctions. In addition, the difference in the electric permittivities, ϵ_p and ϵ_n, respectively, of the p and n regions can be significant for a p–N or P–n heterojunction but is practically negligible for a p–n homojunction. This issue is minor because it can be easily taken care of by considering ϵ_p and ϵ_n to be generally different, even for a p–n homojunction. Other than these differences, these junctions have similar electrical characteristics. Therefore, the discussions and the mathematical relations regarding the depletion layer, the carrier distribution, the current–voltage characteristics, and the capacitance are treated generally and are valid for both homojunctions and heterojunctions.

Energy bands and electrostatic potential

As discussed in the preceding section, a semiconductor in thermal equilibrium is characterized by a spatially constant Fermi level. This statement is true for both homojunctions and heterojunctions. Therefore, as shown in Figs. 12.6, 12.7, and 12.8 for p–n, p–N, and P–n junctions, respectively, $E_{Fp} = E_{Fn} = E_F$ for a junction in thermal equilibrium, where E_{Fp} and E_{Fn} are the Fermi levels in the p and n regions, respectively. Because E_{Fp} lies close to the valence-band edge in the p region but E_{Fn} lies close to the conduction-band edge in the n region, a constant Fermi level throughout the semiconductor in thermal equilibrium leads to bending of the energy bands across the junction, as shown in Fig. 12.6 for a p–n homojunction and in Figs. 12.7 and 12.8 for p–N and P–n heterojunctions, respectively. As we shall see later, this band bending occurs primarily within the depletion layer. The energy bands remain relatively flat outside the depletion layer on both p and n sides.

For a homojunction, shown in Fig. 12.6, the energy bands remain continuous and smooth across the junction because the semiconductors on the two sides of the junction have the same bandgap. For a heterojunction, the semiconductors on the two sides of the junction have different bandgaps. At the junction where these two semiconductors

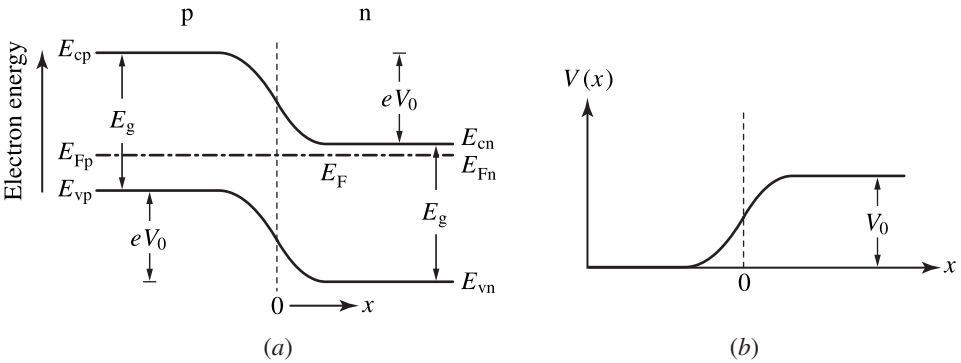


Figure 12.6 (*a*) Energy bands and (*b*) built-in electrostatic potential for a p–n homojunction in thermal equilibrium, where $E_{gp} = E_{gn} = E_g$.

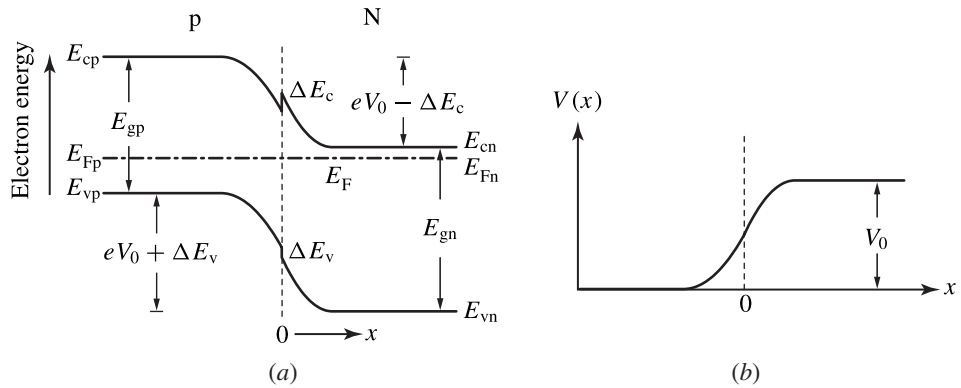


Figure 12.7 (*a*) Energy bands and (*b*) built-in electrostatic potential for a p–N heterojunction in thermal equilibrium, where $E_{gp} < E_{gn}$ and $\Delta E_g = E_{gn} - E_{gp} = \Delta E_c + \Delta E_v$. The slope of $V(x)$ has a discontinuity at $x = 0$ caused by the abrupt change of bandgap at the junction.

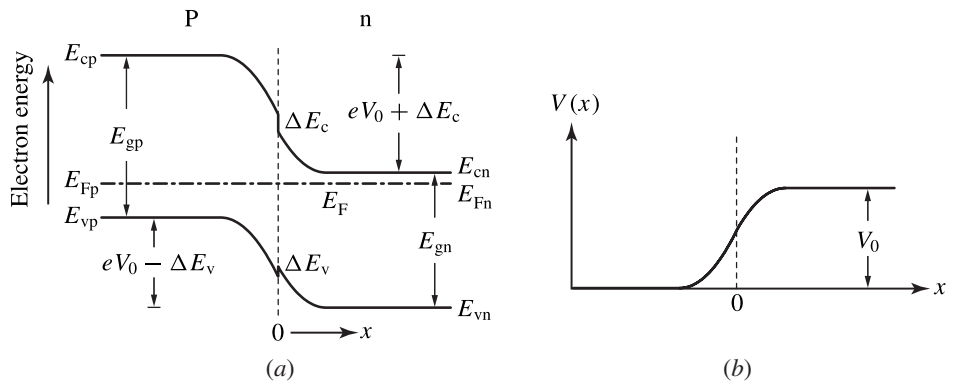


Figure 12.8 (*a*) Energy bands and (*b*) built-in electrostatic potential for a P–n heterojunction in thermal equilibrium, where $E_{gp} > E_{gn}$ and $\Delta E_g = E_{gp} - E_{gn} = \Delta E_c + \Delta E_v$. The slope of $V(x)$ has a discontinuity at $x = 0$ caused by the abrupt change of bandgap at the junction.

are joined together, the disparity in their bandgaps results in a discontinuity of ΔE_c in the conduction-band edge and a discontinuity of ΔE_v in the valence-band edge, as shown in Figs. 12.7 and 12.8. Therefore,

$$\Delta E_c + \Delta E_v = \Delta E_g, \tag{12.71}$$

where ΔE_g is the difference between the bandgaps of the two dissimilar semiconductors. The value of ΔE_g and those of the band offsets ΔE_c and ΔE_v are intrinsic properties of a specific pair of semiconductors. The conduction-band offset ΔE_c is determined by the difference in the electron affinities of the two semiconductors; the valence-band offset ΔE_v is then fixed by (12.71). In practice, these parameters are measured experimentally for each given pair of semiconductors. For GaAs–$Al_xGa_{1-x}As$ heterojunctions, $\Delta E_c \approx 65\%\ \Delta E_g$ and $\Delta E_v \approx 35\%\ \Delta E_g$.

p–n homojunction in thermal equilibrium

Because the bandgap remains constant across a semiconductor homojunction, at any given location the conduction- and valence-band edges have the same gradient. According to (12.62), this spatially varying band-edge gradient creates a spatially varying built-in electric field that is seen by both electrons and holes: $\boldsymbol{E}_e = \boldsymbol{E}_h$ at any given location. In thermal equilibrium, this built-in electric field results in a built-in electrostatic potential across the p–n junction, as shown in Fig. 12.6(b). The height, V_0, of this built-in potential is called the *contact potential* of the junction. Note that the energy bands plotted in Fig. 12.6(a) refer to the energy of an electron. The n region has a lower energy for an electron than the p region, whereas the converse is true for a hole. Because an electron carries a negative charge of $q = -e$, the built-in electrostatic potential is higher on the n side than on the p side. As shown in Fig. 12.6(a), we have, for a p–n homojunction in thermal equilibrium,

$$E_{cp} - E_{cn} = E_{vp} - E_{vn} = eV_0, \tag{12.72}$$

where E_{cp} and E_{cn} are, respectively, the conduction-band edges in the homogeneous p and n regions, and E_{vp} and E_{vn} are, respectively, the valence-band edges in the homogeneous p and n regions. Therefore, eV_0 is the same energy barrier for an electron on the n side to move to the p side as that for a hole on the p side to move to the n side.

In the case when both p and n regions are nondegenerate so that (12.26) and (12.27) are valid, we can use (12.72) to obtain the following relation for the carrier concentrations in the homogeneous p and n regions:

$$\frac{p_{p0}}{p_{n0}} = \frac{n_{n0}}{n_{p0}} = e^{eV_0/k_BT}. \tag{12.73}$$

By using the law of mass action, $p_{p0}n_{p0} = p_{n0}n_{n0} = n_i^2$, we find that the contact potential of a nondegenerate p–n homojunction in a nondegenerate semiconductor is

given by

$$V_0 = \frac{k_B T}{e} \ln \frac{p_{p0}}{p_{n0}} = \frac{k_B T}{e} \ln \frac{n_{n0}}{n_{p0}} = \frac{k_B T}{e} \ln \frac{p_{p0} n_{n0}}{n_i^2}. \tag{12.74}$$

Under the condition that $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$, the contact potential can be found as

$$V_0 = \frac{k_B T}{e} \ln \frac{N_a N_d}{n_i^2} = \frac{k_B T}{e} \ln \frac{N_a N_d}{N_v N_c} + \frac{E_g}{e}. \tag{12.75}$$

EXAMPLE 12.8 An abrupt GaAs p–n homojunction is formed by creating a uniform p region on one side and a uniform n region on the other side. The p region is doped with fully ionized acceptors of a concentration $N_a = 1 \times 10^{23}$ m^{-3}, and the n region is doped with fully ionized donors of a concentration $N_d = 1 \times 10^{22}$ m^{-3}. Find the contact potential of this junction at 300 K.

Solution From Example 12.2, we have $n_i = 2.33 \times 10^{12}$ m^{-3} for GaAs at 300 K. At the given doping levels for the p and n regions, $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$. At 300 K, $k_B T/e = 25.9$ mV $= 0.0259$ V. Therefore, we can use the first relation in (12.75) to find the contact potential for this junction at 300 K as

$$V_0 = 0.0259 \text{ V} \times \ln \frac{1 \times 10^{23} \times 1 \times 10^{22}}{(2.33 \times 10^{12})^2} = 1.209 \text{ V}.$$

p–N heterojunction in thermal equilibrium

For a p–N heterojunction, the semiconductor on the n side has a larger bandgap than that on the p side: $E_{gn} > E_{gp}$. Therefore,

$$\Delta E_c + \Delta E_v = E_{gn} - E_{gp} = \Delta E_g, \tag{12.76}$$

according to (12.71). Because of the presence of band offsets at the junction, (12.72) is not valid for a p–N junction. Instead, as can be seen in Fig. 12.7(a), we have

$$E_{cp} - E_{cn} = eV_0 - \Delta E_c, \tag{12.77}$$

$$E_{vp} - E_{vn} = eV_0 + \Delta E_v, \tag{12.78}$$

where V_0 is the contact potential of the p–N junction. In contrast to the case of a p–n homojunction, where electrons and holes have the same energy barrier of eV_0, electrons and holes have different energy barriers in the case of a p–N junction. According to (12.77) and (12.78), the energy barrier for an electron on the n side is lowered from eV_0 by ΔE_c due to the conduction-band offset, but that for a hole on the p side is raised by ΔE_v due to the valence-band offset. Therefore, in a p–N heterojunction the energy barrier for an electron on the n side is lower than that for a hole on the p

side by the amount of the bandgap difference of $\Delta E_g = E_{gn} - E_{gp}$ between the two semiconductors.

Though the energy barriers for electrons and holes to cross a p–N junction are different, electrons and holes see the same spatially varying built-in electric field because the conduction- and valence-band edges are parallel to each other and have the same gradient at every location except at the junction where the discontinuities of the energy bands take place. Therefore, similarly to the situation in a homojunction, we still have a common electrostatic field, $\boldsymbol{E}_e = \boldsymbol{E}_h$, at any given location for both electrons and holes. As a result, there is a common built-in electrostatic potential, shown in Fig. 12.7(*b*), across a p–N junction. For an abrupt p–N junction, there is a sudden change of slope in $V(x)$ at the junction because of the sudden change in electric permittivity from one semiconductor to the other.

For a p–N junction, (12.73) is not valid. Instead, we can use (12.77) and (12.78) to obtain the following relation in the case when both p and n regions are nondegenerate:

$$\frac{p_{p0}}{p_{n0}}\frac{N_{vn}}{N_{vp}}e^{-\Delta E_v/k_BT} = \frac{n_{n0}}{n_{p0}}\frac{N_{cp}}{N_{cn}}e^{\Delta E_c/k_BT} = e^{eV_0/k_BT}, \tag{12.79}$$

where N_{cp} and N_{vp} are the effective densities of states, as defined in (12.24), for the semiconductor on the p side, and N_{cn} and N_{vn} are the effective densities of states for the semiconductor on the n side. From this relation, we find that the contact potential for a nondegenerate p–N junction can be expressed as

$$\begin{aligned} V_0 &= \frac{k_BT}{e}\ln\left(\frac{p_{p0}}{p_{n0}}\frac{N_{vn}}{N_{vp}}\right) - \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{n_{n0}}{n_{p0}}\frac{N_{cp}}{N_{cn}}\right) + \frac{\Delta E_c}{e} \\ &= \frac{k_BT}{e}\ln\left(\frac{p_{p0}n_{n0}}{n_{in}^2}\frac{N_{vn}}{N_{vp}}\right) - \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{p_{p0}n_{n0}}{n_{ip}^2}\frac{N_{cp}}{N_{cn}}\right) + \frac{\Delta E_c}{e}, \end{aligned} \tag{12.80}$$

where n_{ip} and n_{in} are the intrinsic carrier concentrations for the semiconductors on the p and n sides, respectively. Under the condition that $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$, the contact potential can be found as (see Problem 12.5.8)

$$\begin{aligned} V_0 &= \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_{in}^2}\frac{N_{vn}}{N_{vp}}\right) - \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_{ip}^2}\frac{N_{cp}}{N_{cn}}\right) + \frac{\Delta E_c}{e} \\ &= \frac{k_BT}{e}\ln\frac{N_aN_d}{N_{vp}N_{cn}} + \frac{E_{gn}}{e} - \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\frac{N_aN_d}{N_{vp}N_{cn}} + \frac{E_{gp}}{e} + \frac{\Delta E_c}{e}. \end{aligned} \tag{12.81}$$

P–n heterojunction in thermal equilibrium

For a P–n heterojunction, the semiconductor on the p side has a larger bandgap than that on the n side: $E_{gp} > E_{gn}$. Therefore,

$$\Delta E_c + \Delta E_v = E_{gp} - E_{gn} = \Delta E_g. \tag{12.82}$$

As shown in Fig. 12.8(a), the band offsets at the junction lead to

$$E_{cp} - E_{cn} = eV_0 + \Delta E_c, \tag{12.83}$$

$$E_{vp} - E_{vn} = eV_0 - \Delta E_v, \tag{12.84}$$

where V_0 is the contact potential of the P–n junction. Similarly to the case of the p–N junction discussed above, electrons and holes do not have the same energy barrier. However, because $E_{gp} > E_{gn}$ in the case of a P–n heterojunction, the energy barrier for an electron on the n side is now higher than that for a hole on the p side by the amount of the bandgap difference of $\Delta E_g = E_{gp} - E_{gn}$ between the two semiconductors. Figure 12.8(b) shows the built-in electrostatic potential across a P–n junction. For an abrupt P–n junction, there is also a sudden change of slope in $V(x)$ at the junction because of the sudden change in electric permittivity across the junction.

For a P–n junction, we can use (12.83) and (12.84) to obtain the following relation in the case when both p and n regions are nondegenerate:

$$\frac{p_{p0}}{p_{n0}}\frac{N_{vn}}{N_{vp}}e^{\Delta E_v/k_BT} = \frac{n_{n0}}{n_{p0}}\frac{N_{cp}}{N_{cn}}e^{-\Delta E_c/k_BT} = e^{eV_0/k_BT}. \tag{12.85}$$

From this relation, we find the following contact potential for a nondegenerate P–n junction:

$$\begin{aligned} V_0 &= \frac{k_BT}{e}\ln\left(\frac{p_{p0}}{p_{n0}}\frac{N_{vn}}{N_{vp}}\right) + \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{n_{n0}}{n_{p0}}\frac{N_{cp}}{N_{cn}}\right) - \frac{\Delta E_c}{e} \\ &= \frac{k_BT}{e}\ln\left(\frac{p_{p0}n_{n0}}{n_{in}^2}\frac{N_{vn}}{N_{vp}}\right) + \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{p_{p0}n_{n0}}{n_{ip}^2}\frac{N_{cp}}{N_{cn}}\right) - \frac{\Delta E_c}{e}. \end{aligned} \tag{12.86}$$

Under the condition that $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$, the contact potential can be found as (see Problem 12.5.8)

$$\begin{aligned} V_0 &= \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_{in}^2}\frac{N_{vn}}{N_{vp}}\right) + \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_{ip}^2}\frac{N_{cp}}{N_{cn}}\right) - \frac{\Delta E_c}{e} \\ &= \frac{k_BT}{e}\ln\frac{N_aN_d}{N_{vp}N_{cn}} + \frac{E_{gn}}{e} + \frac{\Delta E_v}{e} = \frac{k_BT}{e}\ln\frac{N_aN_d}{N_{vp}N_{cn}} + \frac{E_{gp}}{e} - \frac{\Delta E_c}{e}. \end{aligned} \tag{12.87}$$

EXAMPLE 12.9 An abrupt AlGaAs/GaAs P–n heterojunction is formed with a uniform p region of $Al_{0.3}Ga_{0.7}As$ and a uniform n region of GaAs. The p $Al_{0.3}Ga_{0.7}As$ region is doped with fully ionized acceptors of a concentration $N_a = 1 \times 10^{23}$ m^{-3}, and the n GaAs region is doped with fully ionized donors of a concentration $N_d = 1 \times 10^{22}$ m^{-3}. The density of states effective masses for $Al_{0.3}Ga_{0.7}As$ are $m_e^* = 0.092m_0$ and $m_h^* = 0.62m_0$, as compared to $m_e^* = 0.067m_0$ and $m_h^* = 0.52m_0$ for GaAs. Find the contact potential of this junction at 300 K by taking $\Delta E_c = 65\%\ \Delta E_g$ and $\Delta E_v = 35\%\ \Delta E_g$. Compare this contact potential to that of the GaAs homojunction considered in Example 12.8, which has the same doping profile.

Solution We find from (12.3) with $x = 0.3$ that $E_g = 1.798$ eV for $Al_{0.3}Ga_{0.7}As$. Because $E_g = 1.424$ eV for GaAs, we thus have $E_{gp} = 1.798$ eV, $E_{gn} = 1.424$ eV, and $\Delta E_g = E_{gp} - E_{gn} = 0.374$ eV for this heterostructure. Following the procedures in Example 12.2 but taking $m_e^* = 0.092m_0$ and $m_h^* = 0.62m_0$, we find that $N_{cp} = 7.00 \times 10^{23}$ m^{-3}, $N_{vp} = 1.22 \times 10^{25}$ m^{-3}, and $n_{ip} = 2.46 \times 10^9$ m^{-3} for the p $Al_{0.3}Ga_{0.7}As$ at 300 K. Also from Example 12.2, we have $N_{cn} = 4.35 \times 10^{23}$ m^{-3}, $N_{vn} = 9.41 \times 10^{24}$ m^{-3}, and $n_{in} = 2.33 \times 10^{12}$ m^{-3} for the n GaAs at 300 K. Because $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$, we can use any one of the relations in (12.86) and (12.87) to find the contact potential. By using the first relation in (12.87), the contact potential for this junction at 300 K is found as

$$\begin{aligned} V_0 &= \frac{k_B T}{e} \ln\left(\frac{N_a N_d}{n_{in}^2}\frac{N_{vn}}{N_{vp}}\right) + \frac{\Delta E_v}{e} \\ &= 0.0259 \text{ V} \times \ln\frac{1 \times 10^{23} \times 1 \times 10^{22} \times 9.41 \times 10^{24}}{(2.33 \times 10^{12})^2 \times 1.22 \times 10^{25}} + 0.35 \times 0.374 \text{ V} \\ &= 1.333 \text{ V}. \end{aligned}$$

The same result is obtained by using any other relation in (12.86) or (12.87).

The contact potential of this AlGaAs/GaAs heterojunction is different from that of the GaAs homojunction of the same doping profile considered in Example 12.8 for two reasons: (1) the density-of-states effective masses are different for AlGaAs and GaAs, and (2) the band offset due to the difference in the bandgaps between AlGaAs and GaAs causes an additional adjustment for the contact potential.

Junctions under bias

A bias voltage, V, is defined as the voltage applied to the p side of a junction with respect to the n side. A bias voltage changes the electrostatic potential between the p and n regions, thus changing the difference between E_{cp} and E_{cn} and that between E_{vp} and E_{vn}. In the case of a p–n homojunction under a bias voltage V, we have

$$E_{cp} - E_{cn} = E_{vp} - E_{vn} = e(V_0 - V). \tag{12.88}$$

For a p–N junction, we have

$$E_{cp} - E_{cn} = e(V_0 - V) - \Delta E_c, \tag{12.89}$$

$$E_{vp} - E_{vn} = e(V_0 - V) + \Delta E_v. \tag{12.90}$$

For a P–n junction, we have

$$E_{cp} - E_{cn} = e(V_0 - V) + \Delta E_c, \tag{12.91}$$

$$E_{vp} - E_{vn} = e(V_0 - V) - \Delta E_v. \tag{12.92}$$

Clearly from these relations, a bias voltage has similar effects on p–n, p–N, and P–n junctions. A junction is under forward bias if $V > 0$. A forward bias voltage raises the

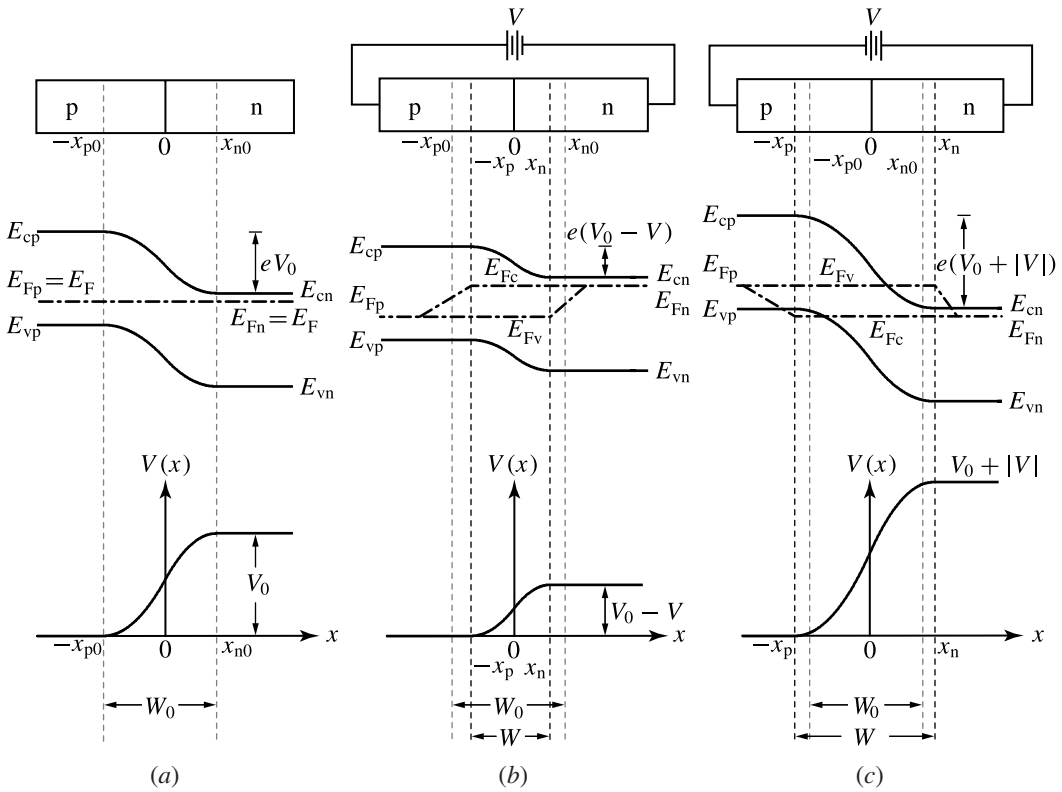


Figure 12.9 Spatial distributions of the p and n regions, the energy bands, and the electrostatic potential of an abrupt p–n homojunction (*a*) in thermal equilibrium, (*b*) under forward bias, and (*c*) under reverse bias.

potential on the p side with respect to that on the n side, resulting in a lower potential barrier of $V_0 - V$. Consequently, the energy barrier between the homogeneous p and n regions is reduced by the amount of eV. A junction is under reverse bias if $V < 0$. A reverse bias voltage lowers the potential on the p side with respect to that on the n side, thus raising the potential barrier to $V_0 - V = V_0 + |V|$. The consequence is an increase in the energy barrier between the homogeneous p and n regions by the amount of $e|V|$. Figure 12.9 shows the energy bands and the electrostatic potential of a p–n homojunction (*a*) in thermal equilibrium, (*b*) under forward bias, and (*c*) under reverse bias. Except for the presence of band offsets at the junction, the characteristics of p–N and P–n heterojunctions under bias are similar to those shown in Fig. 12.9.

A bias voltage causes an electric current to flow in a semiconductor. The bias voltage splits the Fermi level into separate quasi-Fermi levels, E_{Fc} and E_{Fv}, for electrons and holes, respectively, and creates spatial gradients in them to support an electric current in the semiconductor. According to (12.69), when an electric current flows in a semiconductor, the gradients in E_{Fc} and E_{Fv} exist throughout the semiconductor but they vary from one location to another in conjunction with the variations in the local

concentrations of electrons and holes. The spatial variations of the quasi-Fermi levels are shown, along with the spatial variations of the energy bands, in Fig. 12.9(*b*) for the case of forward bias and in Fig. 12.9(*c*) for the case of reverse bias. The largest splitting of E_{Fc} and E_{Fv} occurs in the *depletion layer*. The largest gradients in E_{Fc} and E_{Fv} exist in the *diffusion regions* just outside the depletion layer. The quasi-Fermi levels gradually merge into E_{Fp} in the homogeneous p region and into E_{Fn} in the homogeneous n region. In the presence of a bias voltage, the Fermi levels in the homogeneous p and n regions are not aligned any more:

$$E_{Fn} - E_{Fp} = eV. \tag{12.93}$$

In the case of forward bias, $E_{Fn} > E_{Fp}$ and $E_{Fc} > E_{Fv}$, as shown in Fig. 12.9(*b*). In the case of reverse bias, $E_{Fn} < E_{Fp}$ and $E_{Fc} < E_{Fv}$, as shown in Fig. 12.9(*c*). The relation in (12.93) and the characteristics of the quasi-Fermi levels shown in Fig. 12.9 are valid for p–N and P–n heterojunctions as well.

EXAMPLE 12.10 Find the bias voltage that lines up the band edges of the p and n regions on the two sides of the p–n homojunction considered in Example 12.8. What is the bias voltage that is needed to line up the conduction-band edges of the p and n regions on the two sides of the P–n heterojunction considered in Example 12.9? What is the bias voltage needed to line up the valence-band edges?

Solution Because the bandgap is the same on both sides of a homojunction, a bias voltage that lines up the conduction-band edge also lines up the valence-band edge. From (12.88), we find that $V = V_0$ for $E_{cp} = E_{cn}$ and $E_{vp} = E_{vn}$ across a homojunction. Therefore, we need a forward bias voltage of $V = V_0 = 1.209$ V to line up the band edges of the p and n regions on the two sides of the p–n homojunction considered in Example 12.8.

For a heterojunction, a single bias voltage does not line up conduction-band edges and valence-band edges simultaneously because the bandgaps are different on the two sides of the junction. For the P–n heterojunction considered in Example 12.9, we find from (12.91) that the forward bias voltage required for $E_{cp} = E_{cn}$ to line up the conduction-band edges is

$$V = V_0 + \frac{\Delta E_c}{e} = 1.333 \text{ V} + 0.65 \times 0.374 \text{ V} = 1.576 \text{ V}.$$

We then find from (12.92) that the forward bias voltage required for $E_{vp} = E_{vn}$ to line up the valence-band edges is

$$V = V_0 - \frac{\Delta E_v}{e} = 1.333 \text{ V} - 0.35 \times 0.374 \text{ V} = 1.202 \text{ V}.$$

We find that these two bias voltages are quite different because of the existence of conduction-band and valence-band offsets caused by the bandgap difference on the two sides of the heterojunction.

Depletion layer

The depletion layer is created by the diffusion of holes from the p side, where the hole concentration is high, to the n side, where the hole concentration is low, and the diffusion of electrons from the n side, where the electron concentration is high, to the p side, where the electron concentration is low. The depletion layer has a width of

$$W = x_p + x_n, \tag{12.94}$$

where x_p and x_n are the penetration depths of the depletion layer into the p and n regions, respectively.

The depletion layer is also known as the *space-charge region* because depletion of the majority of carriers in this region leaves the immobile negatively charged acceptor ions on the p side and the immobile positively charged donor ions on the n side as space charges in this region. As shown in Fig. 12.10, the p side has a negative space charge density of $-eN_a$ over a penetration depth of x_p, and the n side has a positive space charge density of eN_d over a penetration depth of x_n. Because of the overall neutrality of the semiconductor, the total negative space charges on the p side must be equal to the total positive space charges on the n side. Therefore, we have

$$N_a x_p = N_d x_n. \tag{12.95}$$

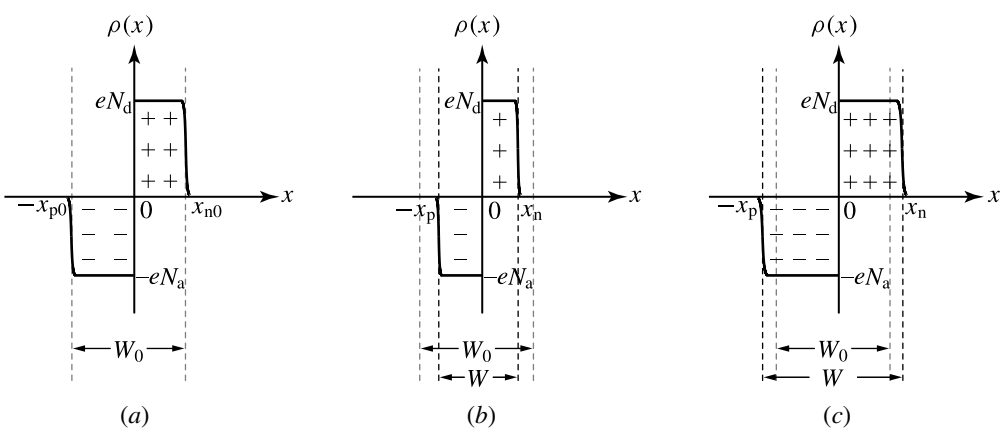


Figure 12.10 Spatial distribution of the space-charge density in the depletion layer of an abrupt p–n junction (*a*) in thermal equilibrium, (*b*) under forward bias, and (*c*) under reverse bias. Abrupt p–N and P–n junctions also have these same characteristics. It is assumed in this illustration that $N_d > N_a$ so that the n region is more heavily doped than the p region.

By combining (12.94) and (12.95), we find that

$$x_p = \frac{N_d}{N_a + N_d} W \quad \text{and} \quad x_n = \frac{N_a}{N_a + N_d} W. \tag{12.96}$$

Clearly, the depletion layer can penetrate p and n regions unevenly, depending on the relative doping on the two sides. The depletion layer penetrates deeper into the region that has a lighter doping concentration. This characteristic is demonstrated in Fig. 12.10.

The electric field associated with the gradient of the band edges is created by the space charges in the depletion layer. Therefore, most of the band-edge gradient exists in the depletion layer and most of the potential difference between the n and p regions is distributed across this layer. Because $\boldsymbol{E} = -\boldsymbol{\nabla} V$ and $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho(x)/\epsilon(x)$, where $\epsilon(x)$ is the spatially varying electric permittivity of the semiconductor, we have the following Poisson equation to describe the potential variations across the depletion layer:

$$\nabla^2 V = \frac{\mathrm{d}^2 V}{\mathrm{d}x^2} = -\frac{\rho(x)}{\epsilon(x)} = \begin{cases} \dfrac{eN_a}{\epsilon_p}, & \text{for } -x_p < x < 0, \\ -\dfrac{eN_d}{\epsilon_n}, & \text{for } 0 < x < x_n, \end{cases} \tag{12.97}$$

where ϵ_p and ϵ_n are the electric permittivities of the p and n regions, respectively. The boundary conditions are $\boldsymbol{E}(-x_p) = \boldsymbol{E}(x_n) = 0$ and $\epsilon_p \boldsymbol{E}(0_-) = \epsilon_n \boldsymbol{E}(0_+)$, meaning that $\mathrm{d}V/\mathrm{d}x = 0$ at $x = -x_p$ and $x = x_n$, and $\epsilon_p \mathrm{d}V/\mathrm{d}x|_{x=0_-} = \epsilon_n \mathrm{d}V/\mathrm{d}x|_{x=0_+}$. By integrating (12.97) through the depletion layer and applying these boundary conditions, we find that, in the presence of a bias voltage V,

$$V_0 - V = V(x_n) - V(-x_p) = \frac{e}{2\epsilon} \frac{N_a N_d}{N_a + N_d} W^2, \tag{12.98}$$

where ϵ is an effective electric permittivity defined as

$$\epsilon = \frac{\epsilon_p \epsilon_n (N_a + N_d)}{\epsilon_p N_a + \epsilon_n N_d}. \tag{12.99}$$

Therefore, the width of the depletion layer can be expressed as a function of the applied bias voltage V in the following form:

$$W = \left[\frac{2\epsilon}{e} \left(\frac{N_a + N_d}{N_a N_d} \right) (V_0 - V) \right]^{1/2}. \tag{12.100}$$

When a junction is in thermal equilibrium without bias, $x_p = x_{p0}$, $x_n = x_{n0}$, and $W = W_0$, as shown in Fig. 12.10(a). From (12.100), we see that $W < W_0$ for $V > 0$ and $W > W_0$ for $V < 0$. Therefore, the depletion layer narrows with forward bias, as shown in Fig. 12.10(b), and broadens with reverse bias, as shown in Fig. 12.10(c).

EXAMPLE 12.11 The static dielectric constant of GaAs is $\epsilon/\epsilon_0 = 13.18$. Find the width of the depletion layer, W_0, and the penetration depths, x_{p0} and x_{n0}, for the GaAs p–n

homojunction described in Example 12.8 when it is in thermal equilibrium without bias at 300 K.

Solution The width of the depletion layer in thermal equilibrium without bias can be found from (12.100) with $V = 0$. With $\epsilon = 13.18\epsilon_0$, $\epsilon_0 = 8.854 \times 10^{-12}$ F m^{-1}, $N_a = 1 \times 10^{23}$ m^{-3}, $N_d = 1 \times 10^{22}$ m^{-3}, and $V_0 = 1.209$ V from Example 12.8, we find that

$$W_0 = \left[\frac{2 \times 13.18 \times 8.854 \times 10^{-12}}{1.6 \times 10^{-19}} \times \left(\frac{1 \times 10^{23} + 1 \times 10^{22}}{1 \times 10^{23} \times 1 \times 10^{22}}\right) \times 1.209\right]^{1/2} \text{ m}$$
$$= 440 \text{ nm}$$

From (12.96), we find the following penetration depths:

$$x_{p0} = \frac{N_d}{N_a + N_d} W_0 = \frac{1 \times 10^{22}}{1 \times 10^{23} + 1 \times 10^{22}} \times 440 \text{ nm} = 40 \text{ nm},$$

$$x_{n0} = \frac{N_a}{N_a + N_d} W_0 = \frac{1 \times 10^{23}}{1 \times 10^{23} + 1 \times 10^{22}} \times 440 \text{ nm} = 400 \text{ nm}.$$

We see that $x_{n0} = 10x_{p0}$ for this junction because $N_a = 10N_d$.

Carrier distribution

As mentioned above, the majority and minority carrier concentrations are p_{p0} and n_{p0}, respectively, in the homogeneous p region and are n_{n0} and p_{n0}, respectively, in the homogeneous n region. The depletion layer is not completely devoid of free carriers. In thermal equilibrium, the electron and hole concentrations in the depletion layer are determined by (12.22) and (12.23), respectively, with spatially varying band edges, $E_c(x)$ and $E_v(x)$; therefore, $n_{p0} \ll n_0(x) \ll n_{n0}$ and $p_{n0} \ll p_0(x) \ll p_{p0}$ for $-x_{p0} < x < x_{n0}$. The distributions of the electron and hole concentrations across a p–n junction in thermal equilibrium are illustrated in Fig. 12.11(*a*).

A bias voltage can cause substantial changes in the minority carrier concentrations at $x = -x_p$ and $x = x_n$, where the edges of the depletion layer are located. From Figs. 12.9(*b*) and (*c*), we find that $E_{Fc}(-x_p) - E_{Fp} \approx E_{Fn} - E_{Fp} = eV$ and $E_c(-x_p) \approx E_{cp}$ at $x = -x_p$, and $E_{Fn} - E_{Fv}(x_n) \approx E_{Fn} - E_{Fp} = eV$ and $E_v(x_n) \approx E_{vn}$ at $x = x_n$. Therefore, using (12.41) and (12.42), we find that the minority carrier concentrations at the edges of the depletion layer are

$$n_p(-x_p) = n_{p0} e^{eV/k_B T}, \tag{12.101}$$

$$p_n(x_n) = p_{n0} e^{eV/k_B T}. \tag{12.102}$$

At the edges of the depletion layer, the bias voltage creates the following changes in the minority carrier concentrations from their equilibrium values:

$$\Delta n_p = n_p(-x_p) - n_{p0} = n_{p0}\left(e^{eV/k_B T} - 1\right), \tag{12.103}$$

$$\Delta p_n = p_n(x_n) - p_{n0} = p_{n0}\left(e^{eV/k_B T} - 1\right). \tag{12.104}$$

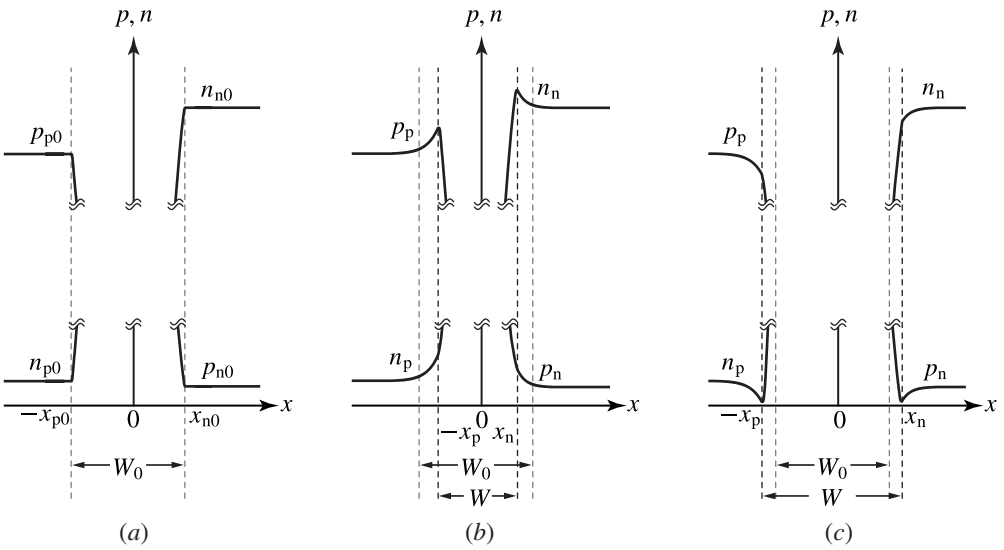


Figure 12.11 Spatial distributions of the electron and hole concentrations of an abrupt p–n junction (*a*) in thermal equilibrium, (*b*) under forward bias, and (*c*) under reverse bias. Abrupt p–N and P–n junctions also have these same characteristics. It is assumed in this illustration that $N_d > N_a$ so that the n region is more heavily doped than the p region.

As shown in Fig. 12.11(*b*), a forward bias creates excess minority carriers on both p and n sides, $\Delta n_p > 0$ and $\Delta p_n > 0$ for $V > 0$, as a result of *minority carrier injection*. In contrast, as shown in Fig. 12.11(*c*), a reverse bias depletes minority carriers on both p and n sides, $\Delta n_p < 0$ and $\Delta p_n < 0$ for $V < 0$, as a result of *minority carrier extraction*.

Because of the diffusion of minority carriers, changes in the minority carrier concentrations caused by a bias voltage are not localized at the edges of the depletion layer. Instead, the minority carrier concentrations have the following spatially dependent variations across the diffusion regions:

$$\begin{aligned} n_p(x) - n_{p0} &= \Delta n_p e^{(x+x_p)/L_e} \\ &= n_{p0}\left(e^{eV/k_BT} - 1\right) e^{(x+x_p)/L_e}, \qquad \text{for } x < -x_p, \end{aligned} \tag{12.105}$$

$$\begin{aligned} p_n(x) - p_{n0} &= \Delta p_n e^{-(x-x_n)/L_h} \\ &= p_{n0}\left(e^{eV/k_BT} - 1\right) e^{-(x-x_n)/L_h}, \qquad \text{for } x > x_n, \end{aligned} \tag{12.106}$$

where $L_e = (D_e\tau_e)^{1/2}$ is the *electron diffusion length* in the p region and $L_h = (D_h\tau_h)^{1/2}$ is the *hole diffusion length* in the n region. Here τ_e is the lifetime of the minority electrons in the p region, and τ_h is the lifetime of the minority holes in the n region. They are not subject to the condition given in (12.54) for excess electrons and holes at the same location because they are the minority carrier lifetimes in separate regions on the two opposite sides of the junction. Therefore, τ_e and τ_h that define L_e and L_h here are independent of each other. Because of charge neutrality in the diffusion regions, the

concentrations of majority carriers also vary in space correspondingly:

$$p_{\mathrm{p}}(x) - p_{\mathrm{p0}} = n_{\mathrm{p}}(x) - n_{\mathrm{p0}}, \qquad \text{for } x < -x_{\mathrm{p}}, \tag{12.107}$$

$$n_{\mathrm{n}}(x) - n_{\mathrm{n0}} = p_{\mathrm{n}}(x) - p_{\mathrm{n0}}, \qquad \text{for } x > x_{\mathrm{n}}. \tag{12.108}$$

The relations in (12.105)–(12.108) that describe the carrier distributions are valid for p–N and P–n heterojunctions as well as for p–n homojunctions. The distributions of majority and minority carrier concentrations for the cases when a junction is subject to a forward bias and when it is subject to a reverse bias are shown in Figs. 12.11(*b*) and (*c*), respectively. Both diffusion regions on p and n sides are clearly defined by the diffusion lengths of minority carriers because the spatial distributions of both minority and majority carriers are determined by the diffusion lengths of the minority carriers. In the case when $p_{\mathrm{p0}} \gg n_{\mathrm{p0}}$ and $n_{\mathrm{n0}} \gg p_{\mathrm{p0}}$, as is the situation in many practical junction devices, the equal amount of local changes in the majority and minority carrier concentrations in the diffusion regions is relatively insignificant for the total majority carrier concentration, but it can be substantial for the total minority carrier concentration.

EXAMPLE 12.12 In this example, we consider the minority carrier concentrations at 300 K for the GaAs p–n homostructure described in Example 12.8. Find n_{p0} and p_{n0} first. Then, find the changes in the minority carrier concentrations, Δn_{p} and Δp_{n}, at the two edges of the depletion layer for two different forward bias voltages of $V = 1$ V and $V = V_0$, respectively.

Solution From Example 12.8, we know that $p_{\mathrm{p0}} \approx N_{\mathrm{a}} = 1 \times 10^{23}\ \mathrm{m}^{-3}$, $n_{\mathrm{n0}} \approx N_{\mathrm{d}} = 1 \times 10^{22}\ \mathrm{m}^{-3}$, and $n_{\mathrm{i}} = 2.33 \times 10^{12}\ \mathrm{m}^{-3}$. Because both p and n regions are nondegenerate, we can use the law of mass action given in (12.31) for nondegenerate semiconductors to find that

$$n_{\mathrm{p0}} = \frac{n_{\mathrm{i}}^2}{p_{\mathrm{p0}}} = \frac{(2.33 \times 10^{12})^2}{1 \times 10^{23}}\ \mathrm{m}^{-3} = 54.3\ \mathrm{m}^{-3},$$

$$p_{\mathrm{n0}} = \frac{n_{\mathrm{i}}^2}{n_{\mathrm{n0}}} = \frac{(2.33 \times 10^{12})^2}{1 \times 10^{22}}\ \mathrm{m}^{-3} = 543\ \mathrm{m}^{-3}.$$

At $T = 300$ K, $k_{\mathrm{B}}T = 0.0259$ eV. For a forward bias voltage of $V = 1$ V, we find from (12.103) and (12.104) that

$$\Delta n_{\mathrm{p}} = 54.3\ \mathrm{m}^{-3} \times \left(\mathrm{e}^{1/0.0259} - 1\right) = 3.18 \times 10^{18}\ \mathrm{m}^{-3},$$

$$\Delta p_{\mathrm{n}} = 543\ \mathrm{m}^{-3} \times \left(\mathrm{e}^{1/0.0259} - 1\right) = 3.18 \times 10^{19}\ \mathrm{m}^{-3}.$$

For a forward bias voltage of $V = V_0 = 1.209$ V, we have

$$\Delta n_{\mathrm{p}} = 54.3\ \mathrm{m}^{-3} \times \left(\mathrm{e}^{1.209/0.0259} - 1\right) = 1 \times 10^{22}\ \mathrm{m}^{-3} = n_{\mathrm{n0}},$$

$$\Delta p_{\mathrm{n}} = 543\ \mathrm{m}^{-3} \times \left(\mathrm{e}^{1.209/0.0259} - 1\right) = 1 \times 10^{23}\ \mathrm{m}^{-3} = p_{\mathrm{p0}}.$$

We find that though n_{p0} and p_{n0} are extremely small in this example, Δn_p and Δp_n are still quite substantial at a reasonable forward bias voltage because of their exponential dependence on bias voltage. We also find that $\Delta n_p = n_{n0}$ and $\Delta p_n = p_{p0}$ when the junction is forward biased at $V = V_0$.

Current–voltage characteristics

The electric current flowing in a semiconductor under bias consists of an electron current and a hole current, each having both drift and diffusion components. The total current is the vectorial sum of the individual current components, which may flow in different directions. The total current is constant throughout the semiconductor under a constant bias voltage, but the electron and hole currents, as well as their drift and diffusion components, vary from one location to another because of spatial variations in the carrier distribution and in the electric field distribution. In the depletion layer, there are drift and diffusion currents for both electrons and holes because in this layer a large electric field exists and the carrier concentration gradients for both electrons and holes are large. In the diffusion regions, both majority and minority carrier diffusion currents are significant because both majority and minority carriers have large concentration gradients here. Furthermore, there is an appreciable majority carrier drift current because the majority carrier concentration is high though the electric field is small. In homogeneous regions, almost the entire current is carried by majority carrier diffusion because both the carrier distribution gradients and the minority carrier concentration are negligibly small in these regions.

There are negligible generation and recombination of carriers in the depletion layer because the large electric field in the depletion layer sweeps the carriers across this layer very swiftly. In this situation, the total electron current density, $J_e(x)$, and the total hole current density, $J_h(x)$, are constant for $-x_p < x < x_n$ across the depletion layer. Consequently, the total current density in a semiconductor can be evaluated as

$$J = J_e(-x_p) + J_h(x_n), \tag{12.109}$$

where $J_e(-x_p)$ is the minority carrier current density at the boundary between the depletion layer and the diffusion region on the p side, and $J_h(x_n)$ is the minority carrier current density at the boundary between the depletion layer and the diffusion region on the n side. Because the minority carrier currents in the diffusion regions are purely diffusive, we have, using (12.105) and (12.106),

$$J_e(-x_p) = eD_e \left.\frac{dn_p}{dx}\right|_{x=-x_p} = \frac{eD_e}{L_e} n_{p0} \left(e^{eV/k_B T} - 1\right), \tag{12.110}$$

$$J_h(x_n) = -eD_e \left.\frac{dp_n}{dx}\right|_{x=x_n} = \frac{eD_h}{L_h} p_{n0} \left(e^{eV/k_B T} - 1\right). \tag{12.111}$$

Consequently, the total current density varies with the bias voltage V as

$$J = J_{\text{sat}}(e^{eV/k_B T} - 1), \tag{12.112}$$

where J_{sat} is the *saturation current density* given by

$$J_{\text{sat}} = \frac{eD_e}{L_e} n_{p0} + \frac{eD_h}{L_h} p_{n0}. \tag{12.113}$$

The minority carrier currents given in (12.110) and (12.111) are contributed by the minority carrier injection across the depletion layer to the diffusion regions. They have the following ratio:

$$\frac{J_e}{J_h} = \frac{D_e L_h}{D_h L_e} \frac{n_{p0}}{p_{n0}} = \left(\frac{D_e \tau_h}{D_h \tau_e}\right)^{1/2} \frac{n_{p0}}{p_{n0}}. \tag{12.114}$$

By using (12.79) for a p–N junction and (12.85) for a P–n junction, this ratio can be expressed as

$$\frac{J_e}{J_h} = \frac{D_e L_h}{D_h L_e} \frac{n_{n0}}{p_{p0}} \frac{N_{cp} N_{vp}}{N_{cn} N_{vn}} e^{(E_{gn} - E_{gp})/k_B T} = \frac{D_e L_h}{D_h L_e} \frac{N_d}{N_a} \frac{N_{cp} N_{vp}}{N_{cn} N_{vn}} e^{(E_{gn} - E_{gp})/k_B T}. \tag{12.115}$$

This relation is valid for p–n homojunctions as well as for p–N and P–n heterojunctions.

In the case of a homojunction, $J_e/J_h = D_e L_h N_d / D_h L_e N_a$ because $N_{cp} = N_{cn}$, $N_{vp} = N_{vn}$, and $E_{gn} = E_{gp}$. Therefore, the relative importance of electron and hole injection is determined by the diffusion parameters of the minority carriers and the doping concentrations in the n and p regions, respectively. It can be seen by examining the numerical values listed in Table 12.2 that $D_e > D_h$ because electrons have a higher mobility than holes in the same semiconductor. We thus reach the following important conclusion: *unless the p side is much more heavily doped than the n side, the injection current through a homojunction is predominantly carried by the electrons injected from the n side into the p side*.

In the case of a heterojunction, the exponential dependence on bandgap difference in (12.115) can be significant if $\Delta E_g > k_B T$. Because $k_B T = 25.9$ meV for $T = 300$ K, this exponential dependence dominates in most practical heterojunctions where the value of ΔE_g is many times this value. Consequently, $J_e \gg J_h$ for a p–N junction where $E_{gn} > E_{gp}$, whereas $J_h \gg J_e$ for a P–n junction where $E_{gp} > E_{gn}$. An important conclusion is reached for heterojunctions: *in the case of a p–N junction the diffusion current is mainly contributed by the injection of electrons from the wide-gap n-type semiconductor to the narrow-gap p-type semiconductor, whereas in the case of a P–n junction it is mainly contributed by the injection of holes from the wide-gap p-type semiconductor to the narrow-gap n-type semiconductor.* This important characteristic can be understood from the observation in earlier discussions on energy bands that the energy barrier across a heterojunction is lower for majority carriers of the

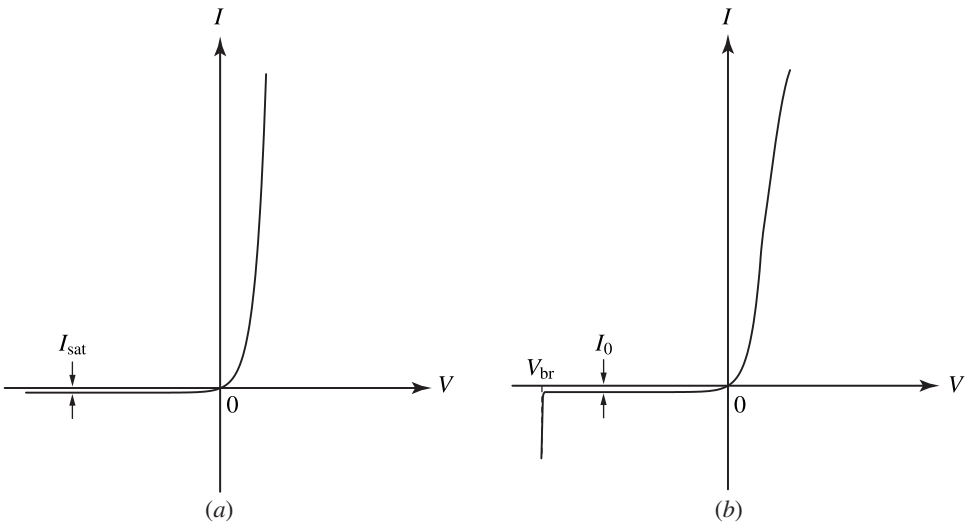


Figure 12.12 Current–voltage characteristics of (*a*) an ideal junction diode and (*b*) a realistic junction diode.

wide-gap semiconductor than for majority carriers of the narrow-gap semiconductor by the amount of the bandgap difference ΔE_g between the two semiconductors.

For a junction that has a cross-sectional area $\mathcal{A}$, the total current is $I = J\mathcal{A}$. Therefore, we have the following *diode equation* between the current and the bias voltage for an ideal p–n junction:

$$I = I_{sat}\left(e^{eV/k_BT} - 1\right), \tag{12.116}$$

where $I_{sat} = J_{sat}\mathcal{A}$ is the *saturation current*. This relation is valid for both forward and reverse bias conditions. It also applies to p–N and P–n heterojunctions. Figure 12.12(*a*) shows the *current–voltage characteristics*, also known simply as the *I–V characteristics*, described by (12.116) for an ideal diode. Deviations from these ideal characteristics are found in all realistic p–n, p–N, and P–n junctions. Some important characteristics of realistic junctions are shown in Fig. 12.12(*b*) and are summarized in the following.

1. At a certain critical reverse bias voltage, known as the *breakdown voltage*, V_{br}, a realistic junction breaks down with a sharp increase of reverse breakdown current.
2. The ideal diode equation is obtained by ignoring carrier recombination and generation in the depletion layer. When the effects of carrier recombination and generation in the depletion layer are considered, the reverse current does not saturate at $-I_{sat}$ but slightly increases in magnitude with the reverse bias voltage before a sudden change takes place at the breakdown voltage. Correspondingly, the forward current depends on the bias voltage with a modified factor in the exponent. Therefore, a

realistic diode has the following current–voltage relation for $V > V_{\text{br}}$:

$$I = I_0 \left(\mathrm{e}^{eV/ak_{\text{B}}T} - 1\right), \tag{12.117}$$

where I_0 is a constant current different from I_{sat} and a is a factor that has a value between 1 and 2.

3. At high injection levels when the minority carrier concentrations in the diffusion regions become comparable to the majority carrier concentrations, (12.105) and (12.106) are not valid. A detailed analysis of this situation results in the current–voltage characteristic described by (12.117) with $a = 2$ at these high injection levels.
4. The exponential rise of the current with forward bias voltage does not continue at high current levels as the voltage drop associated with finite resistivity in the neutral regions becomes significant at high currents.

EXAMPLE 12.13 For the GaAs p–n homostructure described in Example 12.8, the p region is doped 10 times more heavily than the n region. Take the minority carrier lifetimes to be $\tau_{\text{e}} = 10$ ns for electrons in the p region and $\tau_{\text{h}} = 100$ ns for holes in the n region. (a) By using the electron and hole diffusion coefficients for GaAs listed in Table 12.2, find the diffusion lengths that define the diffusion regions on the p and n sides, respectively. (b) Find the saturation current density. (c) Compare the relative importance of electron and hole injection currents. (d) Consider a junction that has a 100 μm × 100 μm cross section. Find the saturation current in reverse bias. Use the ideal diode equation to find the current under forward bias voltages of $V = 1$ V and $V = V_0$, respectively.

Solution (a) From Table 12.2, we find that $D_{\text{e}} = 220\ \text{cm}^2\ \text{s}^{-1} = 2.2 \times 10^{-2}\ \text{m}^2\ \text{s}^{-1}$ and $D_{\text{h}} = 10\ \text{cm}^2\ \text{s}^{-1} = 1 \times 10^{-3}\ \text{m}^2\ \text{s}^{-1}$. Therefore,

$$L_{\text{e}} = (D_{\text{e}}\tau_{\text{e}})^{1/2} = (2.2 \times 10^{-2} \times 10 \times 10^{-9})^{1/2}\ \text{m} = 14.8\ \mu\text{m},$$
$$L_{\text{h}} = (D_{\text{h}}\tau_{\text{h}})^{1/2} = (1 \times 10^{-3} \times 100 \times 10^{-9})^{1/2}\ \text{m} = 10\ \mu\text{m}.$$

The diffusion region on the p side is defined by $L_{\text{e}} = 14.8$ μm, and that on the n side is defined by $L_{\text{h}} = 10$ μm.

(b) From Example 12.12, we find that $n_{\text{p0}} = 54.3\ \text{m}^{-3}$ and $p_{\text{n0}} = 543\ \text{m}^{-3}$. Therefore, from (12.113), the saturation current density is

$$J_{\text{sat}} = \left(\frac{1.6 \times 10^{-19} \times 2.2 \times 10^{-2}}{14.8 \times 10^{-6}} \times 54.3 + \frac{1.6 \times 10^{-19} \times 1 \times 10^{-3}}{10 \times 10^{-6}} \times 543\right) \text{A m}^{-2}$$
$$= 2.16 \times 10^{-14}\ \text{A m}^{-2}.$$

(c) By using the relation in (12.114), we find that

$$\frac{J_{\text{e}}}{J_{\text{h}}} = \frac{D_{\text{e}}L_{\text{h}}}{D_{\text{h}}L_{\text{e}}}\frac{n_{\text{p0}}}{p_{\text{n0}}} = \frac{2.2 \times 10^{-2} \times 10}{1 \times 10^{-3} \times 14.8} \times \frac{54.3}{543} = 1.49.$$

Therefore, despite the fact that the doping concentration in the p region is 10 times that in the n region and τ_h is 10 times τ_e, we still find that the electron injection current is 1.49 times the hole injection current. The electron current would be even more important if the p region were not so heavily doped compared to the n region or if the hole lifetime were not so much longer than the electron lifetime.

(d) We have a cross-sectional area of $\mathcal{A} = 100\ \mu\text{m} \times 100\ \mu\text{m} = 1 \times 10^{-8}\ \text{m}^2$. Thus the saturation current is $I_{sat} = J_{sat}\mathcal{A} = 2.16 \times 10^{-22}$ A. At a sufficiently high reverse bias voltage of $V \ll -k_B T/e = -25.9$ mV at $T = 300$ K, $I = -I_{sat} = -2.16 \times 10^{-22}$ A. At a forward bias voltage of $V = 1$ V, we find that

$$I = 2.16 \times 10^{-22}\ \text{A} \times \left(e^{1/0.0259} - 1\right) = 12.7\ \mu\text{A}.$$

At a forward bias voltage of $V = V_0 = 1.209$ V, the current is

$$I = 2.16 \times 10^{-22}\ \text{A} \times \left(e^{1.209/0.0259} - 1\right) = 40.5\ \text{mA}.$$

Comparing these results, we find that the current saturates with voltage in reverse bias but increases very quickly with voltage in forward bias.

Capacitance

There are two types of capacitance associated with a p–n, p–N, or P–n junction: (1) the *junction capacitance*, C_j, also known as the *depletion-layer capacitance*, and (2) the *diffusion capacitance*, C_d, also known as the *charge-storage capacitance*. In a junction under reverse bias, only the junction capacitance is important. In a junction under forward bias, however, the diffusion capacitance dominates.

The depletion layer acts as a capacitor by holding negative space charges on the p side and positive space charges on the n side of the following magnitude:

$$Q = eN_a x_p \mathcal{A} = eN_d x_n \mathcal{A} = e\frac{N_a N_d}{N_a + N_d} W\mathcal{A}, \tag{12.118}$$

where $\mathcal{A}$ is the cross-sectional area of the junction. By using (12.100) and (12.118), we find that the junction capacitance associated with the depletion layer is given by

$$C_j = \left|\frac{dQ}{dV}\right| = \frac{\epsilon\mathcal{A}}{W}. \tag{12.119}$$

Because the width of the depletion layer decreases with forward bias but increases with reverse bias, the junction capacitance increases when the junction is subject to a forward bias voltage but decreases when it is subject to a reverse bias voltage.

Because the diffusion capacitance, C_d, is associated with the storage of minority carrier charges in the diffusion region, it exists only when a junction is under forward bias. This capacitance is a complicated function of the minority carrier lifetime and the modulation frequency of the bias voltage, but it is directly proportional to the injection current. When a junction is under forward bias, C_d can be significantly larger than

C_j at high injection currents though C_j can already be large in this situation. When a junction is under reverse bias, C_j is the only capacitance of significance though it can be small. Consequently, the capacitance of a junction can be substantially smaller when it is under reverse bias than when it is under forward bias.

EXAMPLE 12.14 A GaAs p–n homojunction as described in Example 12.8 and considered in Examples 12.11–12.13 has a 100 μm × 100 μm cross section. Consider the junction in thermal equilibrium without bias at 300 K. Find the amount of the positive and negative space charges stored in the depletion layer. Find the junction capacitance.

Solution The cross-sectional area of the junction is $\mathcal{A} = 100\ \mu\text{m} \times 100\ \mu\text{m} = 1 \times 10^{-8}\ \text{m}^2$. An equal amount of positive and negative space charges is stored on the n and p sides, respectively, of the junction in the depletion layer. By using $N_a = 1 \times 10^{23}\ \text{m}^{-3}$ given in Example 12.8 and $x_{p0} = 40$ nm found in Example 12.11, we find from (12.118) that

$$Q = eN_a x_{p0}\mathcal{A} = 1.6 \times 10^{-19} \times 1 \times 10^{23} \times 40 \times 10^{-9} \times 1 \times 10^{-8}\ \text{C} = 6.4\ \text{pC}.$$

From Example 12.11, we know that $\epsilon = 13.18\epsilon_0$ for GaAs and $W_0 = 440$ nm for the junction under consideration. Therefore, from (12.119), the junction capacitance is

$$C_j = \frac{13.18 \times 8.854 \times 10^{-12} \times 1 \times 10^{-8}}{440 \times 10^{-9}}\ \text{F} = 2.65\ \text{pF}.$$

PROBLEMS

12.1.1 Does the ternary compound $Al_{0.3}Ga_{0.7}As$ have a direct or an indirect bandgap? What are its bandgap E_g and the corresponding optical wavelength λ_g? What is its refractive index at $\lambda = 900$ nm?

12.1.2 Answer the questions asked in Problem 12.1.1 for $Al_{0.7}Ga_{0.3}As$.

12.1.3 The quaternary compound $In_{0.61}Ga_{0.39}As_{0.83}P_{0.17}$ is lattice matched to InP at 300 K. Is it a direct-gap or an indirect-gap semiconductor? What are its bandgap E_g and the corresponding optical wavelength λ_g? What is its refractive index at $\lambda = 1.55$ μm?

12.1.4 Find the compositions of the two InGaAsP quaternary compounds that are both lattice matched to InP at 300 K and have bandgap optical wavelengths of $\lambda_g = 1.007$ and 1.095 μm, respectively.

12.1.5 InGaAsP can be lattice matched to either InP or GaAs. Find from Fig. 12.2 the range of bandgaps and the corresponding range of bandgap wavelengths covered by the InGaAsP quaternary compounds that are lattice matched to GaAs at 300 K. What is the composition range of such compounds?

12.2.1 We have $m_e^* = 1.08m_0$ and $m_h^* = 0.56m_0$ for Si, and $m_e^* = 0.55m_0$ and $m_h^* = 0.31m_0$ for Ge. Use these data to calculate the effective densities of states N_c

and N_v for Si and Ge, respectively. Compare them with those of GaAs found in Example 12.2.

12.2.2 From Table 12.1, we find that $E_g = 1.12$ eV for Si and $E_g = 0.66$ eV for Ge at $T = 300$ K. Use the values of N_c and N_v found for Si and Ge in Problem 12.2.1 to find the intrinsic carrier concentration n_i for Si and Ge at 300 K. Also find the Fermi level for intrinsic Si and Ge at 300 K. Compare the results with those found in Example 12.2 for GaAs. Explain the significant differences in the intrinsic carrier concentration among Si, Ge, and GaAs at the same temperature of 300 K.

12.2.3 Find the intrinsic carrier concentration and the Fermi level of GaAs at 400 K and those at 500 K. Note that the bandgap of GaAs is 1.424 eV at 300 K, 1.391 eV at 400 K, and 1.357 eV at 500 K because it shrinks with increasing temperature. Compare the results with those found in Example 12.2 for GaAs at 300 K to appreciate how n_i and E_{Fi} vary with temperature.

12.2.4 Answer the questions in Example 12.3 for n-type Si and n-type Ge, respectively, for the same impurity concentration of $N_d^+ - N_a^- = 5 \times 10^{18}$ m^{-3}. Compare the results with those found for n-type GaAs.

12.2.5 Find the impurity doping concentrations required for n-type Si and n-type Ge, respectively, to become degenerate at 300 K. Compare them with that found in Example 12.4 for n-type GaAs.

12.2.6 Answer the questions in Example 12.5 for Si and Ge, respectively. Compare the results with those found for GaAs.

12.2.7 A piece of p-type GaAs is doped with a net impurity concentration of $N_a^- - N_d^+ = 5 \times 10^{18}$ m^{-3}. Is it degenerate or nondegenerate? Find its electron and hole concentrations and its Fermi level at 300 K. How much is the shift of the Fermi level, measured from the intrinsic Fermi level, caused by the doping of the impurity? Compare the results obtained in this problem for the p-type GaAs with those found in Example 12.3 for the n-type GaAs of the same impurity concentration.

12.2.8 What is the impurity doping concentration required for p-type GaAs to become degenerate at 300 K? Compare the result obtained in this problem for p-type GaAs with that found in Example 12.4 for n-type GaAs.

12.3.1 In the situation where excess electron and hole concentrations are equal, $\Delta n = \Delta p = N$, electrons and holes have the same lifetime, $\tau_e = \tau_h = \tau_s$, according to (12.54). Consider a semiconductor in which the Shockley–Read recombination process completely dominates the bimolecular and Auger processes so that the net carrier recombination rate is basically the net Shockley–Read recombination rate: $R = R_{SR}$. Assume, for simplicity, that the recombination centers are so located that $n_1 = p_1 = n_i$.

a. Use (12.48) for R_{SR} to obtain a general expression for the carrier lifetime τ_s in terms of τ_{e0}, τ_{h0}, and the carrier concentrations.

b. Show that $\tau_s = \tau_e = \tau_h = \tau_{e0} + \tau_{h0}$ for an intrinsic semiconductor.
c. For an extrinsic semiconductor, the carrier lifetime is primarily determined by the recombination of the minority carriers. For an n-type semiconductor, show that $\tau_s > \tau_{h0}$ and $\tau_s \to \tau_{h0}$ when $n \gg p$. Similarly, for a p-type semiconductor, show that $\tau_s > \tau_{e0}$ and $\tau_s \to \tau_{e0}$ when $p \gg n$.
d. Show that for both intrinsic and extrinsic semiconductors, $\tau_s = \tau_e = \tau_h \approx \tau_{e0} + \tau_{h0}$ at a high excess carrier concentration when $N \gg n_0, p_0$.

12.3.2 Answer the questions in Example 12.6 for a p-type GaAs sample that is doped with $N_a^- - N_d^+ = 2 \times 10^{22}$ m^{-3} while maintaining the same values for coefficients A, B, and C as those given in Example 12.6.

12.3.3 An InGaAsP sample has a bandgap at 1.3 μm wavelength. It is not intentionally doped, so that its equilibrium electron and hole concentrations are $n_0 \approx p_0 \approx n_i \approx 1.7 \times 10^{22}$ m^{-3} at 300 K. For an excess carrier concentration of $N \gg n_0, p_0$, the terms in (12.57) that contain n_0 and p_0 can be neglected. Then, (12.57) reduces to

$$\frac{1}{\tau_s} = A + BN + CN^2. \tag{12.120}$$

Coefficients A, B, and C in this relation can be determined experimentally by measuring the carrier lifetime τ_s as a function of the injected excess carrier concentration N. This experiment can be carried out by measuring the decay time of the photoluminescence when the sample is injected with excess electron and hole pairs either optically or electrically. Such an experiment yields the following data: $\tau_s = 30$, 17.1, 3.13, and 1.25 ns for $N = 5 \times 10^{23}$ m^{-3}, 1×10^{24} m^{-3}, 5×10^{24} m^{-3}, and 1×10^{25} m^{-3}, respectively.
a. Find coefficients A, B, and C from these experimental data.
b. Use the results from (a) to find the excess carrier concentration for a carrier lifetime of $\tau_s = 5$ ns.
c. What is the carrier lifetime at an excess carrier concentration of $N = 2 \times 10^{25}$ m^{-3}?

12.3.4 For the InGaAsP sample considered in Problem 12.3.3, the bimolecular recombination process, characterized by the coefficient B, is radiative but the other two recombination processes, characterized by coefficients A and C, are nonradiative. If this sample is to be used for the fabrication of a semiconductor laser or light-emitting diode, what is the range of injection carrier concentrations that will lead to the most efficient operation of the device?

12.4.1 For nondegenerate semiconductors, the Einstein relations given in (12.61) between the diffusion coefficients and the mobilities are valid. For degenerate semiconductors, they have to be generalized. However, the relations for the electron and hole current densities given in (12.67) and (12.68), respectively, are generally valid for both nondegenerate and degenerate semiconductors.

a. Verify the relations for $\boldsymbol{J}_e$ and $\boldsymbol{J}_h$ given in (12.67) and (12.68), respectively, using the Einstein relation given in (12.61) for a nondegenerate semiconductor.
b. Starting from the relations for $\boldsymbol{J}_e$ and $\boldsymbol{J}_h$ given in (12.67) and (12.68), derive the relation between the diffusion coefficients and the mobilities for a degenerate semiconductor.
c. Show that the relation obtained in (b) is a generalized relation for the Einstein relation by showing that it reduces to (12.61) when proper approximations for a nondegenerate semiconductor are taken.

12.4.2 A graded-gap structure can be fabricated by varying the composition of a ternary or quaternary compound semiconductor, such as $Al_xGa_{1-x}As$ or $In_{1-x}Ga_xAs_{1-y}P_y$, or that of an alloy semiconductor, such as Si_xGe_{1-x}. A graded-gap structure of a length l that has a linearly graded bandgap from a small bandgap of $E_{g1} = E_{c1} - E_{v1}$ to a large bandgap of $E_{g2} = E_{c2} - E_{v2}$, for a bandgap change of $\Delta E_g = E_{g2} - E_{g1}$ from one end to the other, is shown in Fig. 12.13. The change in the conduction-band edge over the structure is $\Delta E_c = E_{c2} - E_{c1}$, and that in the valence-band edge is $\Delta E_v = E_{v2} - E_{v1}$. Assume that the electron and hole mobilities and concentrations are uniform across the entire structure.

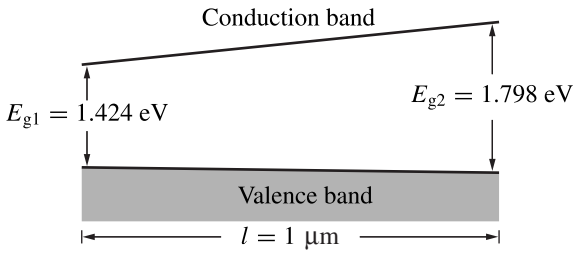


Figure 12.13 Graded-gap $Al_xGa_{1-x}As$ structure with a linearly graded bandgap. The gradients of the band edges are a function of doping type and concentration. They are arbitrarily drawn here for a p-type structure.

a. Show that, under this condition,

$$\Delta E_c = \frac{\mu_h p}{\mu_e n + \mu_h p}\Delta E_g \quad \text{and} \quad \Delta E_v = -\frac{\mu_e n}{\mu_e n + \mu_h p}\Delta E_g. \tag{12.121}$$

b. What are the built-in electric fields seen by electrons and holes, respectively?
c. How are the built-in electric fields changed by heavy n-type or p-type doping?

12.4.3 A graded-gap $Al_xGa_{1-x}As$ structure with a linearly graded bandgap as shown in Fig. 12.13 has its composition varying from $x = 0$ to 0.3 for a bandgap change of $\Delta E_g = 0.374$ eV from $E_{g1} = 1.424$ eV to $E_{g2} = 1.798$ eV at 300 K. It has a length of $l = 1$ μm. Consider an undoped intrinsic structure. Use the electron and hole mobilities, $\mu_e = 8500$ cm^2 V^{-1} s^{-1} and $\mu_h = 400$ cm^2 V^{-1} s^{-1}, of

GaAs for the entire structure to find the built-in electric fields seen by electrons and holes, respectively.

12.4.4 Find the intrinsic conductivity and the intrinsic resistivity of Si at 300 K.

12.4.5 Find the intrinsic conductivity and the intrinsic resistivity of Ge at 300 K.

12.4.6 The intrinsic carrier concentration of GaAs at 300 K is $n_i = 2.33 \times 10^{12}$ m^{-3}, found in Example 12.2. From Table 12.2, $\mu_e = 8500$ cm^2 V^{-1} s^{-1} and $\mu_h = 400$ cm^2 V^{-1} s^{-1}. Find the dark conductivity σ_0 and the dark resistivity ρ_0 for the n-type GaAs considered in Example 12.3, which is lightly doped with $N_d^+ - N_a^- = 5 \times 10^{18}$ m^{-3}. Find σ_0 and ρ_0 for p-type GaAs doped with $N_a^- - N_d^+ = 5 \times 10^{18}$ m^{-3}. What is the reason for the difference in the conductivity between the n-type and p-type GaAs of the same doping concentration considered here?

12.5.1 An abrupt GaAs/AlGaAs p–N heterojunction is formed with a uniform p region of GaAs and a uniform n region of $Al_{0.3}Ga_{0.7}As$. The p GaAs region is doped with fully ionized acceptors of a concentration $N_a = 1 \times 10^{23}$ m^{-3}, and the n $Al_{0.3}Ga_{0.7}As$ region is doped with fully ionized donors of a concentration $N_d = 1 \times 10^{22}$ m^{-3}. The density of states effective masses for $Al_{0.3}Ga_{0.7}As$ are $m_e^* = 0.092m_0$ and $m_h^* = 0.62m_0$, as compared with $m_e^* = 0.067m_0$ and $m_h^* = 0.52m_0$ for GaAs. Find the contact potential of this junction at 300 K by taking $\Delta E_c = 65\%\ \Delta E_g$ and $\Delta E_v = 35\%\ \Delta E_g$. Compare this contact potential with that of the AlGaAs/GaAs P–n heterojunction considered in Example 12.9, which has the same doping profile.

12.5.2 Find the bias voltage required to line up the conduction-band edges of the p and n regions on the two sides of the p–N heterojunction described in Problem 12.5.1. Find also the bias voltage needed to line up the valence-band edges. Compare these voltages with those found in Example 12.10 for the P–n heterojunction.

12.5.3 The static dielectric constant of GaAs is $\epsilon/\epsilon_0 = 13.18$ and that of $Al_{0.3}Ga_{0.7}As$ is $\epsilon/\epsilon_0 = 12.24$. Find the width of the depletion layer, W_0, and the penetration depths, x_{p0} and x_{n0}, in thermal equilibrium without bias at 300 K for (a) the AlGaAs/GaAs P–n heterojunction described in Example 12.9 and (b) the GaAs/AlGaAs p–N heterojunction described in Problem 12.5.1. Compare the results with those found in Example 12.11 for the GaAs p–n homojunction.

12.5.4 Answer the questions in Example 12.12 for (a) the AlGaAs/GaAs P–n heterojunction described in Example 12.9 and (b) the GaAs/AlGaAs p–N heterojunction described in Problem 12.5.1. Compare the results with those found in Example 12.12 for the GaAs p–n homojunction.

12.5.5 Answer the questions in Example 12.13 for the AlGaAs/GaAs P–n heterojunction described in Example 12.9. Take electron and hole diffusion coefficients to be those of GaAs and the minority carrier lifetimes to be $\tau_e = 10$ ns for electrons in the p region and $\tau_h = 100$ ns for holes in the n region. Compare the results with those found in Example 12.13 for the GaAs p–n homojunction.

12.5.6 Answer the questions in Example 12.13 for the GaAs/AlGaAs p–N heterojunction described in Problem 12.5.1. Take electron and hole diffusion coefficients to be those of GaAs and the minority carrier lifetimes to be $\tau_e = 10$ ns for electrons in the p region and $\tau_h = 100$ ns for holes in the n region. Compare the results with those found in Example 12.13 for the GaAs p–n homojunction and those found in Problem 12.5.5 for the AlGaAs/GaAs P–n heterojunction.

12.5.7 Answer the questions in Example 12.14 for (a) the AlGaAs/GaAs P–n heterojunction described in Example 12.9 and (b) the GaAs/AlGaAs p–N heterojunction described in Problem 12.5.1. Compare the results with those found in Example 12.14 for the GaAs p–n homojunction.

12.5.8 Show that when $p_{p0} \approx N_a \gg n_{p0}$ and $n_{n0} \approx N_d \gg p_{n0}$, the contact potential of a p–N heterojunction can be expressed in the form of (12.81) and that of a P–n heterojunction can be expressed in the form of (12.87), both of which reduce to the form of (12.75) when the heterojunctions reduce to homojunctions with $E_{gp} = E_{gn} = E_g$.

12.5.9 Show, by solving (12.97) with proper boundary conditions, that the width of the depletion layer for a junction under a bias voltage V is that given by (12.100) with an effective electric permittivity defined in (12.99), which is valid for both homojunctions and heterojunctions.

12.5.10 Discuss why a p–n junction has a larger capacitance when it is under forward bias than when it is under reverse bias.

SELECT BIBLIOGRAPHY

Agrawal, G. P. and Dutta, N. K., *Semiconductor Lasers*, 2nd edn. New York: Van Nostrand Reinhold, 1993.

Bhattacharya, P., *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall, 1997.

Chuang, S. L., *Physics of Optoelectronic Devices*. New York: Wiley, 1995.

Ebeling, K. J., *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*. Berlin: Springer-Verlag, 1993.

Gowar, J., *Optical Communication Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Hunsperger, R. G., *Integrated Optics: Theory and Technology*, 5th edn. New York: Springer-Verlag, 2002.

Kressel, H. and Butler, J. K., *Semiconductor Lasers and Heterojunction LEDs*. New York: Academic Press, 1977.

Madelung, O., ed., *Semiconductors: Basic Data*, 2nd edn. Berlin: Springer, 1996.

Semiconductors: Other than Group IV Elements and III–V Compounds. Berlin: Springer-Verlag, 1992.

Pankove, J. I., *Optical Processes in Semiconductors*. New York: Dover, 1975.

Rosencher, E. and Vinter, B., *Optoelectronics*. Cambridge: Cambridge University Press, 2002.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Streetman, B. G. and Banerjee, S., *Solid State Electronic Devices*, 5th edn. Upper Saddle River, NJ: Prentice-Hall, 2000.

Sze, S. M., *Physics of Semiconductor Devices*, 2nd edn. New York: Wiley, 1981.

Wolfe, C. M., Holonyak, N., Jr., and Stillman, G. E., *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

ADVANCED READING LIST

Adachi, S., "Material parameters of $In_{1-x}Ga_xAs_yP_{1-y}$ and related binaries," *Journal of Applied Physics* **53**(12): 8775–8792, Dec. 1982.

"GaAs, AlAs, and $Al_xGa_{1-x}As$: material parameters for use in research and device applications," *Journal of Applied Physics* **58**(3): R1–R29, Aug. 1985.

Alferov, Z. I., "The double heterostructure concept and its applications in physics, electronics and technology," *International Journal of Modern Physics B* **16**(5): 647–675, Feb. 2002.

"Nobel lecture: the double heterostructure concept and its applications in physics, electronics, and technology," *Reviews of Modern Physics* **73**(3): 767–782, July 2001.

Blakemore, J. S., "Semiconducting and other major properties of gallium arsenide," *Journal of Applied Physics* **53**(10): R123–R181, Oct. 1982.

Kroemer, H., "Nobel lecture: quasielectric fields and band offsets – teaching electrons new tricks," *Reviews of Modern Physics* **73**(3): 783–793, July 2001.

Miles, A. G., "Semiconductor heterojunction topics: introduction and overview," *Solid-State Electronics* **29**(2): 99–121, 1986.

Notte, D. D., "Semi-insulating semiconductor heterostructures: optoelectronic properties and applications," *Journal of Applied Physics* **85**(9): 6259–6289, May 1999.

Vurgaftman, I., Meyer, J. R., and Ram-Mohan, L. R., "Band parameters for III–V compound semiconductors and their alloys," *Journal of Applied Physics* **89**(11/1–2): 5815–75, June 2001.

13 Semiconductor lasers and light-emitting diodes

In this chapter, we discuss semiconductor lasers and light-emitting diodes (LEDs). Both semiconductor lasers and LEDs are semiconductor light sources based on electroluminescence, which results from the radiative recombination of electrons and holes in a semiconductor. A semiconductor laser emits coherent laser light with a relatively small divergence, whereas the emission of an LED is incoherent and divergent. These semiconductor devices have several unique properties. They are rugged devices that are reliable and have long operating lifetimes because of their very small, compact sizes with integrated solid-state structures. They have very high efficiencies and consume very little power in comparison with other light sources of similar brightness because they are cold light sources operating at temperatures that are much lower than the equilibrium temperatures of their emission spectra. They can be electrically pumped by current injection at relatively low current and voltage levels and can be directly current modulated with very fast response for high-speed applications, including broadband optical communications. Their compatibility with semiconductor fabrication and processing technologies allows them to take advantage of semiconductor electronics technology for easy integration into electronic systems. Furthermore, the mature nature of semiconductor electronics technology allows them to be mass produced at a low cost. These unique properties make semiconductor lasers and LEDs the light sources of choice in many practical applications.

13.1 Radiative recombination

The general characteristics of electron–hole recombination processes in a semiconductor are discussed in Section 12.3. The net result of any recombination process is the transition of an electron from an occupied state at a higher energy to an empty state at a lower energy, accompanied by the release of the energy that is the difference between these two states. An electron–hole recombination process in a semiconductor can be either radiative or nonradiative. In a radiative recombination process, the released energy is emitted as electromagnetic radiation. In a nonradiative recombination process, no radiation is emitted, and the released energy is eventually converted to thermal energy

in the form of lattice vibrations. Only radiative processes are useful to the function of semiconductor lasers and LEDs.

There are primarily three different radiative recombination processes: (1) band-to-band recombination; (2) exciton recombination, through either a free exciton or a bound exciton; and (3) recombination through impurity states. The most important radiative recombination process is the bimolecular band-to-band recombination process, the details of which are discussed in the following section. Free exciton recombination is radiative, but it is not important for practical device applications at room temperature because free excitons can form only at very low temperatures due to their small ionization energies. Radiative recombination of certain types of bound excitons can be useful. Certain radiative recombination processes associated with impurities in a semiconductor are important in the operation of some semiconductor lasers or LEDs. A photon emitted by band-to-band recombination has an energy slightly higher than the bandgap, whereas one that is emitted through a process involving the impurities has an energy lower than the bandgap of the semiconductor.

Certain impurities in a semiconductor can form *isoelectronic centers*. An isoelectronic center is normally neutral but introduces a local potential that can trap an electron or a hole, depending on the type of impurity that creates the isoelectronic center. An isoelectronic center that traps an electron becomes negatively charged. The negatively charged center can then capture a hole from the valence band to form a bound exciton. Similarly, an isoelectronic center that traps a hole becomes positively charged and is able to capture an electron from the conduction band to form a bound exciton. Subsequent annihilation of the electron–hole pair in the bound exciton is a radiative process that results in the emission of a photon of an energy equal to the bandgap minus the binding energy of the center. Because the momentum of a trapped, localized electron or hole is highly diffused according to the uncertainty principle of quantum mechanics, conservation of momentum can easily be satisfied in the radiative recombination process through an isoelectronic center no matter whether the host semiconductor is a direct-gap or an indirect-gap material. Consequently, this mechanism of radiative recombination is important in indirect-gap semiconductors, in which band-to-band radiative recombination probabilities are very low. In particular, this process is responsible for improving the luminescence efficiency of the indirect-gap semiconductors GaP, $GaAs_xP_{1-x}$, and $In_xGa_{1-x}P$ for their applications as materials for LEDs. As an example, the energy levels of the isoelectronic traps created by the impurities N and Zn,O in GaP are illustrated in Fig. 13.1. The N and Zn,O impurities in GaP both act as electron traps. At room temperature, a photon emitted through a N center in GaP : N has an energy of about 2.20 eV, and that emitted through a Zn,O center in GaP : Zn,O has an energy of about 1.79 eV.

A high impurity concentration in a semiconductor can lead to the formation of conduction and valence *bandtail states*, which in effect extend the conduction- and valence-band edges into the gap. Optical transitions associated with such bandtail states

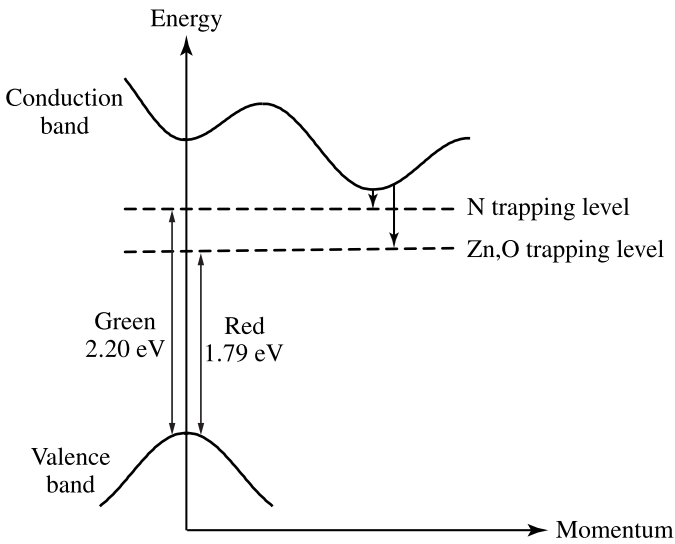


Figure 13.1 Isoelectronic trapping levels of N and Zn,O centers in GaP.

result in the absorption or emission of photons of energies less than the bandgap of a semiconductor. This bandtail effect is important only in a direct-gap semiconductor that is doped with a high concentration of impurities.

Radiative efficiency

The total recombination rate for the excess carriers in a semiconductor can be expressed as the sum of radiative and nonradiative recombination rates:

$$R = R_{\text{rad}} + R_{\text{nonrad}}. \tag{13.1}$$

The lifetime of an excess electron–hole pair associated with radiative recombination is called the *radiative carrier lifetime*, τ_{rad}, and that associated with nonradiative recombination is called the *nonradiative carrier lifetime*, τ_{nonrad}. They are related to the total *spontaneous carrier recombination lifetime*, τ_{s}, of the excess carriers by

$$\frac{1}{\tau_{\text{s}}} = \frac{1}{\tau_{\text{rad}}} + \frac{1}{\tau_{\text{nonrad}}}. \tag{13.2}$$

The *spontaneous carrier recombination rate*, γ_{s}, is defined as

$$\gamma_{\text{s}} = \frac{1}{\tau_{\text{s}}}. \tag{13.3}$$

This parameter is the total rate of carrier recombination including the contributions from all, radiative and nonradiative, spontaneous recombination processes but excluding the contribution from the stimulated recombination process. In the presence of stimulated

emission, the effective recombination rate of the carriers can be much higher than that given by γ_s because of stimulated recombination.

The *radiative efficiency*, or the *internal quantum efficiency*, of a semiconductor is defined as

$$\eta_i = \frac{R_{rad}}{R} = \frac{\tau_s}{\tau_{rad}}. \tag{13.4}$$

Bimolecular radiative lifetime

In a practical operating condition of a semiconductor laser or LED, the radiative recombination rate is almost entirely contributed by bimolecular recombination, including band-to-band and exciton recombination processes. In thermal equilibrium, bimolecular recombination is balanced by bimolecular thermal generation. The bimolecular thermal generation rate, G_0, is the same for the generation of electrons and holes. Therefore, the net radiative recombination rate in the presence of excess electron–hole pairs is given by

$$R_{rad} = Bnp - G_0 = Bnp - Bn_0p_0, \tag{13.5}$$

where G_0 is identified with Bn_0p_0 because $R_{rad} = 0$ in the state of thermal equilibrium when $n = n_0$ and $p = p_0$. In contrast to the bimolecular recombination rate, which depends on the total electron and hole concentrations, the thermal generation rate is largely independent of the carrier concentrations because the bound electrons in the valence bands and the empty states in the conduction bands that are available for thermal generation of free electrons and free holes are always much more numerous than the values of n and p. Consequently, even in a semiconductor that has a high concentration of excess carriers generated by external excitation, the bimolecular thermal generation rate remains $G_0 = Bn_0p_0$.

With the excess carrier density $N = n - n_0 = p - p_0$ as expressed in (12.55), the radiative lifetime of the excess carriers is then given by

$$\tau_{rad} = \frac{N}{R_{rad}} = \frac{1}{B(N + n_0 + p_0)}. \tag{13.6}$$

A short radiative lifetime corresponds to a high radiative recombination rate. In the case when the excess carrier density is low so that $N \ll n_0$, p_0, the radiative lifetime is a constant that is independent of the density of the excess carriers:

$$\tau_{rad} \approx \frac{1}{B(n_0 + p_0)}. \tag{13.7}$$

In the case when the excess carrier density is high so that $N \gg n_0$, p_0, the radiative lifetime varies inversely with the excess carrier density:

$$\tau_{rad} \approx \frac{1}{BN}. \tag{13.8}$$

EXAMPLE 13.1 Find the radiative carrier lifetime and the internal quantum efficiency for the optically excited n-type GaAs considered in Example 12.6 if both the Shockley–Read and the Auger recombination processes in this semiconductor are nonradiative while the bimolecular process is purely radiative. Plot them as a function of excess carrier concentration N for N in the range between 10^{18} and 10^{26} m^{-3}. In what range of carrier densities is high radiative efficiency found? What is the peak internal quantum efficiency?

Solution From Example 12.6, we have the following spontaneous carrier lifetime:

$$\frac{1}{\tau_s} = A + B(N + n_0 + p_0) + C\left[N^2 + \frac{3}{2}(n_0 + p_0)N + \frac{1}{2}(n_0^2 + p_0^2) + 2n_0 p_0\right].$$

Because the bimolecular process is purely radiative while the Shockley–Read and the Auger recombination processes are nonradiative, the radiative carrier lifetime is that given in (13.6):

$$\tau_{rad} = \frac{N}{R_{rad}} = \frac{1}{B(N + n_0 + p_0)}.$$

From Example 12.6, we have $A = 5.0 \times 10^5$ s^{-1}, $B = 8.0 \times 10^{-17}$ m^3 s^{-1}, and $C = 5.0 \times 10^{-42}$ m^6 s^{-1}. From Example 12.3, we have $n_0 = 5.0 \times 10^{18}$ m^{-3} and $p_0 = 1.1 \times 10^6$ m^{-3}. Using these parameters, we can find τ_s and τ_{rad} as a function of the carrier concentration N. Then the internal quantum efficiency can be found by using (13.4) as $\eta_i = \tau_s/\tau_{rad}$. The results are plotted in Fig. 13.2. We find from these results that $\eta_i > 0.5$ for carrier concentrations in the range of 6.25×10^{21} m^{-3} $< N < 1.6 \times 10^{25}$ m^{-3}

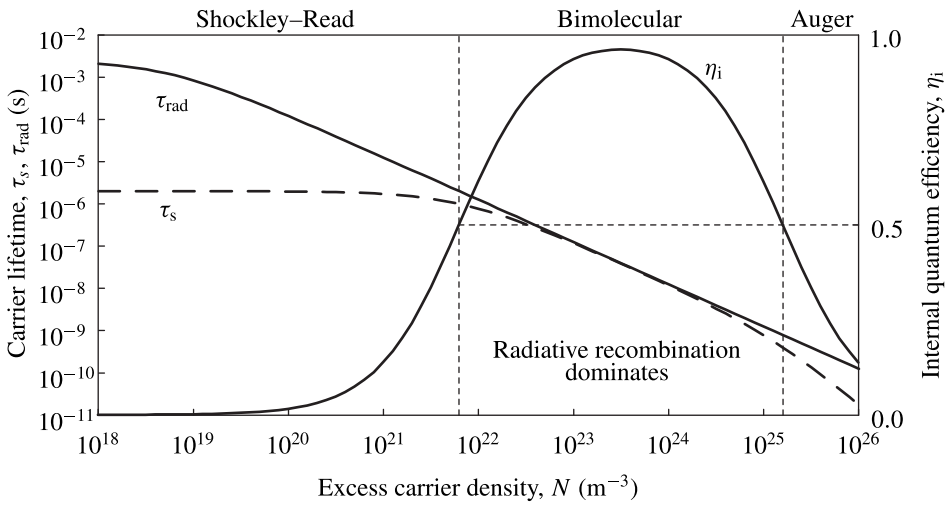


Figure 13.2 Spontaneous carrier lifetime τ_s, radiative carrier lifetime τ_{rad}, and internal quantum efficiency η_i as a function of excess carrier density. The carrier lifetimes are referred to the left axis while the quantum efficiency is referred to the right axis.

where the bimolecular recombination process dominates, according to Example 12.6. We also find from Fig. 13.2 that the peak internal quantum efficiency is 96.2% for $N = 3.16 \times 10^{23}$ m^{-3}.

13.2 Band-to-band optical transitions

In the discussions of the characteristics of optical transitions between the energy levels of an individual atom or molecule in Section 10.1, each active atom or molecule is considered a separate system in the sense that it has its own energy levels and it can reside in a particular state independently of the states of other active atoms or molecules. For the electrons and holes in a semiconductor, the situation is quite different. The states of all of the valence electrons in a semiconductor collectively form energy bands. Because the electron population in the band states is governed by the Fermi–Dirac distribution function, the state of a given electron in a semiconductor is not independent of other electrons. A band-to-band transition in a semiconductor takes place through the transition of such an electron between a valence band and a conduction band. Consequently, not every concept discussed in Section 10.1 regarding optical transitions between the energy levels of an individual atom or molecule is directly applicable to band-to-band optical transitions in a semiconductor. In particular, the concepts of transition cross section and population inversion have to be modified. When considering a band-to-band optical transition, the characteristics of the band structure have to be considered.

There are two types of band-to-band transitions in a semiconductor. A *direct transition* takes place when an electron makes an upward or downward transition without the participation of a phonon. In contrast, when an electron makes an *indirect transition*, it has to absorb or emit a phonon, thereby exchanging energy and momentum with the crystal lattice, in order to complete the transition. The transition probability differs significantly between a direct process and an indirect process.

When an electron makes a band-to-band transition between a state $|1\rangle$ of energy E_1 and wavevector $\mathbf{k}_1$ in a valence band and a state $|2\rangle$ of energy E_2 and wavevector $\mathbf{k}_2$ in a conduction band, both the conservation of energy and the conservation of wavevector have to be satisfied among all parties involved, including any participating photon and phonon. The magnitude of the electron wavevector in a crystal is of the order of $2\pi/a$, where the lattice constant a is smaller than 1 nm, but the wavevector of a photon is $2\pi/\lambda$, where the wavelength λ is on the order of 1 μm. Clearly, the photon wavevector is negligibly small in comparison to the electron wavevector. Consequently, the conditions for direct band-to-band transition with the absorption or emission of a photon are

$$E_2 - E_1 = h\nu \qquad \text{and} \qquad \mathbf{k}_2 = \mathbf{k}_1 + \mathbf{k}_{\text{photon}} \approx \mathbf{k}_1. \tag{13.9}$$

The requirement of the conservation of wavevector is akin to the requirement of phase matching in the interaction among optical waves and the requirement of the conservation

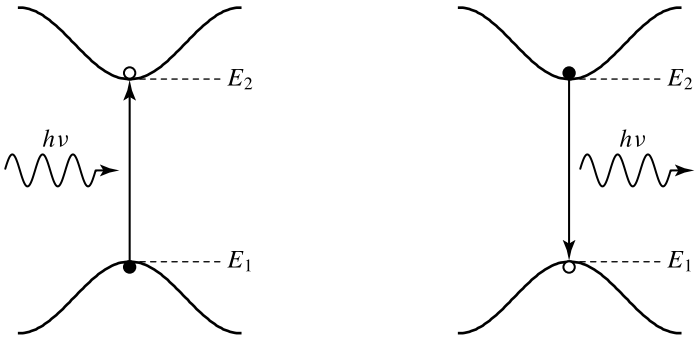


Figure 13.3 Direct optical transitions in a direct-gap semiconductor.

of momentum in the interaction among particles. The vector quantity $\hbar\mathbf{k}$ is known as the *crystal momentum* of an electron in a band state of wavevector $\mathbf{k}$, but it is not really the momentum of the electron in the usual sense.

As illustrated in Fig. 13.3, the conditions for direct transition can be satisfied in a direct-gap semiconductor for transitions between states near the conduction- and valence-band edges. Band-to-band absorption in a direct-gap semiconductor normally occurs through a direct absorption process for a photon energy of $h\nu \geq E_g$. As a result, the absorption spectrum of a direct-gap semiconductor shows a sharp edge at $h\nu = E_g$ and rises quickly when the photon energy increases above the bandgap. Band-to-band recombination in a direct-gap semiconductor can also take place through a direct recombination process with emission of a photon of an energy of $h\nu \geq E_g$. Because the conditions in (13.9) for a direct transition process can be easily satisfied, the probability of radiative recombination in a direct-gap semiconductor is very high, leading to a short radiative lifetime and a high radiative efficiency.

In an indirect-gap semiconductor, the requirement of conservation of wavevector for direct transition cannot be satisfied for transitions between states near the band edges, as illustrated in Fig. 13.4(*a*). An indirect optical transition between two such states is possible, however, if the process is assisted by the absorption or emission of a phonon of an energy $\hbar\Omega$ and a wavevector $\mathbf{K}$ that satisfy the following conditions:

$$E_2 - E_1 = h\nu \pm \hbar\Omega \qquad \text{and} \qquad \mathbf{k}_2 = \mathbf{k}_1 + \mathbf{k}_{\text{photon}} \pm \mathbf{K} \approx \mathbf{k}_1 \pm \mathbf{K}. \tag{13.10}$$

An indirect transition process has a much lower probability than a direct transition process because, in comparison to a direct process that involves only a photon and an electron, an indirect process is a high-order process that requires the participation of a phonon.

Near the band edges of an indirect-gap semiconductor, both optical absorption and radiative carrier recombination can take place only through an indirect transition process. Direct optical transition between a state near the valence-band edge and a state high above the conduction-band edge and that between a state well below the valence-band edge and a state near the conduction-band edge are possible in an indirect-gap

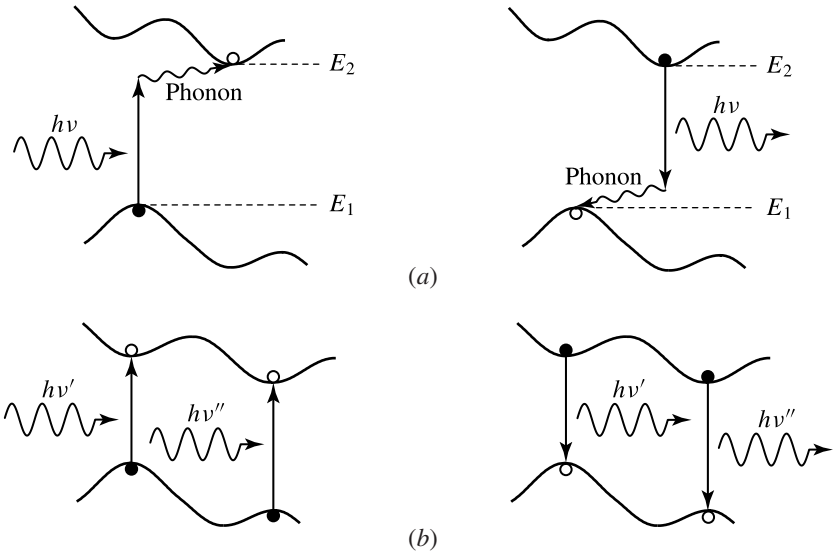


Figure 13.4 (*a*) Indirect optical transitions and (*b*) direct optical transitions in an indirect-gap semiconductor.

semiconductor, as illustrated in Fig. 13.4(*b*). If the photon energy is sufficiently larger than E_g, direct optical absorption occurs readily in an indirect-gap semiconductor. Therefore, the absorption coefficient of an indirect-gap semiconductor first increases gradually as the photon energy increases just above the bandgap where only indirect absorption takes place. It then has a sharp increase when the photon energy reaches the threshold for direct absorption to occur. Carrier recombination through a direct optical transition process in an indirect-gap semiconductor is highly unlikely, however, because a state high above the conduction-band edge is normally not occupied by an electron while a state deep down below the valence-band edge is generally occupied. Consequently, band-to-band carrier recombination in an indirect-gap semiconductor is generally an indirect process, which has a low radiative recombination probability and a long radiative lifetime. Because of this long radiative lifetime, competing nonradiative recombination processes can easily take place, resulting in a low radiative efficiency for an indirect-gap semiconductor. This is the reason why the important semiconductors Si and Ge are not useful for fabricating lasers and LEDs, though they are good for making photodetectors.

Direct transition rates

To evaluate the transition rates of direct band-to-band optical transitions in a semiconductor, a few conditions imposed by the band structure have to be considered. As a result, the formulation for the direct transition rates of a semiconductor is different from the transition rates obtained in Section 10.1 for individual atoms or molecules.

First, the conditions in (13.9) that dictate the conservation of energy and momentum for a direct optical transition have to be satisfied. By taking $|\mathbf{k}_2| \approx |\mathbf{k}_1| = k$ for momentum conservation and considering the fact that the electron and hole energies vary with the value of k quadratically near the band edges, we have

$$E_2 = E_c + \frac{\hbar^2 k^2}{2m_e^*}, \tag{13.11}$$

$$E_1 = E_v - \frac{\hbar^2 k^2}{2m_h^*}. \tag{13.12}$$

By applying the condition $E_2 - E_1 = h\nu$ for energy conservation and using the relation of $E_c - E_v = E_g$, (13.11) and (13.12) can be used to find E_2 and E_1 in terms of the photon energy as (see Problem 13.2.3)

$$E_2 = E_c + \frac{m_r^*}{m_e^*}(h\nu - E_g), \tag{13.13}$$

$$E_1 = E_v - \frac{m_r^*}{m_h^*}(h\nu - E_g), \tag{13.14}$$

where m_r^* is the *reduced effective mass* defined as

$$m_r^* = \frac{m_e^* m_h^*}{m_e^* + m_h^*}. \tag{13.15}$$

To satisfy the conservation of energy and momentum simultaneously, a band-to-band optical transition associated with a photon of an energy $h\nu$ can occur only between a conduction-band state of energy E_2 given by (13.13) and a valence-band state of energy E_1 given by (13.14).

Next, we have to consider the density of states in the conduction and valence bands that satisfy the conservation of energy and momentum for the optical transition. This can be done by considering the states in the conduction and valence bands that satisfy (13.13) and (13.14), respectively. The *density of states for band-to-band optical transitions* corresponding to optical frequencies in the range from ν to $\nu + \mathrm{d}\nu$ can be evaluated as

$$\rho(\nu)\mathrm{d}\nu = \rho_c(E_2)\mathrm{d}E_2 = -\rho_v(E_1)\mathrm{d}E_1, \tag{13.16}$$

where the minus sign in front of $\rho_v(E_1)$ is introduced because the sign of $\mathrm{d}E_1$ is opposite to that of $\mathrm{d}\nu$, as can be seen in (13.14). Using (12.16) for $\rho_c(E)\mathrm{d}E$ and (13.13) for E_2, or (12.16) for $\rho_v(E)\mathrm{d}E$ and (13.14) for E_1, we find that (see Problem 13.2.3)

$$\rho(\nu)\mathrm{d}\nu = \frac{4\pi(2m_r^*)^{3/2}}{h^2}(h\nu - E_g)^{1/2}\mathrm{d}\nu \qquad (\mathrm{m}^{-3}) \tag{13.17}$$

for direct band-to-band optical transitions associated with absorption or emission of photons in the frequency range between ν and $\nu + \mathrm{d}\nu$.

Finally, the probabilities of occupancy for the states that are involved in an optical transition have to be considered. For an optical transition from a valence-band state $|1\rangle$

of energy E_1 to a conduction-band state $|2\rangle$ of energy E_2, state $|1\rangle$ has to be occupied and state $|2\rangle$ has to be empty before the transition takes place. Therefore, the probability of the transition associated with optical absorption is $f_v(E_1)(1 - f_c(E_2))$. For optical emission, the probability is $f_c(E_2)(1 - f_v(E_1))$ because it involves the transition from an occupied conduction-band state $|2\rangle$ to an empty valence-band state $|1\rangle$.

Based on the above discussions, we can easily write down the transition rates for direct band-to-band optical transitions in a semiconductor by following the line of reasoning employed to derive the atomic transition rates in Section 10.1. In the presence of an optical radiation field that has a spectral energy density of $u(\nu)$, the induced transition rates *per unit volume* of the semiconductor in the spectral range between ν and $\nu + d\nu$ are

$$R_a(\nu)d\nu = B_{12}u(\nu)f_v(E_1)[1 - f_c(E_2)]\rho(\nu)d\nu \qquad (m^{-3}\ s^{-1}) \tag{13.18}$$

for optical absorption associated with upward transitions of electrons from the valence band to the conduction band and

$$R_e(\nu)d\nu = B_{21}u(\nu)f_c(E_2)[1 - f_v(E_1)]\rho(\nu)d\nu \qquad (m^{-3}\ s^{-1}) \tag{13.19}$$

for stimulated emission resulting from downward transitions of electrons from the conduction band to the valence band. The spontaneous emission rate is independent of $u(\nu)$ and can be expressed as

$$R_{sp}(\nu)d\nu = A_{21}f_c(E_2)[1 - f_v(E_1)]\rho(\nu)d\nu \qquad (m^{-3}\ s^{-1}). \tag{13.20}$$

The A and B coefficients in (13.18)–(13.20) are the Einstein A and B coefficients, which are evaluated in the following through a procedure similar to that used in Section 10.1.

We consider a semiconductor in thermal equilibrium at a temperature T with blackbody radiation, which has a spectral energy density of $u(\nu)$ given by (10.20). In thermal equilibrium, the electrons in both conduction and valence bands follow the same distribution function $f(E)$ given by (12.1) that is characterized by a single Fermi level, E_F. Therefore, $f_c(E_2) = f(E_2)$ and $f_v(E_1) = f(E_1)$. For the semiconductor to maintain thermal equilibrium with blackbody radiation, the total absorption rate in any given frequency range has to be equal to the total emission rate in the same frequency range:

$$R_a(\nu)d\nu = R_e(\nu)d\nu + R_{sp}(\nu)d\nu. \tag{13.21}$$

By substituting (13.18), (13.19), and (13.20) in (13.21) and using the fact that $f_c(E_2) = f(E_2)$ and $f_v(E_1) = f(E_1)$ in this situation, we have

$$\frac{B_{12}u(\nu)}{B_{21}u(\nu) + A_{21}} = \frac{f(E_2)[1 - f(E_1)]}{f(E_1)[1 - f(E_2)]} = e^{-(E_2 - E_1)/k_B T} = e^{-h\nu/k_B T}. \tag{13.22}$$

This result can be rearranged to yield the following relation:

$$u(\nu) = \frac{A_{21}/B_{21}}{(B_{12}/B_{21})e^{h\nu/k_B T} - 1}. \tag{13.23}$$

Similarly to what is done in (10.30), the coefficient A_{21} can be expressed in terms of a spontaneous time constant as

$$A_{21} = \frac{1}{\tau_{sp}}. \tag{13.24}$$

Note, however, that τ_{sp} is not the same as the radiative carrier lifetime τ_{rad} or the total spontaneous carrier recombination lifetime τ_s defined in the preceding section. The physical meaning and the characteristics of τ_{sp} are further discussed in Section 13.4. By identifying $u(\nu)$ in (13.23) with the spectral energy density given in (10.20) for blackbody radiation, we find that

$$B_{12} = B_{21} = \frac{c^3}{8\pi n^3 h\nu^3 \tau_{sp}}, \tag{13.25}$$

where n is the refractive index of the semiconductor. Though the relation in (13.25) for the coefficients B_{12} and B_{21} was obtained by considering the interaction of a semiconductor with blackbody radiation in thermal equilibrium, it is an intrinsic property of the semiconductor material that is independent of the source and characteristics of the optical radiation.

Using the results obtained above and the relation given by (10.15) between the spectral intensity $I(\nu)$ and the spectral energy density $u(\nu)$ of an optical field at a frequency ν, we obtain the following relations for direct band-to-band optical transitions:

$$\begin{aligned} R_a(\nu) &= \frac{c^3}{8\pi n^3 h\nu^3 \tau_{sp}} u(\nu) f_v(E_1)[1 - f_c(E_2)]\rho(\nu) \\ &= \frac{c^2}{8\pi n^2 h\nu^3 \tau_{sp}} I(\nu) f_v(E_1)[1 - f_c(E_2)]\rho(\nu) \qquad (\mathrm{m}^{-3}) \end{aligned} \tag{13.26}$$

for optical absorption,

$$\begin{aligned} R_e(\nu) &= \frac{c^3}{8\pi n^3 h\nu^3 \tau_{sp}} u(\nu) f_c(E_2)[1 - f_v(E_1)]\rho(\nu) \\ &= \frac{c^2}{8\pi n^2 h\nu^3 \tau_{sp}} I(\nu) f_c(E_2)[1 - f_v(E_1)]\rho(\nu) \qquad (\mathrm{m}^{-3}) \end{aligned} \tag{13.27}$$

for stimulated emission, and

$$R_{sp}(\nu) = \frac{1}{\tau_{sp}} f_c(E_2)[1 - f_v(E_1)]\rho(\nu) \qquad (\mathrm{m}^{-3}) \tag{13.28}$$

for spontaneous emission. The validity of these relations is quite general. When the carriers in the conduction and valence bands of a semiconductor are in thermal equilibrium, they are governed by the same Fermi–Dirac distribution with f_c and f_v characterized by the same Fermi level E_F. When the carriers are in quasi-equilibrium, f_c and f_v are characterized by different quasi-Fermi levels, E_{Fc} and E_{Fv}, respectively. In either situation, the relations in (13.26)–(13.28) are valid. The relations in (13.26)

and (13.27) for the transitions induced by a radiation field are also valid regardless of whether the interaction optical field is a coherent field like a laser field or an incoherent field like blackbody radiation.

Note that $R_a(\nu)d\nu$, $R_e(\nu)d\nu$, and $R_{sp}(\nu)d\nu$ given in (13.18), (13.19), and (13.20), respectively, represent the transition rates per unit volume of a semiconductor and thus have units of cubic meters per second. In contrast, $W_{12}(\nu)d\nu$, $W_{21}(\nu)d\nu$, and $W_{sp}(\nu)d\nu$ given in (10.17), (10.18), and (10.19), respectively, represent the transition rates of a single atom or molecule and are measured in units per second. As mentioned at the beginning of this section, it is not possible to consider the transition rates of each individual electron separately from other electrons in the band structure of a semiconductor. Consequently, the transition rates obtained in this section for the band-to-band transitions in a semiconductor already account for the distribution and density of the carriers in the energy bands of a semiconductor. The concept of transition cross section defined in Section 10.1 is not directly applicable to the band-to-band transitions in a semiconductor though an equivalent gain cross section can be obtained, as defined later in (13.40). Instead, $R_a(\nu)$, $R_e(\nu)$, and $R_{sp}(\nu)$ for the band-to-band transitions in a semiconductor are respectively equivalent to $N_1W_{12}(\nu)$, $N_2W_{21}(\nu)$, and $N_2W_{sp}(\nu)$ for the transitions between the energy levels of active atoms or molecules in a material.

EXAMPLE 13.2 In this example, we consider direct band-to-band optical transitions in GaAs at $\lambda = 850$ nm wavelength at 300 K. (a) Find the reduced effective mass m_r^* for GaAs. (b) Find the energy levels, E_2 and E_1, for the optical transitions at this wavelength. (c) Calculate the value of the density of states $\rho(\nu)$ for these transitions. (d) By taking $\tau_{sp} = 500$ ps, find the spontaneous emission rate $R_{sp}(\nu)$ for intrinsic GaAs at this optical wavelength.

Solution (a) From Table 12.2, we have $m_e^* = 0.067m_0$ and $m_h^* = 0.52m_0$ for GaAs. We then find from (13.15) that

$$m_r^* = \frac{0.067 \times 0.52}{0.067 + 0.52} m_0 = 0.0594m_0.$$

(b) The photon energy for $\lambda = 850$ nm $= 0.85$ μm is

$$h\nu = \frac{1.2398}{0.85} \text{ eV} = 1.459 \text{ eV}.$$

At 300 K, the bandgap of GaAs is $E_g = 1.424$ eV. We find by using (13.13) and (13.14) that

$$E_2 = E_c + \frac{0.0594}{0.067} \times (1.459 - 1.424) \text{ eV} = E_c + 31 \text{ meV},$$

$$E_1 = E_v - \frac{0.0594}{0.52} \times (1.459 - 1.424) \text{ eV} = E_v - 4 \text{ meV}.$$

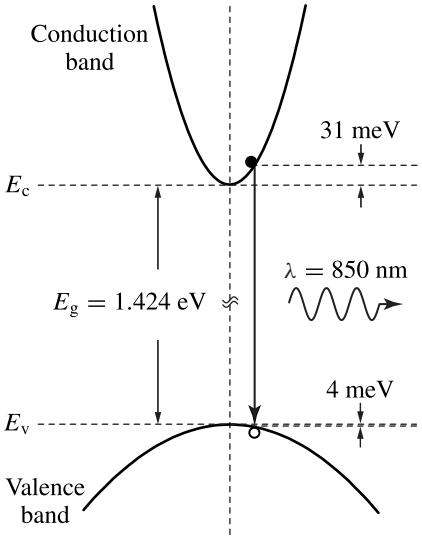


Figure 13.5 Direct band-to-band optical transition at 850 nm optical wavelength in intrinsic GaAs at 300 K. The conduction band is plotted for $m_e^* = 0.067m_0$, and the valence band is plotted for $m_h^* = 0.52m_0$ by combining the heavy- and light-hole bands into one band.

Therefore, the direct band-to-band absorption or emission of a photon at 850 nm wavelength in GaAs at 300 K takes place between a conduction-band state that is located at 31 meV above the conduction-band edge and a valence-band state that is located at 4 meV below the valence-band edge, as shown in Fig. 13.5.

(c) The density of states can be calculated directly using (13.17):

$$\begin{aligned}\rho(\nu) &= \frac{4\pi \times (2 \times 0.0594 \times 9.11 \times 10^{-31})^{3/2}}{(6.626 \times 10^{-34})^2} \\ &\quad \times [(1.459 - 1.424) \times 1.6 \times 10^{-19}]^{1/2} \text{ m}^{-3} \text{ Hz}^{-1} \\ &= 7.63 \times 10^{10} \text{ m}^{-3} \text{ Hz}^{-1}.\end{aligned}$$

(d) To find $R_{sp}(\nu)$ from (13.28), we have to calculate $f_c(E_2)(1 - f_v(E_1))$. It can be calculated by plugging the values of E_2 and E_1 found above into the Fermi–Dirac distribution function given in (12.1) with $E_F = E_{Fi}$ for the intrinsic GaAs found in Example 12.2. Alternatively, we know from Example 12.2 that $(E_c - E_{Fi})/k_BT \approx (E_{Fi} - E_v)/k_BT \approx E_g/(2k_BT) = 27.49$ for GaAs at $T = 300$ K. Because $E_2 > E_c$ and $E_v > E_1$, we have

$$f_c(E_2) \approx e^{(E_{Fi}-E_2)/k_BT} \quad \text{and} \quad 1 - f_v(E_1) \approx e^{(E_1-E_{Fi})/k_BT}.$$

Therefore, for intrinsic GaAs at 300 K, we have

$$f_c(E_2)[1 - f_v(E_1)] \approx e^{(E_1-E_2)/k_BT} = e^{-h\nu/k_BT}. \tag{13.29}$$

Note that this relation is valid only for an unexcited intrinsic semiconductor with $E_g \gg k_B T$ and with E_{Fi} located near the center of its bandgap. It is not valid when the bandgap is small or when the Fermi level lies close to one of the band edges. With this relation and with $\tau_{sp} = 500$ ps, we can calculate $R_{sp}(\nu)$, using (13.28), as

$$R_{sp}(\nu) = \frac{1}{500 \times 10^{-12}} \times e^{-1.459/0.0259} \times 7.63 \times 10^{10}\ \text{m}^{-3} = 5.23 \times 10^{-5}\ \text{m}^{-3}.$$

This is the spontaneous emission rate per unit spectral bandwidth of unexcited intrinsic GaAs in the thermal equilibrium state at 300 K at $h\nu = 1.459$ eV for $\lambda = 850$ nm.

13.3 Optical gain

By following a line of reasoning similar to that used in Section 10.2 while associating $R_a(\nu)$ with $N_1 W_{12}(\nu)$ and $R_e(\nu)$ with $N_2 W_{21}(\nu)$, we can write down the absorption and gain coefficients contributed by direct band-to-band transitions in a semiconductor as

$$\alpha(\nu) = \frac{h\nu}{I(\nu)}[R_a(\nu) - R_e(\nu)] = \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}}[f_v(E_1) - f_c(E_2)]\rho(\nu) \tag{13.30}$$

and

$$g(\nu) = \frac{h\nu}{I(\nu)}[R_e(\nu) - R_a(\nu)] = \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}}[f_c(E_2) - f_v(E_1)]\rho(\nu), \tag{13.31}$$

respectively. By definition, $g(\nu) = -\alpha(\nu)$. The relations in (13.30) and (13.31) are valid for carriers in either an equilibrium state or a quasi-equilibrium state because their validity follows from that of the relations in (13.26) and (13.27).

If the carriers in the conduction and valence bands are in thermal equilibrium, both f_c and f_v are characterized by the same Fermi level E_F. For an intrinsic semiconductor with no impurity doping, this Fermi level is located very close to the middle of the bandgap. For band-to-band absorption to occur, the photon energy must be $h\nu = E_2 - E_1 \geq E_g$, implying that $E_2 - E_F \geq E_g/2$ and $E_F - E_1 \geq E_g/2$. At $T = 300$ K, we have $k_B T = 25.9$ meV, which is at least one order of magnitude smaller than the bandgaps of most semiconductors, except those that have very small bandgaps such as some alloys containing HgSe or HgTe. For an intrinsic semiconductor in thermal equilibrium, $f(E_1) \approx 1$ and $f(E_2) \approx 0$ if the bandgap of the semiconductor is significantly larger than $k_B T$. Then, according to (13.30), its intrinsic absorption spectrum is given by

$$\alpha_0(\nu) = \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}}\rho(\nu). \tag{13.32}$$

Consequently, the gain and absorption spectra of a semiconductor in an equilibrium or a quasi-equilibrium state at a given temperature T can be expressed in terms of its

intrinsic absorption spectrum at the same temperature as

$$g(\nu) = -\alpha(\nu) = \alpha_0(\nu)[f_c(E_2) - f_v(E_1)]. \tag{13.33}$$

This relation is valid for gain and absorption contributed by direct band-to-band transitions under the condition that $h\nu \geq E_g \gg k_B T$. Though it is obtained for an intrinsic semiconductor, it can be generalized to an extrinsic semiconductor doped with impurities by taking $\alpha_0(\nu)$ to be the absorption coefficient measured when the semiconductor is in thermal equilibrium with its environment at a given temperature T. Because both $f_c(E_2)$ and $f_v(E_1)$ have a minimum value of 0 and a maximum value of 1, this relation implies that $-\alpha_0(\nu) \leq g(\nu) \leq \alpha_0(\nu)$ and $-\alpha_0(\nu) \leq \alpha(\nu) \leq \alpha_0(\nu)$ at any frequency in any condition, which can be clearly seen in Fig. 13.7(*a*) discussed below.

Population inversion

The concept of population inversion in a semiconductor cannot be simply defined as that the conduction band is more populated than the valence band because for a positive optical gain coefficient it is neither necessary nor possible to have more electrons in the conduction band than in the valence band. The population of electrons in an energy band is subject to the requirement of Fermi distribution and the availability of energy states in the band structure. Therefore, a practical definition of population inversion in a semiconductor is when the electron and hole concentrations in the semiconductor lead to a positive optical gain coefficient.

Because $\alpha_0(\nu) \geq 0$ for any frequency ν, the sign of $g(\nu)$ given in (13.33) is determined by that of the quantity $f_c(E_2) - f_v(E_1)$. It can be shown that (see Problem 13.3.2)

$$f_c(E_2) - f_v(E_1) = f_c(E_2)[1 - f_v(E_1)]\left[1 - e^{(h\nu - \Delta E_F)/k_B T}\right], \tag{13.34}$$

where $\Delta E_F = E_{Fc} - E_{Fv}$. Because $1 \geq f_c(E_2)(1 - f_v(E_1)) \geq 0$, the sign of the quantity $f_c(E_2) - f_v(E_1)$ is solely determined by the sign of the quantity $h\nu - \Delta E_F$. Therefore, the condition for a positive gain coefficient that $g(\nu) > 0$ at any given optical frequency ν is that the separation between the quasi-Fermi levels be larger than the photon energy at the frequency ν (see Problem 13.3.2):

$$\Delta E_F = E_{Fc} - E_{Fv} > h\nu > E_g. \tag{13.35}$$

This condition dictates the distributions of electrons and holes for a positive optical gain coefficient. It can thus be considered as the *condition for population inversion in a semiconductor*.

As mentioned in Section 12.1, the quasi-Fermi levels completely quantify the electron and hole concentrations in a semiconductor that is maintained in a quasi-equilibrium state by electrical or optical pumping. Furthermore, (12.43) and (12.46) indicate that the product np strongly depends on the value of ΔE_F. Therefore, the condition given in (13.35) determines the electron and hole concentrations needed for a positive gain

coefficient. When a semiconductor is pumped to have a positive gain coefficient, the pumped electron and hole concentrations are normally a few orders of magnitude higher than the intrinsic electron and hole concentrations so that $n \gg n_0$ and $p \gg p_0$. In this situation, we consider the electron and hole concentrations in a semiconductor gain medium to be practically the same as the excess carrier density N defined in (12.55) so that

$$n \approx p \approx N. \tag{13.36}$$

The minimum carrier density required for a gain is known as the *transparency carrier density*, $N_{\rm tr}$. According to (13.35), it is determined by the condition that $\Delta E_{\rm F} = E_{\rm g}$, which implies that $E_{\rm Fc} - E_{\rm c} = E_{\rm Fv} - E_{\rm v}$. By using (12.41), (12.42), and (13.36), we find that the transparency carrier density is given by

$$N_{\rm tr} = N_{\rm c}(T)F_{1/2}(\xi_{\rm tr}) = N_{\rm v}(T)F_{1/2}(-\xi_{\rm tr}), \tag{13.37}$$

where $\xi_{\rm tr} = (E_{\rm Fc} - E_{\rm c})/k_{\rm B}T = (E_{\rm Fv} - E_{\rm v})/k_{\rm B}T$. With known values of $N_{\rm c}$ and $N_{\rm v}$ at a given temperature, the value of $N_{\rm tr}$ can be found by solving this relation to find the parameter $\xi_{\rm tr}$. *In terms of the carrier density, the condition for population inversion in a semiconductor can be regarded as* $N > N_{\rm tr}$.

EXAMPLE 13.3 Find the transparency carrier density for GaAs at 300 K. When GaAs is injected with this concentration of electron–hole pairs, where are its quasi-Fermi levels located?

Solution To find $N_{\rm tr}$, we have to find the value of $\xi_{\rm tr}$ first by solving the second relation in (13.37). From (12.24), $N_{\rm c} \propto (m_{\rm e}^*)^{3/2}$ and $N_{\rm v} \propto (m_{\rm h}^*)^{3/2}$. Therefore, we find the following relation from (13.37):

$$\frac{F_{1/2}(\xi_{\rm tr})}{F_{1/2}(-\xi_{\rm tr})} = \left(\frac{m_{\rm h}^*}{m_{\rm e}^*}\right)^{3/2}. \tag{13.38}$$

For GaAs, $m_{\rm e}^* = 0.067m_0$ and $m_{\rm h}^* = 0.52m_0$. Then,

$$\frac{F_{1/2}(\xi_{\rm tr})}{F_{1/2}(-\xi_{\rm tr})} = \left(\frac{0.52}{0.067}\right)^{3/2} = 21.62.$$

The solution of this relation is found from the value of $F_{1/2}(\xi)$ as a function of ξ plotted in Fig. 12.3 to be $\xi_{\rm tr} = 1.99$ for $F_{1/2}(\xi_{\rm tr}) = F_{1/2}(1.99) \approx 2.81$ and $F_{1/2}(-\xi_{\rm tr}) = F_{1/2}(-1.99) \approx 0.13$. From Example 12.2, we have $N_{\rm c} = 4.35 \times 10^{23}$ m^{-3} and $N_{\rm v} = 9.41 \times 10^{24}$ m^{-3} for GaAs at 300 K. We then find from (13.37) that the transparency carrier density for GaAs at 300 K is

$$N_{\rm tr} = 2.81 \times 4.35 \times 10^{23}\ {\rm m}^{-3} = 0.13 \times 9.41 \times 10^{24}\ {\rm m}^{-3} = 1.22 \times 10^{24}\ {\rm m}^{-3}.$$

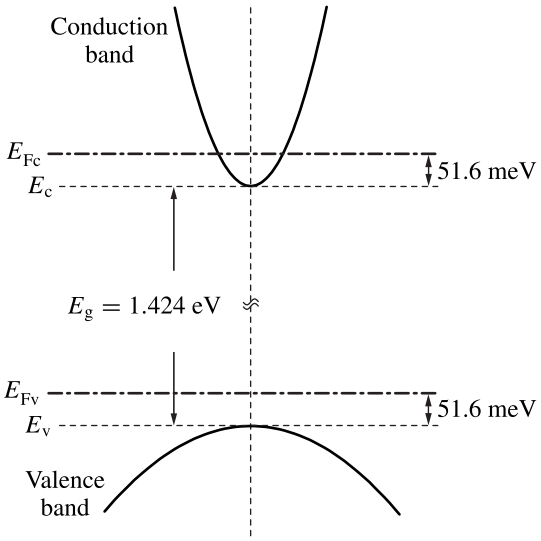


Figure 13.6 Quasi-Fermi levels of GaAs at 300 K at transparency with an injected electron–hole pair concentration of the transparency carrier density. The conduction band is plotted for $m_e^* = 0.067m_0$, and the valence band is plotted for $m_h^* = 0.52m_0$ by combining the heavy- and light-hole bands into one band.

Because $\xi_{tr} = 1.99$ at this injected carrier concentration, the quasi-Fermi levels are located at

$$E_{Fc} = E_c + 1.99k_BT = E_c + 51.6 \text{ meV},$$
$$E_{Fv} = E_v + 1.99k_BT = E_v + 51.6 \text{ meV}.$$

Though $E_{Fc} - E_{Fv} = E_g$ for a semiconductor at transparency with $N = N_{tr}$, we find that E_{Fc} and E_{Fv} do not respectively lie at the conduction-band and valence-band edges due to the fact that $m_e^* \neq m_h^*$. Because $m_h^* > m_e^*$ for GaAs, E_{Fc} lies above the conduction-band edge and E_{Fv} lies above the valence-band edge while they are subject to the condition $E_{Fc} - E_{Fv} = E_g$, as shown in Fig. 13.6.

Carrier dependence of gain

Both the spectrum and the magnitude of the optical gain coefficient in a semiconductor are a function of the excess carrier density N. To find the carrier dependence of the gain spectrum, one starts with a given value of $N \approx n \approx p$ to find the corresponding values of the quasi-Fermi levels E_{Fc} and E_{Fv} from (12.41) and (12.42) through the following relation:

$$N = N_c(T)F_{1/2}(\xi_c) = N_v(T)F_{1/2}(\xi_v), \tag{13.39}$$

where $\xi_c = (E_{Fc} - E_c)/k_BT$ and $\xi_v = (E_v - E_{Fv})/k_BT$. Note that $\xi_c = -\xi_v = \xi_{tr}$ in the case when $N = N_{tr}$, as seen in (13.37). In general, however, $\xi_c \neq -\xi_v$ when

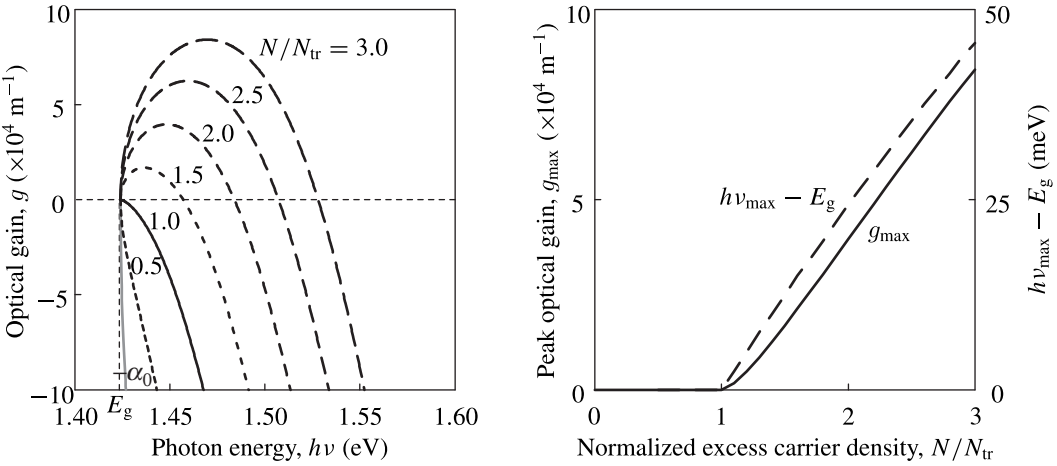


Figure 13.7 (*a*) Gain and absorption spectra of GaAs as a function of photon energy at various levels of normalized excess carrier density, N/N_{tr}. Also shown for comparison is the intrinsic absorption spectrum, α_0, of the semiconductor. (*b*) Peak optical gain coefficient (solid curve) and gain-peak photon energy (dashed curve) as a function of carrier density.

$N \neq N_{tr}$. Therefore, the values of ξ_c and ξ_v have to be found separately from (13.39) in order to find E_{Fc} and E_{Fv}. According to (13.35), we find that $E_{Fc} - E_{Fv} > E_g$ when $N > N_{tr}$, but $E_{Fc} - E_{Fv} < E_g$ when $N < N_{tr}$. Then, $g(\nu)$ as a function of the optical frequency ν can be found using (13.31) or (13.33), and $\alpha(\nu)$ can be similarly calculated. Figure 13.7(*a*) shows the gain and absorption spectra of GaAs as a function of photon energy at various levels of excess carrier density N.

It can be seen that an optical gain associated with direct band-to-band transitions occurs in a range of photon energies larger than the bandgap only when $N > N_{tr}$. Both the spectral range and the peak value of the optical gain coefficient increase with the carrier density. A semiconductor gain medium has a very broad gain bandwidth. In typical operating conditions of semiconductor lasers and amplifiers, the gain bandwidth is in the range of $2k_BT$ to $4k_BT$ in photon energy. At room temperature, this gain bandwidth is 50–100 meV in photon energy, corresponding to a frequency range of 12–24 THz. In most cases, the peak gain coefficient increases almost linearly with carrier density, as shown in Fig. 13.7(*b*). Therefore, we can express the peak value of the optical gain coefficient approximately as

$$g_{max}(N) = \sigma(N - N_{tr}), \tag{13.40}$$

where the coefficient σ is an equivalent *gain cross section* similar to the transition cross section described in Section 10.1. The value of this gain coefficient depends on the composition of the semiconductor material and the operating temperature. It is typically in the range of 1 to 5×10^{-20} m^2 for the common semiconductor laser materials of GaAs and InGaAsP at room temperature. The value of N_{tr} is on the

order of 10^{24} m^{-3} for such semiconductors. Note that, as can be seen in Fig. 13.7(a), the peak of the gain spectrum does not occur at a fixed frequency but moves to a higher frequency as the carrier density is increased. Therefore, $g_{\max}$ as expressed in (13.40) represents the gain coefficient at different optical frequencies as the value of N varies. The linear relation in (13.40) is normally a very good approximation, but it does not hold strictly. In particular, as the carrier density increases to a certain level, the peak gain coefficient tends to increase less than linearly with carrier density.

EXAMPLE 13.4 A GaAs sample at 300 K is injected with excess electron–hole pairs of a concentration $N = 2.83 \times 10^{24}$ m^{-3}. Take $\tau_{sp} = 500$ ps for GaAs at 300 K. (a) Find the sample's quasi-Fermi levels. (b) Find its optical gain coefficient at 850 nm wavelength. The refractive index of GaAs at 850 nm is $n = 3.65$. (c) Use this gain coefficient to calculate an equivalent gain cross section. At this injection level, the gain peak occurs very close to, though not exactly at, 850 nm (see Problem 13.3.4).

Solution (a) By using $N_c = 4.35 \times 10^{23}$ m^{-3} and $N_v = 9.41 \times 10^{24}$ m^{-3} found in Example 12.2 for GaAs at 300 K, we find that (13.39) leads to

$$F_{1/2}(\xi_c) = \frac{N}{N_c} = 6.51 \quad \text{and} \quad F_{1/2}(\xi_v) = \frac{N}{N_v} = 0.30$$

for $N = 2.83 \times 10^{24}$ m^{-3}. From the value of $F_{1/2}(\xi)$ as a function of ξ plotted in Fig. 12.3, we find that $\xi_c = 4$ and $\xi_v = -1.1$. Therefore, $E_{Fc} - E_c = \xi_c k_B T = 4 \times 25.9$ meV $= 103.6$ meV and $E_{Fv} - E_v = -\xi_v k_B T = 1.1 \times 25.9$ meV $= 28.5$ meV. As shown in Fig. 13.8, both quasi-Fermi levels lie above their respective band edges with E_{Fc} located at 103.6 meV above E_c and E_{Fv} located at 28.5 meV above E_v. We also find that $E_{Fc} - E_{Fv} > E_g$ because $N > N_{tr}$.

(b) According to the results obtained in Example 13.2(b), the optical transition at $\lambda = 850$ nm takes place between $E_2 = E_c + 31$ meV and $E_1 = E_c - 4$ meV. Using the values of E_{Fc} and E_{Fv} found above, we find that $E_2 - E_{Fc} = -72.6$ meV and $E_1 - E_{Fv} = -32.5$ meV. These relations are shown in Fig. 13.8. We then find that

$$f_c(E_2) = \frac{1}{e^{(E_2 - E_{Fc})/k_B T} + 1} = \frac{1}{e^{-72.6/25.9} + 1} = 0.9428,$$
$$f_v(E_1) = \frac{1}{e^{(E_1 - E_{Fv})/k_B T} + 1} = \frac{1}{e^{-32.5/25.9} + 1} = 0.7781.$$

We have found from Example 13.2 that $\rho(\nu) = 7.63 \times 10^{10}$ m^{-3} Hz^{-1} for GaAs at 850 nm. By using (13.31), the optical gain coefficient at 850 nm can then be

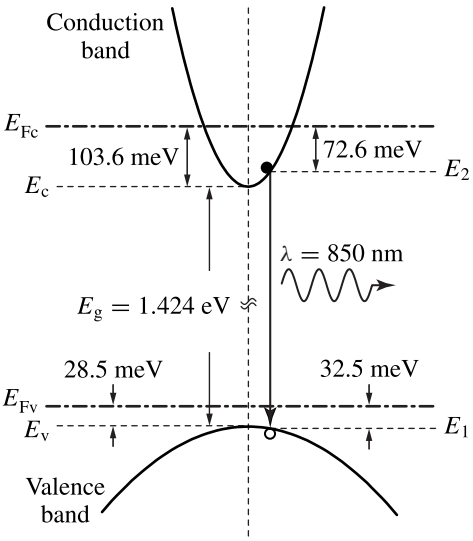


Figure 13.8 Quasi-Fermi levels of GaAs at 300 K with an injected electron–hole pair concentration of $N = 2.83 \times 10^{24}$ m^{-3}. The optical transition for 850 nm wavelength takes place between E_2 and E_1, which lie below E_{Fc} and E_{Fv}, respectively. The conduction band is plotted for $m_e^* = 0.067m_0$, and the valence band is plotted for $m_h^* = 0.52m_0$ by combining the heavy- and light-hole bands into one band.

found:

$$\begin{aligned} g(\nu) &= \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}} [f_c(E_2) - f_v(E_1)]\rho(\nu) \\ &= \frac{\lambda^2}{8\pi n^2 \tau_{sp}} [f_c(E_2) - f_v(E_1)]\rho(\nu) \\ &= \frac{(850 \times 10^{-9})^2}{8\pi \times 3.65^2 \times 500 \times 10^{-12}} \times (0.9428 - 0.7781) \times 7.63 \times 10^{10} \text{ m}^{-1} \\ &= 5.42 \times 10^4 \text{ m}^{-1}. \end{aligned}$$

(c) By taking the value of $N_{tr} = 1.22 \times 10^{24}$ m^{-3} found in Example 13.3 and using (13.40), we find that

$$\sigma = \frac{g}{N - N_{tr}} = \frac{5.42 \times 10^4}{2.83 \times 10^{24} - 1.22 \times 10^{24}} \text{ m}^2 = 3.37 \times 10^{-20} \text{ m}^2.$$

13.4 Spontaneous emission

The spontaneous emission spectrum of a semiconductor can be explicitly related to the absorption and gain spectra of the semiconductor. By using (13.32) to eliminate $\rho(\nu)$

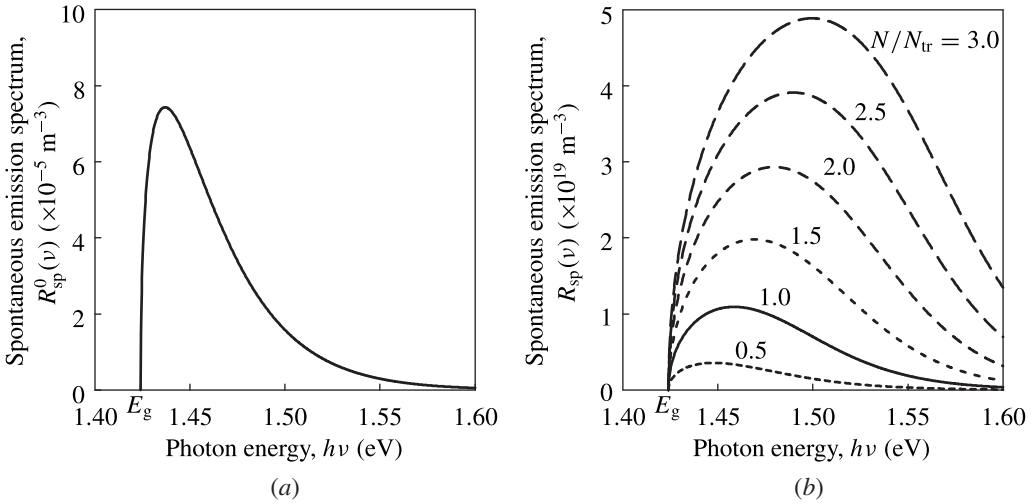


Figure 13.9 Spontaneous emission spectra of GaAs (*a*) in thermal equilibrium and (*b*) at various levels of normalized excess carrier density N/N_{tr}. Note that there is a difference of 24 orders of magnitude between the vertical scales of (*a*) and (*b*). These spectra are to be compared with the gain and absorption spectra shown in Fig. 13.7(*a*).

in (13.28), we find

$$R_{sp}(\nu) = \frac{8\pi n^2 \nu^2}{c^2} \alpha_0(\nu) f_c(E_2)[1 - f_v(E_1)]. \tag{13.41}$$

Using (13.33), (13.34), and (13.41), we can express the spontaneous emission spectrum $R_{sp}(\nu)$ in terms of the absorption spectrum $\alpha(\nu)$ and the gain spectrum $g(\nu)$ as follows:

$$R_{sp}(\nu) = \frac{8\pi n^2 \nu^2}{c^2} \frac{\alpha(\nu)}{e^{(h\nu - \Delta E_F)/k_B T} - 1} = \frac{8\pi n^2 \nu^2}{c^2} \frac{g(\nu)}{1 - e^{(h\nu - \Delta E_F)/k_B T}}. \tag{13.42}$$

In the case when $N < N_{tr}$, there is no optical gain but only absorption. In this situation, $\alpha(\nu) > 0$ and $R_{sp}(\nu)$ has positive values for photon energies larger than the bandgap because $h\nu > E_g > \Delta E_F$ when $N < N_{tr}$. In the case when $N > N_{tr}$, the semiconductor has an optical gain with $g(\nu) > 0$ and $\Delta E_F > h\nu > E_g$. From (13.42), we find that $R_{sp}(\nu)$ again has positive values. Therefore, as should be intuitively expected, $R_{sp}(\nu) \geq 0$ for all frequencies no matter whether the semiconductor has a positive gain or not. Figure 13.9 shows the spontaneous emission spectra of GaAs (*a*) in thermal equilibrium and (*b*) at various levels of excess carrier density.

For a semiconductor in thermal equilibrium, $\Delta E_F = 0$ and $\alpha(\nu) = \alpha_0(\nu)$. Then, its spontaneous emission spectrum is given by

$$R_{sp}^0(\nu) = \frac{8\pi n^2 \nu^2}{c^2} \frac{\alpha_0(\nu)}{e^{h\nu/k_B T} - 1}. \tag{13.43}$$

This relation is known as the *van Roosbroeck–Shockley* relation. Though we have obtained this relation by considering band-to-band transitions, it is a general relation that can be obtained by considering the equilibrium between the spontaneous emission of the semiconductor and the absorption of the surrounding blackbody radiation by the semiconductor. Consequently, it is generally valid for any transitions that contribute to $\alpha_0(\nu)$.

If we consider only band-to-band transitions, then (13.32) can be used to obtain

$$R_{\rm sp}^0(\nu) = \frac{8\pi n^2 \nu^2}{c^2} \frac{\alpha_0(\nu)}{e^{h\nu/k_{\rm B}T} - 1} = \frac{1}{\tau_{\rm sp}} \frac{\rho(\nu)}{e^{h\nu/k_{\rm B}T} - 1}. \tag{13.44}$$

Thus, the total band-to-band spontaneous recombination rate of electron–hole pairs in the thermal equilibrium state is (see Problem 13.4.1)

$$\begin{aligned} R_{\rm sp}^0 &= \int_0^\infty R_{\rm sp}^0(\nu) {\rm d}\nu = \frac{8\pi n^2}{c^2} \int_0^\infty \frac{\nu^2 \alpha_0(\nu)}{e^{h\nu/k_{\rm B}T} - 1} {\rm d}\nu \\ &\approx \frac{2}{\tau_{\rm sp}} \left(\frac{2\pi m_{\rm r}^* k_{\rm B} T}{h^2} \right)^{3/2} e^{-E_{\rm g}/k_{\rm B}T}, \end{aligned} \tag{13.45}$$

where the functional form of $\rho(\nu)$ given in (13.17) is used in carrying out the last integration and the fact that $E_{\rm g} \gg k_{\rm B}T$ for most semiconductors of interest at room temperature is taken. Accordingly, we find that $\tau_{\rm sp}$ is related to the absorption coefficient $\alpha_0(\nu)$ by

$$\frac{1}{\tau_{\rm sp}} = \frac{4\pi n^2}{c^2} \left(\frac{2\pi m_{\rm r}^* k_{\rm B} T}{h^2} \right)^{-3/2} e^{E_{\rm g}/k_{\rm B}T} \int_0^\infty \frac{\nu^2 \alpha_0(\nu)}{e^{h\nu/k_{\rm B}T} - 1} {\rm d}\nu. \tag{13.46}$$

This relation is analogous to that given in (10.45) for atomic transitions.

In thermal equilibrium, the total recombination rate has to be balanced by the total generation rate. Therefore, if only bimolecular band-to-band transitions are considered, we also have

$$R_{\rm sp}^0 = G_0 = B n_0 p_0. \tag{13.47}$$

In a nondegenerate semiconductor where the law of mass action given in (12.31) is valid, we can use (13.45) and (13.47) to express the bimolecular recombination coefficient B in terms of $\tau_{\rm sp}$ as follows (see Problem 13.4.1):

$$B = \frac{1}{2\tau_{\rm sp}} \left(\frac{2\pi k_{\rm B} T}{h^2} \right)^{-3/2} (m_{\rm e}^* + m_{\rm h}^*)^{-3/2}. \tag{13.48}$$

We have mentioned in Section 13.2 that $\tau_{\rm sp}$ is not the same as either $\tau_{\rm rad}$ or $\tau_{\rm s}$. This fact can easily be seen by comparing (13.48) with (13.6). Indeed, both $\tau_{\rm rad}$ and $\tau_{\rm s}$ are functions of the excess carrier density N, but $\tau_{\rm sp}$ is an intrinsic parameter of a semiconductor that is independent of the excess carrier density. Clearly from (13.46) and

(13.48), the values of both τ_{sp} and B vary significantly with temperature. In practice, they also depend on other mundane conditions, such as the impurities and defects in a sample. In the case of GaAs at 300 K, for example, the relation from (13.46) yields a theoretical value of $\tau_{sp} \approx 500$ ps, corresponding to $B \approx 1.8 \times 10^{-16}$ m^3 s^{-1}. Experimentally measured values for the B coefficient of GaAs are typically smaller than the theoretical value; they fall in the range of 0.5 to 2×10^{-16} m^3 s^{-1} at room temperature, for corresponding values of τ_{sp} in the range between 2 ns and 500 ps. In contrast, τ_{rad} can be anywhere from 500 ps to the order of seconds, depending on the excess carrier density and the amount of impurities in the material. From these discussions, it can be clearly seen that τ_{sp} is a parameter that reflects the strength of the coupling between the electrons making optical transitions and the optical radiation emitted or absorbed by electrons. The smaller the value of τ_{sp}, the stronger the coupling is and the more efficient the interaction between the electrons and the radiation is.

EXAMPLE 13.5 Calculate the value of the total spontaneous emission rate R_{sp}^0 and that of the bimolecular recombination coefficient B for GaAs at 300 K with $\tau_{sp} = 500$ ps.

Solution From Example 13.2(a), we know that $m_r^* = 0.0594 m_0$ for GaAs. At 300 K, $E_g = 1.424$ eV for GaAs and $k_B T = 25.9$ meV $= 0.0259$ eV. Using (13.45), the total spontaneous emission rate can be calculated:

$$R_{sp}^0 = \frac{2}{500 \times 10^{-12}} \times \left[\frac{2\pi \times 0.0594 \times 9.11 \times 10^{-31} \times 0.0259 \times 1.6 \times 10^{-19}}{(6.626 \times 10^{-34})^2}\right]^{3/2} \times e^{-1.424/0.0259} \text{ m}^{-3} \text{ s}^{-1}$$

$$= 9.64 \times 10^8 \text{ m}^{-3} \text{ s}^{-1}.$$

Using (13.48) with $m_e^* = 0.067 m_0$ and $m_h^* = 0.52 m_0$ for GaAs, the bimolecular recombination coefficient can be calculated:

$$B = \frac{1}{2 \times 500 \times 10^{-12}} \times \left[\frac{2\pi \times 0.0259 \times 1.6 \times 10^{-19}}{(6.626 \times 10^{-34})^2}\right]^{-3/2} \times \left[(0.067 + 0.52) \times 9.11 \times 10^{-31}\right]^{-3/2} \text{ m}^3 \text{ s}^{-1}$$

$$= 1.77 \times 10^{-16} \text{ m}^3 \text{ s}^{-1}.$$

13.5 Junction structures

A semiconductor junction device can have either a *homostructure* or a *heterostructure*. A basic homostructure simply consists of a p–n homojunction. There are a number of different heterostructures, but the two basic concepts are the *single heterostructure* (SH), which consists of a single heterojunction, and the *double heterostructure* (DH),

which consists of two heterojunctions. When the layer between the junctions of a DH is thin enough, the structure becomes a *quantum well* (QW) because of the quantum size effect in the thin layer.

An electrically pumped LED or semiconductor laser has the basic structure of a semiconductor diode. In normal operation, an LED or semiconductor laser is pumped under forward bias by electric current injection to create an excess concentration of electron–hole pairs in an *active layer*, or an *active region*, where radiative recombination takes place. For this reason, electrically pumped semiconductor lasers are also called *laser diodes, diode lasers*, or *injection lasers*. A practical LED can have either a homostructure or a heterostructure, but all practical semiconductor lasers are made of heterostructures for the significant advantages of semiconductor heterostructures over homostructures in terms of carrier confinement and optical waveguiding. Tight carrier confinement creates a high carrier concentration for a given injection current; tight optical confinement makes stimulated emission efficient. These are two critical issues in lowering the threshold and in improving the efficiency of a semiconductor laser. Because of the many advantages of quantum wells, many semiconductor lasers are *quantum-well lasers* consisting of a single quantum well (SQW) or multiple quantum wells (MQW).

The optical field in an LED or a semiconductor laser can be either horizontally propagating, in a direction parallel to the junction plane, or vertically propagating, in a direction perpendicular to the junction plane. An LED or semiconductor laser can be either *edge emitting* or *surface emitting*. The optical wave in an edge-emitting device propagates horizontally and is emitted from one or two side surfaces of the structure that are perpendicular to the junction plane. The optical wave in a surface-emitting device can propagate either vertically or horizontally in the structure, but it is emitted from a surface that is parallel to the junction interface. Various structures for edge-emitting and surface-emitting devices are discussed in later sections.

Homostructures

The characteristics of a homostructure device under forward bias are shown in Fig. 13.10. Usually both p and n regions of a homostructure device are heavily doped for good conductivity throughout the device. The excess minority electrons on the p side due to electron injection from the n side are distributed over an electron diffusion length of L_e, and the excess minority holes on the n side due to hole injection from the p side are distributed over a hole diffusion length of L_h. In a given semiconductor, $D_e \gg D_h$ because the electron mobility μ_e is generally higher than the hole mobility μ_h. Then, according to (12.114), if the p side is not much more heavily doped than the n side, the injection current is predominantly carried by the electrons injected from the n side to the p side. If the injected minority electrons in the p region and the injected minority holes in the n region have comparable lifetimes, as is normally the case in the operating conditions of LEDs and lasers, the fact that $D_e \gg D_h$ also implies that $L_e \gg L_h$.

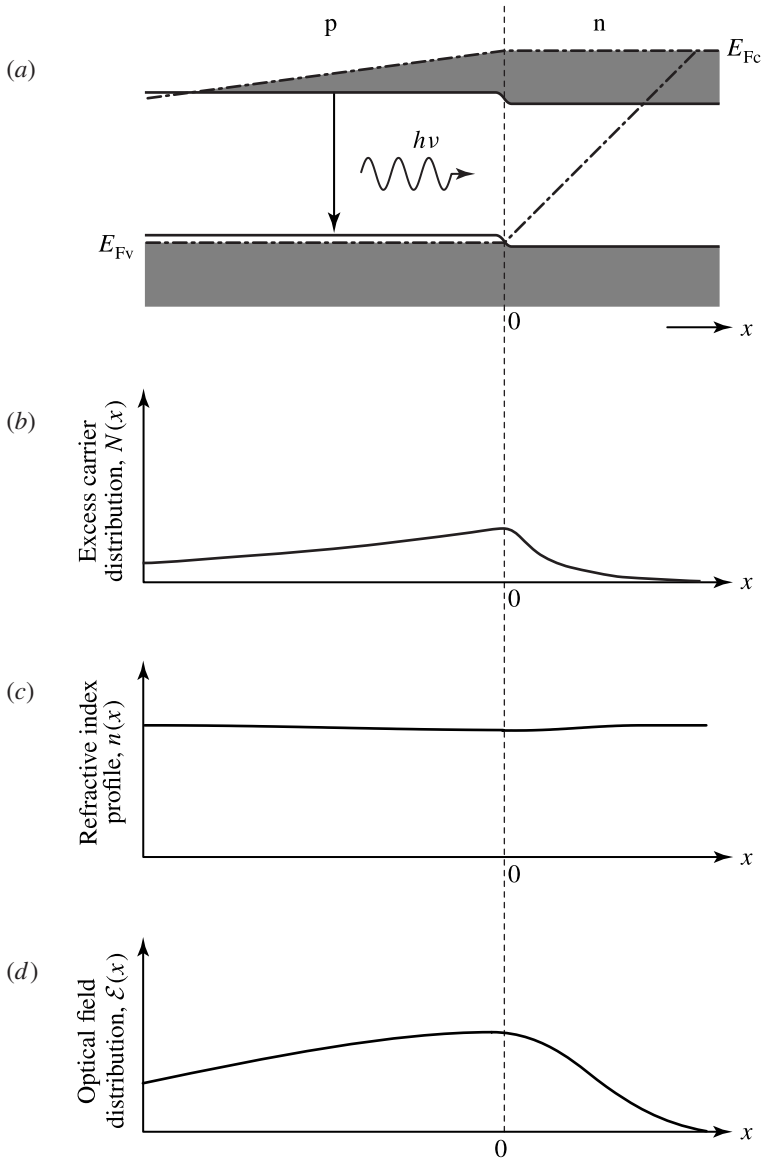


Figure 13.10 (*a*) Energy bands, (*b*) excess carrier distribution, (*c*) refractive index profile, and (*d*) distribution of a horizontally propagating optical field of a p–n homostructure device under forward bias.

Therefore, the total number of excess minority electrons on the p side is much larger than that of excess minority holes on the n side. As expressed in (12.55), (12.107) and (12.108), in normal operating conditions of LEDs and semiconductor lasers, excess electrons and holes appear as excess electron–hole pairs of equal concentration $N(x)$ at any given location x.

From the above discussions, we find that the excess carriers in a homostructure LED or semiconductor laser are distributed mainly in the p region of the device over an electron diffusion length of L_e from the junction, as illustrated in Fig. 13.10(*b*). For an LED, this region is the active layer where radiative carrier recombination for light emission takes place. For a semiconductor laser, an optical gain exists at the locations where $N(x) > N_{tr}$ within this active layer. In a homostructure device, the transverse distribution of a horizontally propagating optical field and the distribution of the excess carriers overlap. There is slight variation of the refractive index in the vicinity of the junction due to spatial variation of the carrier density, but it has little waveguiding effect on the optical field. Therefore, the optical field in a homostructure device is not well confined in the vertical direction but spreads over a distance that can be as large as a few times L_e. These characteristics are illustrated in Figs. 13.10 (*c*) and (*d*).

Clearly, a homostructure has two major deficiencies:

1. The excess carriers are neither confined nor concentrated but are spread by diffusion. For this reason, the thickness of the active layer in a homostructure is normally on the order of one to a few micrometers, depending on the diffusion length of the electrons.
2. There is no waveguiding mechanism in the structure for optical confinement. It is therefore difficult to control the spatial mode characteristics of a homostructure laser diode.

Both problems can be solved with properly designed heterostructures.

Single heterostructures

Because the distribution of the excess carriers in a homostructure is largely determined by the diffusion of the minority electrons in the p region, it is possible to place a P–p heterojunction in the p region to restrict the diffusion of excess electrons that are injected into the p region across the p–n junction. This additional P–p heterojunction results in a P–p–n SH diode, the characteristics of which under forward bias are illustrated in Fig. 13.11.

The n region of a P–p–n SH is heavily doped so that electron injection into the p region completely dominates hole injection into the n region. As illustrated in Fig. 13.11(*b*), the energy barrier at the P–p heterojunction blocks the diffusion of injected electrons so that they are confined within the narrow-gap p region, which thus defines an active layer populated with excess carriers. The thickness of this active layer can be controlled by the location of the P–p heterojunction with respect to that of the p–n junction.

A heterostructure is fabricated with lattice-matched layers of compound semiconductors that have different compositions. One important feature of such materials is that at a given optical wavelength, the refractive index of a narrow-gap composition is

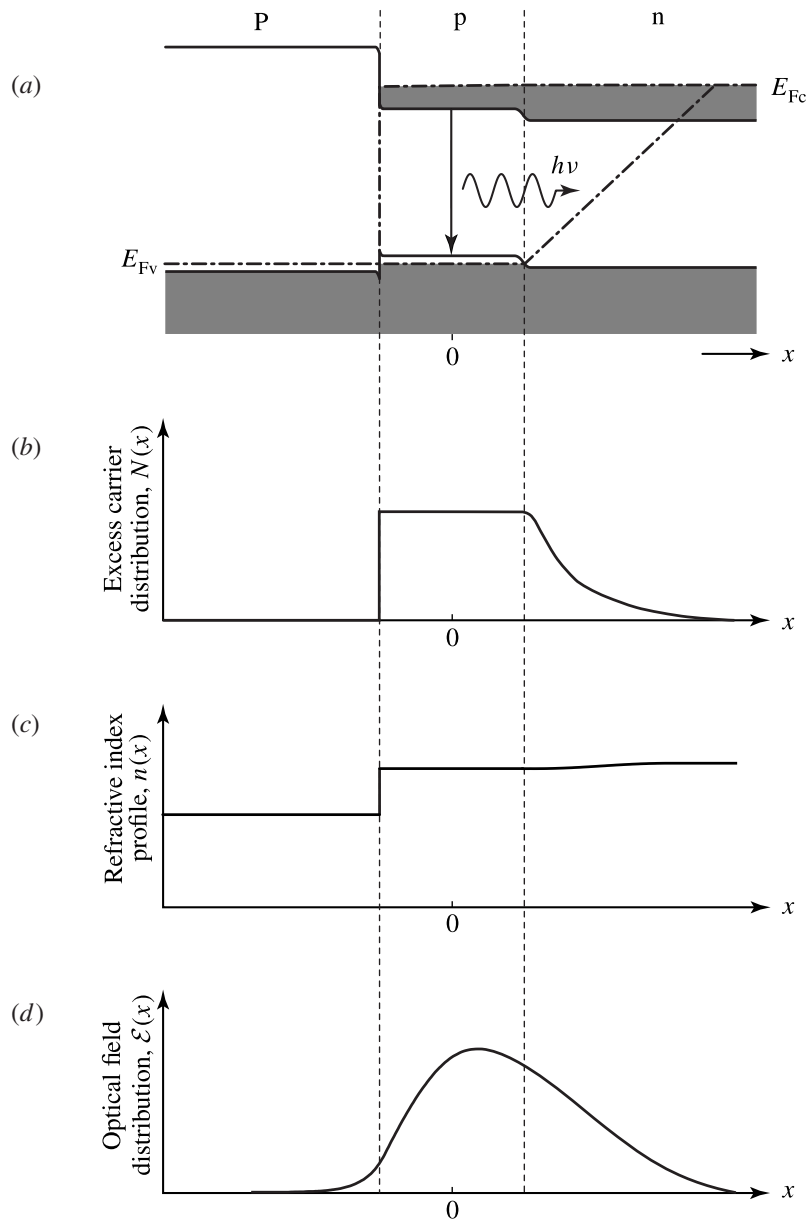


Figure 13.11 (*a*) Energy bands, (*b*) excess carrier distribution, (*c*) refractive index profile, and (*d*) distribution of a horizontally propagating optical field of a P–p–n single heterostructure device under forward bias.

higher than that of a wide-gap composition. Therefore, in addition to carrier confinement, the heterojunction also provides the needed index step, shown in Fig. 13.11(*c*), for optical confinement of a horizontally propagating optical field. However, because no significant index step appears at the p–n junction, optical confinement in the SH geometry is one-sided and is not completely effective, as illustrated in Fig. 13.11(*d*).

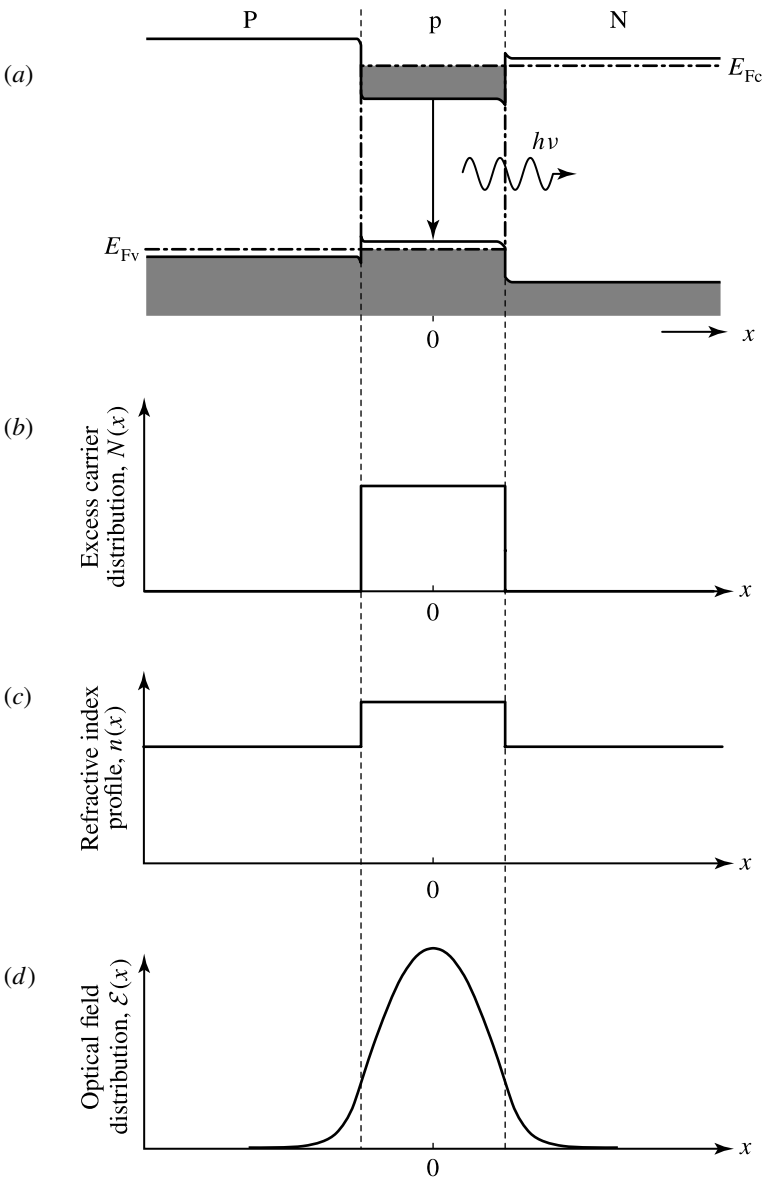


Figure 13.12 (*a*) Energy bands, (*b*) excess carrier distribution, (*c*) refractive index profile, and (*d*) distribution of a horizontally propagating optical field of a P–p–N double heterostructure under forward bias.

Double heterostructures

Very effective carrier and optical confinement can be simultaneously accomplished with DH geometry. A basic DH can be either P–p–N or P–n–N. Figure 13.12 shows the basic characteristics of a P–p–N DH under forward bias.

From the discussions in Section 12.5 regarding carrier injection across a heterojunction, we know that the injection current in a P–p–N structure is primarily carried by the

electrons injected from the wide-gap N region to the narrow-gap p region, whereas the injection of holes from the P region to the n region dominates in a P–n–N structure. In either structure, almost all of the excess carriers created by current injection are injected into the narrow-gap active layer and are confined within this layer by the energy barriers of the heterojunctions on both sides of the active layer, as illustrated in Fig. 13.12(*b*) for a P–p–N structure. Because the narrow-gap active layer has a higher refractive index than the wide-gap outer layers on both sides, an optical waveguide with the active layer being the waveguide core is built into the DH geometry for vertical confinement of a horizontally propagating optical field, as shown in Figs. 13.12(*c*) and (*d*).

Because of the index steps on both sides of the active layer, the waveguide in the DH geometry is much more effective than that in the SH geometry. The active layer in a DH device is typically in the range of 100–300 nm. By properly designing the structure with sufficient index steps, the waveguiding active layer in a DH can be made as thin as 100 nm or less while maintaining effective confinement of the optical wave.

Compared to the active layer of a homostructure that has a thickness of the order of 1 μm or larger, which is determined by electron diffusion, the confinement of carriers in a thin DH active layer greatly increases the concentration of excess carriers at a given injection level, resulting in a high radiative efficiency for a DH LED and a large optical gain for a DH semiconductor laser. One additional advantage of a DH device is that because the optical radiation in the device has a photon energy near the narrow bandgap of the active layer, the wide-gap outer layers are transparent to the optical wave in the device, thus reducing the absorption loss of the device in comparison to that of a homostructure or SH device. Except for some specially designed devices in which intentional strain caused by lattice mismatch is desired, common DH devices are fabricated with lattice-matched layers to reduce the defects and the recombination centers at the heterojunction interfaces so that good current injection efficiency and high radiative efficiency can both be obtained.

Quantum-well structures

In terms of structure, a quantum well is just a very thin DH. As the active layer of a semiconductor DH gets thinner than about 50 nm, the effect of quantum confinement for the electrons and holes in the thin active layer starts to appear in the direction perpendicular to the junction plane. This effect leads to quantization of momentum in the perpendicular direction, resulting in discrete energy levels associated with the motion of electrons and holes in this direction, as shown in Fig. 13.13. In the horizontal dimensions, electrons and holes remain free and form energy bands. As a result, both conduction and valence bands are split into a number of subbands corresponding to the quantized levels, as shown in Fig. 13.13. For a particle of an effective mass m^* in an infinite square quantum well of a width d_{QW}, the energies of the quantized levels are inversely proportional to $m^* d_{QW}^2$ but increase quadratically with an integral quantum

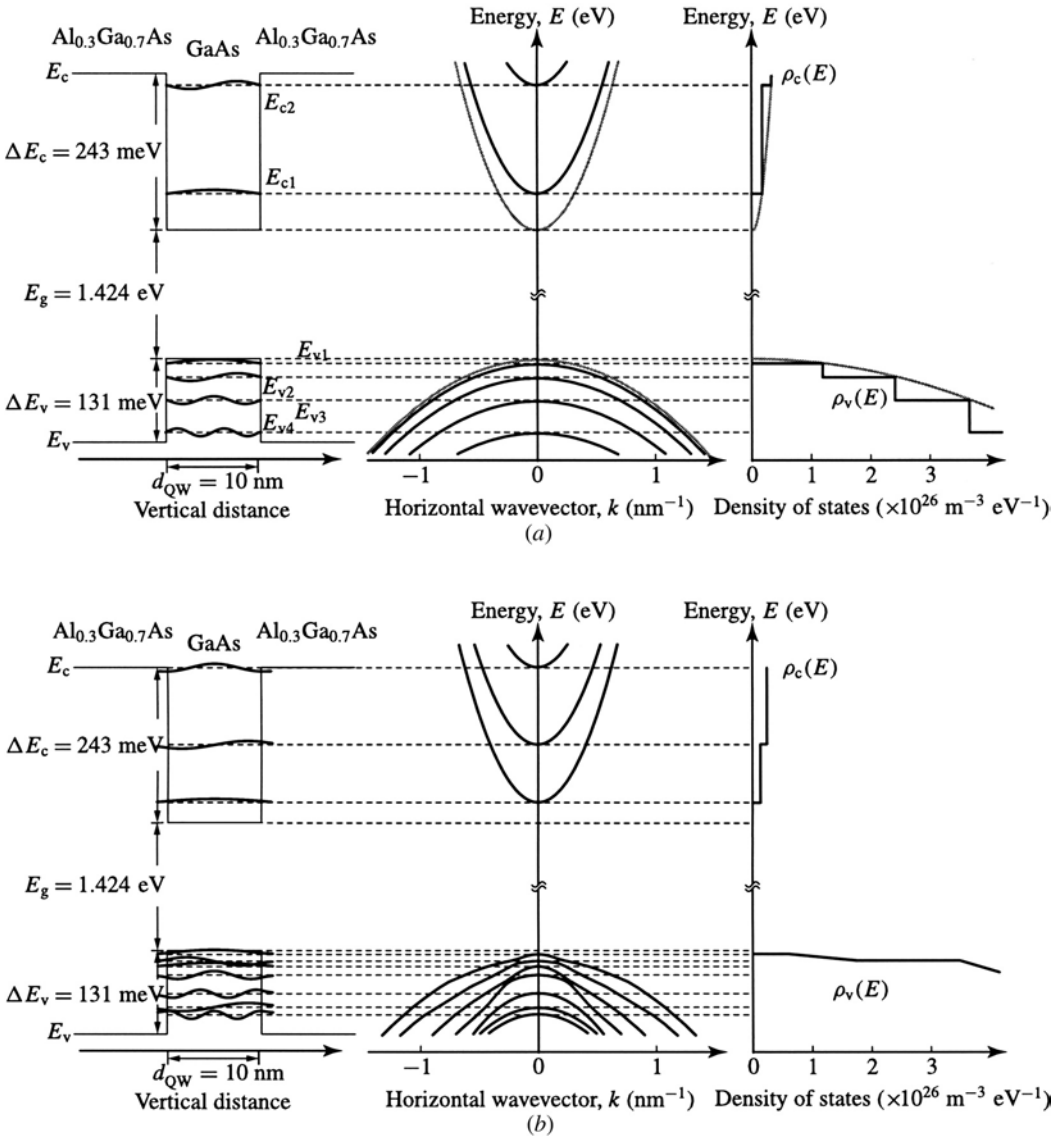


Figure 13.13 Quantized energy levels and corresponding subbands of a semiconductor quantum well. Shown as a quantitative example here is a square $Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}$ As quantum well at 300 K that has a thickness of $d_{QW} = 10$ nm. Shown in (*a*) are simplified, but not completely realistic, characteristics calculated using the quantized states of infinite wells as approximation. The conduction band is plotted for $m_e^* = 0.067m_0$, and the valence band is plotted for $m_h^* = 0.52m_0$ by combining the heavy- and light-hole bands into one band. Thus, separate quantized levels of heavy and light holes are not shown. The gray curves represent bulk properties of GaAs. Shown in (*b*) are characteristics obtained by considering finite well heights, separate heavy- and light-hole bands, and interaction among the quantized hole subbands.

number q, which starts with $q = 1$ for the lowest quantized level. A semiconductor quantum well is not an infinite potential well because the heights of the energy steps at the DH junctions are finite. However, taking the energy quantization of an infinite potential well as an approximation, we can express the band edges of the quantized conduction and valence subbands, respectively, as

$$E_{c,q}^{QW} = E_c + \frac{q^2h^2}{8m_e^*d_{QW}^2} \tag{13.49}$$

and

$$E_{v,q}^{QW} = E_v - \frac{q^2h^2}{8m_h^*d_{QW}^2}, \tag{13.50}$$

where $q = 1, 2, 3, \ldots$ The number of quantized subbands for electrons and that for holes depend on the heights of the potential barriers, ΔE_c in the conduction band and ΔE_v in the valence band, as well as on the well width d_{QW}.

The effective bandgap for a quantum well is no longer E_g of the semiconductor material in the active layer, but is the separation between the lowest subband of the conduction band and the highest subband of the valence band, both associated with the $q = 1$ quantized levels. The *selection rules* for optical transitions of electrons between quantized conduction subbands and quantized valence subbands require that only transitions between a conduction subband and a valence subband of the same quantum number q are allowed. Thus, optical transitions take place only between the $q = 1$ conduction subband and the $q = 1$ valence subband or between the $q = 2$ conduction subband and the $q = 2$ valence subband, and so on, but not between the $q = 1$ conduction subband and the $q = 2$ valence subband or between the $q = 2$ conduction subband and the $q = 1$ valence subband, and so forth. The photon energy required for transition between the conduction subband of quantum number q and the valence subband of quantum number q is

$$h\nu > E_g + \frac{q^2h^2}{8m_e^*d_{QW}^2} + \frac{q^2h^2}{8m_h^*d_{QW}^2} = E_g + \frac{q^2h^2}{8m_r^*d_{QW}^2}, \tag{13.51}$$

where m_r^* is the reduced effective mass defined in (13.15). Clearly, the lowest photon energy required for a transition between the conduction band and the valence band of a quantum well is that with $q = 1$ in (13.51). Therefore, a quantum well has an effective bandgap given by

$$E_g^{QW} = E_g + \frac{h^2}{8m_e^*d_{QW}^2} + \frac{h^2}{8m_h^*d_{QW}^2} = E_g + \frac{h^2}{8m_r^*d_{QW}^2}, \tag{13.52}$$

which is larger than that of the bulk semiconductor by an amount that is inversely proportional to the square of the well width.

EXAMPLE 13.6 A square $Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}As$ quantum well has a thickness of $d_{QW} = 10$ nm, as shown in Fig. 13.13. (a) By approximating the quantization of

this quantum well with that of an infinite potential well, find the quantized electron and hole energy levels that define the edges of the conduction and valence subbands, respectively. (b) By taking band offsets to be $\Delta E_c \approx 65\%\ \Delta E_g$ and $\Delta E_v \approx 35\%\ \Delta E_g$, find the number of conduction subbands and valence subbands for this finite quantum well. (c) What are the photon energies and the corresponding optical wavelengths for optical transitions between the conduction and valence subbands? (d) What is the effective bandgap of this quantum well? (e) What is the effect of the finite energy height of the quantum well in reality?

Solution (a) The well region consists of GaAs, which has $m_e^* = 0.067m_0$ and $m_h^* = 0.52m_0$ from Table 12.2. The subband edges are defined by the quantized energy levels of the quantum well. Thus, according to (13.49), the conduction subband edges are located at

$$\begin{aligned} E_{c,q}^{QW} - E_c &= \frac{q^2h^2}{8m_e^*d_{QW}^2} \\ &= q^2 \times \frac{(6.626 \times 10^{-34})^2}{8 \times 0.067 \times 9.11 \times 10^{-31} \times (10 \times 10^{-9})^2} \times \frac{1}{1.6 \times 10^{-19}}\ \text{eV} \\ &= 56.2q^2\ \text{meV}, \quad \text{for } q = 1, 2, \ldots, \end{aligned}$$

and, according to (13.50), the valence subband edges are located at

$$\begin{aligned} E_{v,q}^{QW} - E_v &= -\frac{q^2h^2}{8m_h^*d_{QW}^2} \\ &= -q^2 \times \frac{(6.626 \times 10^{-34})^2}{8 \times 0.52 \times 9.11 \times 10^{-31} \times (10 \times 10^{-9})^2} \times \frac{1}{1.6 \times 10^{-19}}\ \text{eV} \\ &= -7.2q^2\ \text{meV}, \quad \text{for } q = 1, 2, \ldots \end{aligned}$$

(b) At 300 K, $\Delta E_g = 374$ meV because $E_g = 1.424$ eV for GaAs and $E_g = 1.798$ eV for $Al_{0.3}Ga_{0.7}As$, according to (12.3). This quantum well has an energy height of $\Delta E_c \approx 65\%\Delta E_g = 243$ meV in the conduction band and an energy height of $\Delta E_v \approx 35\%\Delta E_g = 131$ meV in the valence band. We then find that the quantum well can only accommodate two conduction subbands because $3^2 \times 56.2\ \text{meV} = 505.8\ \text{meV} > \Delta E_c = 243\ \text{meV} > 2^2 \times 56.2\ \text{meV} = 224.8\ \text{meV}$. We also find that it can accommodate four valence subbands because $5^2 \times 7.2\ \text{meV} = 180\ \text{meV} > \Delta E_v = 131\ \text{meV} > 4^2 \times 7.2\ \text{meV} = 115.2\ \text{meV}$. Thus, there are two conduction subbands with band edges located at 56.2 meV ($q = 1$) and 224.8 meV ($q = 2$) from E_c of GaAs, and four valence subbands with band edges located at -7.2 meV ($q = 1$), -28.8 meV ($q = 2$), -64.8 meV ($q = 3$), and -115.2 meV ($q = 4$) from E_v of GaAs. These characteristics are shown in Fig. 13.13(*a*).

(c) The selection rules require that optical transitions occur only between conduction and valence subbands of the same quantum number q. Though there are four valence subbands, no corresponding conduction subbands are found for $q = 3$ and $q = 4$

because there are only two conduction subbands. Therefore, there are only two possible transitions for $q = 1$ and $q = 2$, respectively. The photon energy required for the transition between the $q = 1$ subbands is

$$h\nu > 1.424\ \text{eV} + 0.0562\ \text{eV} + 0.0072\ \text{eV} = 1.4874\ \text{eV},$$

which corresponds to an optical wavelength of $\lambda < 834$ nm. The photon energy required for the transition between the $q = 2$ subbands is

$$h\nu > 1.424\ \text{eV} + 0.2248\ \text{eV} + 0.0288\ \text{eV} = 1.6776\ \text{eV},$$

which corresponds to an optical wavelength of $\lambda < 739$ nm.

(d) The effective bandgap of the quantum well is determined by the $q = 1$ conduction- and valence-subband edges. Thus, according to the result found in (c) for the $q = 1$ transition, the effective bandgap $E_g^{QW} = 1.4874$ eV, corresponding to an optical wavelength of $\lambda_g^{QW} = 834$ nm. Compared to the bandgap of $E_g = 1.424$ eV at $\lambda_g = 871$ nm for GaAs, the effective bandgap is increased by 63.4 meV with a corresponding blue-shift of the bandgap optical wavelength by 37 nm due to the quantum confinement effect.

(e) In the above, we have used the formula for the energy levels of an infinite square well to calculate the energies of the subband edges of the quantum well. In reality, the barrier heights of the quantum well for both conduction and valence bands are finite, being $\Delta E_c = 243$ meV for the conduction band and $\Delta E_v = 131$ meV for the valence band in the case considered here. There are two major effects of these finite energy heights. First, each quantized energy level that defines a subband edge is actually somewhat smaller in its absolute magnitude than what is found above using the formula for an infinite potential well. Second, because of the first effect, there can be more quantized levels than those obtained above. A further complication in reality is the fact that there are heavy- and light-hole bands in the valence band. For this quantum well, it is found that there are actually three electron subbands, with band edges at 30.1, 117, and 232 meV from E_c, in the conduction band, and five heavy-hole subbands, at -4, -14.4, -35.9, -63.6, and -97.5 meV from E_v, and two light-hole subbands, at -22.6 and -87.1 meV from E_v, in the valence band. Therefore, the actual bandgap is $E_g^{QW} = 1.4581$ eV, which is smaller than 1.4874 eV found above, and the bandgap wavelength is $\lambda_g^{QW} = 850$ nm, which is longer than 834 nm found above. The quantized heavy- and light-hole subbands can interact with one another, resulting in mixed subbands due to band crossing. These characteristics are shown in Fig. 13.13(*b*).

For a fixed well width of 10 nm, E_g^{QW} decreases while λ_g^{QW} increases correspondingly as the barrier height decreases. This can be accomplished by reducing the Al content of the AlGaAs barrier layer to reduce the energy barrier height. It is a fundamental fact of quantum-mechanical quantization, however, that *there is always at least one quantized energy level in a symmetric potential well such as a square well no matter how shallow the well is*. Therefore, there are always at least one quantized conduction subband and one quantized valence subband no matter how low the well energy heights in the conduction and valence bands are reduced before they completely vanish.

A very important property of the quantum well is that its density of states is different from that of the bulk semiconductor. Because of quantization in the perpendicular direction, the density of states for each subband is that of a two-dimensional system divided by the thickness of the well. Every subband in the conduction band has the same density of states, and every subband in the valence band also has the same density of states:

$$\rho_c(E)\mathrm{d}E = \frac{4\pi m_e^*}{h^2 d_{QW}}\mathrm{d}E, \qquad \rho_v(E)\mathrm{d}E = \frac{4\pi m_h^*}{h^2 d_{QW}}\mathrm{d}E, \tag{13.53}$$

for each conduction subband and each valence subband, respectively.

For band-to-band optical transitions in a quantum well, (13.13) for E_2 and (13.14) for E_1 are still valid except that $E_2 - E_1 = h\nu$ has to satisfy the condition in (13.51). The density of states $\rho(\nu)$, defined in (13.16), for band-to-band optical transitions in a quantum well is then given by

$$\rho(\nu)\mathrm{d}\nu = \frac{4\pi m_r^*}{h d_{QW}} \sum_q H\left(h\nu - E_g - \frac{q^2h^2}{8m_r^* d_{QW}^2}\right)\mathrm{d}\nu, \tag{13.54}$$

where $H(x)$ is the *Heaviside function*, which has a value of $H(x) = 0$ for $x < 0$ and a value of $H(x) = 1$ for $x > 0$. Clearly, this density of states for optical transitions in a quantum well is different from that given in (13.17) for optical transitions in a bulk semiconductor. In a quantum well, dominant optical transitions occur between the first conduction subband and the first valence subband because those subbands are most populated by electrons and holes, respectively.

The density of states for optical transitions, $\rho(\nu)$, obtained in (13.54) is the most significant basic property of a quantum well that sets it apart from bulk semiconductors. Irrespective of this difference, most of the concepts discussed in the preceding four sections are valid for quantum wells too. The relations obtained in those sections can also be applied to a quantum well with the modification that $\rho(\nu)$ given in (13.54), instead of that in (13.17), is used, along with any other modifications that are needed to take account of the relations in (13.51) and (13.53). For example, the optical gain coefficient $g(\nu)$ expressed in (13.31) now takes the form:

$$\begin{aligned} g(\nu) &= \frac{c^2}{8\pi n^2 \nu^2 \tau_{sp}}[f_c(E_2) - f_v(E_1)]\rho(\nu) \\ &= \frac{c^2 m_r^*}{2n^2\nu^2\tau_{sp} h d_{QW}}[f_c(E_2) - f_v(E_1)] \sum_q H\left(h\nu - E_g - \frac{q^2h^2}{8m_r^* d_{QW}^2}\right) \end{aligned} \tag{13.55}$$

for band-to-band optical transitions in a quantum well. As another example, the condition for population inversion given in (13.35) is now modified to

$$\Delta E_F = E_{Fc} - E_{Fv} > h\nu > E_g^{QW} = E_g + \frac{h^2}{8m_r^* d_{QW}^2} \tag{13.56}$$

for a quantum well of a thickness d_{QW}, according to (13.51). When the condition in (13.56) is first reached, the optical gain coefficient of a quantum well is simply that

given in (13.55) for $q = 1$ because optical transitions only take place between the first conduction subband and the first valence subband. As the injected carrier density increases to a level at which the population inversion reaches the second conduction and valence subbands, additional gain contributed by transitions between the second subbands adds to the gain that originally exists from the transitions between the first subbands. If higher subbands exist, further contribution to the gain can come from transitions between higher subbands at higher injection levels. In contrast, the gain coefficient of a bulk semiconductor increases smoothly with carrier density.

Because a quantum well has a constant two-dimensional density of states, the transparency and gain characteristics of a quantum well are dependent on the two-dimensional carrier density, $N_{2D} = Nd_{QW}$, which is measured per square meter, of the quantum well. For a given material, the two-dimensional transparency carrier density, N_{tr}^{2D}, is a constant that is independent of well width. Thus, the three-dimensional transparency carrier density is a function of well width:

$$N_{tr} = \frac{N_{tr}^{2D}}{d_{QW}}. \tag{13.57}$$

The carrier dependence of the peak gain coefficient of a quantum well is not as linear as that of a bulk semiconductor; it can be approximated as

$$g_{max}(N) = \sigma N_{tr} \ln \frac{N}{N_{tr}}, \tag{13.58}$$

where the *gain cross section* σ is the *differential gain* at the transparency carrier density defined as

$$\sigma = \left.\frac{dg}{dN}\right|_{N_{tr}}. \tag{13.59}$$

The value of σ is relatively independent of well width. Therefore, the value of σN_{tr} is inversely proportional to well width. For $N - N_{tr} \ll N_{tr}$, this relation can be approximated as a linear relation in the form of (13.40):

$$g_{max}(N) = \sigma(N - N_{tr}). \tag{13.60}$$

Therefore σ has the meaning of the gain cross section for carrier densities close to the transparency carrier density.

EXAMPLE 13.7 For GaAs quantum wells, $N_{tr}^{2D} = 1.16 \times 10^{16}$ m^{-2} and $\sigma = 2.2 \times 10^{-19}$ m^2. Find the values of N_{tr} and σN_{tr} for (a) a GaAs quantum well of 10 nm thickness and (b) a GaAs quantum well of 8 nm thickness.

Solution (a) For $d_{QW} = 10$ nm, we have

$$N_{tr} = \frac{N_{tr}^{2D}}{d_{QW}} = \frac{1.16 \times 10^{16}}{10 \times 10^{-9}} \text{ m}^{-3} = 1.16 \times 10^{24} \text{ m}^{-3}$$

and

$$\sigma N_{tr} = 2.2 \times 10^{-19} \times 1.16 \times 10^{24}\ \text{m}^{-1} = 2.55 \times 10^{5}\ \text{m}^{-1}.$$

(b) For $d_{QW} = 8$ nm, we have

$$N_{tr} = \frac{N_{tr}^{2D}}{d_{QW}} = \frac{1.16 \times 10^{16}}{8 \times 10^{-9}}\ \text{m}^{-3} = 1.45 \times 10^{24}\ \text{m}^{-3}$$

and

$$\sigma N_{tr} = 2.2 \times 10^{-19} \times 1.45 \times 10^{24}\ \text{m}^{-1} = 3.19 \times 10^{5}\ \text{m}^{-1}.$$

We find that the values of N_{tr} for GaAs quantum wells are not much different from $N_{tr} = 1.22 \times 10^{24}$ m^{-3} found in Example 13.3 for bulk GaAs. In contrast, the value of σ for GaAs quantum wells is much larger than $\sigma = 3.37 \times 10^{-20}$ m^2 found in Example 13.4 for bulk GaAs. However, this value of σ for the quantum wells is evaluated at the transparency carrier density, and it decreases as the carrier density increases high above the transparency density.

Quantum wells have several advantages over bulk semiconductor media. The injected carriers are more concentrated in a quantized subband of a quantum well than in the entire band of a bulk semiconductor. Because the density of states for each subband of a quantum well is a constant that does not vary with energy, there are already a large number of electrons of the same energy near the edge of a conduction subband and a large number of holes of the same energy near the edge of a valence subband. In comparison, near the edges of the conduction and valence bands of a bulk semiconductor, there are very few electrons and holes because the density of states in a bulk medium varies with energy and starts at zero from the band edges. Therefore, a quantum well has a much larger gain cross section σ than a bulk semiconductor, as seen in Example 13.7. The transparency carrier density, N_{tr}, of a quantum well is comparable to that of a bulk semiconductor. This fact implies that a much lower injection current density is required for a quantum well than that required for a DH to reach transparency because the thickness of a quantum well is typically an order of magnitude smaller than that of a DH. At a given injection current density, a quantum well thus has a much higher gain than a DH of the same material. These characteristics lead to low threshold and high modulation speed for a QW laser. In addition, they also help in narrowing the laser linewidth and in reducing the temperature dependence of a QW laser. The gain spectrum of a quantum well can be made larger than a DH at high injection levels because of the constant density of states in each subband and because higher subbands can be successively reached for additional gain as the injection current increases. The gain bandwidth of a typical quantum well is in the range of 20–40 THz, about twice that of a typical bulk semiconductor.

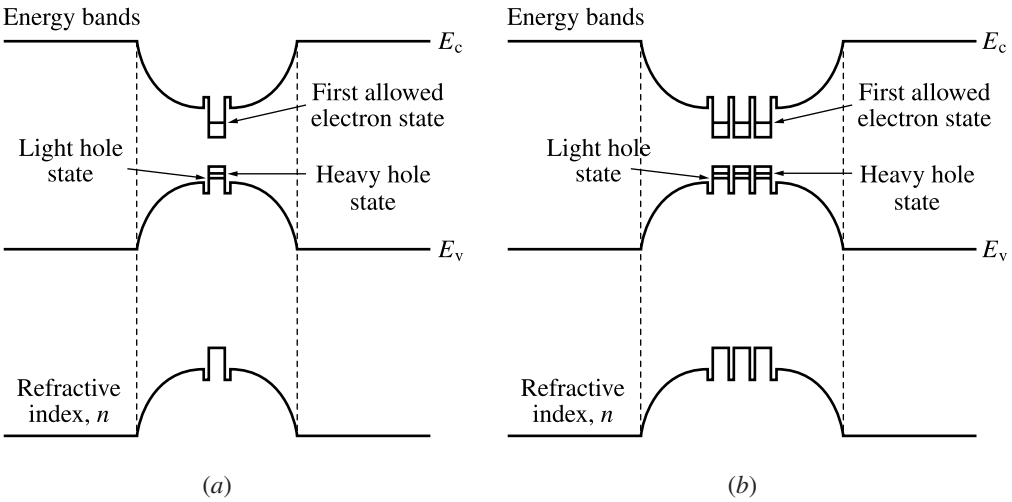


Figure 13.14 Energy bands and refractive index profiles of graded-index separate confinement heterostructures (GRIN-SCH) with (*a*) a single quantum well and (*b*) multiple quantum wells.

The thickness of a typical quantum well is in the range of 5–10 nm though quantum wells as thin as 2 nm and as thick as 20 nm are also used in some devices. Because of the small thickness of a quantum well, strain can be incorporated to create a *strained quantum well* without introducing undesirable defects and dislocations in the structure. Properly designed strained quantum wells can have a higher effective bandgap and a larger gain than unstrained quantum wells. To increase the total thickness of the active region for a larger gain volume while keeping the benefits of a quantum well, multiple quantum wells can be stacked together to make a multiquantum-well device. Compared to a conventional DH, a quantum well has very poor optical waveguiding ability because of its small thickness. Using multiple quantum wells helps to improve optical waveguiding. To have really good optical confinement, however, separate confinement heterostructures are used in QW devices. Among different variations, graded-index separate confinement heterostructures (GRIN-SCH), such as those shown in Fig. 13.14, are most often used.

13.6 Lateral structures

The junction structure of a device determines the carrier and optical field distributions in the vertical direction perpendicular to the junction plane of a device. The carrier and optical field distributions in the transverse directions parallel to the junction plane are determined by the lateral structure.

The lateral structure of a surface-emitting device can have either a *broad active area*, formed with little or no lateral restriction on the injected current, or a *small active area*,

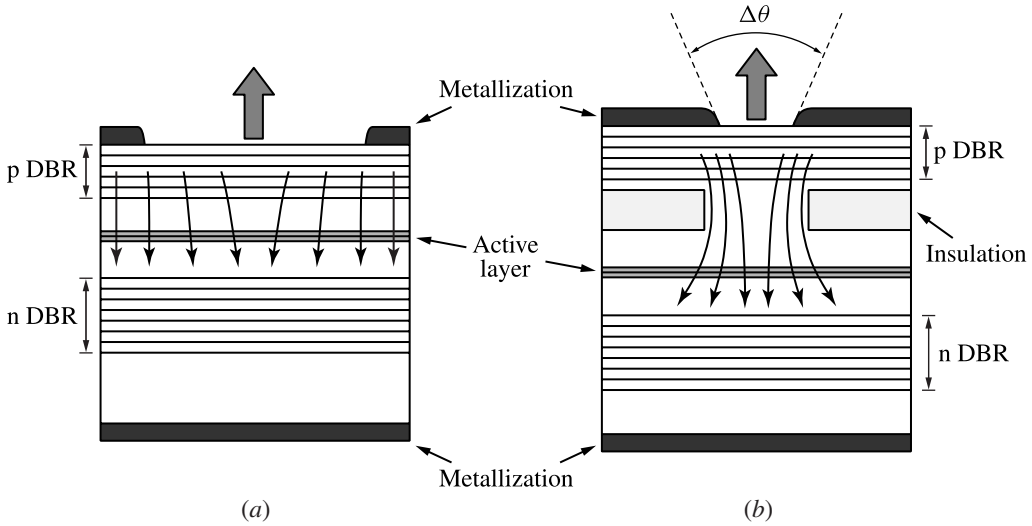


Figure 13.15 (*a*) Broad-area surface-emitting device and (*b*) small-area surface-emitting device.

formed by restricting the current flow into a confined area, as shown in Figs. 13.15(*a*) and (*b*), respectively. Most surface-emitting LEDs are broad-area devices. Some surface-emitting LEDs for fiber-optic applications and the vertical cavity surface-emitting lasers are small-area devices. For a small-area surface-emitting device, an optical waveguiding structure with index steps in lateral directions can be incorporated in conjunction with the small light-emitting active region for lateral optical confinement. The geometry of this structure is normally symmetric in transverse dimensions. For a laser, a small waveguiding area is important for single-transverse-mode emission. As a result, the emission of the device forms a circular beam of symmetric divergence with a divergence angle of $\Delta\theta$ shown in Fig. 13.15(*b*). The divergence angle of a device varies significantly among devices of different structural dimensions and emission wavelengths; it also varies under different operating conditions for a given device. Representative values are $\Delta\theta = 100^\circ$ for incoherent emission of a surface-emitting LED and $\Delta\theta = 10^\circ$ for coherent emission of a surface-emitting laser.

The lateral structures of edge-emitting devices have two basic types of geometry: *broad-area geometry* and *stripe geometry*, shown in Figs. 13.16(*a*) and (*b*), respectively. A broad-area device has no particular structures for restricting current flow to a particular region or for guiding the optical wave in lateral directions parallel to the junction plane. As a result, the broad-area geometry provides neither lateral carrier confinement nor lateral optical waveguiding. These deficiencies of broad-area geometry in lateral directions are similar to those of the homojunction in the vertical direction. Such a geometry leads to multiple transverse modes and filaments in the emission profile of the device. For this reason, almost all practical edge-emitting devices have stripe geometry.

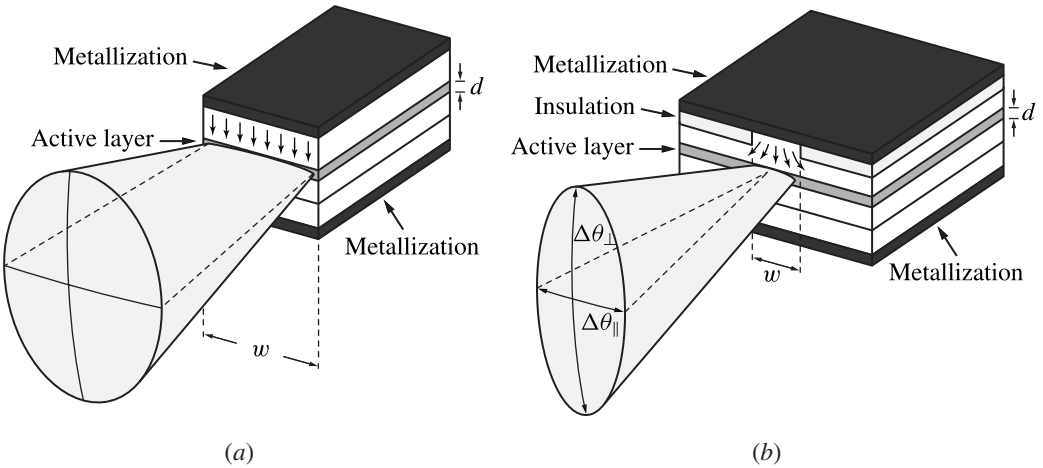


Figure 13.16 (*a*) Broad-area edge-emitting device and (*b*) stripe-geometry edge-emitting device.

Except for some high-power lasers of special interest, all practical edge-emitting semiconductor lasers have narrow stripes to ensure single-transverse-mode emission. In contrast, edge-emitting LEDs typically have broader stripes than lasers. There are two types of structures for the stripe geometry: gain-guiding structure and index-guiding structure. Both types of stripe geometry provide lateral confinement for the carriers and the optical field. Because of the asymmetry between vertical and lateral wave-guiding structures, the emission profile of a stripe-geometry device is asymmetric with different vertical and lateral divergence angles, $\Delta\theta_\perp$ and $\Delta\theta_\parallel$, respectively, as shown in Fig. 13.16(*b*). Because the width w of the stripe is normally much larger than the thickness d of the active layer, $\Delta\theta_\perp > \Delta\theta_\parallel$ for the coherent emission of a laser but $\Delta\theta_\perp < \Delta\theta_\parallel$ for the incoherent emission of an LED. Both vertical and lateral divergence angles vary significantly among devices of different structural dimensions and emission wavelengths, as well as under different operating conditions for a given device. Representative values are $\Delta\theta_\perp = 30^\circ$ and $\Delta\theta_\parallel = 120^\circ$ for the incoherent emission of an edge-emitting LED and $\Delta\theta_\perp = 30^\circ$ and $\Delta\theta_\parallel = 10^\circ$ for the coherent emission of a single-transverse-mode edge-emitting laser.

Gain-guiding stripe geometry

The gain-guiding stripe geometry in a gain-guided device is formed by injecting the current within a narrow stripe, which typically has a width ranging from a few micrometers for a gain-guided laser to a few tens of micrometers for an LED. No additional lateral structure is incorporated in the device. The current stripe defines the longitudinal direction of the edge-emitting device. Figure 13.17 shows the basic characteristics of a gain-guided device. There are a few different structures, shown in Fig. 13.18, that can be used to create the gain-guiding stripe.

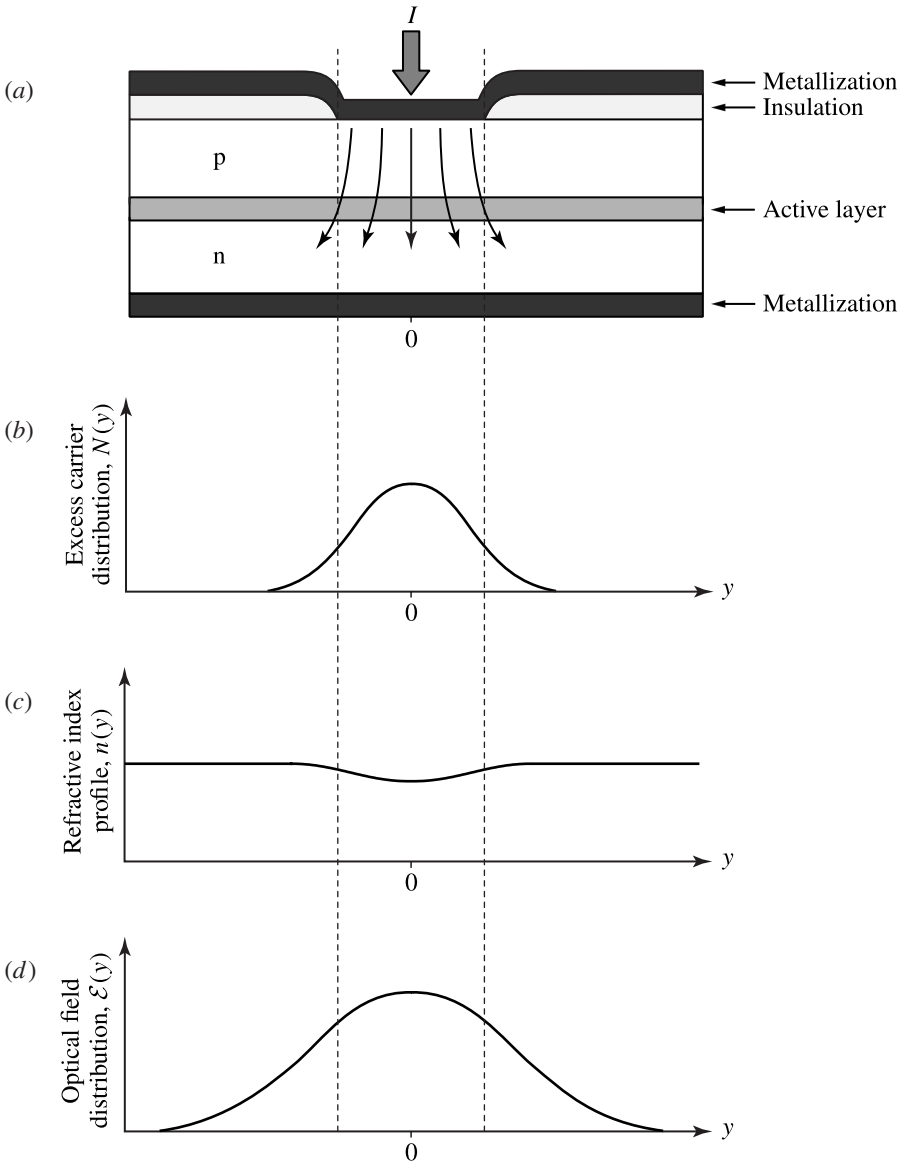


Figure 13.17 (*a*) Basic structure, (*b*) excess carrier distribution, (*c*) refractive index profile, and (*d*) lateral optical field distribution of a gain-guided stripe-geometry device.

As is shown in Fig. 13.17(*b*), a stripe of concentrated carriers in the active layer is formed along the longitudinal direction of a gain-guided device. In a semiconductor, an increase in carrier concentration is generally accompanied by a decrease in refractive index. This phenomenon is known as the *antiguidance effect*. In a semiconductor gain medium, it is often described by an experimentally measurable *antiguidance factor*,

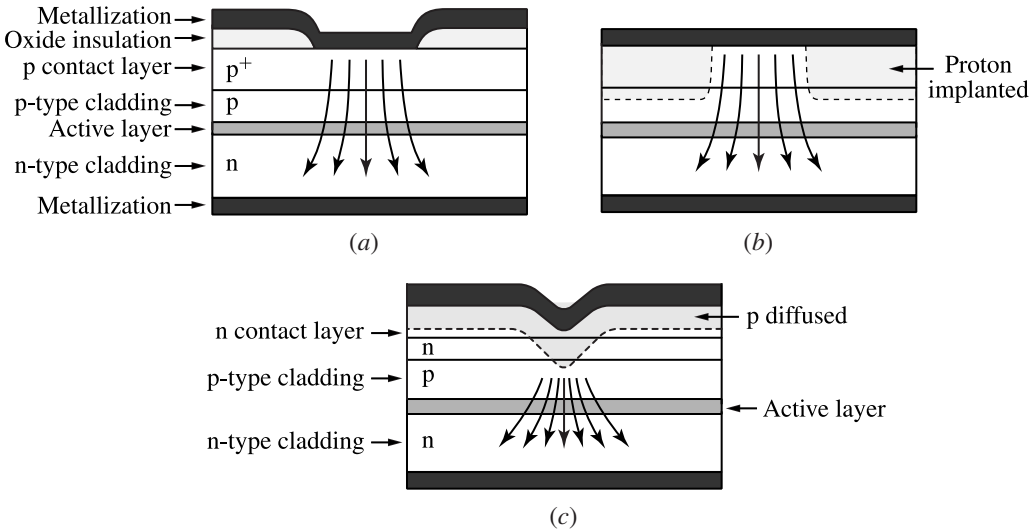


Figure 13.18 Structures of gain-guided devices: (*a*) oxide stripe, (*b*) proton-bombarded stripe, and (*c*) V-groove stripe.

also known as the *linewidth enhancement factor*, defined as

$$b = \frac{\partial n'/\partial N}{\partial n''/\partial N} = -\frac{2\omega}{c}\frac{\partial n'/\partial N}{\partial g/\partial N} = -\frac{4\pi}{\lambda}\frac{\partial n'/\partial N}{\partial g/\partial N}, \tag{13.61}$$

where n' and n'' are the real and imaginary parts of the refractive index and g is the optical gain coefficient. The antiguidance factor is a function of material, optical wavelength, temperature, and other operating conditions. It typically ranges between 2 and 6 for most semiconductor structures, but it can be as small as 1 and as large as 10.

The antiguidance effect results in an antiguiding index profile, shown in Fig. 13.17(*c*), that tends to spread the optical field laterally instead of confining it. This effect is not important for an LED, which does not require an optical gain and has incoherent emission. For a semiconductor laser, this effect is counteracted by the optical gain that is concentrated along the stripe. Stimulated amplification of the coherent laser field by this stripe of optical gain results in a gain-guiding effect for a laser beam to propagate along the longitudinal direction of the stripe. Nevertheless, the antiguiding index profile of a gain-guided laser still causes the beam to expand laterally to a certain degree.

The stripe width of a gain-guiding device varies between 10 and 200 μm, depending on intended applications. A wide stripe allows a large current to be injected for a high-power device, but it also makes the threshold current high in the case of a laser. A gain-guided laser tends to oscillate in multiple transverse modes, making it difficult to focus or collimate. Gain-guided LEDs are common because they are not subject to these limitations. Gain-guided lasers find important applications in the areas where high power and high conversion efficiency are needed but coherence and collimation

of the beam are not very important, such as in pumping solid-state lasers or fiber lasers.

EXAMPLE 13.8 Use the results obtained in Example 13.4 to estimate the carrier-induced index change in GaAs at 850 nm for an injected carrier density of $N = 2.83 \times 10^{24}\ \text{m}^{-3}$ if the antiguidance factor is $b = 3$.

Solution From Example 13.4(c), we find that

$$\frac{\partial g}{\partial N} = \sigma = 3.37 \times 10^{-20}\ \text{m}^2$$

for GaAs at 850 nm at 300 K. To estimate the index change, we assume that the refractive index changes linearly with carrier density. Thus, by using (13.61), we have

$$\frac{\Delta n}{\Delta N} \approx \frac{\partial n'}{\partial N} = -\frac{b\lambda}{4\pi}\frac{\partial g}{\partial N} = -\frac{b\lambda\sigma}{4\pi}.$$

For $\Delta N = N = 2.83 \times 10^{24}\ \text{m}^{-3}$, we have

$$\Delta n \approx -\frac{b\lambda\sigma N}{4\pi} = -\frac{3 \times 850 \times 10^{-9} \times 3.37 \times 10^{-20} \times 2.83 \times 10^{24}}{4\pi} = -1.94 \times 10^{-2}.$$

Compared to the base refractive index of $n = 3.65$ at 850 nm for GaAs in the absence of injected carriers, this carrier-induced index change is small. Nevertheless, such a small index reduction can cause a significant antiguiding effect to spread the distribution of an optical field if the optical field is not otherwise confined by a waveguiding index profile.

Index-guiding stripe geometry

For truly effective lateral optical confinement, an index-guiding structure has to be used. In an index-guiding structure, the lateral waveguide is formed by introducing a lateral index profile around the active region along the stripe where current is injected. As a result, both carrier confinement and optical waveguiding in the lateral direction are accomplished in a manner similar to the way a DH provides confinement for both carriers and optical field in the vertical direction. The basic characteristics of an index-guided device are illustrated in Fig. 13.19. There are many different index-guiding structures. A few examples are shown in Fig. 13.20.

In an index-guiding structure, the antiguidance effect is not important because the physical index steps that create the lateral waveguide are larger than the small changes in the refractive index caused by the injected carriers in the active region. Consequently, as the operating parameters, such as the injection current or the operating temperature, are varied, the output beam characteristics, including its profile, size, and divergence, of an index-guided device are more stable than those of a gain-guided device. In addition,

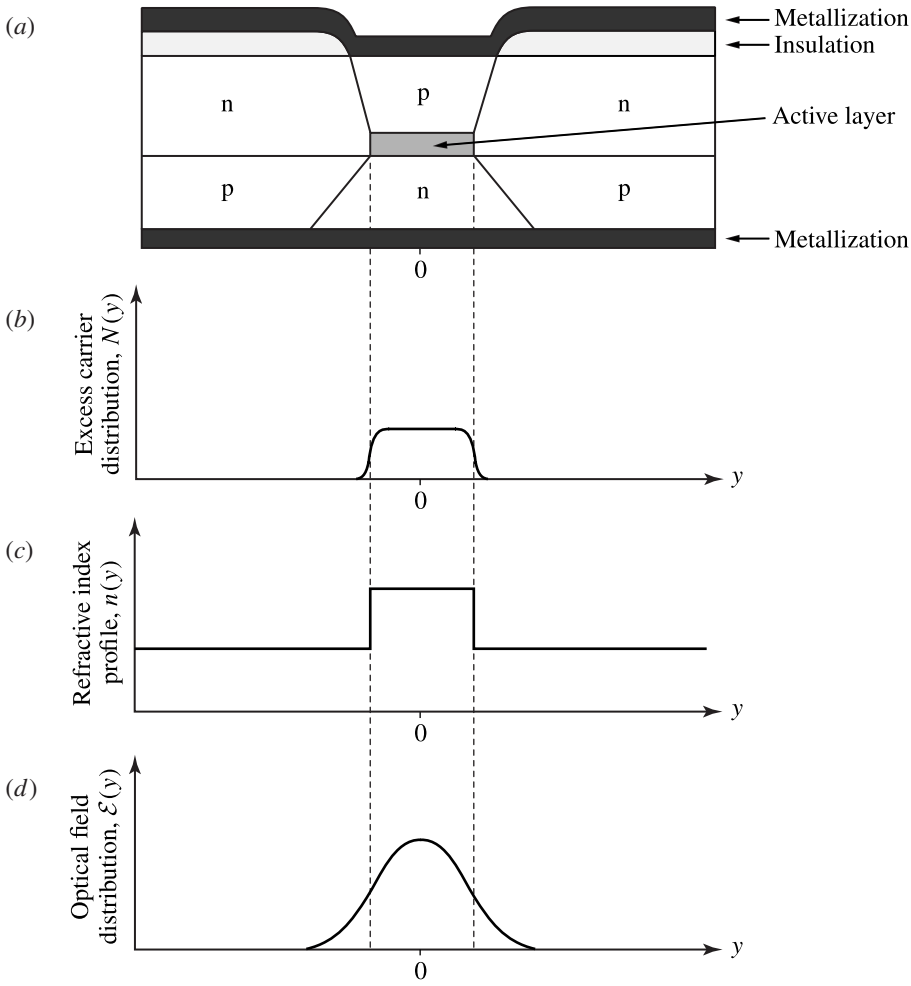


Figure 13.19 (*a*) Basic structure, (*b*) excess carrier distribution, (*c*) refractive index profile, and (*d*) lateral optical field distribution of an index-guided stripe-geometry device.

because the bandgap steps associated with the physical index steps create energy barriers for carriers like those in a DH, a high concentration of carriers can be injected and confined within the active region of an index-guiding device. Whereas the lateral carrier distribution in a gain-guided device is determined by the lateral spread of the current injected into the device and the lateral diffusion of carriers in the active layer, the lateral carrier distribution in an index-guided device is defined by the width of the index-guiding stripe. Therefore, it is possible to increase the carrier concentration by using a narrow index-guiding stripe to increase the efficiency of an LED and to lower the threshold of a laser. For single-transverse-mode lasers, the width of the stripe can be as narrow as 1–2 μm.

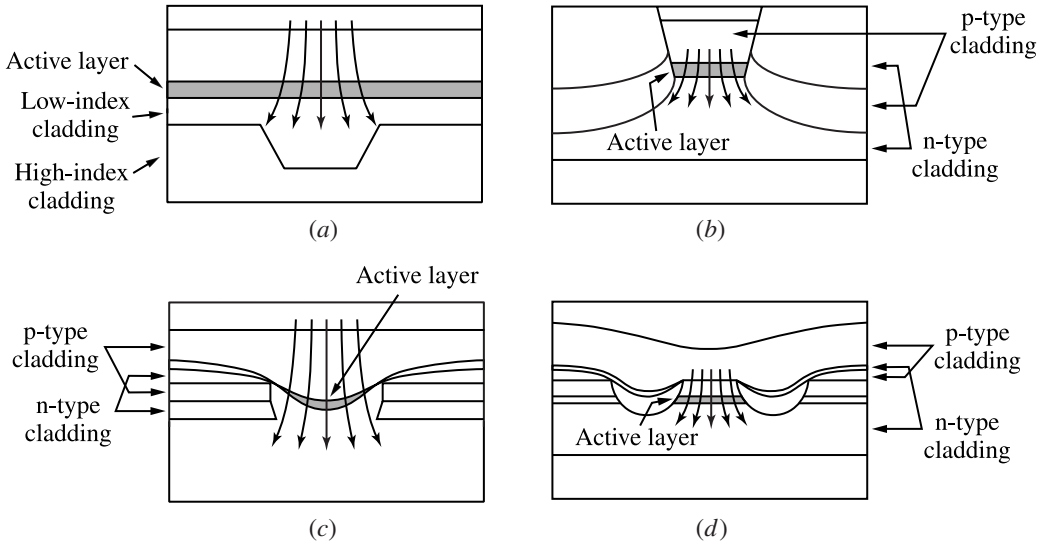


Figure 13.20 Structures of index-guided devices: (*a*) channeled-substrate planar (CSP) structure, (*b*) planar buried heterostructure (PBH), (*c*) buried crescent heterostructure, and (*d*) double-channel planar buried heterostructure (DC-PBH). Note that current flow is defined by the contact geometry, which is not shown in these illustrations.

EXAMPLE 13.9 An edge-emitting GaAs laser with an index-guiding stripe geometry emits at 850 nm wavelength. The vertical and lateral far-field divergence angles of the laser beam are $\Delta\theta_\perp = 30^\circ$ and $\Delta\theta_\parallel = 10^\circ$, respectively. What are the vertical and lateral sizes of the beam at the output facet of the laser?

Solution Because $\Delta\theta_\perp \neq \Delta\theta_\parallel$, the laser spot is elliptical with $w_\perp \neq w_\parallel$. Because the divergence angles are measured in free space, we take the refractive index of $n = 1$ for free space to find the spot size at the beam waist, which is located at the output facet of the laser. We have $\Delta\theta_\perp = 30^\circ = \pi/6$ rad and $\Delta\theta_\parallel = 10^\circ = \pi/18$ rad. Thus, by using (1.137) for Gaussian beam divergence, we find that

$$w_{0\perp} = \frac{2\lambda}{\pi\,\Delta\theta_\perp} = \frac{2 \times 850 \times 10^{-9}}{\pi \times \pi/6}\ \text{m} = 1.03\ \mu\text{m}$$

and

$$w_{0\parallel} = \frac{2\lambda}{\pi\,\Delta\theta_\parallel} = \frac{2 \times 850 \times 10^{-9}}{\pi \times \pi/18}\ \text{m} = 3.1\ \mu\text{m}.$$

The near-field spot at the laser facet is elliptical with $w_{0\parallel} = 3w_{0\perp}$ because $\Delta\theta_\perp = 3\Delta\theta_\parallel$. Such small beam sizes in both vertical and lateral directions can be made possible only by a laser structure with index guiding in both vertical and lateral dimensions.

13.7 Light-emitting diodes

LEDs are simple, but important, solid-state light sources that have a wide range of applications. LEDs that emit light in the visible spectral region are widely used in displays and in fiber-optic illumination. Infrared LEDs are useful for fiber-optic communications in those systems where the coherence, high power, and high speed of semiconductor lasers are not needed. Recent breakthroughs have resulted in LEDs of very high performance, in terms of efficiency and brightness, and have extended the spectral range of these high-brightness LEDs to the blue, violet, and ultraviolet regions. As the luminous performance of LEDs exceeds that of traditional incandescent lamps, LEDs become competitive in various lighting applications. Solid-state white light sources also become available by mixing the emission of red, green, and blue LEDs. These advances have created many new possibilities for the applications of LEDs.

Commercially available LEDs today cover the spectral range from the near ultraviolet to the near infrared, with optical wavelengths ranging from about 370 nm to 1.65 μm. These commercial LEDs are made of III–V compound semiconductors. Blue LEDs based on SiC have been developed, but they have very low efficiencies and are not practically useful. Blue and green LEDs based on II–VI compounds such as ZnTeSe and ZnCdSe have also been developed, but their commercial usefulness is limited. Organic LEDs based on polymers hold great promise, but they are still in the early stage of development. The main characteristics of LEDs based on III–V semiconductors are listed in Table 13.1.

The light output of an LED is the spontaneous emission generated by radiative recombination of electrons and holes in the active region of the diode under forward bias. From the discussions in Section 13.4, we learn that a semiconductor emits spontaneous photons no matter whether its electron and hole populations are in thermal equilibrium, characterized by a common Fermi level, or in quasi-equilibrium, characterized by separate quasi-Fermi levels. Spontaneous emission occurs in both direct-gap and indirect-gap semiconductors though a direct-gap semiconductor generally has a much larger radiative recombination rate and thus a much higher spontaneous emission efficiency than an indirect-gap semiconductor. As can be seen from Table 13.1, many LEDs are made of indirect-gap semiconductors doped with impurities that form isoelectronic centers to improve their luminescence efficiencies. Such LEDs typically have low quantum efficiencies compared with LEDs made of direct-gap semiconductors. Unlike a laser, an LED emits incoherent and unpolarized spontaneous photons that are not amplified by stimulated emission. Therefore, no optical gain is needed, and the condition for population inversion given in (13.35) is not required for the operation of an LED. No resonant optical cavity is needed for an LED, either. As a result, the emission from an LED does not have the coherence, or the directionality, of the emission

Table 13.1 *Basic characteristics of III–V semiconductor LEDs*

LED material	Substrate	Type[a]	Wavelength (nm)	Color	Efficiency[b]
InGaN	Sapphire	D	370–680	UV–Red	Medium–High[c]
AlGaInP	GaAs	D	560	Green	Medium
AlGaInP	GaP	D	570	Green	Medium
AlGaInP	GaP	D	590	Yellow	High
AlGaInP	GaP	D	607	Orange	High
AlGaInP	GaP	D	620–650	Red	High
AlGaAs	GaAs	D	650–675	Red	Medium
GaAsP : N	GaP	I	589	Yellow	Low
GaAsP : N	GaP	I	632	Red	Low
GaAsP	GaAs	D	649	Red	Low
GaP	GaP	I	555	Green	Low
GaP : N	GaP	I	565	Green	Low
GaP : N,N	GaP	I	590	Yellow	Low
GaP : Zn,O	GaP	I	699	Red	Medium
AlGaAs : Si	GaAs	D	820–890	IR	High
GaAs : Si	GaAs	D	920–950	IR	High
InGaAsP	InP	D	1100–1650	IR	High

[a] D, direct gap; I, indirect gap.

[b] External quantum efficiency. High, greater than 10%; medium, between 1 and 10%; low, less than 1%.

[c] Efficiency varies with wavelength, being higher in the blue and green spectral regions and lower toward both ends of the spectral range.

from a laser. Unlike a laser, an LED does not have a threshold, either. It starts emitting light as soon as a forward bias voltage is applied to its junction.

LED efficiency

The *power conversion efficiency* of an LED is defined in the same manner as that of a laser given in (11.89):

$$\eta_c = \frac{P_{out}}{P_p}, \tag{13.62}$$

where P_{out} is the optical output power of the LED and P_p is the electric pump power supplied by the injection current. Because an LED has no threshold, its *external quantum efficiency* is defined as

$$\eta_e = \frac{\Phi_{out}}{\Phi_p}, \tag{13.63}$$

where Φ_{out} is the output photon flux of the LED and Φ_p is the pump electron flux. For an LED that emits at an optical frequency ν, the output photon flux is simply

$\Phi_{out} = P_{out}/h\nu$. If the LED is injected with a current I at a forward bias voltage V, then the pump power is $P_p = IV$ and the pump electron flux is $\Phi_e = I/e = P_p/(eV)$, where e is the electronic charge. Therefore, the power conversion efficiency and the external quantum efficiency have the following relation:

$$\eta_c = \eta_e \frac{h\nu}{eV}. \tag{13.64}$$

In general, $\eta_e \geq \eta_c$ because the law of the conservation of energy requires that $eV \geq h\nu$. The power conversion efficiency of a typical LED that emits photons at an energy close to its bandgap energy is approximately equal to, though slightly less than, its external quantum efficiency.

For an LED that emits in the visible spectral region, a *photometric efficiency*, or *luminous efficiency*, η_l, is also introduced to account for the spectral response of the human eye:

$$\eta_l = K \int \frac{d\eta_c}{d\lambda} V(\lambda) d\lambda \approx K \eta_c V(\lambda_0), \tag{13.65}$$

where $K = 683$ lm W^{-1}, known as peak *efficacy*, is the *photometric radiation equivalent* for photopic vision; $V(\lambda)$ is the *normalized photopic spectral luminous efficiency* that characterizes the relative spectral sensitivity of the human eye; and λ_0 is the peak emission wavelength of the LED. The luminous function $V(\lambda)$, shown in Fig. 13.21, has a peak value of 1 at the green spectral wavelength of $\lambda = 555$ nm where the human eye is most responsive, and it drops to a value of 0.01 at the violet wavelength of $\lambda = 414$ nm and at the red wavelength of $\lambda = 687$ nm near the two edges of the visible

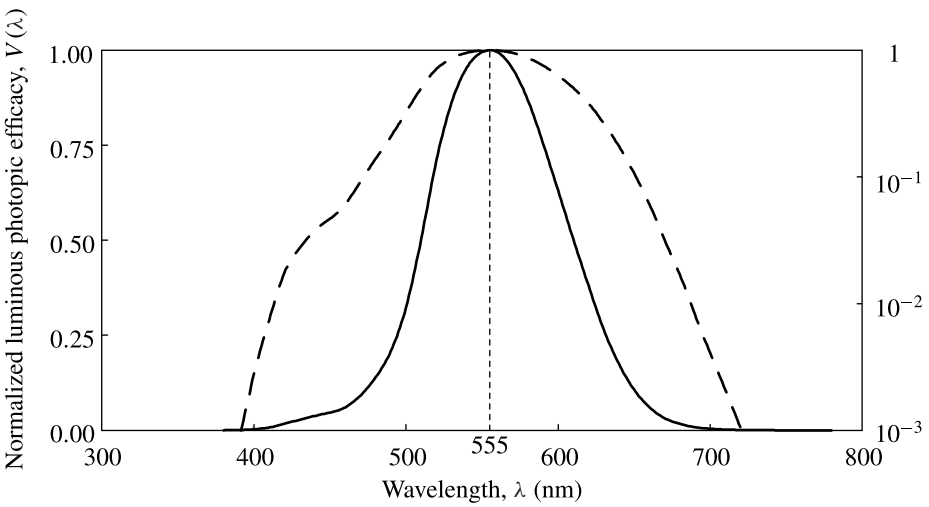


Figure 13.21 Photopic luminous efficiency function, $V(\lambda)$, plotted in linear scale (solid curve referring to the left axis) and logarithmic scale (dashed curve referring to the right axis). This plot is based on CIE 1988 updated data.

region. Therefore, a green LED appears much brighter than a blue or red LED of the same efficiency. The luminous efficiency is the luminous flux emitted by an LED per watt of electric pump power. It has the unit of lumen per watt, lm W^{-1}. The *luminous flux*, Φ_l, which is measured in lumens, of an LED that has an electric pump power of P_p and an optical output power of P_{out} is

$$\Phi_l = \eta_l P_p = \frac{\eta_l}{\eta_c} P_{out} = K V(\lambda_0) P_{out}. \tag{13.66}$$

EXAMPLE 13.10 A transparent substrate (TS) AlGaInP/GaP LED emits at 636 nm. It has an external quantum efficiency of $\eta_e = 23.7\%$ when operated at a forward voltage of 2.02 V with an injection current of 20 mA. (a) Find its power conversion efficiency under the given operating conditions. (b) Find its optical output power. (c) Find its luminous efficiency and luminous flux.

Solution (a) The photon energy at 636 nm is

$$h\nu = \frac{1239.8}{636} \text{ eV}.$$

Using (13.64) with $V = 2.02$ V, we find the following power conversion efficiency:

$$\eta_c = \eta_e \frac{h\nu}{eV} = 23.7\% \times \frac{1239.8}{636} \times \frac{1}{2.02} = 22.9\%.$$

(b) The electric pump power is $P_p = IV = 2.02 \times 20$ mW $= 40.4$ mW. Thus the optical output power

$$P_{out} = \eta_c P_p = 22.9\% \times 40.4 \text{ mW} = 9.25 \text{ mW}.$$

(c) From Fig. 13.21, we find that $V(\lambda) = 0.208\,16$ for $\lambda = 636$ nm. Therefore, the luminous efficiency of this LED is

$$\eta_l = K \eta_c V(\lambda_0) = 683 \times 22.9\% \times 0.208\,16 \text{ lm W}^{-1} = 32.6 \text{ lm W}^{-1}.$$

The luminous flux

$$\Phi_l = \eta_l P_p = 32.6 \times 40.4 \times 10^{-3} \text{ lm} = 1.32 \text{ lm}.$$

The external quantum efficiency of an LED is the probability for each charged carrier that is injected into the LED to give rise to one emitted photon. It is determined by three factors, each measured by a characteristic efficiency of its own. First, the current injected into an LED consists of the diffusion current, which injects carriers into the active region, and other components such as surface recombination current and space-charge recombination current, which do not contribute to the carriers injected into the active region. The fraction of the total injection current that actually contributes to

the injected carriers in the active region of an LED is the *injection efficiency*, η_{inj}. Of the carriers injected into the active region, only those that recombine radiatively are responsible for the generation of photons. The probability of radiative recombination is quantified by the radiative efficiency, or the *internal quantum efficiency*, η_i, which is defined in (13.4). Not all of the photons generated by those carriers that recombine radiatively can be extracted out of the LED, however. A large portion of them is trapped inside the LED and is eventually reabsorbed by the LED material. The *extraction efficiency*, η_t, quantifies the probability that a photon generated in the active region of the LED can successfully escape to the outside and contribute to the optical output of the LED. The external quantum efficiency can thus be expressed as a product of these three characteristic efficiencies:

$$\eta_e = \eta_{inj}\eta_t\eta_i. \tag{13.67}$$

Clearly, the characteristic efficiencies of all three factors have to be maximized in order for an LED to have a high external quantum efficiency.

To have a high injection efficiency, surface recombination and carrier leakage have to be avoided by proper choice of the doping concentration and the thickness of each layer in the LED and by careful design of the electrical contacts. The injection efficiency is generally not a limiting factor for the external quantum efficiency of an LED, however, because it can easily be made higher than 80% for a well-designed LED. The internal quantum efficiency is normally very high for direct-gap semiconductors but is low for indirect-gap semiconductors. Because the radiative recombination rate increases with carrier concentration, according to (13.5), the use of a DH geometry as discussed in the preceding section can significantly improve the internal quantum efficiency of an LED by providing effective carrier confinement in the active layer. For a direct-gap semiconductor LED operating at a properly chosen injection current level, the radiative efficiency can be close to unity and, like the injection efficiency, is not a limiting factor for the external quantum efficiency of the LED, either. For an indirect-gap semiconductor LED, however, the radiative efficiency is a limiting factor though it can be improved with the doping of isoelectronic centers. For example, the radiative efficiency for GaP : Zn,O is about 30%, but that for GaP : N is only about 3%. The most significant limiting factor is normally the extraction efficiency, which depends on the details of the LED structure discussed below. A typical LED can have an extraction efficiency between 3 and 30%, depending on the geometry and the material of the LED. Combining all three factors, the external quantum efficiency of an LED ranges from lower than 1% to higher than 30%.

LED construction

The construction of an LED is determined largely by the consideration of maximizing the extraction efficiency of the LED. For LEDs used in fiber-optic applications,

including the infrared LEDs used in optical communications, the ultimate efficiency that counts includes not only the external quantum efficiency but also the coupling efficiency of the emission into the fiber. Therefore, the coupling efficiency to an optical fiber is also a factor to be considered in the construction of an LED that is used in a fiber-optic application. Because spontaneous emission radiates in all directions, a properly designed surface-emitting LED, which allows emission output in many different directions, generally has a much higher extraction efficiency than a comparable edge-emitting LED, which limits its emission output to a narrow angular spread in only one direction. However, because of the optical waveguiding effect in an edge-emitting device discussed in the preceding section, the emission of an edge-emitting LED is much more collimated, thus allowing for much more efficient direct coupling to an optical fiber, than that of a surface-emitting LED. For these reasons, conventional LEDs are surface-emitting devices, but edge-emitting LEDs are often used in fiber-optic applications.

The limitation on the extraction efficiency of a surface-emitting LED is caused by the absorption of the LED material and the Fresnel reflection between the high-index semiconductor and the low-index air. If nothing is done to optimize the structure of the LED, the extraction efficiency is less than 2%. Clearly, this is an important factor to be considered, and there is plenty of room for it to be improved through careful design of the LED structure. The spontaneous photons generated in the active region of an LED are emitted isotropically in all directions, but only those photons that reach a surface of the LED at angles of incidence smaller than the critical angle, θ_c, can be transmitted through that surface. This critical angle defines an escape cone of a solid angle Ω_{esc} with respect to each surface of the LED. Because spontaneous emission is distributed isotropically over the 4π solid angle, the probability for emitted photons to escape through a given surface is

$$\eta_{esc} = \frac{\Omega_{esc}}{4\pi} T, \tag{13.68}$$

where T is the transmittance of the surface. For a flat interface between the LED material of a refractive index n_1 and the ambient medium of a refractive index n_2, with $n_1 > n_2$, $\Omega_{esc} = 2\pi(1 - \cos\theta_c)$ and (see Problem 13.7.5)

$$\eta_{esc} \approx \frac{n_2^3}{n_1(n_1 + n_2)^2}, \tag{13.69}$$

where we have approximated the transmittance T with that of normal incidence.

EXAMPLE 13.11 In this example, we calculate the escape efficiency of an AlGaInP/GaP LED like the one considered in Example 13.10. The refractive index of AlGaInP is $n = 3.4$. The AlGaInP LED surface is exposed directly to the air without any treatment. (a) Find the critical angle θ_c and the escape solid angle Ω_{esc} for the interface between AlGaInP and air. What is the transmittance of this surface? (b) Find the escape efficiency

η_{esc} using the relation in (13.68). (c) Find η_{esc} using the approximation given in (13.69). Compare the result with that obtained in (b). (d) How does this escape efficiency compare with the external quantum efficiency of the AlGaInP/GaP LED described in Example 13.10?

Solution (a) With $n_1 = 3.4$ and $n_2 = 1$, we find that

$$\theta_c = \sin^{-1} \frac{1}{3.4} = 17.1^\circ.$$

We then find that

$$\Omega_{esc} = 2\pi(1 - \cos 17.1^\circ) = 0.0884\pi.$$

The transmittance

$$T = \frac{4n_1n_2}{(n_1 + n_2)^2} = \frac{4 \times 3.4 \times 1}{(3.4 + 1)^2} = 0.70.$$

(b) Using the parameters obtained in (a) for (13.68), we find that

$$\eta_{esc} = \frac{\Omega_{esc}}{4\pi} T = \frac{0.0884\pi}{4\pi} \times 0.70 = 1.55\%.$$

(c) Using $n_1 = 3.4$ and $n_2 = 1$ for (13.69), we find that

$$\eta_{esc} \approx \frac{n_2^3}{n_1(n_1 + n_2)^2} = \frac{1^3}{3.4 \times (3.4 + 1)^2} = 1.52\%.$$

This approximate result of $\eta_{esc} = 1.52\%$ is 98% of the result of $\eta_{esc} = 1.55\%$ obtained in (b). Therefore, the convenient relation given in (13.69) is a very accurate approximation.

(d) The TS AlGaInP/GaP LED described in Example 13.10 has an external quantum efficiency of $\eta_e = 23.7\%$, which is more than 15 times the escape efficiency of $\eta_{esc} = 1.55\%$ found for the AlGaInP surface to the air. In the face of the small value of η_{esc}, such a high external quantum efficiency looks quite impossible, but it is real. Many techniques can be applied to realize such a high external quantum efficiency. The basic concepts of such techniques are described in the following text.

In a surface-emitting LED, a transparent window layer grown on top of the active layer allows light emission from the top surface. The device can have either an absorbing substrate (AS) or a transparent substrate (TS). For an AS LED with a thin window layer of a thickness less than approximately 10 μm, as shown in Fig. 13.22(*a*), we find that $\eta_t \leq \eta_{esc}$ because only those photons that are emitted directly toward the top surface do not get totally absorbed by the substrate before reaching a surface. If an AS LED has a thick window layer, as shown in Fig. 13.22(*b*), half of the photons emitted toward each side surface of the LED chip can reach the side surface without being absorbed by the substrate, thus increasing the extraction efficiency to $\eta_t \leq 3\eta_{esc}$. In a TS LED with a thick window, as shown in Fig. 13.22(*c*), it is possible for photons emitted in any direction to reach a surface, and the theoretical limit of the extraction efficiency is

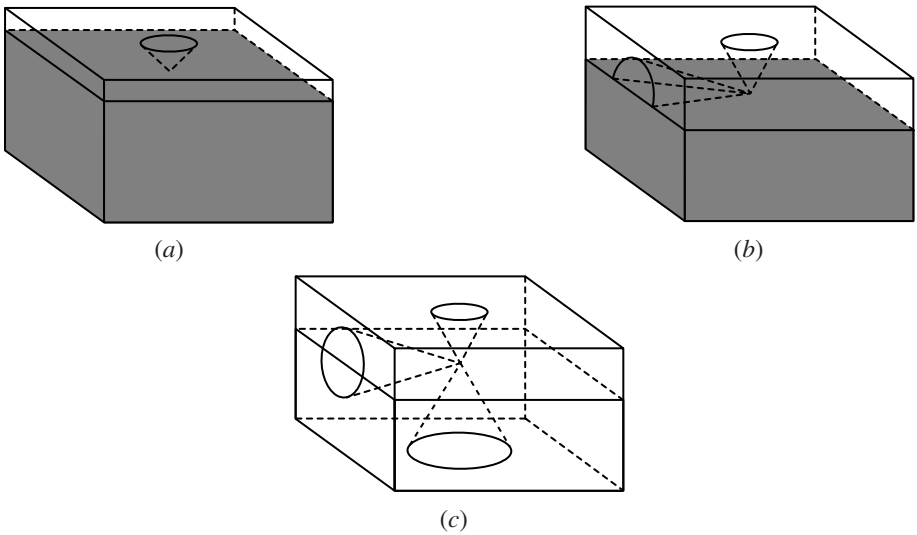


Figure 13.22 (*a*) Surface-emitting LED with a thin window and an absorbing substrate; only one top cone. (*b*) Surface-emitting LED with a thick window and an absorbing substrate; top cone plus four half-side cones. (*c*) Surface-emitting LED with a thick window and a transparent substrate; top and bottom cones plus four side cones. (Adapted from Vanderwater, D. A., Tan, I. H., Höfler, G. E., Defevere, D. C., and Kish, F. A., "High-brightness AlGaInP light emitting diodes," *Proceedings of the IEEE* **85**(11): 1752–1764, Nov. 1997.)

further increased to $\eta_t \leq 6\eta_{esc}$. A highly reflecting mirror surface is normally applied to the bottom surface of a TS LED to reflect light back to the top window surface. By examining Table 13.1, we see that among the AlGaInP LEDs, those that have a transparent GaP substrate have higher efficiencies than those that have an absorbing GaAs substrate. The InGaN/sapphire LEDs also have transparent substrates, which is part of the reason for their high efficiencies.

From (13.69), we find that η_{esc} for a flat semiconductor/air interface has a very small value. In Example 13.11 with $n_1 = 3.4$ for AlGaInP and $n_2 = 1$ for air, we get $\eta_{esc} \approx 1.55\%$. This value is exceedingly small. Even for a TS LED with a thick window, this value only permits a maximum extraction efficiency of less than 10%. Something has to be done to increase the value of η_{esc} if the extraction efficiency is to be increased further. Clearly from (13.68), the value of η_{esc} can be increased by increasing either the value of Ω_{esc}, thus allowing photons reaching the surface at large angles of incidence to be transmitted through the surface, or the value of T, thus allowing a higher probability of transmittance for a photon striking the surface at a given angle within the cone of Ω_{esc}, or both. A few solutions have been developed for this purpose.

One method to increase the value of Ω_{esc} significantly, but not that of T, is to shape the surface of the top window layer of an LED into a hemisphere with a radius much larger than the thickness of the active layer so that all spontaneous photons radiating toward this spherical surface come close to normal incidence, thus completely avoiding

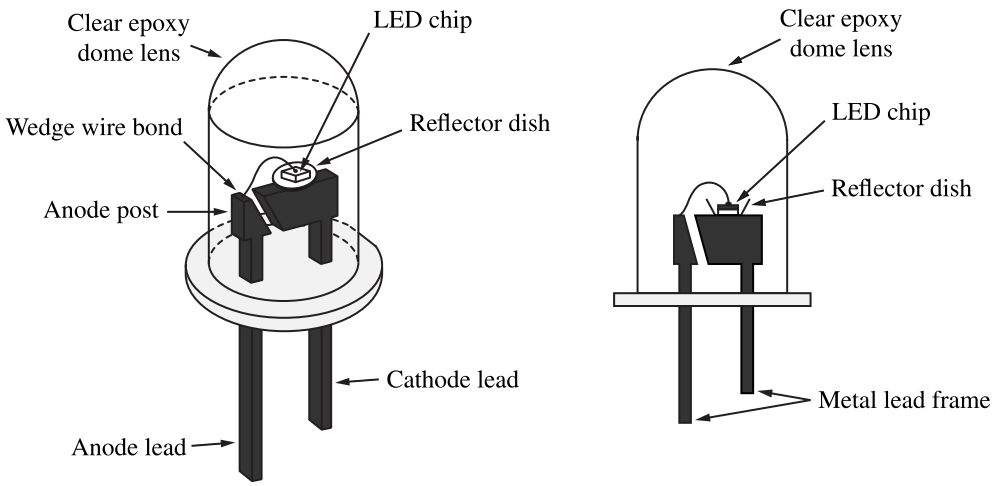


Figure 13.23 Construction of an LED encapsulated in plastic epoxy with a spherical dome lens. If a TS LED is used, it can be placed in a dish-shaped reflector to send side emission to the dome lens.

total reflection. This approach, however, requires polishing the semiconductor into a spherical surface and is therefore expensive and not practically useful. A second technique to increase the effective value of Ω_{esc} is to roughen the surfaces of the LED, thus randomizing the angles of incidence to reduce the probability of total reflection. The needed microscopic surface textures can be easily created by chemical etching or by coating with small polymer spheres. A factor of 2 increase in η_{esc} can be accomplished by proper surface texturing; therefore, this technique is practical. Another practical solution is to encapsulate the LED chip in a transparent plastic epoxy, which normally has a refractive index close to 1.5. In this approach, both the value of Ω_{esc} and that of T are increased because n_2 is increased. For $n_1 = 3.4$ and $n_2 = 1.5$, we find that $\eta_{esc} \approx 4.2\%$, an improvement of nearly three-fold over that of an LED without encapsulation. The last two techniques can be combined by encapsulating an LED that has textured surfaces to improve η_{esc} further. Applying these solutions to a TS LED properly, a high extraction efficiency of $\eta_t > 30\%$ is possible.

Figure 13.23 shows the construction of a surface-emitting LED with plastic encapsulation that is shaped into a spherical dome lens. When a TS LED, such as an AlGaInP/GaP LED, is assembled in this package, the LED can be placed in a miniature dish-shaped reflector that is coined into the top of the cathode post and is used to direct the emission from the side surfaces of the LED toward the dome lens. The shape and size of the plastic dome control the radiation pattern of the LED. Various radiation patterns for different applications can be obtained by tailoring the shape and size of the plastic encapsulation. In addition to the design of the spherical dome lens, aspherical dome lenses and rectangular packages are also used.

LEDs do not couple efficiently into single-mode fibers because the incoherent emission of an LED has a large divergence. Therefore, only multimode fibers are used in

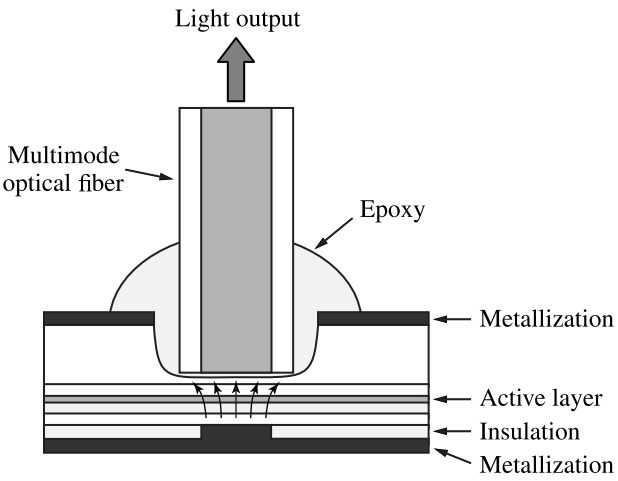


Figure 13.24 Surface-emitting Burrus-type LED for fiber-optic applications.

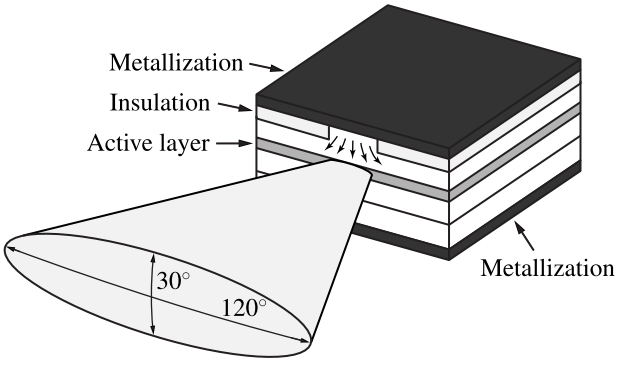


Figure 13.25 Stripe-geometry edge-emitting LED.

fiber-optic applications of LEDs. When a surface-emitting LED is used, special arrangement has to be made to bring the fiber tip into close proximity to the emitting active region of the LED for coupling of its emission into the fiber. A famous design is the Burrus type shown in Fig. 13.24. Other designs feature a microlens on the fiber tip or on the emitting surface of the LED to facilitate efficient coupling.

An edge-emitting LED has a structure similar to that of an edge-emitting semiconductor laser. The LED is prevented from laser oscillation by adding antireflection coating on the emitting facet or by leaving an unpumped absorbing section in the structure. The nonemitting facet is normally coated to be highly reflective to enhance the output from the emitting facet. The emission of an edge-emitting LED is much more collimated than that of a surface-emitting LED. As a result, its radiance is typically about 10 times higher than that of a surface emitting LED. Figure 13.25 shows the structure of a stripe-geometry edge-emitting LED. Its emission has a vertical spread of about 30° and a horizontal spread of about 120°. With its emission radiating from a

very small area on the emitting facet, an edge-emitting LED allows a fiber easy access to its emission output for a good coupling efficiency to the fiber.

Light–current characteristics

An LED is basically a p–n, P–n, or p–N junction diode though it may have a sophisticated DH structure for improved performance, as discussed in the preceding section. Therefore, the general electrical properties of an LED are those of a semiconductor junction diode described in Section 12.5 with the current–voltage characteristics shown in Fig. 12.12. The excess carriers in an LED that has an active layer of a thickness d much smaller than the diffusion length of the injected minority carriers can be considered uniformly distributed in the active region with a uniform density N. This is normally true for a DH device with a thin active layer. In this situation, the temporal variation of the carrier density in response to the variation in the injection current can be expressed as

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{J}{ed} - \frac{N}{\tau_\mathrm{s}}, \tag{13.70}$$

where e is the electronic charge, τ_s is the spontaneous carrier lifetime defined in (13.2), and J is the injection current density in the active region. Taking into consideration the carrier injection efficiency, the current density J that actually contributes to carrier injection is related to the total current supplied to the device as follows:

$$J = \eta_\mathrm{inj} \frac{I}{\mathcal{A}}, \tag{13.71}$$

where η_inj is the carrier injection efficiency defined earlier and $\mathcal{A}$ is the area of the junction.

As a light-emitting device that is pumped by current injection, a very important property of an LED is its output optical power as a function of the injection current, known as the *light–current characteristics*, or simply as the *L–I characteristics*, or the *power–current characteristics*, or simply as the *P–I characteristics*. The steady-state solution with $\mathrm{d}N/\mathrm{d}t = 0$ for (13.70) results in the following ideal power–current relation for an LED (see Problem 13.7.7):

$$P_\mathrm{out} = \eta_\mathrm{e} \frac{h\nu}{e} I, \tag{13.72}$$

which indicates that the output power of an LED increases linearly with the injection current. The L–I characteristics of a typical LED, shown in Fig. 13.26, are not exactly linear throughout the entire range of operation, however. These characteristics have several important features that distinguish an LED from a laser. First, there is no threshold in the L–I characteristics of an LED, indicating that an LED is turned on and starts emitting light once it is forward biased with any amount of injection current. The L–I curve of an LED is indeed quite linear, particularly at moderate current levels,

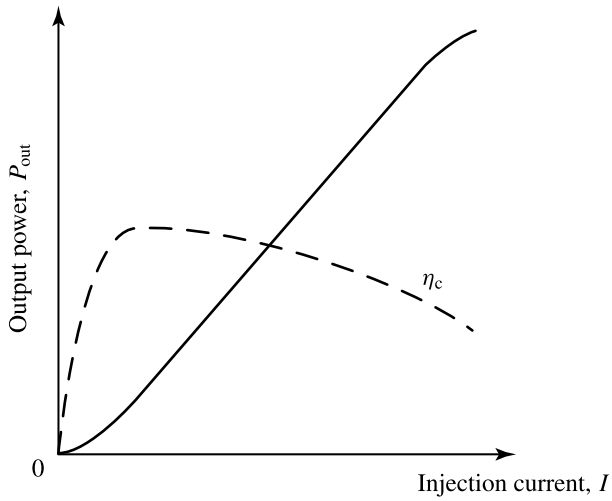


Figure 13.26 Typical light–current characteristics of an LED. The solid curve shows the output power, P_{out}, and the dashed curve shows the power conversion efficiency, η_c.

as indicated by (13.72). This linearity is useful for analog modulation of an LED. Nonlinearities in the L–I relationship are usually found at very low and very high current levels. Because of these nonlinearities, the efficiency of an LED changes as the injection current is varied. The efficiency starts low at low injection levels, increases sharply with an increasing injection current, but saturates or even decreases at high injection levels.

EXAMPLE 13.12 The optical output power and luminous flux of an LED at a given injection current level can be found without knowing the bias voltage if the external quantum efficiency is known. Find them for the AlGaInP/GaP LED described in Example 13.10 without using the knowledge of its bias voltage. Compare the results with those found in Example 13.10.

Solution From Example 13.10, we have $\lambda = 636$ nm, $\eta_e = 23.7\%$, and $I = 20$ mA. First, we note that

$$\frac{h\nu}{e} = \frac{1239.8 \text{ eV}}{636} \times \frac{1}{e} = \frac{1239.8}{636} \text{ V}.$$

Then, when we use (13.72) to find the optical output power, we simply have

$$P_{out} = \eta_e \frac{h\nu}{e} I = 23.7\% \times \frac{1239.8}{636} \times 20 \text{ mW} = 9.25 \text{ mW}.$$

With $K = 683 \text{ lm W}^{-1}$ and $V(\lambda_0) = 0.208\,16$ as found in Example 13.10, we find from (13.66) that

$$\Phi_l = K V(\lambda_0) P_{out} = 683 \times 0.208\,16 \times 9.25 \times 10^{-3} \text{ lm} = 1.32 \text{ lm}.$$

Compared with the results obtained in Example 13.10, we find exactly the same values for both P_{out} and Φ_1, as expected.

Spectral characteristics

The spectral characteristics of an LED include the emission wavelength, the spectral width, and the spectral shape. The emission wavelength of a direct-gap LED is determined by the bandgap of the active layer. Because of the *band-filling effect* of the injected electrons and holes taking up the states near the edges of the conduction and valence bands, respectively, the peak emission wavelength tends to be somewhat shorter than $\lambda_g = hc/E_g$, corresponding to a photon energy somewhat larger than the bandgap energy. However, if the active layer is heavily doped, the formation of bandtail states can lead to a long emission wavelength corresponding to a photon energy smaller than the bandgap energy, as discussed in Section 12.1. For an indirect-gap LED doped with isoelectronic impurities, the emission wavelength is longer than λ_g with a photon energy smaller than the bandgap energy, as is also discussed in Section 12.1 and illustrated in Fig. 12.1.

The peak emission wavelength of an LED varies with injection current and temperature. Because the bandgap of a III–V semiconductor normally decreases with increasing temperature, the peak emission wavelength of an LED becomes longer as the operating temperature increases. The rate of change depends on the specific semiconductor material of the LED. When the injection current increases, the band-filling effect caused by the corresponding increase in the concentration of the injected carriers leads to an increase in the emitted photon energy and a corresponding reduction in the peak emission wavelength. This effect is often abated by the shrinkage of the bandgap due to heating of the junction that accompanies the increase in injection current.

The spectral width and shape of the emission are intrinsically defined by the spontaneous emission spectrum in (13.42). The emission spectra of an LED, however, are often further complicated by frequency-dependent absorption and scattering by impurities and other materials, which have different bandgaps, in the layered structure of an LED. The spectral width in terms of the photon energy is approximately $h\Delta\nu = 3k_B T$, but it can range between $2k_B T$ and $4k_B T$. At room temperature, the spectral width of an LED is approximately 80 meV, but it can be as narrow as 50 meV or as broad as 100 meV in some devices. In terms of optical wavelength, the spectral width $\Delta\lambda$ ranges from approximately 20 nm for InGaN LEDs emitting short-wavelength ultraviolet or blue light to the order of 100 nm for InGaAsP LEDs with long-wavelength infrared emission. The spectral width of an LED normally increases with both temperature and injection current. Because an LED emits spontaneous radiation without an optical cavity, the longitudinal and transverse mode structures that are characteristic of a laser spectrum do not exist in the emission spectrum of an LED. Figure 13.27 shows a representative emission spectrum of an LED.

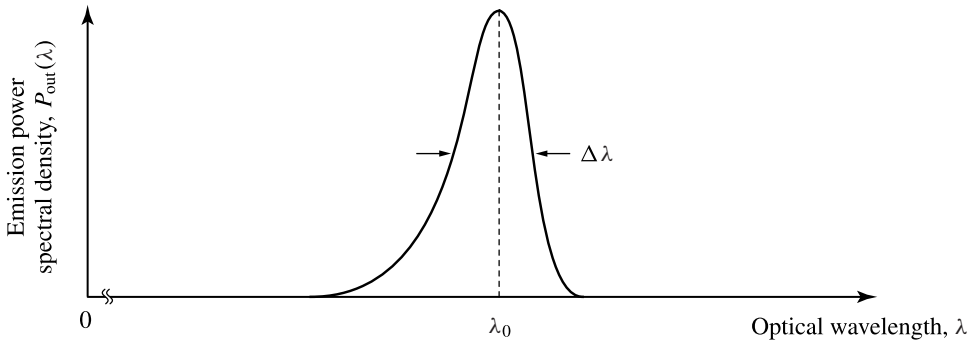


Figure 13.27 Representative emission spectrum of an LED.

Modulation characteristics

An LED can be directly modulated by applying the modulation signal to the injection current, an approach known as *direct current-modulation*. For high-speed applications, a large modulation bandwidth is desired. There are two factors that limit the modulation bandwidth of an LED: the junction capacitance, C_j, and the diffusion capacitance, C_d, both discussed in Section 12.5. Because an LED is operated under a forward bias, the diffusion capacitance is the dominating factor for its frequency response. The diffusion capacitance is a function of the carrier lifetime τ_s because it is associated with the injection or removal of carriers in the diffusion region in response to the modulation on the injection current. Therefore, the intrinsic speed of an LED is primarily determined by the lifetime of the injected carriers in the active region.

For an LED that is biased at a DC injection current level I_0 and is modulated at a frequency $\Omega = 2\pi f$ with a *modulation index* m, we can write the total time-dependent current that is injected to the LED as

$$I(t) = I_0 + I_1(t) = I_0(1 + m\cos\Omega t). \tag{13.73}$$

In the linear response regime under the condition that $m \ll 1$, the output optical power of the LED in response to this modulation can be expressed as

$$P(t) = P_0 + P_1(t) = P_0[1 + |r|\cos(\Omega t - \varphi)], \tag{13.74}$$

where P_0 is the constant optical output power at the bias current level of I_0, $|r|$ is the magnitude of the response to the modulation, and φ is the phase delay of the response to the modulation signal. For an LED modulated in the linear response regime, the complex response as a function of modulation frequency Ω is (see Problem 13.7.10(a))

$$r(\Omega) = |r|e^{i\varphi} = \frac{m}{1 - i\Omega\tau_s}. \tag{13.75}$$

The frequency response and modulation bandwidth of an LED are usually measured in terms of the electrical power spectrum of a broadband, high-speed photodetector that converts the optical output of the LED into an electric current. In the linear operating

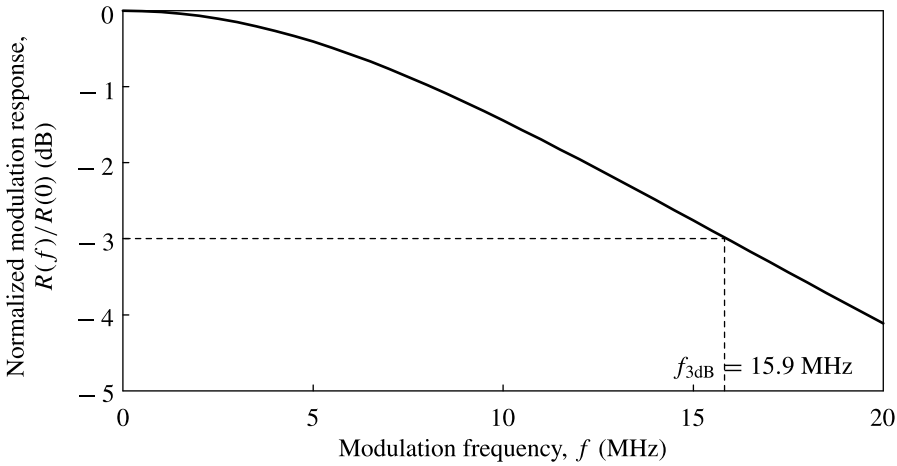


Figure 13.28 Normalized current-modulation frequency response of an LED measured in terms of the electrical power spectrum of a photodetector. The spontaneous carrier lifetime is taken to be $\tau_s = 10$ ns, as in Example 13.13, for this plot.

regime of the detector, the detector current is linearly proportional to the optical power of the LED. Therefore, the electrical power spectrum of the detector output is proportional to $|r|^2$:

$$R(f) = |r(f)|^2 = \frac{m^2}{1 + 4\pi^2 f^2 \tau_s^2}, \tag{13.76}$$

which has a 3-dB modulation bandwidth of

$$f_{3\text{dB}} = \frac{1}{2\pi\tau_s}, \tag{13.77}$$

as shown in Fig. 13.28. The spontaneous carrier lifetime τ_s is normally on the order of a few hundred to 1 ns for an LED. Therefore, the modulation bandwidth of an LED is typically in the range of a few megahertz to a few hundred megahertz. A modulation bandwidth up to 1 GHz can be obtained with a reduction in the internal quantum efficiency of the LED by reducing the carrier lifetime to the subnanosecond range. Aside from this intrinsic response speed determined by the carrier lifetime, the modulation bandwidth of an LED can be further limited by parasitic effects from its electrical contacts and packaging, as well as from its driving circuitry.

At an injection current I, the output optical power and the small-signal modulation bandwidth of an LED have the following *power–bandwidth product* (see Problem 13.7.10(b)):

$$P_{\text{out}} f_{3\text{dB}} = \eta_e \frac{h\nu}{e} \frac{I}{2\pi\tau_s} = \eta_{\text{inj}} \eta_t \eta_i \frac{h\nu}{e} \frac{I}{2\pi\tau_{\text{rad}}}, \tag{13.78}$$

where $h\nu$ is the photon energy of the LED emission. Therefore, at a given injection level, the modulation bandwidth of an LED is inversely proportional to its output power. A high-power LED tends to have a low speed, and vice versa.

EXAMPLE 13.13 The AlGaInP/GaP LED described in Example 13.10 has a spontaneous carrier lifetime of $\tau_s = 10$ ns. Find its 3-dB modulation bandwidth and its power–bandwidth product under the operating conditions described in Example 13.10.

Solution Using (13.77) for $\tau_s = 10$ ns, the 3-dB modulation bandwidth of the LED is easily found:

$$f_{3\text{dB}} = \frac{1}{2\pi\tau_s} = \frac{1}{2\pi \times 10 \times 10^{-9}}\ \text{Hz} = 15.9\ \text{MHz}.$$

Because the output power is $P_{\text{out}} = 9.25$ mW according to Example 13.10, the power–bandwidth product is

$$P_{\text{out}} f_{3\text{dB}} = 9.25\ \text{mW} \times 15.9\ \text{MHz} = 147\ \text{kW Hz}^{-1}.$$

13.8 Semiconductor optical amplifiers

An amplifier requires an optical gain for stimulated amplification of an optical signal, but it does not need a resonant cavity. Thus, a *semiconductor optical amplifier* (SOA), also called a *semiconductor laser amplifier*, can be made by simply eliminating the optical feedback mechanism of a semiconductor laser. For a solitary SOA as shown in Fig. 13.29, the end facets have to be antireflection coated. Meanwhile, no other feedback mechanism, such as a distributed feedback grating, is incorporated into the device structure. In theory, the output coupling loss of an amplifier is infinitely large so that it has an infinitely high laser threshold and thus never oscillates no matter how hard it is pumped. In practice, there is always some residual optical feedback in an SOA,

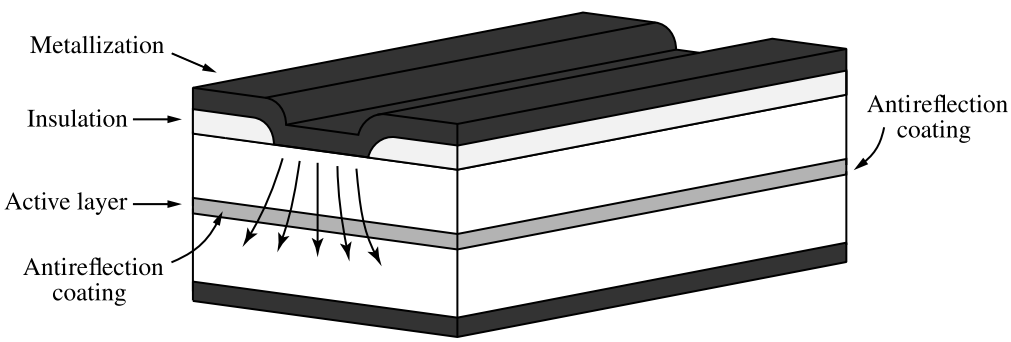


Figure 13.29 Basic structure of a solitary SOA.

but it is made small enough not to cause laser oscillation in the operating conditions of the amplifier. Indeed, a semiconductor laser can be used as an amplifier when it is biased below threshold but above transparency. An SOA is normally pumped with current injection. It has to be pumped to reach population inversion in its active region because an optical gain is required for its operation.

The general characteristics of laser amplifiers described in Section 10.4 apply to SOAs as well and need not be repeated here. Similarly to fiber amplifiers, SOAs can also be used as power amplifiers, optical repeaters, or optical preamplifiers in optical communication systems, as shown in Fig. 10.18. Compared to fiber amplifiers, however, SOAs have certain disadvantages and advantages in these applications because of the differences in physical properties between semiconductors and rare-earth ion-doped fibers. A few of these differences are apparent. For any optical signal emitted by a semiconductor laser at any wavelength, it is always possible to design a matching SOA by simply using the same material and the same structure as the laser. This convenience is not always available to fiber amplifiers. An SOA can be electrically pumped directly, whereas a fiber amplifier needs to be optically pumped. SOAs are compatible with other semiconductor devices, including semiconductor lasers, waveguides, detectors, and other semiconductor electronic devices. Therefore, they can be directly integrated into optoelectronic circuits on the chip level when they are used as power amplifiers or preamplifiers. However, when used in a fiber link, an SOA has a relatively large insertion loss. In contrast, fiber amplifiers have physical compatibility with fiber transmission lines and are ideally suited for all-optical repeater applications. As discussed in Section 10.5, fiber amplifiers have the advantage of being polarization insensitive. Because of the polarization-dependent characteristics of a semiconductor waveguide, the optical gain of an SOA normally depends on the polarization of the incoming signal if no special effort has been made to accommodate a polarization-independent design.

Besides these externally apparent differences, there are other significant differences between SOAs and fiber amplifiers due to fundamental differences between semiconductor and fiber gain media. Among the most important parameters of an optical gain medium are the gain bandwidth $\Delta\nu_g$, the emission cross section σ_e, or the gain cross section σ for a semiconductor, and the fluorescence lifetime τ_2, or the carrier lifetime τ_s for a semiconductor.

The gain bandwidth of an SOA, which is on the order of 10–20 THz, is greater than that of a rare-earth ion-doped fiber amplifier, which is on the order of several terahertz. Both are quite broad, however. Therefore, both semiconductor and fiber amplifiers are capable of amplifying optical signals of broad spectral widths, such as those in the form of ultrashort optical pulses, or optical signals of multiple wavelengths, such as those in a wavelength-division multiplexed system. The other two parameters are very different for SOAs and fiber amplifiers, however. For an SOA, σ is in the range of 1 to 5×10^{-20} m^2 and τ_s is on the order of 500 ps to 5 ns, whereas for a fiber amplifier, σ_e is in the range of 1 to 5×10^{-25} m^2 and τ_2 is on the order of 100 μs to 10 ms,

depending on the particular rare-earth ions. According to these numbers, an SOA has a gain cross section about five orders of magnitude larger and a gain relaxation time about six orders of magnitude shorter than those of a fiber amplifier. A large gain cross section means that the gain changes significantly in response to changes in the population inversion. In such a high-gain medium, the optical gain is easily saturated by the presence of stimulated emission because the saturation intensity of a gain medium is inversely proportional to the emission cross section. For this reason, a large gain cross section is disadvantageous for the application of an SOA as a linear amplifier. It can cause distortion in a given optical signal due to self saturation by the signal, interference between successive bits or symbols in the same channel, and crosstalk among different channels in a multichannel system due to cross saturation. The short carrier lifetime of an SOA is also a disadvantage in terms of its susceptibility to high-frequency noise. A carrier lifetime of 1 ns allows the gain of the amplifier to respond to noise or fluctuations at frequencies up to 1 GHz from the electrical pump source or from the optical signal. Consequently, an SOA tends to be less linear but noisier than a fiber amplifier.

13.9 Semiconductor lasers

Semiconductor lasers, also called *laser diodes* or *diode lasers*, are compact and efficient lasers that have found many important applications in optical communications, optical data storage, optical signal processing, compact-disk players, laser printers, and medical instruments. High-power semiconductor lasers are also used as highly efficient pump sources for other optically pumped lasers and amplifiers, such as solid-state lasers, optical fiber lasers, and fiber amplifiers. Unlike the light-emitting active region of an LED, which can be made of either a direct-gap or an indirect-gap semiconductor, the active region of a semiconductor laser has to be made of a direct-gap semiconductor of high radiative efficiency though indirect-gap semiconductors can be used for its cladding layers. Practical semiconductor lasers are based on III–V semiconductors though lasers based on IV–VI compounds have also been developed. In principle, any direct-gap semiconductor can be used as a laser material. In reality, however, there are many other issues to be considered. For most lasers, the active region has to be lattice matched to the substrate in order to avoid defects in the active layer, which can act as nonradiative recombination centers. In addition, lattice-matched cladding layers that have larger bandgaps and lower indices than the active layer are required for heterostructures. These requirements limit the spectral ranges of certain materials for laser applications. As a consequence, not all ternary and quaternary compounds that have direct bandgaps are successfully used to make lasers. One notable exception is the InGaN lasers fabricated on sapphire substrates. Another exception is strained QW lasers. The wavelength ranges of major lasers based on III–V semiconductors are listed in Table 13.2.

Table 13.2 *Major III–V semiconductor lasers*

Laser material	Substrate	Wavelength	Color
InGaN	Sapphire	370–680 nm	UV–red
InGaP	GaAs	620–700 nm	Red
AlGaAs	GaAs	750–870 nm	Red–IR
InGaAs	GaAs	980–1050 nm	IR
InGaAsP	InP	1.1–1.65 μm	IR
InGaAsSb	GaSb	2–3 μm	IR

There are many significant differences, in fundamental principles, structures, and characteristics, between a semiconductor laser and an LED though they use similar semiconductor materials and share the basic structures discussed in Section 13.5. The fundamental principles of a semiconductor laser differ from those of an LED in that a laser requires an optical gain for stimulated amplification of the emitted photons and a resonant cavity for optical feedback, both of which are not needed for an LED. Because of the need for an optical gain, the active region of a semiconductor laser has to be made of a direct-gap semiconductor of high radiative efficiency. An indirect-gap semiconductor simply cannot be pumped efficiently to reach the condition of population inversion for an optical gain. To reach population inversion, a semiconductor laser requires a higher current density than that required by a typical LED. Therefore, confinement of the current flow for efficiently injecting carriers into the active region and confinement of injected carriers for reducing carrier leakage are important factors to be considered in designing the structure of a semiconductor laser. As discussed in Section 13.5, the structures that serve these purposes well are DH junctions and quantum wells in the vertical direction and, for edge-emitting lasers, index-guiding structures in the lateral direction. For this reason, efficient semiconductor lasers are commonly index-guided devices with DH junctions or quantum wells. The need of a resonant cavity to provide optical feedback for the laser action leads to many different designs of laser structures. The combination of stimulated emission and optical feedback results in many characteristics of a laser, including the presence of a laser threshold, the existence of laser modes, the coherence and narrow linewidth of the laser emission, the high quantum efficiency of a laser, and the large modulation bandwidth of a laser, that are absent from the characteristics of an LED.

In terms of the mechanism for optical feedback, there are two basic types of resonant cavities for semiconductor lasers: the Fabry–Perot cavity and the grating-feedback cavity. Though both types of cavities serve the same purpose of providing optical feedback for laser oscillation, they are based on very different principles and have very different characteristics. Each type of cavity can have a number of different variations. Hybrids of the two types are also used in some lasers. Both types of cavities and their variations and hybrids can be used to make either edge-emitting or surface-emitting

lasers. In terms of structural geometry, the cavity of a semiconductor laser can be a horizontal cavity, formed in a direction parallel to the junction plane, a vertical cavity, formed in the direction perpendicular to the junction plane, or a folded cavity. An edge-emitting laser normally has a horizontal cavity. In contrast, a surface-emitting laser can have a horizontal cavity, a vertical cavity, or a folded cavity.

The general discussions on laser oscillation in Section 11.2, including the concepts of laser threshold, mode pulling, and longitudinal modes, apply to semiconductor lasers as well. For a laser to oscillate at a particular frequency, the general concept is that the round-trip gain has to exactly balance the round-trip loss while the round-trip phase shift is a multiple of 2π. This concept is also applied to determine the threshold gain and the oscillating modes of a semiconductor laser. However, some special considerations are often needed in the application of this concept because of the structural variation among different kinds of semiconductor lasers. Two structural factors are most significant for semiconductor lasers. First, the overlap between the laser field distribution and the active gain medium in a semiconductor can be small; it has to be considered for the gain of the laser. Second, many semiconductor lasers use gratings for their optical feedback; the effects of a grating on the phase and amplitude of a laser field have to be considered in such cases.

In a semiconductor laser, the volume of a laser mode is generally larger than that of the active region where the gain exists. The *gain overlap factor*, or the *gain filling factor*, of a laser mode is thus defined as

$$\Gamma = \frac{\iiint_{\text{active}} |\mathbf{E}|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\mathbf{E}|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z} \approx \frac{\mathcal{V}_{\text{active}}}{\mathcal{V}_{\text{mode}}}, \tag{13.79}$$

where $\mathbf{E}$ is the intracavity laser field, $\mathcal{V}_{\text{active}}$ is the volume of the active region, and $\mathcal{V}_{\text{mode}}$ is the effective volume of the laser mode under consideration.

With the notable exception of the vertical-cavity surface-emitting laser (VCSEL), most semiconductor lasers are basically waveguide lasers of stripe geometry. For a typical stripe-geometry laser that has a thin active layer, the laser waveguide has a thickness much less than its width, $d \ll w$. It can be considered as a single-mode slab waveguide that has a small V number. Then, according to (2.93), the confinement factor, Γ_{mode}, of the laser mode can be approximated by

$$\Gamma_{\text{mode}} = \frac{V^2}{2+V^2} \approx \frac{4\pi^2 n\Delta n d^2/\lambda^2}{1+4\pi^2 n\Delta n d^2/\lambda^2}, \tag{13.80}$$

where n and Δn are, respectively, the refractive index and the index step of the laser waveguide. For a DH semiconductor laser that does not contain quantum wells, the carriers distribute almost uniformly in the thickness d of the active layer, which also

serves as the optical waveguide. Then the overlap factor Γ is the same as the mode confinement factor Γ_{mode}. For a QW laser, the carriers are confined in the width d_{QW} of a quantum well but the laser mode is confined by the waveguide width d. Because $d_{\text{QW}} \ll d$ in general for a QW laser, the overlap factor Γ is not the same as, but is smaller than, the mode confinement factor Γ_{mode}. We thus have

$$\Gamma = \begin{cases} \Gamma_{\text{mode}}, & \text{for DH lasers,} \\ \dfrac{M_{\text{QW}} d_{\text{QW}}}{d} \Gamma_{\text{mode}}, & \text{for QW lasers,} \end{cases} \tag{13.81}$$

where M_{QW} is the number of quantum wells in the active layer of a QW laser.

For a VCSEL, which is generally a QW laser, the overlap factor has the form of the filling factor defined in Section 11.2. It takes the following simple form:

$$\Gamma = a \frac{M_{\text{QW}} d_{\text{QW}}}{l}, \tag{13.82}$$

where l is the length of the laser cavity and a is a factor between 1 and 2. The confinement factor of a stripe-geometry laser is independent of the cavity length, but that of a VCSEL can be increased by reducing the length of the VCSEL cavity.

The threshold gain coefficient for each mode of a semiconductor laser can be found by applying the concept of balancing the gain with the loss of the laser mode. Because some semiconductor lasers, such as DFB lasers, do not use localized cavity mirrors, the threshold gain coefficient of a given laser mode can be generally expressed as

$$\Gamma g_{\text{th}} = \overline{\alpha} + \alpha_{\text{out}}, \tag{13.83}$$

where $\overline{\alpha}$ is the internal distributed loss as defined in (11.56) and α_{out} is the output coupling loss of the laser oscillator.

EXAMPLE 13.14 A GaAs/AlGaAs laser emits at 850 nm wavelength. The refractive index of GaAs at 850 nm is $n = 3.65$. (a) Find the gain overlap factor Γ if the laser is a stripe-geometry DH laser that has an active waveguide thickness of $d = 0.2$ μm defined by an index step of $\Delta n = 0.2$. (b) Find the gain overlap factor Γ if the laser is a stripe-geometry MQW laser that contains three quantum wells each of a thickness of $d_{\text{QW}} = 10$ nm in a waveguide of $d = 0.2$ μm defined by an index step of $\Delta n = 0.2$. (c) Find the gain overlap factor Γ if the laser is an MQW VCSEL that contains three quantum wells each of a thickness of $d_{\text{QW}} = 10$ nm in a cavity of $l = 1$ μm. Take the factor $a = 2$.

Solution (a) For $\lambda = 850$ nm, $d = 0.2$ μm $= 200$ nm, $n = 3.65$, and $\Delta n = 0.2$, we find that

$$\Gamma_{\text{mode}} = \frac{4 \times \pi^2 \times 3.65 \times 0.2 \times (200/850)^2}{1 + 4 \times \pi^2 \times 3.65 \times 0.2 \times (200/850)^2} = 0.61.$$

Thus, $\Gamma = \Gamma_{\text{mode}} = 61\%$ for the DH laser without quantum wells.

(b) For the MQW laser, we have $M_{QW} = 3$ and $d_{QW} = 10$ nm. We also have $\Gamma_{mode} = 61\%$ from (a). From (13.81), we find the following overlap factor:

$$\Gamma = \frac{3 \times 10}{200} \times 61\% = 9.2\%.$$

(c) For the MQW VCSEL, we have $M_{QW} = 3$, $d_{QW} = 10$ nm, $l = 1$ μm, and $a = 2$. From (13.82), we find the following gain overlap factor:

$$\Gamma = 2 \times \frac{3 \times 10 \times 10^{-9}}{1 \times 10^{-6}} = 6\%.$$

We find that the gain overlap factor of the MQW VCSEL and that of the stripe-geometry MQW laser are both less than 10% and much smaller than that of the DH laser. This small gain overlap factor is normally compensated by the much higher gain of quantum wells in comparison to that of an ordinary DH structure.

Edge-emitting lasers

Most edge-emitting lasers are stripe-geometry lasers, though some broad-area edge-emitting lasers are still useful. The cavity of an edge-emitting laser is normally a horizontal cavity with a longitudinal axis defined by a gain-guiding or index-guiding stripe. There are three different kinds of edge-emitting semiconductor lasers: the *Fabry–Perot laser*, the *distributed Bragg reflector laser* (DBR laser), and the *distributed feedback laser* (DFB laser).

Fabry–Perot lasers

A Fabry–Perot resonant cavity for an edge-emitting semiconductor laser, shown in Fig. 13.30, can be realized by simply cleaving end facets. Because the entire structure of a semiconductor laser forms a single crystal, the cleaved facets are guaranteed to be perfectly parallel and vertical if they are cleaved along one of the crystalline planes. Typical III–V semiconductor lasers have end facets cleaved along the (110) plane of

Figure 13.30 Structure of an edge-emitting Fabry–Perot semiconductor laser with cleaved facets.

the crystal. Because a III–V semiconductor has a high refractive index, a cleaved facet in the air has a reflectivity typically in the range of 25–35%, depending on the specific composition of the semiconductor and the polarization and the mode of the laser field. Because of the high optical gain of a typical semiconductor laser, the optical feedback provided by such natural reflectivity from the cleaved facets is normally sufficient for laser oscillation. With such a Fabry–Perot cavity formed by two cleaved facets, the laser emits equally from both ends. To increase the laser output in one direction, the back facet can be coated with total-reflection coating so that all of the laser power is emitted from the uncoated front facet. Without an additional spectrum-filtering or frequency-selecting mechanism incorporated into the device, a Fabry–Perot semiconductor laser tends to oscillate in multiple longitudinal modes with a mode spacing of $\Delta\nu_{\mathrm{L}}$ given in (11.43) but modified by the mode-pulling effect. The threshold gain coefficient for each mode of a Fabry–Perot semiconductor laser is found by taking $\alpha_{\mathrm{out}} = -(\ln\sqrt{R_1 R_2})/l$ in (13.83) for the output-coupling loss to be the mirror loss, thus reducing (13.83) to the form of (11.56).

Distributed Bragg reflector lasers

Both DBR and DFB lasers use built-in gratings for optical feedback, but they have some basic structural differences and thus different characteristics. A DBR laser simply utilizes one or two Bragg reflectors as end mirrors in a manner similar to the mirrors of a Fabry–Perot laser. In contrast, a DFB laser uses a grating not as an end mirror but as a distributed feedback mechanism.

The principle and detailed characteristics of grating waveguide couplers are discussed in Section 5.1. By properly choosing the grating period Λ, a DBR can be designed to have a peak reflectivity at a desired Bragg frequency of

$$\nu_{\mathrm{B}} = \frac{c}{\lambda_{\mathrm{B}}}, \tag{13.84}$$

where λ_{B} is the Bragg wavelength defined in (5.23). At the Bragg frequency, the Bragg reflector has a peak reflectivity of

$$R_{\mathrm{DBR}} = \tanh^2 |\kappa| l_{\mathrm{DBR}}, \tag{13.85}$$

where l_{DBR} is the *actual physical length* of the DBR. From (5.24), we know that the phase shift on reflection from a DBR for a wave that has a propagation constant $\beta(\omega)$ at a frequency ω is

$$\varphi_{\mathrm{DBR}} = \varphi_{\mathrm{B}} + 2\left[\beta(\omega) - \beta_{\mathrm{B}}\right] l_{\mathrm{DBR}}^{\mathrm{eff}}, \tag{13.86}$$

where $\beta_{\mathrm{B}} = \beta(\omega_{\mathrm{B}})$ and

$$l_{\mathrm{DBR}}^{\mathrm{eff}} = \frac{\tanh |\kappa| l_{\mathrm{DBR}}}{2|\kappa|} = \frac{R_{\mathrm{DBR}}^{1/2}}{2|\kappa|} \tag{13.87}$$

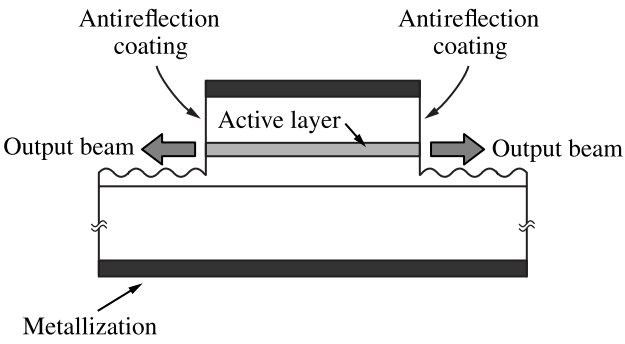


Figure 13.31 Structure of an edge-emitting distributed Bragg reflector (DBR) semiconductor laser with two Bragg reflectors.

is the *effective phase length* of the DBR for its reflection phase shift. We see from (13.86) that $\varphi_{\text{DBR}} = \varphi_{\text{B}}$ at the Bragg frequency. According to (5.29), the frequency bandwidth for high reflectivity of a DBR is approximately

$$\Delta\nu_{\text{DBR}} \approx \frac{|\kappa|c}{\pi N_\beta}, \tag{13.88}$$

where N_β is the effective group index of the mode field at the Bragg frequency. The peak reflectivity and the bandwidth of a DBR can be chosen by properly choosing the coupling coefficient κ and the physical length l_{DBR} of the DBR. The effective phase length $l_{\text{DBR}}^{\text{eff}}$ is then determined by both κ and l_{DBR}. Thus the phase shift φ_{DBR} is only a function of optical frequency once the physical parameters κ and l_{DBR} of the DBR are given.

To make a DBR laser, one or both of the reflective facets of a Fabry–Perot laser are replaced with DBR mirrors. Figure 13.31 shows a DBR laser with two Bragg reflectors. Note that in a DBR laser, the grating is placed outside the active region of the laser. A DBR laser cavity of a length l between the two DBR mirrors such as the one shown in Fig. 13.31 has many longitudinal modes at the frequencies defined by the following resonance condition:

$$2\beta l + \varphi_{\text{DBR1}} + \varphi_{\text{DBR2}} = 2q\pi, \tag{13.89}$$

where β is the propagation constant of the laser field in the laser waveguide; φ_{DBR1} and φ_{DBR2} are the phase shifts of the field upon reflection from the left and the right DBR mirrors, respectively, and q is the integral longitudinal mode number of the resonant cavity (not to be confused with the grating order of the DBR). Using (13.86) for the DBR phase shift, this resonance condition for the longitudinal mode frequencies of a DBR laser with DBR mirrors on both ends can be expressed as

$$2\beta(\omega_q)l_{\text{eff}} = 2q\pi - \varphi_{\text{B1}} - \varphi_{\text{B2}} + 2\beta_{\text{B}}(l_{\text{DBR1}}^{\text{eff}} + l_{\text{DBR2}}^{\text{eff}}), \tag{13.90}$$

where

$$l_{\text{eff}} = l + l_{\text{DBR1}}^{\text{eff}} + l_{\text{DBR2}}^{\text{eff}} \tag{13.91}$$

is the effective phase length of the DBR laser cavity including the effect of the DBR mirrors. Note that all the parameters on the right-hand side of (13.90) are frequency-independent constants.

We see from (13.90) that the Bragg frequency $\nu_B = \omega_B/2\pi$ is not necessarily a longitudinal mode frequency of the DBR laser. Therefore, *a DBR laser does not generally oscillate exactly at the Bragg frequency of its DBR mirrors.* However, in normal situations when the gain spectrum of the semiconductor material peaks near the Bragg frequency with a bandwidth much broader than the spectral bandwidth of the DBR mirrors, the oscillating longitudinal mode frequency must be the one that is closest to the Bragg frequency because the cavity has the lowest loss at the Bragg frequency where the DBR reflectivity has a maximum value. Near the Bragg frequency, $\beta(\omega) = n_\beta\omega/c$ is approximately linearly proportional to the optical frequency. Therefore, the longitudinal mode spacing of a DBR laser cavity can be given approximately by

$$\Delta\nu_L = \frac{c}{2n_\beta l_{\text{eff}}}, \tag{13.92}$$

where n_β is the effective phase index of the mode field. Though a DBR laser cavity has multiple longitudinal modes similarly to a Fabry–Perot cavity, the DBR mirrors are much more frequency selective than the Fabry–Perot cavity mirrors. Therefore, if the DBR bandwidth $\Delta\nu_{\text{DBR}}$ is made sufficiently narrow, a DBR laser will oscillate in a single longitudinal mode at a frequency that is closest to the Bragg frequency ν_B.

To find the threshold gain coefficient of a DBR laser, we consider the balance of the gain with the loss of the laser. The optical gain for a pass through the laser cavity is Γgl because the length of the gain medium is l and the gain overlap factor is Γ. The loss for a pass is $\overline{\alpha}l - \ln\sqrt{R_1R_2}$, where $\overline{\alpha}l$ is the total distributed loss including that contributed by the scattering and absorption, but not the transmission, of the laser field in the DBRs and $-\ln\sqrt{R_1R_2}$ is the transmission loss of the DBRs. Therefore, for a DBR laser, $\overline{\alpha}$ is a weighted average of the distributed loss of the entire structure divided only by the length l. By equating the gain with the loss for the laser threshold, we then find that the threshold gain coefficient of a DBR semiconductor laser can be expressed in the form of (13.83) by taking $\alpha_{\text{out}} = -(\ln\sqrt{R_1R_2})/l$ like that of a Fabry–Perot laser but by using the reflectivities of the Bragg mirrors at the oscillating laser frequency for R_1 and R_2. Note that l_{eff} defined in (13.91) is used to find $\Delta\nu_L$ but is not used to evaluate α_{out} for a DBR laser because l_{eff} is an effective *phase* length, which only determines the phase shift of the laser field but does not determine the amplification or attenuation of the laser intensity.

EXAMPLE 13.15 An InGaAsP DBR laser consists of a gain section of a length $l = 300\ \mu\text{m}$ and two identical DBRs as end mirrors, each of a length $l_{\text{DBR}} = 150\ \mu\text{m}$. The Bragg wavelength of the DBRs is $\lambda_B = 1.530\,00\ \mu\text{m}$. The effective indices for the laser modes are taken to be $n_\beta = N_\beta = 3.45$. The gain overlap factor is $\Gamma = 0.4$. The DBR coupling

coefficient is $|\kappa| = 50\ \text{cm}^{-1}$. The laser has a distributed loss of $\overline{\alpha} = 40\ \text{cm}^{-1}$, which includes the contributions from the DBRs and scattering at the junctions between the gain section and the DBR sections. (a) Find the peak reflectivity and the bandwidth of the two identical DBRs. (b) Find the effective phase length of the DBRs and that of the DBR laser to determine the longitudinal mode spacing of the laser. (c) How many longitudinal modes fall within the DBR bandwidth? If the laser is pumped in such a way that only one longitudinal mode oscillates, what is its wavelength? (d) What is the threshold gain coefficient of the oscillating mode? (e) If the gain medium has a gain cross section of $\sigma = 3 \times 10^{-20}\ \text{m}^2$, what is the required carrier density above transparency for the laser to reach its threshold?

Solution (a) For $|\kappa| = 50\,\text{cm}^{-1} = 5000\,\text{m}^{-1}$ and $l_{\text{DBR}} = 150\ \mu\text{m}$, we have $|\kappa| l_{\text{DBR}} = 0.75$. Therefore, the DBR peak reflectivity at the Bragg wavelength is

$$R_{\text{DBR}} = \tanh^2 |\kappa| l_{\text{DBR}} = \tanh^2 0.75 = 0.403.$$

The bandwidth of the Bragg reflectors is

$$\Delta\nu_{\text{DBR}} \approx \frac{|\kappa| c}{\pi N_\beta} = \frac{5000 \times 3 \times 10^8}{\pi \times 3.45}\ \text{Hz} = 138.4\ \text{GHz}.$$

(b) The effective phase length, $l_{\text{DBR}}^{\text{eff}} = l_{\text{DBR1}}^{\text{eff}} = l_{\text{DBR2}}^{\text{eff}}$, of both DBRs is

$$l_{\text{DBR}}^{\text{eff}} = \frac{R_{\text{DBR}}^{1/2}}{2|\kappa|} = \frac{0.403^{1/2}}{2 \times 5000}\ \text{m} = 63.5\ \mu\text{m}.$$

Thus, the effective phase length of the laser is

$$l_{\text{eff}} = l + 2l_{\text{DBR}}^{\text{eff}} = 427\ \mu\text{m}.$$

We then find the following longitudinal mode spacing:

$$\Delta\nu_{\text{L}} = \frac{c}{2n_\beta l_{\text{eff}}} = \frac{3 \times 10^8}{2 \times 3.45 \times 427 \times 10^{-6}}\ \text{Hz} = 101.8\ \text{GHz}.$$

(c) Because $2\Delta\nu_{\text{L}} > \Delta\nu_{\text{DBR}} > \Delta\nu_{\text{L}}$, there is at least one longitudinal mode, but at most two modes, within the reflector bandwidth. Whether one or two modes fall within the DBR bandwidth depends on where the mode frequencies are located with respect to the Bragg frequency. To answer this question, we need to find the mode number q for the longitudinal mode frequency ν_q that is closest to ν_{B}. Using (13.90), we find that the phase mismatch for a mode at ν_q can be expressed as

$$\delta_q = -\beta(\omega_q) + \beta_{\text{B}} = \frac{1}{2l_{\text{eff}}}\left[2\beta_{\text{B}} l - (2q-1)\pi\right].$$

The mode frequency that is closest to ν_{B} is found by finding the number q that minimizes the value of $|\delta_q|$. Using $n_\beta = 3.45$ and $\lambda_{\text{B}} = 1.530\,00\ \mu\text{m}$ for $\beta_{\text{B}} = 2\pi n_\beta/\lambda_{\text{B}}$, we find

that the value of $|\delta_q|$ is minimized with $q = 1353$ for $\delta = 32.46\ \text{cm}^{-1} = 3246\ \text{m}^{-1}$. Thus

$$\nu - \nu_B = -\frac{c\delta}{2\pi n_\beta} = -\frac{3 \times 10^8 \times 3246}{2 \times \pi \times 3.45}\ \text{Hz} = -44.9\ \text{GHz}.$$

For the two neighboring modes, corresponding to $q + 1 = 1354$ and $q - 1 = 1352$, the $q + 1 = 1354$ mode falls within the DBR bandwidth because $|\nu_{q+1} - \nu_B| = |-44.9 + 101.8|\ \text{GHz} < \Delta\nu_{DBR}/2 = 69.2\ \text{GHz}$, but the $q - 1 = 1352$ mode falls outside the DBR bandwidth because $|\nu_{q-1} - \nu_B| = |-44.9 - 101.8|\ \text{GHz} > \Delta\nu_{DBR}/2 = 69.2\ \text{GHz}$. Therefore, there are two modes that fall within the DBR bandwidth. If the laser is pumped right at the threshold so that only one mode oscillates, the $q = 1353$ mode will oscillate because it has the smallest phase mismatch, thus the lowest threshold. Its wavelength

$$\begin{aligned}\lambda &= \frac{c}{\nu} = \frac{c\lambda_B}{c + (\nu - \nu_B)\lambda_B} \\ &= \frac{3 \times 10^8 \times 1.53 \times 10^{-6}}{3 \times 10^8 - 44.9 \times 10^9 \times 1.53 \times 10^{-6}}\ \text{m} \\ &= 1.530\,35\ \mu\text{m}.\end{aligned}$$

(d) With $|\kappa| = 50\ \text{cm}^{-1}$, $\delta = 32.46\ \text{cm}^{-1}$, and $l_{DBR} = 150\ \mu\text{m}$, we have $|\kappa|l = 0.75$ and $|\delta/\kappa|^2 = (32.46/50)^2$. By using (4.91) for the contradirectional coupling efficiency in the presence of phase mismatch, the DBR reflectivity at the oscillating mode frequency can be found as

$$R = \frac{\sinh^2\left(|\kappa|l\sqrt{1 - |\delta/\kappa|^2}\right)}{\cosh^2\left(|\kappa|l\sqrt{1 - |\delta/\kappa|^2}\right) - |\delta/\kappa|^2} = 0.385,$$

which is somewhat smaller than R_{DBR} because the mode frequency does not fall right at ν_B. Taking $R_1 = R_2 = R = 0.385$, the output coupling loss for the DBR laser is

$$\alpha_{out} = -\frac{\ln\sqrt{R_1 R_2}}{l} = -\frac{\ln 0.385}{300}\ \mu\text{m}^{-1} = 3.18 \times 10^{-3}\ \mu\text{m}^{-1} = 31.8\ \text{cm}^{-1}.$$

Thus the threshold gain coefficient

$$g_{th} = \frac{\overline{\alpha} + \alpha_{out}}{\Gamma} = \frac{40 + 31.8}{0.4}\ \text{cm}^{-1} = 179.5\ \text{cm}^{-1}.$$

(e) For $g_{th} = 179.5\ \text{cm}^{-1} = 1.795 \times 10^4\ \text{m}^{-1}$ with $\sigma = 3 \times 10^{-20}\ \text{m}^2$, the threshold carrier density above transparency is

$$N_{th} - N_{tr} = \frac{g_{th}}{\sigma} = \frac{1.795 \times 10^4}{3 \times 10^{-20}}\ \text{m}^{-3} = 5.98 \times 10^{23}\ \text{m}^{-3}.$$

For this laser, there are two modes within the DBR bandwidth. In this particular case, the threshold of the second mode is not much higher than the first mode considered above

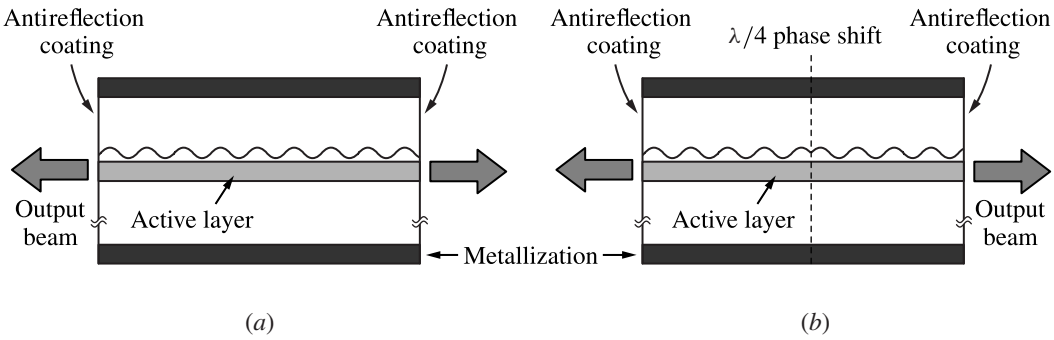


Figure 13.32 Structures of edge-emitting distributed feedback (DFB) semiconductor lasers: (*a*) DFB laser with no phase shift and (*b*) $\lambda/4$ phase-shifted DFB laser.

(see Problem 13.9.4). Therefore, this DBR laser can possibly oscillate in two modes. This problem can be avoided by proper design of a DBR laser (see Problems 13.9.3 and 13.9.5).

Distributed feedback lasers

In a DFB laser, the grating is placed right next to the waveguiding layer along the length of the active region, as shown in Fig. 13.32(*a*). We learn from the discussions in Section 5.1 that it can be placed either above or below the active layer for the same effect. This grating provides all of the optical feedback for laser oscillation. The end facets of a DFB laser are coated with antireflection coating to eliminate any reflection from the facets. Because the grating in a DFB laser runs along the length of the active region where optical gain exists, it does not function as a simple passive reflector like that in a DBR laser but rather as a frequency-selective contradirectionally coupled amplifier for the intracavity laser field. Consequently, the characteristics of a DFB laser are more complicated than those of a DBR laser.

A DFB laser also has multiple longitudinal modes whose frequencies are still determined by the basic requirement that the round-trip phase shift be an integral multiple of 2π. Meanwhile, its threshold gain coefficient is still subject to the relation in (13.83) under the requirement that the gain and loss of an oscillating laser mode exactly balance each other. Because of the distributed nature of the optical feedback in a DFB laser, however, it is not feasible to apply these concepts by simply following the round-trip propagation of a laser field as is done for Fabry–Perot lasers and ring lasers. Instead, we have to consider coupling of the contrapropagating fields in a DFB laser cavity by using the concepts discussed in Sections 4.3 and 5.1.

A DFB laser that has a continuous grating across its entire length l and perfect antireflection coating on its end facets, as shown in Fig. 13.32(*a*), is considered. This structure is basically a DBR. From the discussions in Sections 4.3 and 5.1, we know that its complex reflection coefficient can be found by replacing κ_{ab} with κ and κ_{ba} with

κ^* in (4.72) for α_c and in (4.78) for r, where $\kappa = \kappa_{ab}(q)$ as defined in (5.8) and (5.10). Thus, we have

$$r = \frac{i\kappa^* \sinh \alpha_c l}{\alpha_c \cosh \alpha_c l + i\delta \sinh \alpha_c l}, \tag{13.93}$$

where

$$\alpha_c = (|\kappa|^2 - \delta^2)^{1/2}. \tag{13.94}$$

A DFB laser differs from a passive DBR in that an optical field propagating in a DFB laser sees an optical gain coefficient of Γg and a distributed loss of $\overline{\alpha}$ just like a field in any laser. To account for the effects of such gain to and loss from the laser medium, the laser field at a frequency ω has a complex propagation constant β of the following form:

$$\beta = \frac{n_\beta \omega}{c} - i\frac{\Gamma g - \overline{\alpha}}{2}. \tag{13.95}$$

Then, from (5.21) and (5.22), we have the following frequency-dependent phase mismatch:

$$\delta = -\beta(\omega) + \beta_B = -\frac{n_\beta \Delta\omega}{c} + i\frac{\Gamma g - \overline{\alpha}}{2}, \tag{13.96}$$

where $\Delta\omega = \omega - \omega_B$ and $\omega_B = 2\pi\nu_B$. Note that α_c given in (13.94) is frequency dependent because of the frequency dependence of δ, and α_c is complex when $\Delta\omega \neq 0$. Clearly, the reflection coefficient r given in (13.93) is highly frequency dependent.

The oscillation condition of a DFB laser can be found by considering the fact that when a laser mode is oscillating, there is a laser output without an optical input at that particular mode frequency. This condition is met when $r = \infty$ at the oscillating mode frequency. Because the numerator of r in (13.93) is always finite for finite values of $|\kappa|$, δ, and l, the oscillation condition of a DFB laser is found by setting its denominator to zero:

$$\alpha_c \cosh \alpha_c l + i\delta \sinh \alpha_c l = 0. \tag{13.97}$$

This oscillation condition for a DFB laser can be transformed to the following simple form (see Problem 13.9.6(a)):

$$|\kappa| \sinh \alpha_c l = i\alpha_c. \tag{13.98}$$

The longitudinal mode frequencies and the threshold gain coefficient of each mode for a DFB laser of a given $|\kappa|l$ value can be found by solving the oscillation condition in (13.97) or, equivalently, that in (13.98).

The complex transcendental equations in (13.97) and (13.98) have no simple analytical solutions. Some of their characteristics can be obtained from approximate solutions, but accurate solutions must be obtained numerically. It can be shown analytically, however, that the Bragg frequency ν_B is not a longitudinal mode of the DFB laser and that

the longitudinal modes are symmetrically distributed on both sides of ν_B (see Problem 13.9.6(b)). By solving (13.97), or (13.98), for a longitudinal mode frequency of a DFB laser that has a given $|\kappa|l$ value, the value of $(\Gamma g_{th} - \overline{\alpha})l = \alpha_{out}l$ at the oscillation threshold of that particular mode is obtained simultaneously.

Numerical solutions show that the longitudinal modes have the following longitudinal mode frequencies:

$$\nu_q \approx \nu_B \pm (q + \mu)\frac{c}{2n_\beta l}, \tag{13.99}$$

where $q = 0, 1, 2, \ldots$ is an integral mode number for the DFB laser modes (not to be confused with the grating order) and μ is a constant that is a function of $|\kappa|l$. The longitudinal mode spacing of a DFB laser is approximately, but not exactly,

$$\Delta\nu_L \approx \frac{c}{2n_\beta l}, \tag{13.100}$$

which is similar to that of a Fabry–Perot laser of an effective index of n_β and a cavity length of l. *The optical gain required for a DFB laser to oscillate is lowest at the Bragg frequency, but a DFB laser that does not have a structural phase shift in its grating does not oscillate at its Bragg frequency because ν_B is not one of its mode frequencies.* In an ideal situation, the two lowest-order frequencies on the two sides of ν_B, corresponding to $q = 0$, have the same lowest oscillation threshold. Therefore, in a normal operating condition, the spectral feature of a DFB laser often consists of the two longitudinal modes at

$$\nu = \nu_B \pm \frac{\mu c}{2n_\beta l} = \nu_B \pm \mu\Delta\nu_L, \tag{13.101}$$

which have the following two wavelengths:

$$\lambda \approx \lambda_B \pm \frac{\mu\lambda_B^2}{2n_\beta l}. \tag{13.102}$$

Clearly, there is a *stop band* that is centered at the Bragg frequency between these two fundamental mode frequencies:

$$\Delta\nu_{SB} = \frac{\mu c}{n_\beta l} = 2\mu\Delta\nu_L. \tag{13.103}$$

Figure 13.33 shows the numerically solved value of $\mu = \Delta\nu_{SB}/2\Delta\nu_L$ and the output coupling loss $\alpha_{out}l = (\Gamma g_{th} - \overline{\alpha})l$ of a DFB laser at the threshold of its fundamental mode frequencies, both as a function of the value of $|\kappa|l$. We see that $\mu \to 1/2$ for $|\kappa|l \to 0$, but $\mu > 1/2$ for $|\kappa|l \neq 0$.

Figure 13.34 shows two longitudinal mode spectra of a DFB laser for $|\kappa|l = 1.5$ and $|\kappa|l = 1$, respectively, when the laser oscillates at its fundamental mode frequencies. This spectrum is obtained by plotting $R = |r|^2$ as a function of the frequency difference $\Delta\nu = \nu - \nu_B$ normalized to the mode spacing $\Delta\nu_L$.

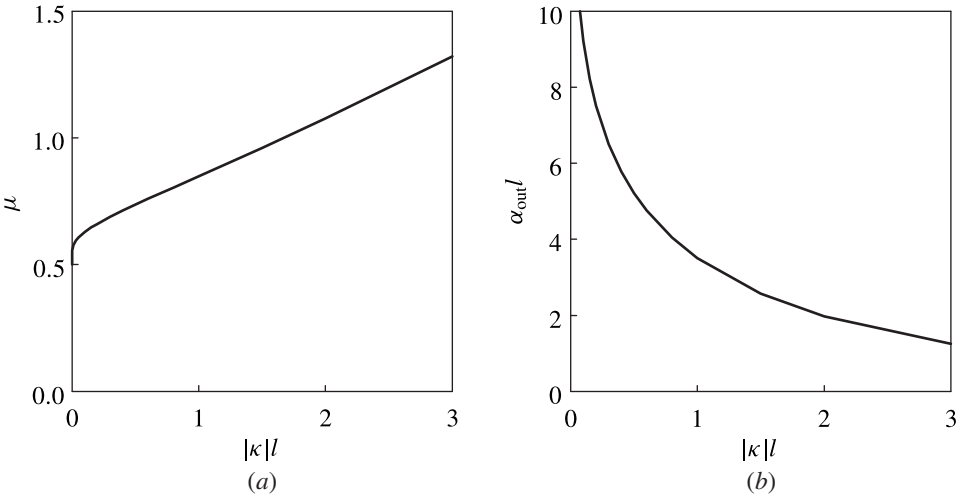


Figure 13.33 (*a*) Value of $\mu = \Delta\nu_{SB}/2\Delta\nu_L$, which defines the stop band and the fundamental mode frequencies, and (*b*) value of $\alpha_{out}l = (\Gamma g_{th} - \overline{\alpha})l$, which defines the fundamental mode threshold, as a function of the value of $|\kappa|l$ for a non-phase-shifted DFB laser.

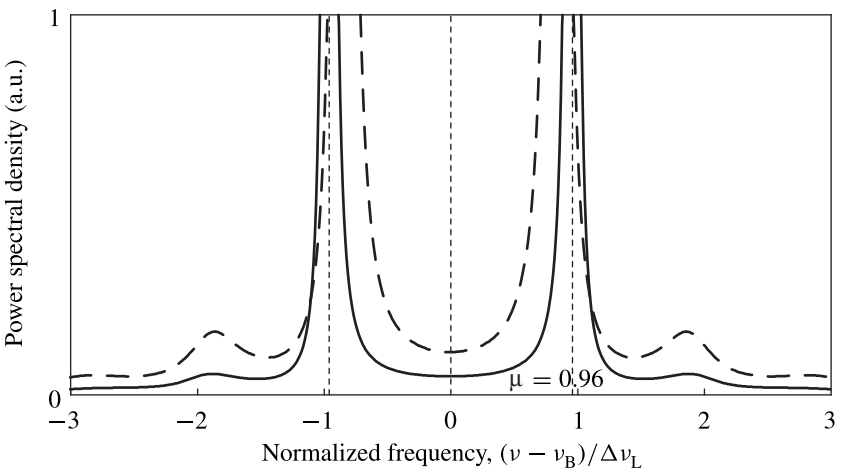


Figure 13.34 Longitudinal mode spectra of a non-phase-shifted DFB laser of $|\kappa|l = 1.5$ (solid curve) and $|\kappa|l = 1$ (dashed curve) when the laser oscillates at its fundamental mode frequencies.

The symmetry of the two lowest-order modes for a DFB laser can be upset by introducing a fixed phase shift in the DFB structure. A phase shift at either end facet, whether introduced intentionally by selective coating or unintentionally by the cleaving process, can remove the degeneracy to result in single-mode, or quasi-single-mode, oscillation. The most effective is to incorporate a $\pi/2$ phase shift in the grating. In such a so-called $\lambda/4$-shifted DFB laser as shown in Fig. 13.32(*b*), a longitudinal mode appears exactly at the Bragg frequency ν_B where the cavity has the lowest loss and the laser has the corresponding lowest threshold. Consequently, *a* $\lambda/4$ *phase-shifted*

DFB laser oscillates in a single longitudinal mode at ν_B. The characteristics of the $\lambda/4$ phase-shifted DFB laser are considered in Problems 13.9.8 and 13.9.9.

EXAMPLE 13.16 An InGaAsP DFB laser has its grating fabricated along its gain section, which has a length of $l = 300$ μm. There is no intentional or unintentional structural phase shift in the grating. It has the same $l = 300$ μm length of the gain section as that of the DBR laser considered in Example 13.15. Its grating length of 300 μm is also the same as the total grating length of the two DBRs of the DBR laser. Most of the parameters of this DFB laser are the same as those of the DBR laser described in Example 13.15 with $\lambda_B = 1.530\,00$ μm, $n_\beta = N_\beta = 3.45$, $\Gamma = 0.4$, and $|\kappa| = 50\ \text{cm}^{-1}$. It has a distributed loss of $\overline{\alpha} = 10\ \text{cm}^{-1}$, which is smaller than that of the DBR laser because the DBR laser has additional losses contributed by the external DBRs. (a) Find the longitudinal mode spacing and the stop band of the DFB laser. (b) How many longitudinal modes will oscillate if the laser is pumped to its lowest threshold? What are their wavelengths? (c) What is the threshold gain coefficient of the oscillating modes? (d) If the gain medium has a gain cross section of $\sigma = 3 \times 10^{-20}\ \text{m}^2$, what is the required carrier density above transparency for this laser to reach its threshold? Compare the characteristics of this DFB laser to those of the DBR laser in Example 13.15 while answering these questions.

Solution (a) The longitudinal mode spacing

$$\Delta\nu_L \approx \frac{c}{2n_\beta l} = \frac{3 \times 10^8}{2 \times 3.45 \times 300 \times 10^{-6}}\ \text{Hz} = 144.9\ \text{GHz}.$$

For this laser, we have $|\kappa|l = 1.5$. From Fig. 13.33, we find that $\mu = 0.96$ for $|\kappa|l = 1.5$. Thus, the stop band

$$\Delta\nu_{SB} = 2\mu\Delta\nu = 278.2\ \text{GHz}.$$

The longitudinal mode spacing $\Delta\nu_L$ of this DFB laser is much larger than that of the DBR laser considered in Example 13.15 because of the larger effective phase length of the DBR laser. The spacing between the two fundamental longitudinal modes of the DFB laser is $\Delta\nu_{SB}$, which is even larger than $\Delta\nu_L$.

(b) Because there is no structural phase shift in the grating, both fundamental modes have the same threshold. Therefore, both of them should oscillate when the laser is pumped to reach its lowest threshold. The wavelengths of these two modes are

$$\lambda \approx \lambda_B \pm \frac{\mu\lambda_B^2}{2n_\beta l} = \left(1.53 \pm \frac{0.96 \times 1.53^2}{2 \times 3.45 \times 300}\right)\ \mu\text{m} = (1.53 \pm 1.09 \times 10^{-3})\ \mu\text{m}.$$

Thus we find two wavelengths at 1.528 91 and 1.531 09 μm. Both of these two modes will oscillate once the DFB laser reaches its threshold because both of them have the same threshold. In comparison, the DBR laser oscillates in only one wavelength at its

threshold, though two modes fall in the DBR bandwidth and the second mode may have a threshold only slightly higher than the first.

(c) From Fig. 13.33, we find that $\alpha_{out}l = 2.574$ at the DFB laser threshold for $|\kappa|l = 1.5$. For $l = 300$ μm, we have

$$\alpha_{out} = \frac{2.574}{300}\ \mu m^{-1} = 8.58 \times 10^{-3}\ \mu m^{-1} = 85.8\ cm^{-1}.$$

Therefore,

$$g_{th} = \frac{\overline{\alpha} + \alpha_{out}}{\Gamma} = \frac{10 + 85.8}{0.4}\ cm^{-1} = 240\ cm^{-1}.$$

This threshold gain coefficient is higher than that of the DBR laser despite the fact that the DBR laser has a much larger distributed loss than this DFB laser.

(d) For $g_{th} = 240\ cm^{-1} = 2.4 \times 10^4\ m^{-1}$ with $\sigma = 3 \times 10^{-20}\ m^2$, the threshold carrier density above transparency is

$$N_{th} - N_{tr} = \frac{g_{th}}{\sigma} = \frac{2.4 \times 10^4}{3 \times 10^{-20}}\ m^{-3} = 8 \times 10^{23}\ m^{-3}.$$

The threshold carrier density above transparency for this DFB laser is higher than that for the DBR laser because of the higher threshold gain coefficient for the DFB laser.

Clearly, a DFB laser without a structural phase shift in its grating has a high threshold and oscillates in two modes. These characteristics are inferior to those of a similar DBR laser. A DFB laser with a proper structural phase shift, such as the $\lambda/4$-shifted DFB laser, has a lower threshold with only one oscillating mode (see Problem 13.9.10). It is then competitive to the DBR laser in performance and indeed is favored over the DBR laser because of its simpler and shorter structure than the DBR laser structure.

Surface-emitting lasers

The common feature for all surface-emitting lasers irrespective of their structures is that the laser output is emitted in a direction perpendicular to the semiconductor substrate. In comparison to edge-emitting lasers, a unique advantage of surface-emitting lasers is that they can be made in a two-dimensional array on a common substrate, which is very useful for applications in parallel optical interconnects and parallel optical signal processing. Nevertheless, a surface-emitting laser can be packaged separately and used as a discrete laser as well.

The cavity of a surface-emitting laser can be a horizontal cavity, a vertical cavity, or a folded cavity. Each type of cavity can have different variations. Different cavity structures lead to different characteristics and different applications for the surface-emitting lasers. Considering only the basic structures, there are three kinds of surface-emitting semiconductor lasers: the *folded-cavity surface-emitting laser* (FCSEL), the *grating-coupled surface-emitting laser* (GCSEL), and the *vertical-cavity surface-emitting laser*

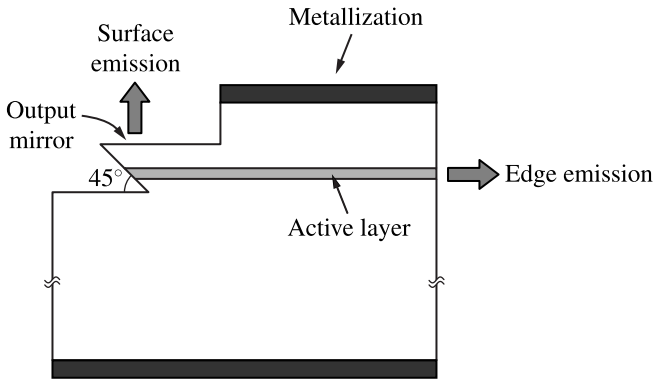


Figure 13.35 Structure of a folded-cavity surface-emitting laser (FCSEL).

(VCSEL). Both FCSELs and GCSELs are typically stripe-geometry lasers similar to the edge-emitting lasers, but VCSELs are quite different from all of them.

Folded-cavity surface-emitting lasers

A FCSEL can be constructed by modifying the output-coupling geometry of an edge-emitting laser with the addition of 45° internal total-reflection mirrors to direct the laser output to the surface-emitting direction, as illustrated in Fig. 13.35. A 45° semiconductor facet serves as an internal total-reflection mirror for the intracavity laser field because the critical angle for internal reflection at a semiconductor–air interface is much smaller than 45°, due to the large refractive index of a semiconductor. In principle, each of the three concepts for edge-emitting lasers can be used for a FCSEL.

Except for the advantages associated with its surface-emitting geometry, the general characteristics of a FCSEL are similar to those of a corresponding edge-emitting laser based on the same concept. Uncoated semiconductor surfaces are sufficient as surface-emitting output mirrors for a Fabry–Perot FCSEL. For a DBR FCSEL, high-reflection DBR surface-facing mirrors for output coupling can be constructed with alternating thin layers of semiconductors that have different compositions, thus different refractive indices. Such multilayer DBR reflectors are also used in the VCSELs discussed below. Because these vertically alternating layers are parallel to the substrate surface, they can be fabricated using crystal growth technology with more ease and control than a horizontal grating.

Grating-coupled surface-emitting lasers

The concept of grating surface output coupling discussed in Section 5.3 can be applied to a horizontal-cavity semiconductor laser for vertical emission through grating coupling. Because of the use of grating coupling, a GCSEL is normally based on a DBR laser.

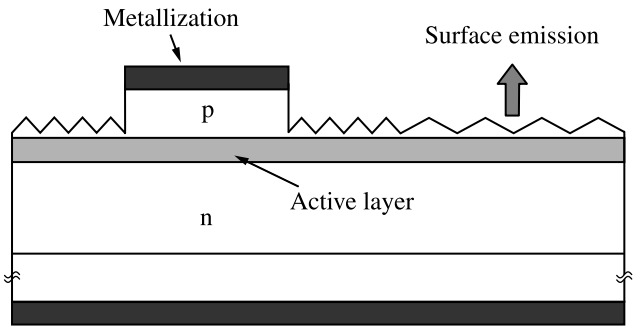


Figure 13.36 Structure of a grating-coupled surface-emitting laser (GCSEL).

It is possible to use the same set of gratings for both optical feedback and surface output coupling. For better control of the output beam characteristics, however, separate gratings are often used to serve the functions of optical feedback and output coupling, respectively. Figure 13.36 shows a GCSEL that is a DBR laser with two first-order DBR mirrors and a separate section of second-order grating for surface output coupling. The surface output-coupling grating does not provide optical feedback and thus does not participate in laser oscillation.

GCSELs have a few unique features because of their structures. As a DBR laser, a GCSEL oscillates in a single longitudinal mode close to the Bragg frequency of its grating. Sophisticated two-dimensional geometry for the output-coupling grating can be used for output beam shaping and control. A curved or circular output-coupling grating can be used to emit a collimated output laser beam with a very small divergence. A large grating can be used to increase the output-coupling efficiency for a high-power laser. To concentrate most of the output power into one surface-emitting beam, a blazed grating can be used to reduce the emission in the substrate direction. On the other hand, as we have learned in Section 5.3, it is also possible to choose a grating for multiple output beams emitting in different directions if desired.

Vertical-cavity surface-emitting lasers

Uniquely among all edge-emitting and surface-emitting lasers, the resonant cavity of a VCSEL is formed in the direction perpendicular to the junction plane and the substrate. The uniqueness of a VCSEL is that it has a very short cavity made possible by its vertical orientation. This feature has several important implications for the structure and the performance characteristics of a VCSEL. Because of the short cavity of a VCSEL, it is required that its gain section be thin but highly efficient and its mirror reflectivities be high in order for the VCSEL to function. A thin and efficient gain section is achieved by using quantum wells. High mirror reflectivities are achieved by using semiconductor DBRs, such as AlAS/GaAs DBRs, or dielectric DBRs, such as SiO_2/TiO_2 or Si/Al_2O_3 DBRs. A metal layer can also be deposited on top or below to

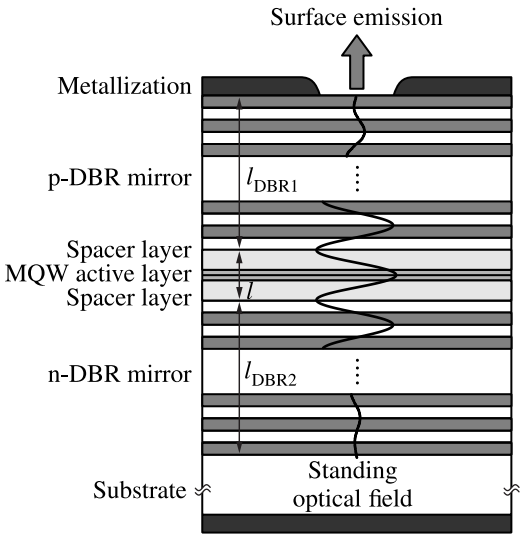


Figure 13.37 Structure of a vertical-cavity surface-emitting laser (VCSEL).

increase the reflectivity of a DBR. Consequently, a VCSEL is normally a QW DBR laser. Figure 13.37 shows the basic structure of a VCSEL. A typical VCSEL contains a very thin active region of one to four quantum wells, each of which has a typical thickness of 5–10 nm. This active region is sandwiched between two spacer layers. Optical feedback is provided by monolithically integrated DBRs.

The DBR mirrors of a VCSEL are index-modulation gratings containing thin layers of alternating compositions and refractive indices. Because of their vertical stacking geometry, the thickness and composition of each layer, as well as the sharp transition between neighboring layers, can be precisely controlled using advanced fabrication technology. Such a square index grating with alternating layers of indices of $\overline{n} \pm \Delta n/2$ has an index step of Δn and an average index of $\overline{n}$ between two neighboring layers. It can be shown that a first-order square grating of a 50% duty factor has the largest coupling coefficient with a magnitude given by (see Problem 13.9.12)

$$|\kappa| = \frac{2\Delta n}{\lambda}, \tag{13.104}$$

where λ is the optical wavelength in free space. The period of a first-order DBR is $\Lambda = \lambda_B/2\overline{n}$ for $q = 1$ in (5.23). Therefore, the thickness of each alternating layer for the first-order square grating of a 50% duty factor is $\lambda_B/4\overline{n}$. The physical length of such a DBR is

$$l_{DBR} = N_{DBR}\Lambda = \frac{N_{DBR}\lambda_B}{2\overline{n}}, \tag{13.105}$$

where N_{DBR} is the number of pairs of alternating layers that define the grating periods. The value of N_{DBR} can only be an integer or half-integer because there is an integral

number of layers. The number of pairs for a semiconductor DBR ranges between 10 and 40, but that for a dielectric DBR is typically less than 10, sometimes only a few.

The peak reflectivity of the DBR at the Bragg wavelength can be expressed as

$$R_{\mathrm{DBR}} = \tanh^2 |\kappa| l_{\mathrm{DBR}} = \tanh^2 \frac{N_{\mathrm{DBR}} \Delta n}{\overline{n}}. \tag{13.106}$$

The DBR reflectivities for VCSELs are required to be very high, normally higher than 98% but often higher than 99%. The two DBRs for a VCSEL are not symmetric. The top DBR has a lower reflectivity to accommodate the output coupling from the top of the laser, whereas the bottom DBR normally has a very high reflectivity as close to 100% as possible to compensate for the loss of the top DBR and to direct all laser energy to the output window on top. These DBRs have very large bandwidths because of their high reflectivities and small lengths that lead to correspondingly large values for $|\kappa|$. Because current is injected vertically through these DBR mirrors, the top DBR is heavily p doped and the bottom DBR is heavily n doped for high conductivity.

In a VCSEL, the laser field propagates perpendicularly to the active layer. This feature has profound implications for the structure and characteristics of a VCSEL. As seen in Fig. 13.37, the cavity length l of a VCSEL is defined by the active layer and two spacer layers around the active layer. From (13.82), the gain overlap factor of a VCSEL is

$$\Gamma = a \frac{M_{\mathrm{QW}} d_{\mathrm{QW}}}{l}. \tag{13.107}$$

The factor a in (13.107) has a value between 1 and 2 that depends on the overlap of the active layer with the field pattern in the cavity. Because the intracavity laser field forms a standing wave in the direction perpendicular to the active layer, it is important to locate the active layer at a crest of the standing wave so that the optical gain in the active layer is most efficiently used for stimulated amplification of the laser field. Furthermore, to ensure that the entire active layer stays within the high-intensity crest region of the standing wave, the thickness d of the active layer, which consists of the quantum wells and the barrier layers between wells, is normally limited to one-quarter of the wavelength in the medium, $\lambda/4n$, where n is the refractive index in the laser cavity and is normally different from $\overline{n}$ of either of the two DBRs. In the case when $d \ll \lambda/4n$, the factor a has a value of 2 if the quantum wells are located properly at a crest of the standing-wave pattern. For a GaAs VCSEL emitting at $\lambda = 850$ nm, the thickness of its active layer is thus restricted to 60 nm or less because $n \approx 3.65$. The overlap between the intracavity field and the gain region, which defines the gain overlap factor Γ of a laser, is independent of the cross section of the cavity but increases as the length of the cavity is reduced. It is advantageous from the standpoint of reducing the laser threshold and increasing the laser efficiency that a VCSEL have a very short cavity. The spacer layers are required, however, for two reasons. First, they are required for the standing-wave pattern to form in the cavity with a peak located at the active layer. Second, they can be tailored to guide current injection into the active layer, thus

improving the efficiency of carrier injection and reducing the laser threshold. These characteristics are illustrated in Fig. 13.37.

For a VCSEL, the total length l for the active layer and the spacer layers between the two DBR mirrors is typically on the order of one or a few optical wavelengths in the medium, depending on the thicknesses of the spacer layers. For example, a 1λ cavity has $l = \lambda/n$, and a 3λ cavity has $l = 3\lambda/n$. In order to maintain a high Q for such a short cavity, the reflectivity of the DBR mirrors has to be very high. The required reflectivity is higher than 99% for a VCSEL with a single quantum well but can be slightly lower for a VCSEL with multiple quantum wells. Thus, a VCSEL is a microcavity QW laser with relatively thick DBR mirrors.

The general characteristics of a DBR laser described earlier are still valid for a VCSEL. The longitudinal mode spacing is given by (13.92), and the threshold gain coefficient is given by (13.83) with $\alpha_{\text{out}} = -(\ln \sqrt{R_1 R_2})/l$. In a VCSEL, both $l_{\text{DBR1}}^{\text{eff}}$ and $l_{\text{DBR2}}^{\text{eff}}$ can be larger than l, but the total effective phase length of $l_{\text{eff}} = l + l_{\text{DBR1}}^{\text{eff}} + l_{\text{DBR2}}^{\text{eff}}$ is still on the order of 1–2 μm. Therefore, the longitudinal mode spacing is very large. It can be either larger or smaller than the bandwidths of the highly reflective DBRs, but it can be easily made comparable to, or even larger than, the entire gain bandwidth of a semiconductor quantum well. Consequently, *a microcavity VCSEL inherently oscillates in a single longitudinal mode*.

The requirement for the confinement of electric current and optical field in a VCSEL is fulfilled by controlling the transverse dimension and geometry of the vertical cavity. Symmetric transverse geometry of circular or square shape is normally used for a VCSEL so that its emission has a nice, round pattern with a symmetric beam divergence, making its coupling to an optical fiber easy and efficient. The transverse dimension of the cavity is typically in the range of 3–10 μm in diameter. One significant problem associated with the symmetric optical guiding structure is polarization instability in the VCSEL emission because it does not provide any polarization discrimination. For polarization control, anisotropy in the structure or in the gain medium has to be introduced.

EXAMPLE 13.17 A GaAs/AlAs QW VCSEL is designed to emit at 850 nm wavelength. It consists of two first-order AlAs/$Al_{0.2}Ga_{0.8}As$ DBR mirrors of a 50% duty factor. The p-side top DBR has 21 pairs of alternating quarter-wavelength AlAs and $Al_{0.2}Ga_{0.8}As$ layers, and the n-side bottom DBR has 24 pairs. The active layer consists of three 8-nm wide GaAs quantum wells separated by 4-nm wide $Al_{0.2}Ga_{0.8}As$ barriers, embedded in a 1λ cavity with $Al_{0.2}Ga_{0.8}As$ spacer layers. The refractive indices at 850 nm are $n = 3.65$ for GaAs, $n = 3.00$ for AlAs, and $n = 3.52$ for $Al_{0.2}Ga_{0.8}As$. (a) Find the coupling coefficient and the reflectivities of the DBRs. (b) Find the length of the cavity and those of the DBRs. What is the total length of the device? (c) If the device has a distributed loss of $\overline{\alpha} = 18\ \text{cm}^{-1}$, what is its threshold gain coefficient? (d) Find the longitudinal mode spacing. Compare it to the DBR bandwidth and the gain bandwidth.

Solution (a) For the first-order AlAs/$Al_{0.2}Ga_{0.8}As$ DBRs, we have $\Delta n = 0.52$ and $\overline{n} = 3.26$ because $n = 3.00$ for AlAs and $n = 3.52$ for $Al_{0.2}Ga_{0.8}As$. Therefore, for $\lambda = \lambda_B = 850\ \text{nm} = 0.85\ \mu\text{m}$, we have

$$|\kappa| = \frac{2\Delta n}{\lambda} = \frac{2 \times 0.52}{0.85}\ \mu\text{m}^{-1} = 1.224\ \mu\text{m}^{-1}.$$

We have $N_{\text{DBR1}} = 21$ and $N_{\text{DBR1}} = 24$. Thus,

$$R_{\text{DBR1}} = \tanh^2 \frac{N_{\text{DBR1}}\Delta n}{\overline{n}} = \tanh^2 \frac{21 \times 0.52}{3.26} = 99.5\%$$

and

$$R_{\text{DBR2}} = \tanh^2 \frac{N_{\text{DBR2}}\Delta n}{\overline{n}} = \tanh^2 \frac{24 \times 0.52}{3.26} = 99.8\%.$$

The bottom DBR has a higher reflectivity because it has three more pairs than the top DBR.

(b) The cavity consists mostly of $Al_{0.2}Ga_{0.8}As$, which has $n = 3.52$. Because it is a 1λ cavity, its length

$$l = \frac{\lambda}{n} = \frac{850}{3.52}\ \text{nm} = 241.5\ \text{nm}.$$

For the DBRs, the average index is $\overline{n} = 3.26$. Therefore,

$$l_{\text{DBR1}} = N_{\text{DBR1}}\Lambda = 21 \times \frac{850}{2 \times 3.26}\ \text{nm} = 2737.7\ \text{nm}$$

and

$$l_{\text{DBR2}} = N_{\text{DBR2}}\Lambda = 24 \times \frac{850}{2 \times 3.26}\ \text{nm} = 3128.8\ \text{nm}.$$

The total length of the structure is $l_{\text{device}} = l + l_{\text{DBR1}} + l_{\text{DBR2}} = 6108\ \text{nm} = 6.108\ \mu\text{m}$. We see that the DBRs occupy 96% of the device structure.

(c) The active layer consists of three quantum wells and two barriers with a total thickness of $d = 3 \times 8\ \text{nm} + 2 \times 4\ \text{nm} = 32\ \text{nm}$. In this case, $d \ll \lambda/4n \approx 60\ \text{nm}$. Thus, we have $a \approx 2$ for the following gain overlap factor:

$$\Gamma = a\frac{M_{\text{QW}}d_{\text{QW}}}{l} = 2 \times \frac{3 \times 8}{241.5} = 20\%.$$

With $R_1 = R_{\text{DBR1}} = 99.5\%$ and $R_2 = R_{\text{DBR2}} = 99.8\%$, the output coupling loss of this device is

$$\alpha_{\text{out}} = -\frac{\ln\sqrt{R_1R_2}}{l} = -\frac{\ln\sqrt{0.995 \times 0.998}}{241.5 \times 10^{-9}}\ \text{m}^{-1} = 1.452 \times 10^4\ \text{m}^{-1} = 145.2\ \text{cm}^{-1}.$$

With $\overline{\alpha} = 18\ \text{cm}^{-1}$, the threshold gain coefficient

$$g_{\text{th}} = \frac{\overline{\alpha} + \alpha_{\text{out}}}{\Gamma} = \frac{18 + 145.2}{0.2}\ \text{cm}^{-1} = 816\ \text{cm}^{-1}.$$

(d) First, we calculate the bandwidth of the DBRs by taking $\overline{n}$ for N_β in (13.88). Because both DBRs have the same $|\kappa| = 1.224\ \mu\text{m}^{-1} = 1.224 \times 10^6\ \text{m}^{-1}$ and $\overline{n} = 3.26$, they have the same bandwidth of

$$\Delta\nu_{\text{DBR}} = \frac{|\kappa|c}{\pi\overline{n}} = \frac{1.224 \times 10^6 \times 3 \times 10^8}{\pi \times 3.26}\ \text{Hz} = 35.9\ \text{THz}.$$

To find $\Delta\nu_{\text{L}}$, we need to find l_{eff} first. Because $R_{\text{DBR1}} \approx R_{\text{DBR2}} \approx 1$, we have

$$l_{\text{DBR1}}^{\text{eff}} \approx l_{\text{DBR2}}^{\text{eff}} \approx \frac{1}{2|\kappa|} = \frac{1}{2 \times 1.224}\ \mu\text{m} = 408.5\ \text{nm}.$$

Therefore, $l_{\text{eff}} = l + l_{\text{DBR1}}^{\text{eff}} + l_{\text{DBR2}}^{\text{eff}} = 241.5\ \text{nm} + 2 \times 408.5\ \text{nm} = 1058.5\ \text{nm} = 1.0585\ \mu\text{m}$. We find that l_{eff} is on the order of 1 μm, which is much smaller than the 6.108 μm length of the device but is much larger than the 241.8 nm length of the cavity. Then,

$$\Delta\nu_{\text{L}} = \frac{c}{2nl_{\text{eff}}} = \frac{3 \times 10^8}{2 \times 3.52 \times 1.0585 \times 10^{-6}}\ \text{Hz} = 40.3\ \text{THz}.$$

As discussed in Section 13.5, the gain bandwidth of a quantum well is typically in the range of 20–40 THz. We thus find that the longitudinal mode spacing of this QW VCSEL is larger than both its DBR bandwidth and its gain bandwidth. Clearly, this VCSEL will oscillate in a single longitudinal mode.

13.10 Semiconductor laser characteristics

Similarly to an LED, a semiconductor laser is also a junction diode, which has the general electrical characteristics discussed in Section 12.5 with its voltage–current characteristics shown in Fig. 12.12. The difference between a laser and an LED is that the active layer of a laser has to be pumped sufficiently to reach the condition in (13.35) for an optical gain. When a junction diode is forward biased with a voltage V, the splitting of its Fermi levels is given by (12.93), which is valid for both homojunctions and heterojunctions. In the active region, $E_{\text{Fc}} = E_{\text{Fn}}$ and $E_{\text{Fv}} = E_{\text{Fp}}$. Therefore, we find that to have a positive optical gain coefficient in the active region, a diode has to be forward biased at a voltage larger than the bandgap of its active layer:

$$eV = E_{\text{Fc}} - E_{\text{Fv}} > h\nu > E_{\text{g}}. \tag{13.108}$$

This condition only specifies the forward voltage required for the active region in a junction diode to reach transparency. To reach the laser threshold, a laser diode still has to be biased somewhat higher to reach a gain that is sufficiently large for overcoming the losses in the laser cavity. The bias voltage of a semiconductor laser remains quite constant when the laser oscillates above threshold because the carrier density is clamped

at its threshold value when the injection current is increased above the laser threshold. In contrast, an LED is normally biased at a lower voltage around $V \geq h\nu/e$.

For most applications, it is desired that a semiconductor laser oscillate in a single transverse mode and a single longitudinal mode. Many practical lasers indeed have such a desirable characteristic. For a single-mode semiconductor laser with a uniformly distributed carrier density in a thin active layer of a thickness d, the temporal characteristics of its carrier density N and its intracavity photon density S can be described by the following coupled rate equations:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{J}{ed} - \frac{N}{\tau_\mathrm{s}} - gS, \tag{13.109}$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\gamma_c S + \Gamma g S, \tag{13.110}$$

where e is the electronic charge, τ_s is the spontaneous carrier lifetime, and γ_c is the cavity decay rate. The current density J in the active region of a junction area $\mathcal{A}$ is related to the injection current I through the relation given in (13.71). The overlap factor Γ appears in the last term of (13.110) because only that fraction of the laser mode intensity overlaps with the gain region to receive stimulated amplification. According to (11.68) and (13.40), the gain parameter g (per second) of a semiconductor laser is related to the gain coefficient g (per meter) of the semiconductor gain medium by

$$g = \frac{c}{n}g = \frac{c\sigma}{n}(N - N_\mathrm{tr}) \tag{13.111}$$

for an ordinary DH laser, and by

$$g = \frac{c}{n}g = \frac{c\sigma}{n}N_\mathrm{tr} \ln \frac{N}{N_\mathrm{tr}} \tag{13.112}$$

for a QW laser.

Laser threshold

The threshold characteristics of a laser and its characteristics in steady-state oscillation above threshold can be obtained by considering the steady-state solutions of (13.109) and (13.110) for $\mathrm{d}N/\mathrm{d}t = \mathrm{d}S/\mathrm{d}t = 0$. From (13.110), we find that the threshold condition for a semiconductor laser is

$$\Gamma g_\mathrm{th} = \gamma_\mathrm{c}, \tag{13.113}$$

which leads to the following *threshold carrier density*:

$$N_\mathrm{th} = N_\mathrm{tr} + \frac{n\gamma_c}{\Gamma c\sigma} = N_\mathrm{tr} + \frac{g_\mathrm{th}}{\sigma} \tag{13.114}$$

for an ordinary DH laser and

$$N_\mathrm{th} = N_\mathrm{tr} \exp\left(\frac{n\gamma_c}{\Gamma c\sigma N_\mathrm{tr}}\right) = N_\mathrm{tr} \exp\left(\frac{g_\mathrm{th}}{\sigma N_\mathrm{tr}}\right) \tag{13.115}$$

for a QW laser. From (13.109), we find that the *threshold current density* is $J_{th} = N_{th}ed/\tau_s$ because $S = 0$ right at the laser threshold. Using this result and the relation between J and I in (13.71) with a carrier injection efficiency η_{inj}, we find the following *threshold injection current*:

$$I_{th} = \frac{eN_{th}}{\eta_{inj}\tau_s}\mathcal{V}_{active}. \tag{13.116}$$

The threshold current of a semiconductor laser is linearly proportional to the threshold carrier density and the volume of its active region. There is a limit in decreasing the threshold carrier density to reduce the threshold current because $N_{th} > N_{tr}$ and N_{tr} is an intrinsic property of a semiconductor gain medium. Reducing N_{th} by increasing the value of Γ does not lead to a lower value for I_{th} because both Γ and I_{th} are proportional to $\mathcal{V}_{active}$. It is only practical to reduce the value of γ_c as much as possible in order to make N_{th} as close to its limit of N_{tr} as possible. Therefore, the value of N_{th} does not vary much among properly optimized stripe-geometry lasers and VCSELs. For a VCSEL, the threshold current can be reduced by reducing its transverse dimension without changing its cavity length because $\mathcal{V}_{active} = \mathcal{A}M_{QW}d_{QW}$ is independent of the cavity length of a QW VCSEL. In contrast, for a stripe-geometry laser, reduction of its threshold current is limited by its cavity length because $\mathcal{V}_{active} = \mathcal{A}d = lwd$. Because the junction area of a VCSEL can be easily made two to three orders of magnitude smaller than that of a typical stripe-geometry laser, the threshold current of a VCSEL can be two to three orders of magnitude less than that of a stripe-geometry laser. The threshold current of a VCSEL can be as low as 1 μA.

EXAMPLE 13.18 A GaAs QW VCSEL like the one described in Example 13.17 has a carrier injection efficiency of $\eta_{inj} = 70\%$ and a cross-sectional diameter of 5 μm. (a) Use the data in Examples 13.7 and 13.17 to find its threshold carrier density. (b) Carrier recombination in an efficient laser is almost purely radiative. Take the radiative recombination coefficient of $B = 1.77 \times 10^{-16}\,\text{m}^3\,\text{s}^{-1}$ found in Example 13.5 for GaAs to find the carrier lifetime at the threshold carrier density. (c) Find the threshold current for this VCSEL.

Solution From Example 17, we have the following parameters for this VCSEL: $g_{th} = 816\ \text{cm}^{-1} = 8.16 \times 10^4\ \text{m}^{-1}$, $M_{QW} = 3$, and $d_{QW} = 8$ nm. From Example 13.7(b), we have $\sigma = 2.2 \times 10^{-19}\ \text{m}^2$, $N_{tr} = 1.45 \times 10^{24}\ \text{m}^{-3}$, and $\sigma N_{tr} = 3.19 \times 10^5\ \text{m}^{-1}$ for a GaAs quantum well of $d_{QW} = 8$ nm.

(a) By using (13.115) for a QW laser, we find that

$$N_{th} = N_{tr}\exp\left(\frac{g_{th}}{\sigma N_{tr}}\right) = 1.45 \times 10^{24} \times \exp\left(\frac{8.16 \times 10^4}{3.19 \times 10^5}\right)\ \text{m}^{-3} = 1.87 \times 10^{24}\ \text{m}^{-3}.$$

For the purpose of comparison, we use (13.114) to find that $N_{th} = 1.77 \times 10^{24}\ \text{m}^{-3}$. This value is very close to the value of $N_{th} = 1.87 \times 10^{24}\ \text{m}^{-3}$ found by using the

relation of (13.115) for a QW laser. Because $N_{th} - N_{tr} \ll N_{tr}$ in this situation, (13.114) is a fairly good approximation to (13.115).

(b) For $N = N_{th} \gg n_0, p_0$ in the situation of purely radiative recombination, we find from (13.8) that

$$\tau_s = \tau_{rad} = \frac{1}{BN_{th}} = \frac{1}{1.77 \times 10^{-16} \times 1.87 \times 10^{24}} \text{ s} = 3.02 \text{ ns}.$$

(c) With a diameter of 5 μm, the active volume of the three quantum wells is

$$\mathcal{V}_{active} = \mathcal{A} M_{QW} d_{QW} = \pi \times \left(\frac{5 \times 10^{-6}}{2}\right)^2 \times 3 \times 8 \times 10^{-9} \text{ m}^3 = 4.71 \times 10^{-19} \text{ m}^3.$$

Therefore, we find the following threshold current for this VCSEL:

$$I_{th} = \frac{eN_{th}}{\eta_{inj}\tau_s}\mathcal{V}_{active} = \frac{1.6 \times 10^{-19} \times 1.87 \times 10^{24}}{0.7 \times 3.02 \times 10^{-9}} \times 4.71 \times 10^{-19} \text{ A} = 66.7 \text{ μA}.$$

As expected, the threshold current of this VCSEL is pretty low.

Laser power

In steady-state oscillation above threshold with an injection current of $I > I_{th}$, the carrier density and the gain are clamped at their respective threshold values, $N = N_{th}$ and $g = g_{th}$ given above, while the intracavity photon density builds up for $S \neq 0$. Most of the concepts developed in Section 11.3 for laser power characteristics are directly applicable to semiconductor lasers.

Taking the relation in (13.111) for an ordinary DH laser, it can be shown that the threshold gain parameter of a semiconductor laser also has the form of (11.72) as follows (see Problem 13.10.1(a)):

$$g_{th} = \frac{c\sigma}{n}(N_{th} - N_{tr}) = \frac{c\sigma}{n}\frac{N_{inj} - N_{tr}}{1 + S/S_{sat}}, \tag{13.117}$$

where $N_{inj} = J\tau_s/ed$ is the injected carrier density and S_{sat} is the saturation photon density that has the form of (11.74):

$$S_{sat} = \frac{n}{c\tau_s\sigma}. \tag{13.118}$$

Comparing (13.118) with (11.74), we find that the spontaneous carrier recombination lifetime τ_s of a semiconductor has exactly the same function as the saturation lifetime of an atomic or molecular system in defining the saturation intensity of a gain medium and the saturation photon density of a laser. Therefore, the spontaneous carrier recombination lifetime is also the saturation lifetime of a semiconductor, as mentioned following (12.56) where it is defined. For a semiconductor laser, the dimensionless pumping ratio r, which is defined in (11.76), is conveniently expressed in terms of the pump current

because the bias voltage of a semiconductor laser remains quite constant:

$$r = \frac{I - I_{\text{tr}}}{I_{\text{th}} - I_{\text{tr}}}. \tag{13.119}$$

Using the steady-state solution of (13.109) for S and the relations in (13.118) and (13.119), the output power of a semiconductor laser can be expressed in the form of (11.78) and (11.87) as (see Problem 13.10.1(b))

$$P_{\text{out}} = (r-1)\mathcal{V}_{\text{mode}} S_{\text{sat}} h\nu \gamma_{\text{out}} = \frac{I - I_{\text{th}}}{I_{\text{th}} - I_{\text{tr}}} P_{\text{out}}^{\text{sat}}, \tag{13.120}$$

where $P_{\text{out}}^{\text{sat}} = \mathcal{V}_{\text{mode}} S_{\text{sat}} h\nu \gamma_{\text{out}}$ as defined in (11.82). This relation is obtained for an ordinary DH laser. A similar, but more complicated, relation can be obtained for a QW laser with the same definitions for S_{sat} and r in (13.118) and (13.119), respectively (see Problems 13.10.2(a) and (b)).

Alternatively, by applying the relation $g_{\text{th}} = \gamma_{\text{c}}/\Gamma$ from (13.113) and the relation $N = J\tau_{\text{s}}/ed$ directly to the steady-state solution of S from (13.109), the output power of a semiconductor laser can be expressed as (see Problem 13.10.1(c))

$$P_{\text{out}} = \eta_{\text{inj}} \frac{\gamma_{\text{out}}}{\gamma_{\text{c}}} \frac{h\nu}{e} (I - I_{\text{th}}). \tag{13.121}$$

For a laser, we have

$$\gamma_{\text{out}} = \frac{c}{n}\alpha_{\text{out}} \quad \text{and} \quad \gamma_{\text{c}} = \frac{c}{n}\Gamma g_{\text{th}}. \tag{13.122}$$

These relations in (13.121) and (13.122) are generally applicable to all semiconductor lasers, including DH and QW lasers (see Problem 13.10.2(c)). All of the following discussions are also generally applicable to all semiconductor lasers.

It can be seen from (13.121) that in an ideal situation, the output power of a semiconductor laser above threshold increases linearly with injection current. This characteristic is indeed observed in most semiconductor lasers over a large range of operating conditions. In (13.121), both η_{inj} and I_{th} are temperature dependent. In general, η_{inj} decreases but I_{th} increases as the temperature increases. In addition, at high injection levels, η_{inj} normally becomes current dependent and decreases with increasing current for a given device, resulting in nonlinearities in the *L–I* characteristics of a laser. Figure 13.38(*a*) shows the power–current characteristics of a typical single-mode semiconductor laser. For a multimode laser, the competition and coexistence of multiple modes can lead to the nonlinearities and kinks, shown in Fig. 13.38(*b*), that are often observed in its *L–I* characteristics.

After optimizing the structure of a laser to reduce power losses by maximizing the values of η_{inj} and $\gamma_{\text{out}}/\gamma_{\text{c}}$ and by minimizing the value of I_{th}, the output power available from a semiconductor laser depends solely on the current that can be injected into the laser. As discussed in Section 12.5, there is a limit to the current density J that can be injected into a junction diode. Further limitation on J comes from the limitation

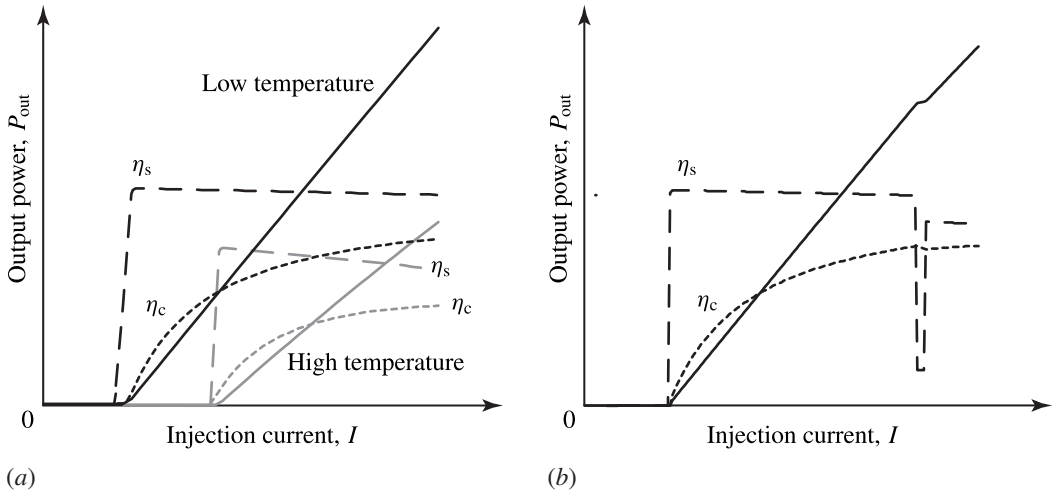


Figure 13.38 Light–current characteristics of (*a*) a single-mode semiconductor laser at different temperatures and (*b*) a multimode semiconductor laser at a given temperature. Also shown are typical characteristics of the power conversion efficiency and the slope efficiency of the laser.

on the carrier density N in the active region before high-order nonradiative processes dominate the recombination process. Because of these limitations on J and because $I = J\mathcal{A}/\eta_{inj}$, the junction area of a laser sets a limit on the output power that the laser can possibly deliver. As the junction area of a VCSEL is made very small to reduce its threshold current, it also limits the VCSEL to a low output power in comparison to a stripe-geometry laser that has a much larger junction area than a VCSEL.

EXAMPLE 13.19 Find the output power of the GaAs QW VCSEL considered in Examples 13.17 and 13.18 when it is operated with an injection current at twice the threshold level.

Solution From Example 13.17, we have the following parameters for this VCSEL: $\lambda = 850$ nm, $\Gamma = 20\%$, $g_{th} = 816$ cm^{-1}, and $\alpha_{out} = 145.2$ cm^{-1}. We then have

$$\frac{h\nu}{e} = \frac{1239.8}{850}\ \text{V} = 1.459\ \text{V},$$

and, from (13.122),

$$\frac{\gamma_{out}}{\gamma_c} = \frac{\alpha_{out}}{\Gamma g_{th}} = \frac{145.2}{0.2 \times 816} = 0.89.$$

The device has $\eta_{inj} = 0.7$ and $I_{th} = 66.7$ μA found in Example 13.18. Thus, when it operates at $I = 2I_{th} = 133.4$ μA, the output power

$$\begin{aligned} P_{out} &= \eta_{inj}\frac{\gamma_{out}}{\gamma_c}\frac{h\nu}{e}(I - I_{th}) \\ &= 0.7 \times 0.89 \times 1.459 \times (133.4 - 66.7) \times 10^{-6}\ \text{W} = 60.7\ \mu\text{W}. \end{aligned}$$

Because the injection current is still very low for this output power, much higher output powers can be obtained at higher injection levels.

Laser efficiency

Because the pump power is $P_p = VI$ in the case of current injection, the power conversion efficiency of a semiconductor laser is

$$\eta_c = \frac{P_{out}}{VI} = \eta_{inj}\frac{\gamma_{out}}{\gamma_c}\frac{h\nu}{eV}\left(1 - \frac{I_{th}}{I}\right) = \eta_e\frac{h\nu}{eV}\left(1 - \frac{I_{th}}{I}\right), \tag{13.123}$$

where η_e is the external quantum efficiency defined in (13.125) below. The slope efficiency of a semiconductor laser operating above threshold is

$$\eta_s = \frac{dP_{out}}{VdI} = \eta_{inj}\frac{\gamma_{out}}{\gamma_c}\frac{h\nu}{eV} = \eta_e\frac{h\nu}{eV}. \tag{13.124}$$

The external quantum efficiency of a semiconductor laser operating above threshold is

$$\eta_e = \frac{P_{out}/h\nu}{I/e - I_{th}/e} = \eta_{inj}\frac{\gamma_{out}}{\gamma_c}. \tag{13.125}$$

Above threshold, the voltage across the junction of a laser diode remains fairly constant because the carrier density is clamped at its threshold value of N_{th}. Therefore, if all of the bias voltage drops across the junction, the slope efficiency of the laser is a constant that is independent of the injection level. In reality, however, there is always some series resistance, which may be added intentionally to protect a laser diode or caused by parasitic effects, in a laser diode. Thus, the bias voltage is always increased by the series resistance as $V = V_j + IR_s$, where V_j is the junction voltage. Clearly, this increase in the bias voltage will reduce both the slope efficiency and the power conversion efficiency of the laser as the injection current increases. Such characteristics are schematically shown in Fig. 13.38. As the additional voltage drop across the resistance is unimportant when $IR_s \ll V_j$, it is clear that the efficiencies of a laser can be improved by reducing the series resistance and the laser threshold (see Problem 13.10.3).

Comparing (13.124) and (13.125), we find that $\eta_e > \eta_s$ for a semiconductor laser because $eV > h\nu$. In addition, if we identify the photon extraction efficiency of a semiconductor laser as

$$\eta_t = \frac{\gamma_{out}}{\gamma_c}, \tag{13.126}$$

we can express η_e as

$$\eta_e = \eta_{inj}\eta_t. \tag{13.127}$$

Comparing this result to η_e of an LED defined in (13.67), we find that the internal quantum efficiency η_i of a semiconductor laser is

$$\eta_i = 1. \tag{13.128}$$

In practice, η_i of a semiconductor laser is not exactly 100% but is often higher than 90%. Such a high internal quantum efficiency for a semiconductor laser reflects the fact that almost all of the injected carriers recombine radiatively through the stimulated recombination process when a laser oscillates above threshold.

We see from the above discussions that a semiconductor laser typically has a very high external quantum efficiency, as well as a very high slope efficiency. They can be as high as 80–90% if internal losses and diffraction losses are minimized to make $\gamma_{out} \approx \gamma_c$ while the injection efficiency η_{inj} is maximized. The power conversion efficiency, however, is normally much lower because of the existence of a laser threshold. A typical laser has a power conversion efficiency of 10–20%. Some lasers have power conversion efficiencies as high as 50%. Clearly, it is important to reduce the laser threshold as much as possible.

EXAMPLE 13.20 Find the various efficiencies of the GaAs QW VCSEL considered in the preceding examples if it has a bias voltage of $V = 2.2$ V when operating at an injection level twice the threshold.

Solution From the data in Example 13.19, we find the following external quantum efficiency for this VCSEL:

$$\eta_e = \eta_{inj}\frac{\gamma_{out}}{\gamma_c} = 0.7 \times 0.89 = 62.3\%.$$

The photon extraction efficiency

$$\eta_t = \frac{\gamma_{out}}{\gamma_c} = 89\%.$$

The photon extraction efficiency is much higher than the external quantum efficiency because the device suffers 30% loss of the injected carriers with an injection efficiency of $\eta_{inj} = 70\%$. This is where efficiency improvement can be targeted for this device.

The power conversion efficiency of this device operating at $I = 2I_{th}$ with a bias voltage of $V = 2.2$ V is

$$\eta_c = \eta_e\frac{h\nu}{eV}\left(1 - \frac{I_{th}}{I}\right) = 62.3\% \times \frac{1.459}{2.2} \times \left(1 - \frac{1}{2}\right) = 20.7\%.$$

The slope efficiency

$$\eta_s = \eta_e\frac{h\nu}{eV} = 62.3\% \times \frac{1.459}{2.2} = 41.3\%.$$

We see that $\eta_s < \eta_e$ because $h\nu < eV$. This reduction of efficiency cannot be avoided because a bias voltage of 2.2 V, which is significantly higher than the photon energy, is required for the laser to reach the desired level of population inversion. Of course, any series resistance that can further increase the bias voltage will further reduce the slope efficiency and the power conversion efficiency of the laser. The power conversion

efficiency is only half that of the slope efficiency because the laser is operated at twice its threshold. At a given injection current, η_c can be increased by reducing the laser threshold. For a laser of a given threshold, η_c can be increased by operating the laser at a level high above its threshold.

Laser spectrum

A semiconductor laser has the general spectral characteristics of a laser, which are very different from those of an LED. The basic difference between a semiconductor laser and other classes of lasers, such as fiber lasers, is that a semiconductor laser has a very short cavity and a high optical gain. As a result, a semiconductor laser has a larger longitudinal mode spacing and a larger linewidth than most other lasers.

As discussed in the preceding section, a VCSEL normally oscillates in a single longitudinal mode because of its large mode spacing. For a semiconductor laser that has a horizontal or folded cavity, the cavity length is typically in the range of 200–500 μm with a corresponding longitudinal mode spacing in the range of 100–200 GHz. Because the gain bandwidth of a semiconductor is typically in the range of 10–20 THz and can be as large as 40 THz for a highly pumped QW laser, a multimode semiconductor laser easily oscillates in 10–20 longitudinal modes. The linewidth of each longitudinal mode is typically on the order of 10 MHz, but can be as narrow as 1 MHz or as broad as 100 MHz. The linewidth narrows, but the number of oscillating modes increases, as the laser is injected at a current level high above its threshold. Figure 13.39(*a*) shows a typical spectrum of a multimode semiconductor laser.

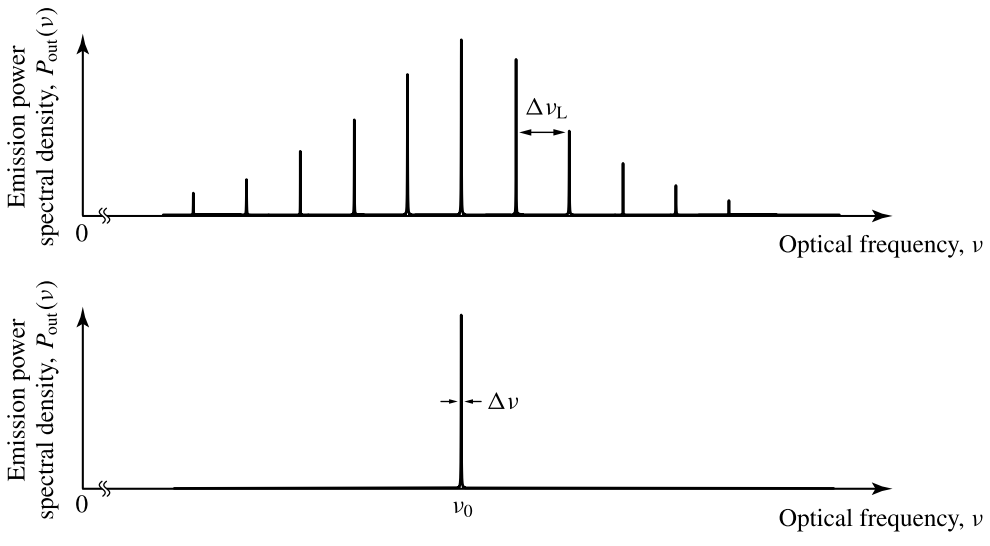


Figure 13.39 Representative emission spectra of (*a*) a multimode semiconductor laser and (*b*) a single-frequency semiconductor laser.

In many applications, a laser oscillating in a single frequency is desired. There are many different approaches to making a semiconductor laser oscillate in a single longitudinal frequency. Some of the most important and practical concepts are already discussed in the preceding section. They include the use of a very short cavity, as is the case of a VCSEL, and the use of a frequency-selective grating, as is the cases of the DBR laser, the phase-shifted DFB laser, and the GCSEL. For these single-frequency lasers, the linewidth is still in the typical range of 1–100 MHz as mentioned above. It is possible to obtain single-frequency output with a linewidth on the order of 100 kHz or less by injection locking with a narrow-linewidth, single-frequency master laser source or by using a highly frequency-selective external grating as one optical-feedback element. Figure 13.39(*b*) shows the spectrum of a single-frequency semiconductor laser. Tuning of the laser frequency, in some cases over a large range close to the entire gain bandwidth of the laser, is possible.

Modulation characteristics

A semiconductor laser can be directly current modulated like an LED. Unlike an LED, however, the modulation speed of a semiconductor laser is not limited by the spontaneous carrier lifetime τ_s in the active region of the laser. This difference is caused by the fact that there is strong coupling between the carrier density and the intracavity laser field. The effective carrier lifetime in an oscillating laser is much shorter than the spontaneous lifetime because of stimulated recombination in a laser. The modulation speed of a semiconductor laser is primarily determined by the intracavity photon lifetime and the effective carrier lifetime. Because both the photon lifetime and the effective carrier lifetime of a semiconductor laser are generally much shorter than the spontaneous carrier lifetime, a semiconductor laser has a higher modulation speed than an LED. Because the stimulated recombination rate increases with the intracavity photon density, the modulation speed of a semiconductor laser increases with laser power.

When a laser is in steady-state oscillation at a bias point with a DC current of $I_0 > I_{th}$ in the absence of modulation, the laser gain and the carrier density are both clamped at their respective threshold values of g_{th} and N_{th}, but the photon density has a value of S_0 corresponding to the laser output power P_0 at the bias point. Under the dynamical perturbation of a modulation signal, the gain can deviate from g_{th} due to the variations in the carrier and photon densities caused by the external perturbation. The dependence of the gain parameter on the carrier and photon densities can be expressed as

$$g = g_{th} + g_n(N - N_{th}) + g_p(S - S_0), \tag{13.129}$$

where $g_n = c\sigma/n$ is the *differential gain parameter* characterizing the dependence of the gain parameter on the carrier density as seen in (13.111) and g_p is the *nonlinear gain parameter* characterizing the effect of *gain compression* due to the saturation of gain by intracavity photons. It has been found empirically that both g_n and g_p stay quite

constant over large ranges of carrier density and photon density in a given laser. For most practical purposes, they can be treated as constants over the operating range of a laser. These parameters are normally measured experimentally though they can also be calculated theoretically. Note that $g_n > 0$ but $g_p < 0$.

It is convenient to define a *differential carrier relaxation rate*, γ_n, and a *nonlinear carrier relaxation rate*, γ_p, as

$$\gamma_n = g_n S_0, \qquad \gamma_p = -\Gamma g_p S_0. \tag{13.130}$$

In addition, we have the cavity decay rate, $\gamma_c = 1/\tau_c$, and the spontaneous carrier relaxation rate, $\gamma_s = 1/\tau_s$. These four relaxation rates, together with the linewidth enhancement factor, b, defined in (13.61), are the intrinsic dynamical parameters of a semiconductor laser that completely determine the dynamical behavior of the laser. All five of these parameters can be directly measured for a given laser. The current-modulation characteristics of a laser, however, are independent of the linewidth enhancement factor but are determined only by the four rate parameters. Note that, for a given laser, γ_c and γ_s are constants that are independent of laser power, but γ_n and γ_p are linearly proportional to laser power.

Because a semiconductor laser has a threshold, the modulation index m for a laser that is biased at a DC current of $I_0 > I_{th}$ and is modulated at a frequency of $\Omega = 2\pi f$ is defined as

$$I(t) = I_0 + I_1(t) = I_{th} + (I_0 - I_{th})(1 + m\cos\Omega t), \tag{13.131}$$

which is different from that defined in (13.73) for an LED. In the regime of linear response, the output power of the laser can be expressed in the same form as that in (13.74):

$$P(t) = P_0 + P_1(t) = P_0[1 + |r|\cos(\Omega t - \varphi)]. \tag{13.132}$$

For small-signal modulation with $m \ll 1$, the complex response function of a laser is (see Problem 13.10.4(b))

$$r(\Omega) = |r|e^{i\varphi} = -\frac{m\gamma_c\gamma_n}{\Omega^2 - \Omega_r^2 + i\Omega\gamma_r}, \tag{13.133}$$

where Ω_r is the *relaxation resonance frequency* and γ_r is the *total carrier relaxation rate* for the relaxation oscillation of the coupling between the carriers and the intracavity laser field in the semiconductor laser. They are related to the intrinsic dynamical parameters of the laser through

$$\Omega_r^2 = 4\pi^2 f_r^2 = \gamma_c\gamma_n + \gamma_s\gamma_p \tag{13.134}$$

and

$$\gamma_r = \gamma_s + \gamma_n + \gamma_p. \tag{13.135}$$

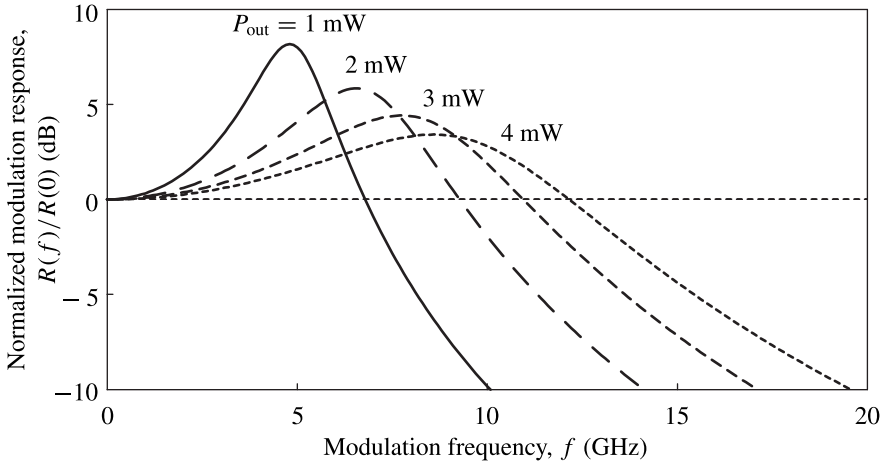


Figure 13.40 Normalized current-modulation frequency response of a semiconductor laser measured in terms of the electrical power spectrum of a photodetector. The frequency response of a semiconductor laser depends on the output laser power, with its 3-dB bandwidth increasing approximately with the square root of the output power. These curves are generated with the following relations: f_r (GHz) $= 5\sqrt{P}$ (mW) and γ_r (ns^{-1}) $= 1.5 + 11P$ (mW).

Clearly, Ω_r and f_r are proportional to the square root of the laser power, whereas γ_r is linearly dependent on, but not proportional to, the laser power. The relation between the relaxation resonance frequency and the carrier relaxation rate is often characterized by a *K factor* defined as

$$K = \frac{\gamma_r - \gamma_s}{f_r^2}. \tag{13.136}$$

The modulation power spectrum of a semiconductor laser is

$$R(f) = |r(f)|^2 = \frac{m^2\gamma_c^2\gamma_n^2}{16\pi^4(f^2 - f_r^2)^2 + 4\pi^2 f^2 \gamma_r^2}. \tag{13.137}$$

As shown in Fig. 13.40, this spectrum has a resonance peak at (see Problem 13.10.4(c))

$$f_{pk} = \left(f_r^2 - \frac{\gamma_r^2}{8\pi^2}\right)^{1/2}, \tag{13.138}$$

and a 3-dB modulation bandwidth (see Problem 13.10.4(d))

$$f_{3dB} = (1 + \sqrt{2})^{1/2}\left(f_r^2 - \frac{\gamma_r^2}{8\sqrt{2}\pi^2}\right)^{1/2} \approx 1.554 f_{pk}. \tag{13.139}$$

Because $f_r \gg \gamma_r/2\pi$ for most lasers and because $f_r \propto P_0^{1/2}$, the modulation bandwidth of a semiconductor laser increases with laser power and scales roughly as $f_{3dB} \propto P_0^{1/2}$. An intrinsic modulation bandwidth on the order of a few gigahertz is common for a semiconductor laser. A high-speed semiconductor laser can have a bandwidth larger than 20 GHz. Because the intrinsic modulation bandwidth of a semiconductor laser

is significantly larger than that of an LED, it is very important to reduce the parasitic effects from electrical contacts and packaging for high-frequency modulation of a semiconductor laser.

EXAMPLE 13.21 The GaAs QW VCSEL considered in the preceding examples can be modulated at very high frequencies because it has a very short cavity and is a QW laser, two important factors that lead to the high speed of the device. The data given or found in the preceding examples are sufficient to find all parameters for the modulation characteristics of the laser, with the exception of the parameter g_p. Here we simply take $g_p = -g_n$. (a) Find the values of γ_s and γ_c. What are the corresponding carrier lifetime and photon lifetime? (b) Find the values of g_n and g_p. Then find the values of γ_n and γ_p. (c) Find the values of f_r and γ_r. What is the value of the K factor of this device? (d) Find the resonance peak of the modulation spectrum. What is the 3-dB modulation bandwidth of this VCSEL?

Solution (a) We have $\tau_s = 3.02$ ns found in Example 13.18. Thus

$$\gamma_s = \frac{1}{\tau_s} = \frac{1}{3.02 \times 10^{-9}} \text{ s}^{-1} = 3.31 \times 10^8 \text{ s}^{-1}.$$

From Example 13.17, we have $n = 3.52$, $\Gamma = 0.2$, and $g_{th} = 8.16 \times 10^4$ m. Thus

$$\gamma_c = \frac{c}{n}\Gamma g_{th} = \frac{3 \times 10^8}{3.52} \times 0.2 \times 8.16 \times 10^4 \text{ s}^{-1} = 1.39 \times 10^{12} \text{ s}^{-1}.$$

We already have $\tau_s = 3.02$ ns for the carrier lifetime. The photon lifetime

$$\tau_c = \frac{1}{\gamma_c} = \frac{1}{1.21 \times 10^{12}} \text{ s} = 719 \text{ fs}.$$

This laser has a very small photon lifetime because of its very short cavity.

(b) Using $\sigma = 2.2 \times 10^{-19}$ m^2, we find

$$g_n = \frac{c\sigma}{n} = \frac{3 \times 10^8 \times 2.2 \times 10^{-19}}{3.52} \text{ m}^3 \text{ s}^{-1} = 1.875 \times 10^{-11} \text{ m}^3 \text{ s}^{-1}.$$

Based on the assumption we have made, $g_p = -g_n = -1.875 \times 10^{-11}$ m^3 s^{-1}. To find the values of γ_n and γ_p, we need to find the intracavity photon density S_0 at the operating point. We have $P_{out} = 60.6$ μW and $h\nu = 1.459$ eV, both found in Example 13.19. To find S_0, we also need the following two parameters:

$$\mathcal{V}_{mode} \approx \mathcal{A}l = \pi \times \left(\frac{5 \times 10^{-6}}{2}\right)^2 \times 241.5 \times 10^{-9} \text{ m}^3 = 4.74 \times 10^{-18} \text{ m}^3$$

for $l = 241.5$ nm found in Example 13.17, and, from Example 13.20,

$$\gamma_{out} = \eta_t \gamma_c = 89\% \times 1.39 \times 10^{12} \text{ s}^{-1} = 1.24 \times 10^{12} \text{ s}^{-1}.$$

Then, the intracavity photon density can be found as

$$\begin{aligned} S_0 &= \frac{P_{\text{out}}}{\mathcal{V}_{\text{mode}} h\nu \gamma_{\text{out}}} \\ &= \frac{60.6 \times 10^{-6}}{4.74 \times 10^{-18} \times 1.459 \times 1.6 \times 10^{-19} \times 1.24 \times 10^{12}} \text{ m}^{-3} \\ &= 4.42 \times 10^{19} \text{ m}^{-3}. \end{aligned}$$

We then find that

$$\gamma_{\text{n}} = g_{\text{n}} S_0 = 1.875 \times 10^{-11} \times 4.42 \times 10^{19} \text{ s}^{-1} = 8.29 \times 10^{8} \text{ s}^{-1}$$

and

$$\gamma_{\text{p}} = -\Gamma g_{\text{p}} S_0 = 0.2 \times 1.875 \times 10^{-11} \times 4.42 \times 10^{19} \text{ s}^{-1} = 1.66 \times 10^{8} \text{ s}^{-1}.$$

(c) We now find that

$$\begin{aligned} f_{\text{r}} &= \frac{(\gamma_{\text{c}}\gamma_{\text{n}} + \gamma_{\text{s}}\gamma_{\text{p}})^{1/2}}{2\pi} \\ &= \frac{(1.39 \times 10^{12} \times 8.29 \times 10^{8} + 3.02 \times 10^{8} \times 1.66 \times 10^{8})^{1/2}}{2\pi} \text{ Hz} \\ &= 5.403 \text{ GHz} \end{aligned}$$

and

$$\gamma_{\text{r}} = \gamma_{\text{s}} + \gamma_{\text{n}} + \gamma_{\text{p}} = (3.02 \times 10^{8} + 8.29 \times 10^{8} + 1.66 \times 10^{8}) \text{ s}^{-1} = 1.29 \times 10^{9} \text{ s}^{-1}.$$

The value of the K factor is

$$K = \frac{\gamma_{\text{r}} - \gamma_{\text{s}}}{f_{\text{r}}^2} = \frac{1.29 \times 10^{9} - 3.02 \times 10^{8}}{(5.403 \times 10^{9})^2} \text{ s} = 33.8 \text{ ps}.$$

(d) The resonance peak of the modulation spectrum is

$$f_{\text{pk}} = \left(f_{\text{r}}^2 - \frac{\gamma_{\text{r}}^2}{8\pi^2} \right)^{1/2} = \left(5.403^2 - \frac{1.29^2}{8\pi^2} \right)^{1/2} \text{ GHz} = 5.401 \text{ GHz}.$$

We see that f_{pk} is very close to f_{r} but slightly lower. We now find the following 3-dB modulation bandwidth of this VCSEL:

$$\begin{aligned} f_{\text{3dB}} &= (1 + \sqrt{2})^{1/2} \left(f_{\text{r}}^2 - \frac{\gamma_{\text{r}}^2}{8\sqrt{2}\pi^2} \right)^{1/2} \\ &= (1 + \sqrt{2})^{1/2} \left(5.403^2 - \frac{1.29^2}{8\sqrt{2}\pi^2} \right)^{1/2} \text{ GHz} \\ &= 8.39 \text{ GHz}. \end{aligned}$$

This VCSEL indeed has a large modulation bandwidth, as expected.

PROBLEMS

13.1.1 Describe the radiative recombination processes that can possibly take place in semiconductors. Discuss which processes can be utilized for LEDs and which ones are useful to semiconductor lasers.

13.1.2 Explain why and how some indirect-gap semiconductors can be used to make LEDs. Are they used for semiconductor lasers? Why?

13.1.3 Answer the questions in Example 13.1 for the p-type GaAs considered in Problem 12.3.2.

13.1.4 Verify the answer to Problem 12.3.4 by showing that high radiative efficiencies for the InGaAsP sample take place in the range of carrier concentrations found as the answer to that problem. What is the highest radiative efficiency? At what carrier density does it occur?

13.2.1 Explain why in an indirect-gap semiconductor both direct absorption and indirect absorption are possible but direct recombination is highly improbable. Under what condition does direct absorption take place in an indirect-gap semiconductor? What is the difference between a direct-absorption process that takes place in an indirect-gap semiconductor and one that takes place in a direct-gap semiconductor?

13.2.2 Explain why Si and Ge are important materials for photodetectors but are not useful as materials for LEDs and semiconductor lasers.

13.2.3 Show that a direct band-to-band optical transition in a semiconductor involving the absorption or emission of a photon of energy $h\nu$ takes place between a conduction-band state of energy E_2 given by (13.13) and a valence-band state of energy E_1 given by (13.14). Use these relations to show that the density of states for direct band-to-band optical transitions can be expressed as a function of optical frequency in the form of (13.17).

13.2.4 Answer questions (b), (c), and (d) in Example 13.2 for direct band-to-band optical transitions in intrinsic GaAs at $\lambda = 800$ nm wavelength at 300 K. Compare the results with those of Example 13.2.

13.3.1 Use the relation in (13.17) for $\rho(\nu)$ and that in (13.32) for $\alpha_0(\nu)$ to find and plot the absorption coefficient of intrinsic GaAs in thermal equilibrium at 300 K as a function of photon energy over a range of photon energies from 1.424 to 1.8 eV, corresponding to optical wavelengths from 871 to 689 nm. Take $\tau_{sp} = 500$ ps and $n = 3.65$ over the entire range considered here.

13.3.2 Show, by verifying the relation in (13.34) first, that the condition for population inversion in a semiconductor can be defined as that expressed in (13.35).

13.3.3 Ignoring the temperature dependence of electron and hole effective masses, find the temperature dependence of the transparency carrier density of a direct-gap semiconductor and that of the quasi-Fermi levels E_{Fc} and E_{Fv} in the transparency condition. Use the results found in Example 13.3 for GaAs at 300 K

to find N_{tr} and the corresponding E_{Fc} and E_{Fv} for GaAs at the liquid-nitrogen temperature of $T = 77$ K and at a high temperature of $T = 350$ K, respectively. What are the implications of the values found at these different temperatures? Note that the bandgap of GaAs is 1.510 eV at 77 K, 1.424 eV at 300 K, and 1.402 eV at 350 K, according to (12.5).

13.3.4 In this problem, we consider the gain bandwidth of the GaAs sample described in Example 13.4 that is injected with excess electron–hole pairs of a concentration of $N = 2.83 \times 10^{24}$ m^{-3} at 300 K. Take $\tau_{sp} = 500$ ps for GaAs at 300 K. For simplicity, ignore wavelength-dependent variations of the refractive index by taking $n = 3.65$ over the wavelength range of interest.

a. Find the spectral range for which $g > 0$ by finding the two photon energies for which $g = 0$ at the two ends of this range. What is the gain bandwidth? What is the wavelength range covered by the gain spectrum?

b. Find and plot the gain coefficient $g(\nu)$ as a function of photon energy over the gain spectral range found in (a). At what photon energy and corresponding optical wavelength does the gain peak occur?

c. What is the gain cross section at the gain peak?

13.3.5 Find the gain or absorption coefficient at 800 nm wavelength for the GaAs sample considered in Example 13.4 that is injected with excess electron–hole pairs of a concentration $N = 2.83 \times 10^{24}$ m^{-3} at 300 K. Take $\tau_{sp} = 500$ ps for GaAs at 300 K. The refractive index of GaAs at 800 nm is $n = 3.68$.

13.4.1 Use the relation in (13.17) for $\rho(\nu)$ and that in (13.32) for $\alpha_0(\nu)$ to carry out the integration in (13.45) for the total band-to-band spontaneous recombination rate of electron–hole pairs in a semiconductor with $E_g \gg k_B T$ in thermal equilibrium at a temperature T. Then, use the result to show that the bimolecular recombination coefficient can be expressed by the relation given in (13.48).

13.4.2 Use the absorption spectrum $\alpha_0(\nu)$ found in Problem 13.3.1 to find and plot the spontaneous emission spectrum $R_{sp}^0(\nu)$ of intrinsic GaAs in thermal equilibrium at 300 K.

13.4.3 Find and plot the spontaneous emission spectrum $R_{sp}(\nu)$ of a GaAs sample injected with $N = 2.83 \times 10^{24}$ m^{-3} at 300 K over its gain spectral range found in Problem 13.3.4. Where is the peak of this spontaneous emission spectrum?

13.5.1 What are the two most important considerations in lowering the threshold and in improving the efficiency of a semiconductor laser? What are the basic strategies used in the structural design of semiconductor lasers to address these two issues?

13.5.2 Explain why SH structures can be used for LEDs but not for semiconductor lasers.

13.5.3 Compare the three junction structures: SH, DH, and QW. What are their structural differences? What are their respective advantages or disadvantages?

13.5.4 Explain why a basic SH structure has to be P–p–n not p–n–N but a basic DH structure can be either P–p–N or P–n–N.

13.5.5 Find the density of states for band-to-band optical transitions in the quantum well described in Example 13.6 as a function of photon energy in the range from the bandgap of GaAs at 1.424 eV to that of $Al_{0.3}Ga_{0.7}As$ at 1.798 eV.

13.5.6 Answer the questions in Example 13.6 for a square $Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}As$ quantum well that has a different well thickness of $d_{QW} = 5$ nm. Compare the results with those obtained in Example 13.6.

13.5.7 Answer the questions in Example 13.6 for a square $Al_{0.45}Ga_{0.55}As/GaAs/Al_{0.45}Ga_{0.55}As$ quantum well that has the same well thickness of $d_{QW} = 10$ nm as the one consider in Example 13.6 but has a different composition for the AlGaAs barrier layers. Compare the results with those obtained in Example 13.6.

13.6.1 Describe the basic types of geometry used for the lateral structures of (a) surface-emitting devices and (b) edge-emitting devices. What are the differences in the characteristics and purposes of the different types of geometry?

13.6.2 Why do most edge-emitting LEDs and semiconductor lasers have stripe geometry for their lateral structures? Why is index-guiding stripe geometry used for almost all practical semiconductor lasers? When is gain-guiding stripe geometry used for a semiconductor laser?

13.6.3 A GaAs VCSEL emitting at 830 nm has a circular beam spot of 3 μm diameter at the emitting surface of the laser. Find the far-field divergence angle of the laser beam.

13.6.4 Discuss the consequence of carrier-induced index changes associated with the antiguidance effect on the far-field divergence of an index-guided DH semiconductor laser as the injection current to the laser is continously increased while the laser is oscillating above threshold.

13.7.1 Find the photon flux Φ for a beam at 555 nm wavelength that has a luminous flux of $\Phi_l = 1$ lm.

13.7.2 A TS AlGaInP/GaP LED emits at 607 nm wavelength. It has an external quantum efficiency of $\eta_e = 14\%$ when operated at a forward voltage of 2.10 V with an injection current of 20 mA. At $\lambda = 607$ nm, the photopic luminous efficiency function has a value of $V(\lambda) = 0.541\,13$. (a) Find its power conversion efficiency under the given operating conditions. (b) Find its optical output power. (c) Find its luminous efficiency and luminous flux. (d) Compare the results for this LED with those for the LED considered in Example 13.10.

13.7.3 What are the limiting factors of the external quantum efficiencies of direct-gap and indirect-gap LEDs, respectively? What can be done to improve their respective efficiencies?

13.7.4 Why are most LEDs surface-emitting devices? In what situation are edge-emitting LEDs used?

13.7.5 Verify the approximate relation given in (13.69) for the escape probability.

13.7.6 An AlGaInP LED with a flat surface is to be encapsulated in a plastic epoxy that has a refractive index of 1.5. Take the refractive index of AlGaInP to be 3.4 and approximate the transmittance T in all cases with that of normal incidence. The critical angle, the solid angle Ω_{esc} of the escape cone, and the escape efficiency η_{esc} before encapsulation when the LED surface is exposed to the air are already found in Example 13.11.

a. After the LED is encapsulated in the plastic epoxy, what is the critical angle between the LED and the epoxy? Show that if the epoxy has a slab geometry with a flat epoxy/air interface parallel to the LED/epoxy interface, there is no improvement on Ω_{esc} for the emission to enter the air with or without the epoxy encapsulation. However, there is still some improvement on η_{esc}. How much is this improvement?

b. If the epoxy/air interface has a spherical dome shape, what is Ω_{esc} from inside the LED to the air through the epoxy encapsulation? Find η_{esc} in this case by taking into account reflections at all interfaces.

13.7.7 Consider an LED that has a DH structure with uniformly distributed excess carriers of a density N in its active layer of a thickness d and an area $\mathcal{A}$ when it is injected with a current I at an efficiency η_{inj}. Answer the following questions by solving (13.70) for the LED in steady-state operation.

a. Express the output optical power, P_{out}, of the LED as a function of the carrier density N, the internal quantum efficiency η_i, and the extraction efficiency η_t of the LED.

b. Using the result obtained in (a), express the LED output power, P_{out}, in terms of the injection current I and the external quantum efficiency η_e of the LED to verify the relation in (13.72).

13.7.8 A blue InGaN SQW LED emitting at 470 nm and a green InGaN SQW LED emitting at 530 nm are found to have the same external quantum efficiency of $\eta_e = 12\%$ when both are injected with a current of $I = 20$ mA. The photopic luminous efficiency function has a value of $V(\lambda) = 0.091$ at $\lambda = 470$ nm and a value of $V(\lambda) = 0.862$ at $\lambda = 530$ nm.

a. Find the optical output power and the luminous flux for each of these two LEDs.

b. Which LED has a higher output power? Which one looks brighter to a human eye?

13.7.9 Explain the nonlinear *L–I* characteristics of an LED at very low and very high injection levels, respectively. Assuming a constant junction voltage for all injection current levels, verify the power conversion efficiency as a function of injection current for an LED that has the *L–I* characteristics shown in Fig. 13.26.

13.7.10 An LED that has an excess carrier density of N from current injection as described by (13.70) is subject to direct current modulation.

a. With the time-dependent injection current expressed as (13.73) and the output optical power expressed as (13.74), show that the frequency-dependent response to the current modulation is that given in (13.75).
b. Verify the relation between the output power and the modulation bandwidth of an LED given in (13.78).

13.7.11 An LED initially has a purely radiative spontaneous carrier lifetime of $\tau_s = \tau_{rad} = 5$ ns. It emits at 620 nm with an optical output power of 10 mW when injected with a current of 30 mA. In order to increase its modulation bandwidth, nonradiative recombination centers in the form of impurities are introduced to the LED to reduce its spontaneous carrier lifetime to $\tau_s = 1$ ns while maintaining its radiative recombination lifetime unchanged at $\tau_{rad} = 5$ ns.
a. Find its modulation bandwidth before the nonradiative recombination centers are incorporated into the device. What is its external quantum efficiency in this condition?
b. Find its modulation bandwidth and output power for the same injection current of 30 mA after the nonradiative recombination centers are incorporated into the device. What is its external quantum efficiency in this condition?

13.8.1 Compare SOAs and fiber amplifiers based on their physical differences and performance characteristics. Discuss the advantages and disadvantages of SOAs versus fiber amplifiers in practical applications.

13.8.2 What are the three most important parameters to be considered for a laser amplifier? What are the implications of each of them for the performance of a laser amplifier?

13.8.3 Why does an SOA tend to be more nonlinear and noisier than a fiber amplifier? What can be done to reduce the nonlinearity and noise of an SOA?

13.9.1 Describe and compare the basic structures, principles, and characteristics of the three different types of edge-emitting lasers.

13.9.2 What are the basic differences between DBR and DFB lasers?

13.9.3 Single-longitudinal-mode oscillation of a DBR laser can be ensured by requiring that $\Delta\nu_L > \Delta\nu_{DBR}$. Show that this condition can be met if a DBR laser using two identical DBRs as end mirrors is designed so that $|\kappa|l + R_{DBR}^{1/2} < \pi/2$.

13.9.4 There are two longitudinal modes that fall within the DBR bandwidth of the DBR laser considered in Example 13.15. The one with the lower threshold has been considered in Example 13.15. Find the wavelength, threshold gain coefficient, and threshold carrier density for the second mode. Compare its threshold with that of the first mode.

13.9.5 An InGaAsP DBR laser consists of a gain section of a length $l = 250$ μm and two identical DBRs as end mirrors, each of a length $l_{DBR} = 250$ μm. The Bragg wavelength of the DBRs is $\lambda_B = 1.530\,00$ μm. The effective indices for the laser modes are taken to be $n_\beta = N_\beta = 3.45$. The gain overlap factor

is $\Gamma = 0.2$. The DBR coupling coefficient is $|\kappa| = 40\ \text{cm}^{-1}$. The laser has a distributed loss of $\overline{\alpha} = 60\ \text{cm}^{-1}$.

a. Find the peak reflectivity and the bandwidth of the two identical DBRs.
b. Find the effective phase length of the DBRs and that of the DBR laser to determine the longitudinal mode spacing of the laser.
c. How many longitudinal modes fall within the DBR bandwidth? If the laser is pumped such that only one longitudinal mode oscillates, what is its wavelength?
d. What is the threshold gain coefficient of the oscillating mode?
e. If the gain medium has a gain cross section of $\sigma = 5 \times 10^{-20}\ \text{m}^2$, what is the required carrier density above transparency for the laser to reach its threshold?

13.9.6 The oscillation condition and longitudinal mode characteristics of a DFB laser that does not have a structural phase shift in its grating are considered.

a. Verify that the oscillation condition of such a DFB laser found in (13.97) can be expressed in the form of (13.98).
b. Show by using the condition in (13.98) without numerical solution that the Bragg frequency ν_B is not a longitudinal mode of such a DFB laser. Show also that the longitudinal modes of such a DFB laser are symmetrically distributed on both sides of ν_B.

13.9.7 An InGaAsP DFB laser has its grating fabricated along its gain section, which has a length of $l = 250\ \mu\text{m}$. There is no intentional or unintentional phase shift in the grating. This DFB laser has all of the same parameters as those of the DBR laser described in Problem 13.9.5 with $\lambda_B = 1.530\,00\ \mu\text{m}$, $n_\beta = N_\beta = 3.45$, $\Gamma = 0.2$, $|\kappa| = 40\ \text{cm}^{-1}$, and $\overline{\alpha} = 20\ \text{cm}^{-1}$.

a. Find the longitudinal mode spacing and the stop band of the DFB laser.
b. How many longitudinal modes will oscillate if the laser is pumped to its lowest threshold? What are their wavelengths?
c. What is the threshold gain coefficient of the oscillating modes?
d. If the gain medium has a gain cross section of $\sigma = 5 \times 10^{-20}\ \text{m}^2$, what is the required carrier density above transparency for the laser to reach its threshold?

13.9.8 The $\lambda/4$ phase-shifted DFB laser shown in Fig. 13.32(b) has a total length of l and a $\pi/2$ phase shift right at the middle point $l/2$ of its grating. Both of its end facets have no reflection and no phase shift. The coupling coefficient of the grating is κ.

a. Use the concept of the $\mathbf{S}(z; z_0)$ matrix obtained in (4.95) for contradirectional coupling to show that the reflection coefficient viewed from one end of this DFB laser is

$$r = \frac{\kappa^* \delta (1 - \cosh \alpha_c l)}{|\kappa|^2 - \delta^2 \cosh \alpha_c l + \text{i}\alpha_c \delta \sinh \alpha_c l} \tag{13.140}$$

so that its oscillation condition is

$$|\kappa|^2 - \delta^2 \cosh \alpha_c l + i\alpha_c \delta \sinh \alpha_c l = 0, \tag{13.141}$$

where α_c and δ are defined in (13.94) and (13.96), respectively.

b. Show by using the condition in (13.141) without numerical solution that the Bragg frequency ν_B is a longitudinal mode of this $\lambda/4$ phase-shifted DFB laser and that other longitudinal modes of this laser are symmetrically distributed on both sides of ν_B.

13.9.9 Plot the value of $\alpha_{out} l = (\Gamma g_{th} - \overline{\alpha}) l$ of the $\lambda/4$ phase-shifted DFB laser at the threshold of its fundamental mode by numerically solving its oscillation condition given in (13.141). Compare the result with that shown in Fig. 13.33 for a DFB laser without a phase shift.

13.9.10 Answer the questions in Example 13.16 for a $\lambda/4$ phase-shifted DFB laser with all of the same parameters except for the presence of the phase shift. From Problem 13.9.9, we find that for $|\kappa| l = 1.5$ the threshold of a $\lambda/4$ phase-shifted laser is at $\alpha_{out} l = 2.04$. Compare the results with those obtained in Example 13.16.

13.9.11 Describe and compare the basic structures, principles, and characteristics of the three different types of surface-emitting lasers.

13.9.12 Consider a DBR that has the form of a square index grating made of alternating layers of indices of $\overline{n} \pm \Delta n/2$ for an index step of Δn and an average index of $\overline{n}$ between two neighboring layers. Show that the largest coupling coefficient with a magnitude given by (13.104) is found for a first-order square grating of a 50% duty factor.

13.9.13 Answer the questions in Example 13.17 if the thickness of the spacer layers is increased to make the cavity a 2λ cavity while all other parameters of the device remain unchanged. If we want to bring the threshold gain coefficient of this device with a 2λ cavity back to that found in Example 13.17 for the device with a 1λ cavity by only changing the reflectivity of the top DBR, how many pairs of grating layers have to be used instead for this DBR?

13.9.14 Compare the basic requirements for the operation of an LED, an SOA, and a semiconductor laser.

13.10.1 For a semiconductor laser in steady-state oscillation, (13.109) and (13.110) are reduced to two algebraic equations by taking $dN/dt = dS/dt = 0$.

a. Show that the threshold gain parameter g_{th} can be expressed in terms of N and S in the form of (13.117) with the saturation photon density S_{sat} defined as that in (13.118).

b. Show that the output power of the laser in steady-state oscillation can be expressed as (13.120).

c. Show that the output power can alternatively be expressed as (13.121).

d. Verify that (13.120) and (13.121) are identical to each other by transforming one to the other.

13.10.2 For a QW laser, the threshold gain parameter g_{th} cannot be expressed in terms of N_{inj} and S in the form of (13.117), but S can still be expressed as a function of $N_{inj} - N_{th}$. Follow the procedure in Problem 13.10.1 to answer the following questions for a QW laser.

a. Express S as a function of $N_{inj} - N_{th}$ together with the parameters of N_{tr}, N_{th}, and the saturation photon density S_{sat} defined as that in (13.118).

b. Express the output power of a QW laser in steady-state oscillation in a modified form of (13.120) in terms of the parameters in (13.120).

c. Show that the output power of a QW laser can also be expressed as (13.121).

d. Verify that the modified form of (13.120) for a QW laser is also identical to (13.121) by transforming one to the other. What is the difference between an ordinary DH laser and a QW laser in this transformation?

13.10.3 In the presence of a series resistance in a semiconductor laser, both the power conversion efficiency and the slope efficiency of the laser are reduced at a given injection level. Both also vary with the injection current in a less favorable manner than the ideal situation in the absence of the series resistance. Examine the effect of the series resistance by plotting η_c and η_s as a function of I/I_{th} with $I_{th} = 62$ μA for the VCSEL considered in Example 13.20 by taking the junction voltage to be a constant of $V_j = 2.2$ V but by considering the three different values of $R_s = 0, 1$, and 10 kΩ for the series resistance. What happens to the efficiencies versus I/I_{th} if the threshold current is increased by one order of magnitude to 620 μA?

13.10.4 Consider current modulation of a semiconductor laser biased at a DC current of $I_0 > I_{th}$ and modulated at a modulation frequency of $\Omega = 2\pi f$ with a modulation index of m defined in (13.131). In response to this modulation, the total time-dependent carrier density and the total time-dependent photon density can be expressed as $N(t) = N_0 + N_1(t) = N_{th} + N_1(t)$ and $S(t) = S_0 + S_1(t)$, respectively.

a. Taking the gain parameter g in the form given by (13.129) and using γ_n and γ_p, show with $I(t)$ given by (13.131) that the coupled equations in (13.109) and (13.110) can be transformed, through linearizing the two equations by keeping only the linear time-dependent terms on the variables N_1 and S_1, into the following equations for the temporally varying components of the carrier and photon densities:

$$\Gamma \frac{dN_1}{dt} = m\gamma_c S_0 \cos \Omega t - \Gamma(\gamma_s + \gamma_n)N_1 - (\gamma_c - \gamma_p)S_1, \tag{13.142}$$

$$\frac{dS_1}{dt} = \Gamma \gamma_n N_1 - \gamma_p S_1. \tag{13.143}$$

b. Solve the coupled equations obtained in (a) for S_1 and use the fact that $P_1 \propto S_1$ to show that the frequency-dependent response function of the laser is that given in (13.133). Note that it is easier to take the Fourier transform of both equations and then solve for $S_1(\omega)$.

c. Show that the power spectrum given in (13.137) has a resonance peak at f_{pk} given by (13.138).

d. Show that the 3-dB modulation bandwidth f_{3dB} of a laser is that given in (13.139).

13.10.5 About one-half of the characteristic parameters of a laser are independent of the injection current level in the ideal situation, but the other half vary with the injection current. The parameters that are independent of the injection current level in an ideal situation are η_{inj}, η_{t}, η_{e}, γ_{s}, τ_{s}, γ_{c}, τ_{c}, and the K factor. For the other parameters that vary with the injection current, it is convenient to define a dimensionless parameter:

$$\tilde{J} = \frac{J - J_{\text{th}}}{J_{\text{th}}} = \frac{I - I_{\text{th}}}{I_{\text{th}}}, \tag{13.144}$$

where J and I are the injection current density and the injection current, respectively, with their values being J_{th} and I_{th} at the laser threshold. The bias voltage at threshold is $V_{\text{th}} = V_{\text{j}} + I_{\text{th}} R_{\text{s}}$, where V_{j} is the junction voltage assumed to be a constant and R_{s} is the series resistance of the laser. The dimensionless parameter $\tilde{J}$ has the physical meaning of how many times *above* threshold a laser is operated. All of the current-dependent parameters of a laser can then be conveniently expressed in terms of the parameter $\tilde{J}$ and the values of the respective parameters at one time above threshold where $I = 2I_{\text{th}}$ for $\tilde{J} = 1$, indicated with a superscript 0 for each parameter. For example, P_{out}^0 is the output power of a laser operated at $I = 2I_{\text{th}}$ for $\tilde{J} = 1$. Show that

$$P_{\text{out}} = \tilde{J} P_{\text{out}}^0, \tag{13.145}$$

$$\eta_{\text{s}} = \frac{\eta_{\text{s}}^0}{1 + \tilde{J}(I_{\text{th}} R_{\text{s}} / V_{\text{th}})}, \tag{13.146}$$

$$\eta_{\text{c}} = \eta_{\text{s}} \frac{\tilde{J}}{1 + \tilde{J}} = \frac{\eta_{\text{s}}^0}{1 + \tilde{J}(I_{\text{th}} R_{\text{s}} / V_{\text{th}})} \frac{\tilde{J}}{1 + \tilde{J}}, \tag{13.147}$$

$$\gamma_{\text{n}} = \tilde{J} \gamma_{\text{n}}^0, \quad \gamma_{\text{p}} = \tilde{J} \gamma_{\text{p}}^0, \tag{13.148}$$

$$\gamma_{\text{r}} = \tilde{J} \gamma_{\text{r}}^0 + (1 - \tilde{J}) \gamma_{\text{s}}, \tag{13.149}$$

and

$$f_{\text{r}} = \sqrt{\tilde{J}} f_{\text{r}}^0, \quad f_{\text{pk}} \approx \sqrt{\tilde{J}} f_{\text{pk}}^0, \quad f_{\text{3dB}} = \sqrt{\tilde{J}} f_{\text{3dB}}^0. \tag{13.150}$$

13.10.6 Find the following performance parameters for the VCSEL considered in Example 13.17 by using the results obtained in Examples 13.18–13.21 when it operates at three times the threshold current with $I = 3I_{\text{th}}$: (a) the output power

P_{out}; (b) the efficiencies η_e, η_c, and η_s; (c) the relaxation rates γ_c, γ_s, γ_n, and γ_p; and (d) the parameters γ_r, f_r, K, f_{pk}, and f_{3dB} for its frequency response.

13.10.7 Summarize the basic differences between the characteristics of an LED and a semiconductor laser.

SELECT BIBLIOGRAPHY

Agrawal, G. P. and Dutta, N. K., *Semiconductor Lasers*, 2nd edn. New York: Van Nostrand Reinhold, 1993.

Long-Wavelength Semiconductor Lasers. New York: Van Nostrand Reinhold, 1986.

Bhattacharya, P., *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall, 1997.

Cheo, P. K., *Fiber Optics and Optoelectronics*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Chuang, S. L., *Physics of Optoelectronic Devices*. New York: Wiley, 1995.

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge, University Press, 1996.

Ebeling, K. J., *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*. Berlin: Springer-Verlag, 1993.

Gillessen, K. and Schairer, W., *Light Emitting Diodes*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

Gooch, C. H., *Injection Electroluminescent Devices*. New York: Wiley, 1973.

Gowar, J., *Optical Communication Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Iizuka, K., *Elements of Photonics for Fiber and Integrated Optics*, Vol. II. New York: Wiley, 2002.

Kapon, E., ed., *Semiconductor Lasers I: Fundamentals*. San Diego, CA: Academic Press, 1999.

Semiconductor Lasers II: Materials and Structures. San Diego, CA: Academic Press, 1999.

Kasap, S. O., *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice-Hall, 2001.

Kressel, H. and Butler, J. K., *Semiconductor Lasers and Heterojunction LEDs*. New York: Academic Press, 1977.

Mroziewicz, B., Bugajski, M., and Nakwaski, W., *Physics of Semiconductor Lasers*. Amsterdam: North-Holland, 1991.

Nakamura, S. and Fasol, G., *The Blue Laser Diode*: *GaN Based Light Emitters and Lasers*. Berlin: Springer, 1997.

Pankove, J. I., *Optical Processes in Semiconductors*. New York: Dover, 1975.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Powers, J., *An Introduction to Fiber Optic Systems*. 2nd edn. Chicago, IL: Irwin, 1997.

Rao, R. P., ed., *Luminescence: Phenomena, Materials, and Devices*. New York: Nova Science Publishers, 1992.

Rosencher, E. and Vinter, B., *Optoelectronics*. Cambridge: Cambridge University Press, 2002.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Thompson, G. H. B., *Physics of Semiconductor Laser Devices*. New York: Wiley, 1980.

Verdeyen, J. T., *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Willardson, R. K. and Beer, A. C., eds., *Semiconductors and Semimetals*, Vol. 22, W. T. Tsang, ed., *Lightwave Communications Technology, Part B, Semiconductor Injection Lasers, I*. New York: Academic Press, 1985.

Semiconductors and Semimetals, Vol. 22, W. T. Tsang, ed., *Lightwave Communications Technology, Part C, Semiconductor Injection Lasers, II*. New York: Academic Press, 1985.

Semiconductors and Semimetals, Vol. 22, W. T. Tsang, ed., *Lightwave Communications Technology, Part E, Integrated Optoelectronics*. New York: Academic Press, 1985.

Williams, E. W. and Hall, R., *Luminescence and the Light Emitting Diode*. New York: Pergamon Press, 1978.

Wilson, J. and Hawkes, J. F. B., *Optoelectronics: An Introduction*, 3rd edn. London: Prentice-Hall Europe, 1998.

Yariv, A., *Optical Electronics in Modern Communications*, 5th edn. Oxford: Oxford University Press, 1997.

Yeh, C., *Applied Photonics*. San Diego, CA: Academic Press, 1994.

ADVANCED READING LIST

Alferov, Z. I., "Quantum wells and superlattices come of age," *III–Vs Review* **10**(7): 26–31, Nov. 1997.

"The double heterostructure: concept and its applications in physics, electronics and technology," *International Journal of Modern Physics B* **16**(5): 647–675, Feb. 2002.

"Nobel lecture: the double heterostructure concept and its applications in physics, electronics, and technology," *Reviews of Modern Physics* **73**(3): 767–782, July 2001.

Bruce, E., "Tunable lasers," *IEEE Spectrum* **39**(2): 35–39, Feb. 2002.

Chow, W. W., Choquette, K. D., Crawford, M. H., Lear, K. L., and Hadley, G. R., "Design, fabrication, and performance of infrared and visible vertical-cavity surface-emitting lasers," *IEEE Journal of Quantum Electronics* **33**(10): 1810–1824, Oct. 1997.

Coleman, J. J., "Strained-layer InGaAs quantum-well heterostructure lasers," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1008–1013, Nov.–Dec. 2000.

Craford, M. G., "LEDs challenge the incandescents," *IEEE Circuits and Devices Magazine* **8**(5): 24–29, Sep. 1992.

Delbeke, D., Bockstaele, R., Bienstman, P., Baets, R., and Benisty, H., "High-efficiency semiconductor resonant-cavity light-emitting diodes: a review," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(2): 189–206, Mar.–Apr. 2002.

Geels, R. S., Corzine, S. W., and Coldren, L. A., "InGaAs vertical-cavity surface-emitting lasers," *IEEE Journal of Quantum Electronics* **27**(6): 1359–1367, June 1991.

Gmachl, C., Capasso, F., Sivco, D. L., and Cho, A. Y., "Recent progress in quantum cascade lasers and applications," *Reports on Progress in Physics* **64**(11): 1533–1601, Nov. 2001.

Göbel, E. O. and Ploog, K., "Fabrication and optical properties of semiconductor quantum wells and superlattices," *Progress in Quantum Electronics* **14**(4): 289–356, 1990.

Hagberg, M., Eriksson, N., and Larsson, A., "Investigation of high-efficiency surface-emitting lasers with blazed grating outcouplers," *IEEE Journal of Quantum Electronics* **32**(9): 1596–1605, Sep. 1988.

Holonyak, N., Jr., "Quantum-well semiconductor lasers (review)," *Soviet Physics–Semiconductors* **19**(9): 943–958, Sep. 1985.

"The semiconductor laser: a thirty-five-year perspective," *Proceedings of the IEEE* **85**(11): 1678–1693, Nov. 1997.

Iga, K., "Surface emitting semiconductor lasers and surface-operating functional devices," *Electronics and Communications in Japan*, *Part 2* **75**(10): 12–26, 1992.

"Surface emitting lasers," *Electronics and Communications in Japan*, *Part 2* **82**(10): 70–81, 1999.

Iga, K., Koyama, F., and Kinoshita, S., "Surface emitting semiconductor lasers," *IEEE Journal of Quantum Electronics* **24**(9): 1845–1855, Sep. 1988.

Jewell, J. L., Harbison, J. P., Schere, A., Lee, Y. H., and Florez, L. T., "Vertical-cavity surface-emitting lasers: design, growth, fabrication, characterization," *IEEE Journal of Quantum Electronics* **27**(6): 1332–1346, June 1991.

Karim, A., Bjorlin, S., Piprek, J., and Bowers, J. E., "Long-wavelength vertical-cavity lasers and amplifiers," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1244–1253, Nov.–Dec. 2000.

Krames, M. R., Amano, H., Brown, J. J., and Heremans, P. L., "Introduction to the issue on high-efficiency light-emitting diodes," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(2): 185–188, Mar.–Apr. 2002.

Kroemer, H., "Nobel lecture: quasielectric fields and band offsets: teaching electrons new tricks," *Reviews of Modern Physics* **73**(3): 783–793, July 2001.

Kunii, T. and Matsui, Y., "Narrow spectral linewidth semiconductor lasers," *Optical and Quantum Electronics* **24**(7): 719–735, July 1992.

Lee, T. P., "Recent advances in long-wavelength semiconductor lasers for optical fiber communication," *Proceedings of the IEEE* **79**(3): 253–276, Mar. 1991.

Liu, J. M. and Simpson, T. B., "Four-wave mixing and optical modulation in a semiconductor laser," *IEEE Journal of Quantum Electronics* **30**(4): 957–965, Apr. 1994.

Mitsunaga, K., Kameya, M., Kojima, K., Noda, S., Kyuma, K., Hamanaka, K., and Nakayama, T., "CW surface-emitting grating-coupled GaAs/AlGaAs distributed feedback laser with very narrow beam divergence," *Applied Physics Letters* **50**(25): 1788–1790, June 1987.

Mukai, T., Nagahama, S., Iwasa, N., Senoh, M., and Yamada, T., "Nitride light-emitting diodes," *Journal of Physics: Condensed Matter* **13**(32): 7089–7098, Aug. 2001.

Murata, S. and Mito, I., "Frequency-tunable semiconductor lasers," *Optical and Quantum Electronics* **22**(1): 1–15, Jan. 1990.

Nakamura, S., "InGaN-based violet laser diodes," *Semiconductor Science and Technology* **14**(6): R27–R40, June 1999.

Nurmikko, A. and Gunshor, R. L., "Blue and green semiconductor lasers: a status report," *Semiconductor Science and Technology* **12**(11): 1337–1347, Nov. 1997.

Orton, J. W. and Foxon, C. T., "Group III nitride semiconductors for short wavelength light-emitting devices," *Reports on Progress in Physics* **61**(1): 1–75, Jan. 1998.

Patel, N. K., Cina, S., and Burroughes, J. H., "High-efficiency organic light-emitting diodes," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(2): 346–361, Mar.–Apr. 2002.

Ponce, F. A. and Bour, D. P., "Nitride-based semiconductors for blue and green light-emitting devices," *Nature* **386**: 351–359, Mar. 1997.

Rediker, R. H., "Semiconductor diode luminescence and lasers – a perspective," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1355–1362, Nov.–Dec. 2000.

Shtengel, G. E., Kazarinov, R. F., Belenky, G. L., Hybertsen, M. S., and Ackerman, D. A., "Advances in measurements of physical parameters of semiconductor lasers," *International Journal of High Speed Electronics and Systems* **9**(4): 901–940, Dec. 1998.

Simon, J. C., "Semiconductor laser amplifier for single mode optical fiber communications," *Journal of Optical Communications* **4**(2): 51–62, 1983.

Spencer, R. M., Greenberg, J., Eastman, L. F., Tsai, C. Y., and O'Keefe, S. S., "High-speed direct modulation of semiconductor lasers," *International Journal of High Speed Electronics and Systems* **8**(3): 417–456, Sep. 1997.

Streubel, K., Linder, N., Wirth, R., and Jaeger, A., "High brightness AlGaInP light-emitting diodes," *IEEE Journal of Selected Topics in Quantum Electronics* **8**(2): 321–332, Mar.–Apr. 2002.

Suematsu, Y. and Arai, S., "Single-mode semiconductor lasers for long-wavelength optical fiber communications and dynamics of semiconductor lasers," *IEEE Journal of Selected Topics in Quantum Electronics* **6**(6): 1436–1449, Nov.–Dec. 2000.

Takamori, T., Coldren, L. A., and Merz, J. L., "Lasing characteristics of a continuous-wave operated folded-cavity surface-emitting laser," *Applied Physics Letters* **56**(23): 2267–2269, June 1990.

Thijs, P. J. A., Tiemeijer, L. F., Binsma, J. J. M., and Van Dongen, T., "Strained-layer InGaAs(P) quantum well semiconductor lasers and semiconductor laser amplifiers," *Philips Journal of Research* **49**(3) 187–224, 1995.

Van Ruyven, L. J., "Double heterojunction lasers and quantum well lasers," *Journal of Luminescence* **29**: 123–161, 1984.

Vanderwater, D. A., Tan, I. H., Höfler, G. E., Defevere, D. C., and Kish, F. A., "High-brightness AlGaInP light emitting diodes," *Proceedings of the IEEE* **85**(11): 1752–1764, Nov. 1997.

14 Photodetectors

A photodetector is a device that converts an optical signal into a signal of another form. Most photodetectors convert optical signals into electrical signals, in the form of either current or voltage, that can be further processed or stored. All photodetectors are square-law detectors that respond to the power or intensity, rather than the field amplitude, of an optical signal. Based on the difference in the conversion mechanisms, there are two classes of photodetectors: *photon detectors* and *thermal detectors*. Photon detectors are *quantum detectors* based on the *photoelectric effect*, which converts a photon into an emitted electron or an electron–hole pair; a photon detector responds to the number of photons absorbed by the detector. Thermal detectors are based on the *photothermal effect*, which converts optical energy into heat; a thermal detector responds to the optical energy, rather than the number of photons, absorbed by the detector. Because of the difference in their fundamental mechanisms, there are a number of important differences in the general characteristics of these two classes of detectors.

The response of a photon detector is a function of optical wavelength with a long-wavelength cutoff, whereas that of a thermal detector is wavelength independent. A photon detector can be much more responsive than a thermal detector in a particular spectral region, which typically falls somewhere within the range from the near ultraviolet to the near infrared. In comparison, a thermal detector normally covers a wide spectral range from the deep ultraviolet to the far infrared with a nearly constant response. Photon detectors can be made extremely sensitive. Some of them have a photon-counting capability that is not possible for a thermal detector. A photon detector can be designed to have a high response speed capable of following very fast optical signals. Most thermal detectors are relatively slow in response because the speed of a thermal detector is limited by thermalization through heat diffusion and by heat dissipation when the power of an optical signal varies. For these reasons, photon detectors are suitable for detecting optical signals in photonic systems, whereas thermal detectors are most often used for optical power measurement or infrared imaging. In this chapter, only photon detectors are discussed because our major concern is with devices for photonics applications.

Photon detectors can be classified into two groups: one based on the *external photoelectric effect* and another based on the *internal photoelectric effect*. Photodetectors

based on the external photoelectric effect are *photoemissive devices*, such as vacuum photodiodes and the photomultiplier tubes, in which photoelectrons are ejected from the surface of a photocathode. Photodetectors based on the internal photoelectric effect are semiconductor devices, in which electron–hole pairs are generated through absorption of incident photons. A host of such devices have been developed, such as photoconductors, junction photodiodes, many photovoltaic devices, phototransistors, and charge-coupled devices.

14.1 Photodetector noise

Noise is one of the most fundamental phenomena in nature. It is ubiquitous. Noise in a photodetector sets the fundamental limit on the detectivity of the detector, thus determining the usefulness of a detector for a particular application. In terms of the physical nature, there are a few different types of noise for a photodetector. Two types of noise, *quantum noise* and *thermal noise*, originate from the basic physical laws of nature. Quantum noise, described as *shot noise* of electrons or photons in electronics and photonics, results from the statistical nature of a quantum event dictated by the uncertainty principle. Thermal noise, known as *Johnson noise* or *Nyquist noise* in electronics and photonics, is the consequence of thermal fluctuations and is directly associated with thermal radiation. Noise of such fundamental nature can only be minimized but can never be completely eliminated. In terms of physical sources, the noise of a photodetector can come from the following: the detector itself, the possible amplifier used in conjunction with the detector, and the circuit used to extract the electrical signal from the detector.

Noise appears in a signal as random fluctuations about the mean value of the signal. A measured signal s has a mean value of $\overline{s}$ defined as

$$\overline{s} = \sum_s p(s)s, \tag{14.1}$$

where $p(s)$ is the probability of the measured signal having a value s and the sum is carried out over all possible values obtained from measuring the signal. This mean value $\overline{s}$ is the expected value, or the ensemble average, of the variable s. The variance, or the mean square deviation, of the signal s is

$$\sigma_s^2 = \overline{(s - \overline{s})^2} = \overline{s^2} - \overline{s}^2. \tag{14.2}$$

The noise in a signal s can be expressed by a random variable s_n defined as

$$s_\mathrm{n} = s - \overline{s}. \tag{14.3}$$

The noise represented by the random variable s_n has a few general characteristics. As

can be seen clearly from (14.3), it has a zero mean value:

$$\overline{s_{\mathrm{n}}} = 0. \tag{14.4}$$

From (14.2) and (14.3), we find that the *mean square value* of s_{n} is equal to the variance of s:

$$\overline{s_{\mathrm{n}}^2} = \sigma_s^2 = \overline{s^2} - \overline{s}^2. \tag{14.5}$$

The mean square value of the noise in a signal is simply the mean square deviation of the signal. Because $\overline{s_{\mathrm{n}}} = 0$ but $\overline{s_{\mathrm{n}}^2} \neq 0$, the average amplitude of the noise vanishes but the power of the noise does not. Therefore, the magnitude of the noise is not measured by its average value but rather by its root mean square (rms) value defined as

$$\mathrm{rms}(s_{\mathrm{n}}) = \overline{s_{\mathrm{n}}^2}^{1/2}. \tag{14.6}$$

Noise characterized by random fluctuations is incoherent. If two or more independent noise sources, s_{n1}, $s_{\mathrm{n2}}, \ldots$, are simultaneously present in a signal s, their combined effect is not found by adding their amplitudes but is obtained by adding their mean square values, or their powers:

$$\overline{s_{\mathrm{n}}^2} = \overline{s_{\mathrm{n1}}^2} + \overline{s_{\mathrm{n2}}^2} + \cdots \tag{14.7}$$

The total noise from different independent sources then has an rms value of

$$\mathrm{rms}(s_{\mathrm{n}}) = \overline{s_{\mathrm{n}}^2}^{1/2} = \left(\overline{s_{\mathrm{n1}}^2} + \overline{s_{\mathrm{n2}}^2} + \cdots\right)^{1/2}. \tag{14.8}$$

One important figure of merit for a detection system is the *signal-to-noise ratio* (SNR or S/N). It is defined as the ratio of the power of a signal to the power of its noise or, equivalently, the ratio of the mean square of a signal to the mean square of its noise:

$$\mathrm{SNR} = \frac{\overline{s^2}}{\overline{s_{\mathrm{n}}^2}} = \frac{\overline{s^2}}{\sigma_s^2}, \qquad \text{or} \qquad \mathrm{SNR} = 10 \log \frac{\overline{s^2}}{\overline{s_{\mathrm{n}}^2}} \ \text{(dB)}. \tag{14.9}$$

The SNR defined above is also known as the *signal-to-noise power ratio* to be distinguished from the *signal-to-noise current ratio* defined as

$$\mathrm{SNR}_{\mathrm{current}} = \frac{\overline{s}}{\overline{s_{\mathrm{n}}^2}^{1/2}} = \frac{\overline{s}}{\sigma_s}. \tag{14.10}$$

Without specification, however, the SNR of a detection system generally refers to the signal-to-noise power ratio defined in (14.9).

In a photodetection system, a signal can take the form of photon number or photon flux as the input optical signal. It can also take the form of photocurrent or photovoltage as the output electrical signal. Therefore, the signal s can represent photon number, photon flux, photocurrent, or photovoltage. The general characteristics discussed above for the noise s_{n} apply to every case.

In the following discussions of photodetector noise, we consider an input optical signal with an optical power P_s. The detection system has an electrical response bandwidth of $\Delta f = B$, which can effectively sample the optical signal within a rectangular time interval of

$$T = \frac{1}{2B}. \tag{14.11}$$

The total number of photons received by the photodetector within this time interval is

$$\mathcal{S} = \frac{P_s}{h\nu}T = \frac{P_s}{2Bh\nu}. \tag{14.12}$$

If the photodetector has a *quantum efficiency* η_e, the total number of charge carriers generated in the detector by the photoelectric effect upon receiving the photons within the time interval T is

$$\mathcal{N} = \eta_e \mathcal{S} = \eta_e \frac{P_s}{2Bh\nu}, \tag{14.13}$$

where $0 \leq \eta_e \leq 1$. Consequently, the photocurrent in the detector is

$$i_{ph} = \frac{e\mathcal{N}}{T} = 2eB\mathcal{N} = \eta_e \frac{eP_s}{h\nu}, \tag{14.14}$$

where e is the electronic charge. For a detector without an internal gain, the signal current is simply $i_s = i_{ph}$. For a detector with an internal gain G, the signal current is $i_s = Gi_{ph}$.

Shot noise

The shot noise in a photodetector results from the quantum nature of the photons in the optical input and that of the charge carriers generated in the detector. Due to the quantum-mechanical probabilistic nature of photons, the photons in an optical signal are not distributed uniformly in time but arrive at the detector randomly in time. Therefore, both the power P_s of the optical signal and the number of photons $\mathcal{S}$ received in a given time interval T fluctuate randomly around their respective average values of $\overline{P_s}$ and $\overline{\mathcal{S}}$. The random fluctuations of the photons are characterized by the Poisson statistics. In any given time interval T, the probability of receiving $\mathcal{S}$ photons is given by the following Poisson probability distribution:

$$p(\mathcal{S}) = \frac{\overline{\mathcal{S}}^{\mathcal{S}} e^{-\overline{\mathcal{S}}}}{\mathcal{S}!}. \tag{14.15}$$

The mean square noise in the photon number fluctuations can then be calculated as (see Problem 14.1.3)

$$\overline{\mathcal{S}_n^2} = \sigma_{\mathcal{S}}^2 = \sum_{\mathcal{S}} p(\mathcal{S})(\mathcal{S} - \overline{\mathcal{S}})^2 = \overline{\mathcal{S}}. \tag{14.16}$$

This photon contribution to the noise of a photodetector is independent of the physical properties of the detector because it is external to the detector. It is the ultimate lower limit of the noise in an optical detection system. It sets the fundamental limit on the detectivity of a photodetector.

The photons received by a photodetector are converted to photoelectrons or electron–hole pairs, depending on the type of the detector, through the photoelectric effect. With a quantum efficiency η_e, which has a value between 0 and 1, the number of photoelectrons generated is only a fraction of that of the photons received by the detector. Because a given photon can only generate either one or no electron, but not a fraction of an electron, the photoelectric process is clearly quantum mechanical and probabilistic. The shot noise associated with this process has to be considered if the quantum efficiency is less than unity. This effect is fully accounted for by considering the statistics of the number $\mathcal{N}$ of charge carriers given in (14.13) that are generated with a quantum efficiency η_e of the photodetector. The random fluctuations of the charge carriers generated by the photoelectric effect are also characterized by the Poisson statistics with the following probability distribution for generating a number $\mathcal{N}$ in a time interval T:

$$p(\mathcal{N}) = \frac{\overline{\mathcal{N}}^{\mathcal{N}} e^{-\overline{\mathcal{N}}}}{\mathcal{N}!}, \tag{14.17}$$

where $\mathcal{N} = \eta_e \mathcal{S}$. We find, through a procedure similar to that used in (14.16), that the mean square noise in the number of photogenerated carriers is

$$\overline{\mathcal{N}_n^2} = \sigma_{\mathcal{N}}^2 = \overline{\mathcal{N}}. \tag{14.18}$$

Because $\overline{\mathcal{N}} < \overline{\mathcal{S}}$ if $\eta_e < 1$, the noise is actually reduced by an imperfect quantum efficiency. This result seems odd. However, what really counts in a detection system is not the noise alone, but the SNR. While the noise is reduced by an imperfect quantum efficiency of $\eta_e < 1$, the signal is reduced even more. As a result, the SNR is lower for a detector that has a poorer quantum efficiency (see Problem 14.1.4).

We consider here a detector without an internal gain, such that $i_s = i_{ph}$. Using (14.14) and (14.18), we find the following shot current noise in the photodetector:

$$\overline{i_{n,sh}^2} = 4e^2 B^2 \overline{\mathcal{N}_n^2} = 4e^2 B^2 \overline{\mathcal{N}} = 2eB\overline{i_s}. \tag{14.19}$$

We then have the following mean square current fluctuations for the shot noise of a photodetector that receives an optical power P_s from an input optical signal:

$$\overline{i_{n,sh}^2} = 2eB\overline{i_s} = 2\eta_e e^2 B \frac{\overline{P_s}}{h\nu}. \tag{14.20}$$

From this relation, we have (see Problem 14.1.3)

$$\overline{i_s^2} = \overline{i_s}^2 + 2eB\overline{i_s}. \tag{14.21}$$

In practice, there are other sources that also contribute to the shot noise of a photodetector. One important source is the photons from the background radiation that impinge on the detector. The contribution of this noise source can be minimized by reducing the aperture of the detector to the minimum needed for receiving the optical signal. It cannot be completely eliminated, however, because at the very minimum there is still background thermal radiation, which can only be reduced by reducing the temperature of the environment surrounding the detector. Another important source of shot noise is the *dark current* of the detector. The dark current is the current in a detector when it is not illuminated with any optical input. In a semiconductor device, it is normally caused by thermal generation of electron–hole pairs and by leakage currents due to surface defects of the device. When these additional noise sources are considered, the total shot noise of a photodetector is given by

$$\overline{i_{\mathrm{n,sh}}^2} = 2eB\bar{i} = 2eB(\overline{i_{\mathrm{s}}} + \overline{i_{\mathrm{b}}} + \overline{i_{\mathrm{d}}}), \tag{14.22}$$

where i_{b} is the photocurrent generated by background radiation and i_{d} is the dark current of the detector.

Excess shot noise

In a photodetector, such as a photomultiplier, a photoconductor, or an avalanche photodiode, that has an internal gain, both signal and noise are amplified. For a detector that has a gain G, the signal current, the background radiation current, and the dark current are all amplified by the factor G:

$$i_{\mathrm{s}} = Gi_{\mathrm{ph}} = G\eta_{\mathrm{e}}\frac{eP_{\mathrm{s}}}{h\nu} \tag{14.23}$$

and

$$i_{\mathrm{b}} = Gi_{\mathrm{b0}}, \qquad i_{\mathrm{d}} = Gi_{\mathrm{d0}}, \tag{14.24}$$

where i_{b0} and i_{d0} are unamplified background and dark currents, respectively, and i_{b} and i_{d} are amplified currents that can be directly measured externally. The shot noise is also amplified through a process of random multiplication of the noise electrons. The statistical nature of this random multiplication process results in an *excess noise factor*, F, which is a function of the material, the structure, and the gain of a detector. As a consequence, the mean square shot noise current for a detector with an internal gain can be expressed as

$$\overline{i_{\mathrm{n,sh}}^2} = 2eBG^2F(\overline{i_{\mathrm{ph}}} + \overline{i_{\mathrm{b0}}} + \overline{i_{\mathrm{d0}}}) = 2eBGF(\overline{i_{\mathrm{s}}} + \overline{i_{\mathrm{b}}} + \overline{i_{\mathrm{d}}}), \tag{14.25}$$

where the excess noise factor $F = \overline{G^2}/\overline{G}^2$ is a function of the gain. For a detector without an internal gain, we find that $G = 1$ and $F = 1$; then, the shot noise given in (14.25) reduces to that in (14.22), as expected.

Thermal noise

Thermal noise results from random thermal motions of the electrons in a conductor. It is associated with the blackbody radiation of a conductor at the radio or microwave frequency range of the signal. Because only materials that can absorb and dissipate energy can emit blackbody radiation, thermal noise is generated only by the resistive components of the detector and its circuit. Capacitive and inductive components do not generate thermal noise because they neither dissipate nor emit energy.

The energy of the thermal noise generated by a resistive element is independent of the detailed physical properties of the resistor but is dictated only by the law of blackbody radiation. At a temperature T, the thermal noise power in a small frequency interval of $\mathrm{d}f$ centered around f is

$$P_{\mathrm{n,th}}(f)\mathrm{d}f = \frac{4hf}{\mathrm{e}^{hf/k_{\mathrm{B}}T} - 1}\mathrm{d}f. \tag{14.26}$$

In normal operation of most photodetectors, $f \ll k_{\mathrm{B}}T/h$. Thus, the frequency dependence of the thermal noise power is negligible, resulting in

$$P_{\mathrm{n,th}}(f)\mathrm{d}f \approx 4k_{\mathrm{B}}T\mathrm{d}f. \tag{14.27}$$

Then, the total thermal noise power for a detection system of a bandwidth B is simply

$$P_{\mathrm{n,th}} = 4k_{\mathrm{B}}TB. \tag{14.28}$$

For a resistor that has a resistance R, the thermal noise can be treated as either current noise or voltage noise through the relation of $P_{\mathrm{n,th}} = \overline{i_{\mathrm{n,th}}^2}R = \overline{v_{\mathrm{n,th}}^2}/R$. Then, we have

$$\overline{i_{\mathrm{n,th}}^2} = \frac{4k_{\mathrm{B}}TB}{R} \tag{14.29}$$

and

$$\overline{v_{\mathrm{n,th}}^2} = 4k_{\mathrm{B}}TBR. \tag{14.30}$$

For an optical detection system, the resistance R is the total equivalent resistance, including the internal resistance of the detector and the load resistance from the circuit, at the output of the detector. For a detector that has a current signal, (14.29) is used. In this case, the thermal noise is determined by the lowest shunt resistance to the detector, which is often the load resistance of the detector. The thermal noise can be reduced by increasing this resistance at the expense of reducing the response speed of the system. For a detector that has a voltage signal, (14.30) is used. In this situation, the thermal noise is determined by the largest series resistance to the detector, which again is often the load resistance of the detector. The thermal noise can now be reduced by decreasing this resistance, but at the expense of reducing the output voltage signal.

Signal-to-noise ratio

There are other noise sources, such as the $1/f$ noise, but they are usually not important for the normal operation of photodetectors. Therefore, the total noise of a photodetector, whether it has an internal gain or not, is basically the sum of its shot noise and thermal noise:

$$\overline{i_n^2} = \overline{i_{n,sh}^2} + \overline{i_{n,th}^2}. \tag{14.31}$$

A photodetector is said to function in the *quantum regime* if $\overline{i_{n,sh}^2} > \overline{i_{n,th}^2}$. A photodetector operating in the quantum regime is *shot-noise limited* because shot noise is the primary source of noise in this regime. A photodetector is in the *thermal regime* if $\overline{i_{n,th}^2} > \overline{i_{n,sh}^2}$. A photodetector operating in the thermal regime is *thermal-noise limited* because its thermal noise dominates its shot noise in this regime.

For a photodetector that has no internal gain, the SNR is given by

$$\begin{aligned} \text{SNR} = \frac{\overline{i_s^2}}{\overline{i_n^2}} &= \frac{\overline{i_{ph}^2}}{2eB(\overline{i_{ph}} + \overline{i_b} + \overline{i_d}) + 4k_B T B/R} \\ &= \frac{\overline{P_s^2}\mathcal{R}^2}{2eB(\overline{P_s}\mathcal{R} + \overline{i_b} + \overline{i_d}) + 4k_B T B/R}, \end{aligned} \tag{14.32}$$

where $\mathcal{R} = \eta_e e/h\nu$ is the *responsivity* of a photodetector without an internal gain, defined in the following section. For a photodetector that has an internal gain G, the SNR is

$$\begin{aligned} \text{SNR} = \frac{\overline{i_s^2}}{\overline{i_n^2}} &= \frac{G^2\overline{i_{ph}^2}}{2eBG^2F(\overline{i_{ph}} + \overline{i_{b0}} + \overline{i_{d0}}) + 4k_B T B/R} \\ &= \frac{\overline{P_s^2}\mathcal{R}^2}{2eBGF(\overline{P_s}\mathcal{R} + \overline{i_b} + \overline{i_d}) + 4k_B T B/R}, \end{aligned} \tag{14.33}$$

where $\mathcal{R} = G\eta_e e/h\nu$ is the responsivity of a photodetector with an internal gain, also defined in the following section.

The relations in (14.32) and (14.33) apply to photodetectors that have current signals at the output. For a photodetector that has an output voltage signal, the SNR is defined as

$$\text{SNR} = \frac{\overline{v_s^2}}{\overline{v_n^2}} = \frac{\overline{P_s^2}\mathcal{R}^2}{\overline{v_n^2}}, \tag{14.34}$$

where $\mathcal{R}$ is the responsivity of a photodetector that has an output voltage signal defined in the following section.

EXAMPLE 14.1 A photodetector that responds to an optical signal with a photocurrent has a load resistance of $R = 50\ \Omega$ and a bandwidth of $B = 100$ MHz. It has a negligible

background radiation current and a dark current of $i_d = 10$ nA. (a) Find its shot noise, thermal noise, and signal-to-noise ratio when it generates a signal photocurrent of 1 μA. (b) What are its shot noise, thermal noise, and signal-to-noise ratio when it generates a signal photocurrent of 1 mA?

Solution (a) For $\overline{i_s} = \overline{i_{ph}} = 1$ μA, we find that the shot noise

$$\overline{i_{n,sh}^2} = 2eB(\overline{i_s} + \overline{i_d}) = 2 \times 1.6 \times 10^{-19} \times 100 \times 10^6 \times (1 \times 10^{-6} + 10 \times 10^{-9})\ \text{A}^2$$
$$= 3.23 \times 10^{-17}\ \text{A}^2.$$

At $T = 300$ K, $k_B T = 25.9$ meV. The thermal noise for $R = 50\ \Omega$ is

$$\overline{i_{n,th}^2} = \frac{4k_B T B}{R} = \frac{4 \times 25.9 \times 10^{-3} \times 1.6 \times 10^{-19} \times 100 \times 10^6}{50}\ \text{A}^2$$
$$= 3.32 \times 10^{-14}\ \text{A}^2.$$

Thus, the total noise

$$\overline{i_n^2} = \overline{i_{n,sh}^2} + \overline{i_{n,th}^2} = 3.23 \times 10^{-17}\ \text{A}^2 + 3.32 \times 10^{-14}\ \text{A}^2 = 3.32 \times 10^{-14}\ \text{A}^2.$$

We see that in this example the shot noise is mainly contributed by the signal photocurrent, but the shot noise is negligible compared to thermal noise. From (14.21), we have

$$\overline{i_s^2} = \overline{i_s}^2 + 2eB\overline{i_s} = (1 \times 10^{-6})^2\ \text{A}^2 + 3.23 \times 10^{-17}\ \text{A}^2 = 1 \times 10^{-12}\ \text{A}^2.$$

We find that $\overline{i_s^2} \approx \overline{i_s}^2$ in this example. Thus, the SNR is

$$\text{SNR} = \frac{\overline{i_s^2}}{\overline{i_n^2}} = \frac{1 \times 10^{-12}}{3.32 \times 10^{-14}} = 30,$$

which is 14.8 dB.

(b) For $\overline{i_s} = \overline{i_{ph}} = 1$ mA, the shot noise

$$\overline{i_{n,sh}^2} = 2eB(\overline{i_s} + \overline{i_d}) = 2 \times 1.6 \times 10^{-19} \times 100 \times 10^6 \times (1 \times 10^{-3} + 10 \times 10^{-9})\ \text{A}^2$$
$$= 3.2 \times 10^{-14}\ \text{A}^2.$$

The thermal noise is the same as that found in (a): $\overline{i_{n,th}^2} = 3.32 \times 10^{-14}\ \text{A}^2$. In this example, the shot noise is contributed almost entirely by the signal photocurrent and is comparable to the thermal noise. The total noise

$$\overline{i_n^2} = \overline{i_{n,sh}^2} + \overline{i_{n,th}^2} = 3.2 \times 10^{-14}\ \text{A}^2 + 3.32 \times 10^{-14}\ \text{A}^2 = 6.52 \times 10^{-14}\ \text{A}^2.$$

We also have

$$\overline{i_s^2} = \overline{i_s}^2 + 2eB\overline{i_s} = (1 \times 10^{-3})^2 \text{ A}^2 + 3.2 \times 10^{-14} \text{ A}^2 = 1 \times 10^{-6} \text{ A}^2.$$

The SNR is

$$\text{SNR} = \frac{\overline{i_s^2}}{\overline{i_n^2}} = \frac{1 \times 10^{-6}}{6.52 \times 10^{-14}} = 1.53 \times 10^7,$$

which is 71.8 dB. We see that the SNR is significantly increased by 57 dB when the signal photocurrent is increased from 1 μA to 1 mA. The reason for this significant improvement is that the detector is limited by thermal noise at low photocurrents.

14.2 Photodetector performance parameters

Several parameters are commonly used to define the performance characteristics of photodetectors. These parameters can be considered as the figures of merit of a photodetector. They are used for comparing one photodetector with another and for determining the suitability of a photodetector for a particular application. In this section, the basic concepts of these parameters are defined and discussed.

Spectral response

Because the response of a photon detector is wavelength dependent, a given photodetector is responsive only within a finite, specific range of the optical spectrum. The spectral range of response for a photodetector is determined by the material, the structure, and the packaging of the detector. The spectral response of a photodetector is usually specified in terms of the spectral responsivity or the spectral detectivity of the detector. In choosing a photodetector for an application, the match between the spectral content of the optical signal and the spectral response of the detector is the first thing to be verified.

Quantum efficiency

Quantum efficiency is the probability of generating a charge carrier in a photodetector for each photon that is incident on the detector. Similarly to the *external quantum efficiency*, η_e, of an LED or a semiconductor laser, the external quantum efficiency of a photodetector is reduced from its *internal quantum efficiency*, η_i, by the *transmission efficiency*, η_t, of the incident optical beam into the active region of the detector and by the *collection efficiency*, η_{coll}, of the photogenerated electrical carriers into a photocurrent. Thus, we can express the external quantum efficiency of a photodetector as

$$\eta_e = \eta_{coll}\eta_t\eta_i. \tag{14.35}$$

Comparing this relation with that in (13.67) for an LED and that in (13.127) for a semiconductor laser, we find that the carrier collection efficiency η_{coll} of a photodetector is equivalent to the carrier injection efficiency η_{inj} of an LED or a laser, and that the optical transmission efficiency η_{t} of a photodetector is equivalent to the photon extraction efficiency η_{t} of an LED or a laser.

As expressed in (14.13), the external quantum efficiency can be defined as the ratio of the number of photogenerated charge carriers, in the form of either photoelectrons or electron–hole pairs, that actually contribute to the photocurrent to the number of incident photons: $\eta_{\text{e}} = \mathcal{N}/\mathcal{S}$. According to (14.14), the external quantum efficiency of a detector can then be expressed in terms of the incident optical power and the photocurrent as

$$\eta_{\text{e}} = \frac{i_{\text{ph}}/e}{P_{\text{s}}/h\nu} = \frac{h\nu i_{\text{ph}}}{eP_{\text{s}}}. \tag{14.36}$$

The quantum efficiency of a photodetector is a function of the wavelength of the incident photons because of the spectral response of the detector. Its wavelength dependence arises not only from its explicit dependence on the optical frequency ν seen in (14.36) but also from the wavelength dependence of the ratio $i_{\text{ph}}/P_{\text{s}}$ defined below as the responsivity of the detector.

EXAMPLE 14.2 A Si photodetector responds to an optical signal at 850 nm of 1 mW power with a photocurrent of 500 μA. What is its external quantum efficiency?

Solution At $\lambda = 850$ nm, we have

$$\frac{h\nu}{e} = \frac{1239.8}{850}\ \text{V}.$$

Therefore, for $i_{\text{ph}} = 500\ \mu\text{A}$ in response to $P_{\text{s}} = 1$ mW, we find from (14.36) the following external quantum efficiency for this detector:

$$\eta_{\text{e}} = \frac{h\nu i_{\text{ph}}}{eP_{\text{s}}} = \frac{1239.8}{850} \times \frac{500 \times 10^{-6}}{1 \times 10^{-3}} = 72.9\%.$$

Responsivity

Responsivity is an important parameter for a photodetector. It allows one to determine the available output signal of a detector for a given input optical signal. The responsivity of a photodetector is defined as the ratio of the output current or voltage signal to the power of the input optical signal. For a photodetector that has an output current signal, the responsivity is defined as

$$\mathcal{R} = \frac{i_{\text{s}}}{P_{\text{s}}} \qquad (\text{A W}^{-1}). \tag{14.37}$$

For a photodetector that has an output voltage signal, the responsivity is defined as

$$\mathcal{R} = \frac{v_s}{P_s} \qquad (\text{V W}^{-1}). \tag{14.38}$$

Because most of the commonly used photodetectors have output current signals, we consider in further detail the responsivity of such photodetectors in the following. Similar concepts can be extended to photodetectors that have output voltage signals.

For a photodetector without an internal gain, the signal current is simply the photocurrent, $i_s = i_{ph}$. By using (14.36), we find the following expression for its responsivity:

$$\mathcal{R} = \frac{i_{ph}}{P_s} = \eta_e \frac{e}{h\nu}. \tag{14.39}$$

For a photodetector with an internal gain, however, the signal current is amplified by the gain, $i_s = G i_{ph}$, and the responsivity is

$$\mathcal{R} = \frac{G i_{ph}}{P_s} = G\eta_e \frac{e}{h\nu} = G\mathcal{R}_0, \tag{14.40}$$

where $\mathcal{R}_0$ is the *intrinsic responsivity* of the detector defined as

$$\mathcal{R}_0 = \frac{i_{ph}}{P_s} = \eta_e \frac{e}{h\nu}. \tag{14.41}$$

The responsivity of a photodetector without an internal gain is simply its intrinsic responsivity, $\mathcal{R} = \mathcal{R}_0$, whereas one with an internal gain has a responsivity $\mathcal{R} = G\mathcal{R}_0$.

The spectral response of a photodetector is usually characterized by the responsivity of the detector as a function of optical wavelength, $\mathcal{R}(\lambda)$, which is known as the *spectral responsivity*. In addition, the responsivity of a photodetector is also a function of signal frequency f. Its frequency dependence, $\mathcal{R}(f)$, characterizes the frequency response of the detector, as discussed later.

EXAMPLE 14.3 Find the responsivity at 850 nm for the Si photodetector described in Example 14.2.

Solution From (14.39), the responsivity of this detector at 850 nm is simply

$$\mathcal{R} = \frac{i_{ph}}{P_s} = \frac{500 \times 10^{-6}}{1 \times 10^{-3}} \text{ A W}^{-1} = 0.5 \text{ A W}^{-1}.$$

Noise equivalent power

The noise equivalent power (NEP) of a photodetector is defined as the input power required of the optical signal for the signal-to-noise ratio to be unity, SNR = 1, at the detector output. Then, using the relations in (14.32) and (14.33), the NEP for a photodetector, with or without an internal gain, that has an output current signal can be

defined as

$$\text{NEP} = \frac{\overline{i_n^2}^{1/2}}{\mathcal{R}} = \frac{\text{rms}(i_n)}{\mathcal{R}} \qquad \text{(W)}, \tag{14.42}$$

where $\overline{i_n^2}$ is the mean square noise current at an input optical power level for SNR = 1 and $\mathcal{R}$ is the responsivity defined in (14.37). Using the relation in (14.34), the NEP for a photodetector that has an output voltage signal can be defined as

$$\text{NEP} = \frac{\overline{v_n^2}^{1/2}}{\mathcal{R}} = \frac{\text{rms}(v_n)}{\mathcal{R}} \qquad \text{(W)}, \tag{14.43}$$

where $\overline{v_n^2}$ is the mean square noise voltage at an input optical power level for SNR = 1 and $\mathcal{R}$ is the responsivity defined in (14.38).

For most detection systems at the small input signal level for SNR = 1, the shot noise contributed by the input optical signal is negligible compared to both the shot noise from other sources and the thermal noise of the detector. In this situation, the NEP of a photodetector with no internal gain that has an output current signal can be expressed as

$$\text{NEP} = \frac{(2e\overline{i_b} + 2e\overline{i_d} + 4k_B T/R)^{1/2}}{\mathcal{R}} B^{1/2}. \tag{14.44}$$

The most fundamental limit of a photodetector is the noise contributed by the ubiquitous blackbody radiation in the background. This background radiation sets the absolute minimum of NEP for a photodetector. It is often the limitation for photodetectors in mid- and far-infrared spectral regions, but it is normally not important for photodetectors in visible and ultraviolet spectral regions. For most photodetectors responding to optical wavelengths shorter than 3 μm, the noise from background blackbody radiation is dominated by that from the dark current or that from resistive thermal noise, or both. For such a photodetector, the intrinsic NEP is that defined by its dark current by assuming that the load resistance is sufficiently large if the detector generates a photocurrent signal, or sufficiently small if it generates a photovoltage signal. However, in order to reduce its RC time constant, a high-speed photodetector that has a current signal normally has a small area, thus a small dark current, but requires a small load resistance, thus a large thermal noise. Therefore, the NEP of a high-speed photodetector is usually limited by the thermal noise from its external load resistance rather than by the shot noise from its internal dark current.

Because the mean square noise of a detector is proportional to the detector bandwidth, $\overline{i_n^2} \propto B$ and $\overline{v_n^2} \propto B$, the NEP of a photodetector is proportional to the square root of the detector bandwidth: NEP $\propto B^{1/2}$. Therefore, the NEP of a photodetector is often specified in terms of the NEP for a bandwidth of 1 Hz as NEP$/B^{1/2}$, in the unit of W Hz$^{-1/2}$.

EXAMPLE 14.4 The Si photodetector considered in Examples 14.2 and 14.3 has an active area of $\mathcal{A} = 5$ mm^2, a bandwidth of $B = 100$ MHz, and a dark current of $i_d = 10$ nA. (a) Find its shot-noise-limited NEP, its thermal-noise-limited NEP, and its total NEP, all for a bandwidth of 1 Hz. (b) Find its shot-noise-limited NEP, its thermal-noise-limited NEP, and its total NEP, all for its entire bandwidth.

Solution (a) The shot noise from the dark current is

$$\overline{i_{n,sh}^2} = 2eB\overline{i_d} = 2 \times 1.6 \times 10^{-19} \times 10 \times 10^{-9} \times B \text{ A}^2 \text{ Hz}^{-1}$$
$$= 3.2 \times 10^{-27} B \text{ A}^2 \text{ Hz}^{-1}.$$

The thermal noise

$$\overline{i_{n,th}^2} = \frac{4k_B T B}{R} = \frac{4 \times 25.9 \times 10^{-3} \times 1.6 \times 10^{-19}}{50} \times B \text{ A}^2 \text{ Hz}^{-1}$$
$$= 3.32 \times 10^{-22} B \text{ A}^2 \text{ Hz}^{-1}.$$

The total noise $\overline{i_n^2} = \overline{i_{n,sh}^2} + \overline{i_{n,th}^2} = 3.32 \times 10^{-22} B$ A^2 Hz^{-1}, which is completely dominated by thermal noise. From Example 14.3, we have $\mathcal{R} = 0.5$ A W^{-1} for this detector. Thus, the shot-noise-limited NEP for a bandwidth of 1 Hz is

$$\frac{(\text{NEP})_{sh}}{B^{1/2}} = \frac{\overline{i_{n,sh}^2}^{1/2}}{B^{1/2}\mathcal{R}} = \frac{(3.2 \times 10^{-27})^{1/2}}{0.5} \text{ W Hz}^{-1/2} = 113 \text{ fW Hz}^{-1/2}.$$

The thermal-noise-limited NEP for a bandwidth of 1 Hz is

$$\frac{(\text{NEP})_{th}}{B^{1/2}} = \frac{\overline{i_{n,th}^2}^{1/2}}{B^{1/2}\mathcal{R}} = \frac{(3.32 \times 10^{-22})^{1/2}}{0.5} \text{ W Hz}^{-1/2} = 36.4 \text{ pW Hz}^{-1/2}.$$

The total NEP for a bandwidth of 1 Hz is

$$\frac{\text{NEP}}{B^{1/2}} = \frac{\overline{i_n^2}^{1/2}}{B^{1/2}\mathcal{R}} = \frac{(3.32 \times 10^{-22})^{1/2}}{0.5} \text{ W Hz}^{-1/2} = 36.4 \text{ pW Hz}^{-1/2}.$$

(b) For $B = 100$ MHz, we find that the shot-noise-limited NEP for the entire bandwidth is

$$(\text{NEP})_{sh} = 113 \times 10^{-15} \times (100 \times 10^6)^{1/2} \text{ W} = 1.13 \text{ nW}.$$

The thermal-noise-limited NEP for the entire bandwidth is

$$(\text{NEP})_{th} = 36.4 \times 10^{-12} \times (100 \times 10^6)^{1/2} \text{ W} = 364 \text{ nW}.$$

The total NEP for the entire bandwidth is

$$\text{NEP} = 36.4 \times 10^{-12} \times (100 \times 10^6)^{1/2} \text{ W} = 364 \text{ nW}.$$

We see that this detector is completely limited by the thermal noise of its load resistance.

Detectivity

The detectivity characterizes the ability of a photodetector to detect a small optical signal. It is defined as the inverse of the NEP of the detector:

$$D = \frac{1}{\text{NEP}} \qquad (\text{W}^{-1}). \tag{14.45}$$

As discussed above, NEP $\propto B^{1/2}$ when the shot noise contributed by the input optical signal is negligibly small compared to the noise from other sources. In addition, the background radiation current, i_b, and the dark current, i_d, are often proportional to the surface area, $\mathcal{A}$, of a photodetector. Therefore, when i_b and i_d are the dominant sources of noise for a photodetector, the intrinsic noise characteristics of the detector can be better quantified by normalizing NEP to $(\mathcal{A}B)^{1/2}$. A useful intrinsic parameter of a photodetector is the *specific detectivity*, D^*, defined as

$$D^* = \frac{(\mathcal{A}B)^{1/2}}{\text{NEP}} \qquad (\text{cm Hz}^{1/2}\ \text{W}^{-1}). \tag{14.46}$$

Then, for a dark-current-limited photodetector without an internal gain, we have

$$D^* \approx \frac{\mathcal{A}^{1/2}\mathcal{R}}{(2e\overline{i}_d)^{1/2}}. \tag{14.47}$$

The specific detectivity D^* is independent of the area of the detector. It is a measure of the intrinsic detection capability of the material and the structure of the detector.

The detectivity of a photodetector is a function of the wavelength of the optical signal. The spectral characteristics of the detectivity, given as $D(\lambda)$ or $D^*(\lambda)$, reflect the spectral response of a photodetector. The detectivity is also a function of the modulation signal frequency f carried by the optical beam.

EXAMPLE 14.5 Find the detectivity and the specific detectivity of the Si photodetector considered in Example 14.4 for the following two situations: (a) when the detector is shot-noise limited by its dark current with a large load resistance and (b) when the detector has a 50 Ω load resistance.

Solution (a) As given in Example 14.4, the detector has an active area of $\mathcal{A} = 5\ \text{mm}^2 = 5 \times 10^{-2}\ \text{cm}^2$. When the detector is shot-noise limited by its dark current, it has a detectivity

$$D = \frac{1}{(\text{NEP})_{\text{sh}}} = \frac{1}{1.13 \times 10^{-9}}\ \text{W}^{-1} = 8.85 \times 10^{8}\ \text{W}^{-1}$$

and a specific detectivity

$$D^* = \frac{(\mathcal{A}B)^{1/2}}{(\text{NEP})_{\text{sh}}} = \frac{(5 \times 10^{-2} \times 100 \times 10^{6})^{1/2}}{1.13 \times 10^{-9}}\ \text{cm Hz}^{1/2}\ \text{W}^{-1}$$
$$= 1.98 \times 10^{12}\ \text{cm Hz}^{1/2}\ \text{W}^{-1}.$$

(b) When the detector has a 50 Ω load resistance, it has a detectivity

$$D = \frac{1}{\text{NEP}} = \frac{1}{364 \times 10^{-9}}\ \text{W}^{-1} = 2.75 \times 10^{6}\ \text{W}^{-1}$$

and a specific detectivity

$$D^* = \frac{(\mathcal{A}B)^{1/2}}{\text{NEP}} = \frac{(5 \times 10^{-2} \times 100 \times 10^{6})^{1/2}}{364 \times 10^{-9}}\ \text{cm Hz}^{1/2}\ \text{W}^{-1}$$
$$= 6.14 \times 10^{9}\ \text{cm Hz}^{1/2}\ \text{W}^{-1}.$$

We find that when the photodetector is loaded with a 50 Ω resistance, its detectivity and specific detectivity are limited by the resistive thermal noise and are much lower than its intrinsic detectivity and specific detectivity, which are limited by the shot noise from its dark current.

Linearity and dynamic range

Linearity of a photodetector is defined by the response of the detector being linear, meaning that its output current or voltage signal is linearly proportional to its input optical signal. Linear response is required for a photodetector to convert the waveform of an input optical signal faithfully to an output electrical signal without distortion. When a photodetector has a linear response, its quantum efficiency η_e and responsivity $\mathcal{R}$ defined above are constants that are independent of the power P_s of the input optical signal. However, every practical photodetector only has a finite range of linear response, as shown in Fig. 14.1. As the power of the input optical signal reaches a certain level, the response of a photodetector starts to saturate, thereby deviating from linearity.

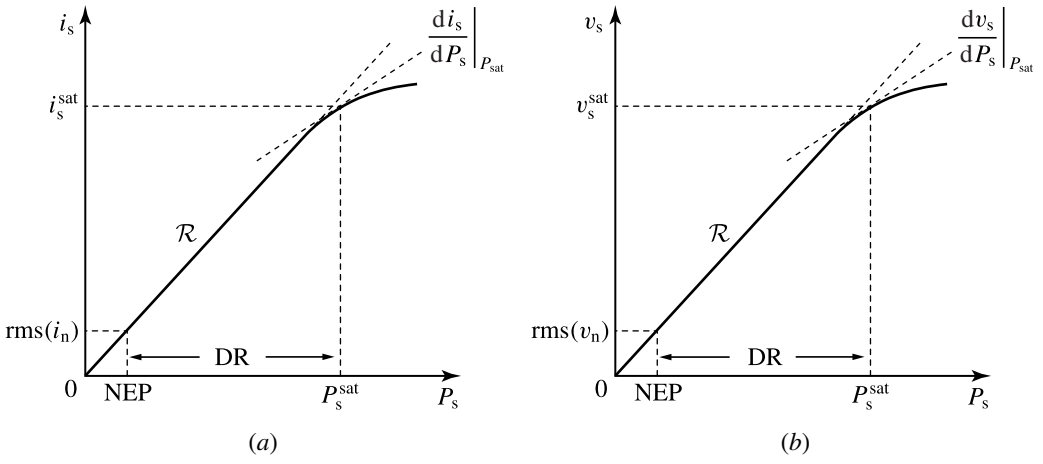


Figure 14.1 Typical response characteristics as a function of the power of the input optical signal for (*a*) a photodetector with an output current signal and (*b*) a photodetector with an output voltage signal.

The maximum input signal power acceptable is determined by the maximum deviation from the linear response of a photodetector that can be tolerated in a particular application. Given the maximum tolerable deviation from linearity to be δ (for $100\delta\%$), the saturation signal power, P_s^{sat}, for the photodetector in the application is the corresponding maximum acceptable input power. As illustrated in Fig. 14.1, the value of P_s^{sat} can be found from

$$\left.\frac{di_s}{dP_s}\right|_{P_s=P_s^{sat}} = (1-\delta)\mathcal{R} \qquad \text{or} \qquad \left.\frac{dv_s}{dP_s}\right|_{P_s=P_s^{sat}} = (1-\delta)\mathcal{R}, \tag{14.48}$$

where $\mathcal{R}$ is the responsivity of the detector in the linear range.

The usefulness of a photodetector for detecting an optical signal is clearly limited by its saturation, which is quantified by P_s^{sat}, at the large-signal end and by its detectivity, which is determined by the NEP of the detector, at the small-signal end. The range of the input signal power above the NEP but below P_s^{sat} in the linear-response region is the useful range of operation for a photodetector. This range is known as the *dynamic range* (DR) of the detector, as indicated in Fig. 14.1. The dynamic range is usually quantified as

$$\text{DR} = 10\log\frac{P_s^{sat}}{\text{NEP}} \qquad \text{(dB)}. \tag{14.49}$$

Alternatively, the dynamic range of a photodetector is also frequently stated in terms of the number of orders of magnitude in the input power from the NEP to P_s^{sat}.

EXAMPLE 14.6 With a load resistance of 50 Ω, the Si photodetector considered in the preceding examples has a saturation current of 10 mA. Find its saturation optical signal power and its dynamic range.

Solution Because $\mathcal{R} = 0.5$ A W^{-1} for this detector, the saturation optical signal power corresponding to $i_s^{sat} = 10$ mA is

$$P_s^{sat} = \frac{i_s^{sat}}{\mathcal{R}} = \frac{10}{0.5}\ \text{mW} = 20\ \text{mW}.$$

The NEP of this detector in the presence of a 50 Ω load resistance is 364 nW from Example 14.4. Therefore, the dynamic range of the detector is

$$\text{DR} = 10\log\frac{20\times10^{-3}}{364\times10^{-9}}\ \text{dB} = 47.4\ \text{dB}.$$

Speed and frequency response

The response speed of a photodetector is directly related to its frequency response. It determines the ability of a photodetector to follow a fast-varying optical signal. To record an optical signal faithfully, a photodetector must have a speed higher than the

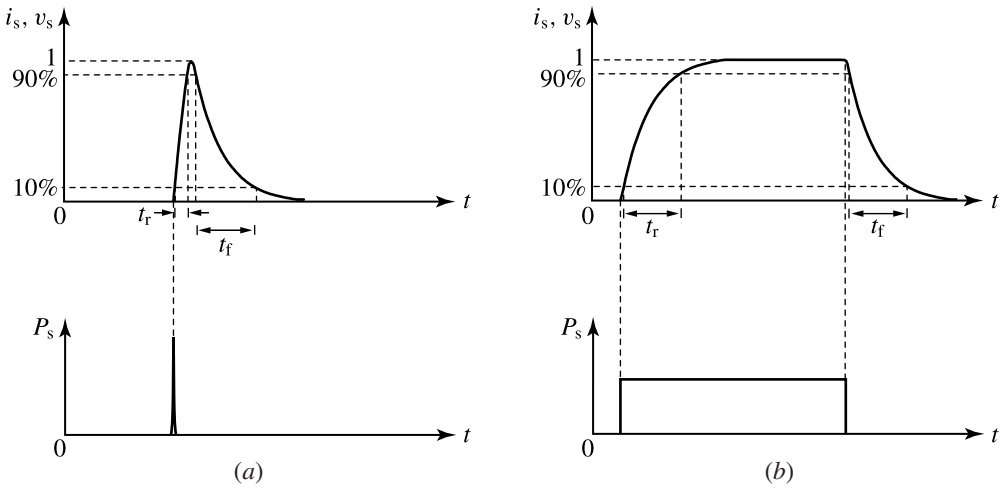


Figure 14.2 Typical responses of a photodetector to (*a*) an impulse signal and (*b*) a square-pulse signal.

fastest temporal variations in the signal or, equivalently, a frequency response that has a bandwidth covering the entire bandwidth of the signal.

In the time domain, the speed of a photodetector is characterized by the *risetime*, t_r, and the *falltime*, t_f, of its response to an impulse signal or to a square-pulse signal, as shown in Fig. 14.2. The risetime is defined as the time interval for the response to rise from 10 to 90% of its peak value, whereas the falltime is defined as the time interval for the response to decay from 90 to 10% of its peak value. Generally, the overall speed of a photodetector is determined by both its intrinsic bandwidth and its RC circuit-limited bandwidth. The risetime of the impulse response is determined by the intrinsic bandwidth of a photodetector, and that of the square-pulse response is determined by the RC circuit-limited bandwidth of the photodetector. The risetime and its corresponding bandwidth have the following relation (see Problem 13.2.3):

$$t_r = \frac{0.35}{f_{3dB}}, \tag{14.50}$$

where f_{3dB} is the 3-dB cutoff frequency defined below.

The frequency response, which is characterized by the frequency dependence of the responsivity $\mathcal{R}(f)$ at a given optical wavelength, can be obtained by simply taking the Fourier transform of the impulse response or by registering the response of the detector at one signal frequency at a time while sweeping the signal frequency. Note that $\mathcal{R}(f)$ is the current or voltage response spectrum of the detector because the responsivity of a photodetector is defined as the output current or voltage signal of the detector. The output electrical power spectrum of the detector is $\mathcal{R}^2(f)$, which defines a 3-dB *cutoff*

frequency, or 3-dB *bandwidth*, for a photodetector as

$$\mathcal{R}^2(f_{3\text{dB}}) = \frac{1}{2}\mathcal{R}^2(0). \tag{14.51}$$

Considering the rectangular time interval used to define the bandwidth B, we have the following relation between $f_{3\text{dB}}$ and B of a photodetector (see Problem 13.2.3):

$$f_{3\text{dB}} = 0.886B = \frac{0.443}{T}. \tag{14.52}$$

The 3-dB bandwidth of a photodetector is a function of the combined effect of a few different physical factors that determine the speed and the frequency response of the detector. These factors and their relative importance depend on the type of photodetector. They are discussed in later sections where the physical properties of various photodetectors are addressed.

EXAMPLE 14.7 Find the 3-dB cutoff frequency and the risetime in response to an impulse signal for the Si photodetector considered in the preceding examples.

Solution From Example 14.4, we find that $B = 100$ MHz for this detector. Therefore, its 3-dB cutoff frequency

$$f_{3\text{dB}} = 0.886B = 88.6 \text{ MHz}.$$

The risetime of its response to an impulse signal is

$$t_\text{r} = \frac{0.35}{f_{3\text{dB}}} = \frac{0.35}{88.6 \times 10^6} \text{ s} = 3.95 \text{ ns}.$$

14.3 Photoemissive detectors

Photoemissive detectors are based on the external photoelectric effect. Photoelectrons are emitted when the surface of a metal or a semiconductor, known as a *photocathode* in this situation, is illuminated with light of a sufficient photon energy. The lowest vacuum energy level, E_vac, for an electron freed from the confinement of a material is higher than the Fermi level in the material. For either a metal or a semiconductor, the energy barrier between the lowest vacuum level and the Fermi level is defined as the *work function*, $e\phi = E_\text{vac} - E_\text{F}$, of the material. For a semiconductor, the difference between the lowest vacuum level and the conduction-band edge is known as the *electron affinity*, $e\chi = E_\text{vac} - E_\text{c}$, of the semiconductor. The quantities ϕ and χ have the physical property of an electric potential measured in volts. The work function and the electron affinity of a material are normally measured in electron volts.

Photoemission from a given material occurs only when the incident photon has an energy higher than a certain *threshold photon energy*, E_{th}, corresponding to an optical wavelength shorter than a *threshold wavelength*, λ_{th}:

$$h\nu \geq E_{\text{th}}, \qquad \text{for} \qquad \lambda \leq \lambda_{\text{th}} = \frac{hc}{E_{\text{th}}} = \frac{1.2398}{E_{\text{th}}} \ \mu\text{m eV}. \tag{14.53}$$

The values of E_{th} and λ_{th} are characteristics of a given material.

1. **Metal.** In a metal, shown in Fig. 14.3(*a*), electrons occupy all of the energy levels below the Fermi level. The threshold photon energy for the emission of a photoelectron from a metal is

$$E_{\text{th}} = e\phi. \tag{14.54}$$

2. **Nondegenerate semiconductor.** In a nondegenerate semiconductor, shown in Fig. 14.3(*b*), not all energy levels below the Fermi level, but only those below the valence-band edge, are occupied by electrons because the Fermi level lies within the bandgap. The threshold photon energy for photoemission from a nondegenerate semiconductor is

$$E_{\text{th}} = e\chi + E_{\text{g}} > e\phi, \tag{14.55}$$

   if $\chi > 0$.
3. **Degenerate semiconductor.** In a degenerate semiconductor, the highest level occupied by electrons is the Fermi level. Therefore, the threshold photon energy for photoemission from a degenerate semiconductor is the work function, just like that given in (14.54) for a metal. For an n-type degenerate semiconductor, $E_{\text{th}} = e\phi < e\chi$, as shown in Fig. 14.3(*c*), because the Fermi level lies in the conduction band. For a p-type degenerate semiconductor, $E_{\text{th}} = e\phi > e\chi + E_{\text{g}}$, as shown in Fig. 14.3(*d*), because the Fermi level lies in the valence band.

The work functions of elemental metals are in the range of 2–5 eV. The lowest is that of Cs at 2.1 eV, corresponding to a threshold wavelength of 590 nm for photoemission. Elemental metals have poor quantum efficiencies. Ordinary group IV and III–V semiconductors, including Si, Ge, GaAs, and InP, have work functions typically in the range of 4–5 eV. Because of their high threshold photon energies and low quantum efficiencies, elemental metals and ordinary semiconductors are not useful for photocathodes in the visible and infrared spectral regions.

There are two groups of practical photocathodes that have both high quantum efficiencies and low threshold photon energies. One group consists of compounds of alkaline metals and cesiated silver oxides that are usually labeled with standard international designation of spectral response and window type, such as S-1 (AgOCs), S-4 (Cs_3Sb), S-10 (AgBiOCs), S-11 (Cs_3Sb), S-20 (Na_2KCsSb), and S-24 (Na_2KSb). These

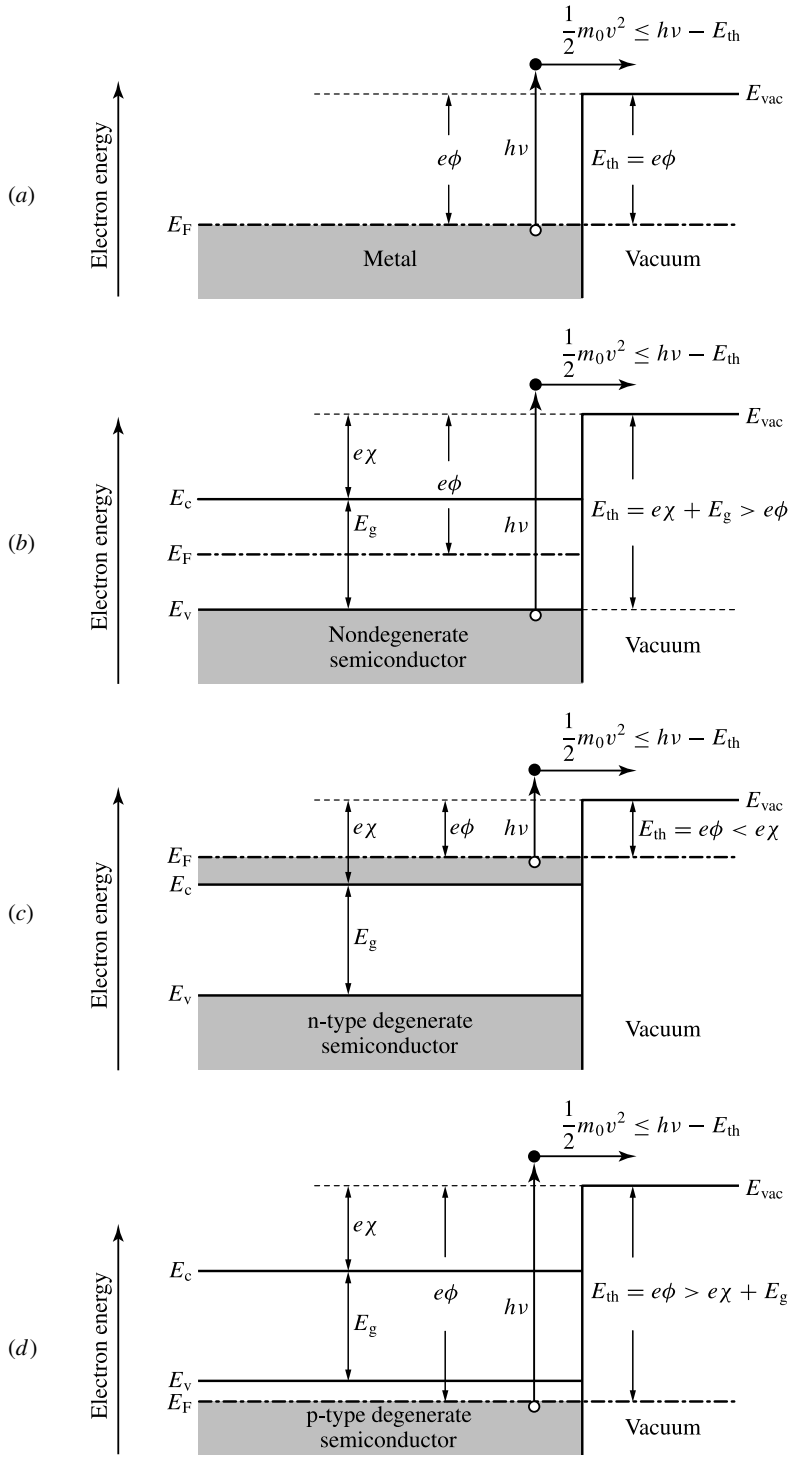


Figure 14.3 Photon energy requirement for photoemission from the surface of (*a*) a metal, (*b*) a nondegenerate semiconductor, (*c*) an n-type degenerate semiconductor, and (*d*) a p-type degenerate semiconductor.

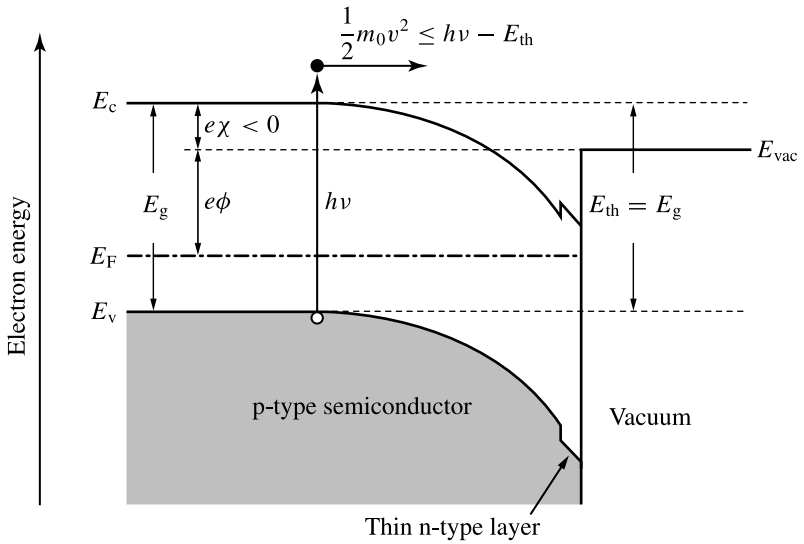


Figure 14.4 Energy levels and photoemission in an NEA photocathode.

compounds are semiconductors that have low threshold photon energies in the range of 1–2 eV because of their small bandgaps and small electron affinities. Another group consists of *negative electron affinity* (NEA) photocathodes. An NEA photocathode is made by depositing a very thin n-type layer on the surface of a p-type semiconductor to cause a large downward band bending at the surface. The photocathode has a negative effective affinity if the band bending is large enough that the conduction-band edge of the p-type semiconductor lies above the vacuum level, as shown in Fig. 14.4. Practical NEA photocathodes have been developed for a few III–V semiconductors by depositing a thin layer of Cs or Cs_2O on the surface, including GaAs : Cs_2O, InGaAs : Cs, and InAsP : Cs. As can be seen in Fig. 14.4, once an electron is excited to the conduction band of an NEA photocathode, it has sufficient energy to be emitted by tunneling through the thin surface layer. Therefore, the threshold photon energy for photoemission from an NEA photocathode is simply the bandgap of the semiconductor:

$$E_{th} = E_g. \tag{14.56}$$

Figure 14.5 shows the spectral responsivity of typical photocathodes. The spectral responsivity of a photoemissive device has a long-wavelength cutoff determined by the threshold wavelength of the photocathode material and a short-wavelength cutoff determined by the window material. The standard international designation with the letter S, such as S-1, includes both the response of the photocathode material and the transmission of the window material. Among all practical photocathodes including alkaline compounds and NEA semiconductors, S-1 has the lowest threshold energy of ~1.1 eV, corresponding to a threshold wavelength of ~1.1 μm. Currently no photocathode can respond at wavelengths longer than 1.2 μm. The spectral response characteristics of the

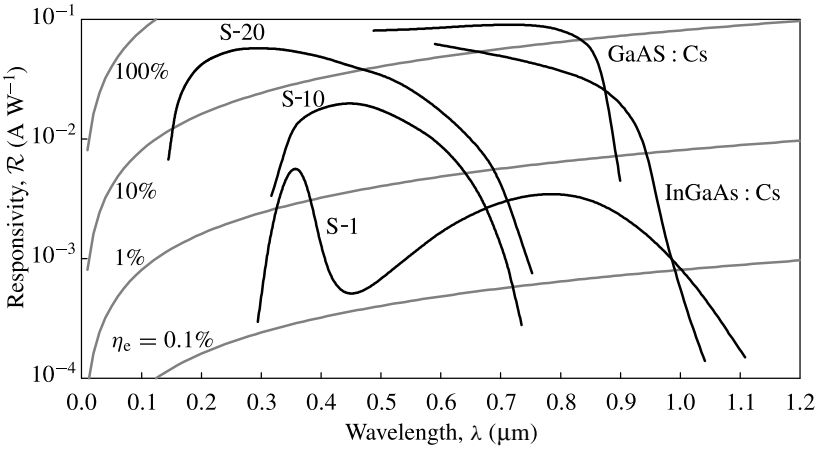


Figure 14.5 Spectral responsivity of representative photocathodes. The quantum efficiency is indicated by the gray curves. (Based on data from assorted sources.)

vacuum photodiodes and photomultipliers discussed in the following are completely determined by the spectral response of their photocathodes. Therefore, no photoemissive detectors exist for the infrared at wavelengths longer than 1.2 μm.

Vacuum photodiodes

The vacuum photodiode is a simple device that consists of a photocathode and an anode enclosed in a vacuum tube. The device can use either a *reflection-mode photocathode*, which is opaque, or a *transmission-mode photocathode*, which is semitransparent. The structure of a vacuum photodiode can have either a *side-on configuration*, with the light incident from the side of the tube, or a *head-on configuration*, with the light incident from the end of the tube. The tube can also be filled with a small amount of inert gas, such as argon, to get a small internal gain through ionization of the gas by the collision of photoelectrons. The gas-filled photodiodes are no longer competitive and therefore are not practically useful because they have a limited gain and a low speed.

Figure 14.6(*a*) shows the basic circuitry of a vacuum photodiode. A voltage, V_{ak}, typically a few hundred volts, is applied between the anode and the photocathode to collect the photoelectrons efficiently when the photocathode is irradiated with an optical signal. Such a high anode voltage is needed to eliminate the space-charge effect between the photocathode and the anode, thus improving the efficiency, and to reduce the electron transit time from the photocathode to the anode, thus increasing the response speed of the device.

The small-signal equivalent circuit, including the noise sources, of a vacuum photodiode is shown in Fig. 14.6(*b*). A photocathode generates a photocurrent in response to an optical signal. A load resistance, R_L, is required to convert the photocurrent into an output voltage signal. The capacitance C in the equivalent circuit is the total

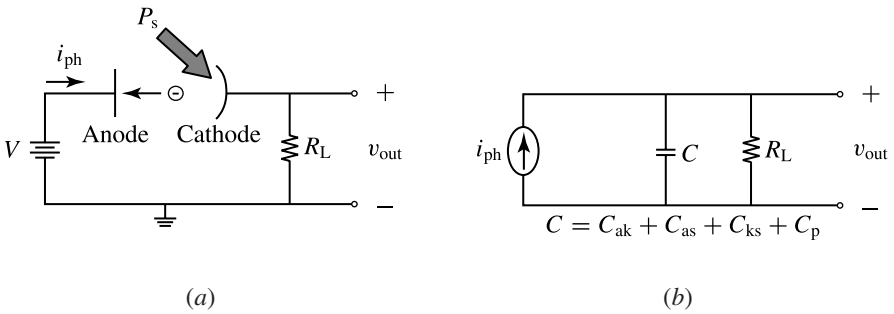


Figure 14.6 (*a*) Basic circuitry and (*b*) small-signal equivalent circuit of a vacuum photodiode.

equivalent capacitance from the anode to the ground, including the effects of the anode-to-photocathode capacitance, C_{ak}, the anode-to-shield capacitance, C_{as}, the photocathode-to-shield capacitance, C_{ks}, and stray capacitance, C_p, from the output wiring to the ground. The minimum value of C is C_{ak}, but the other capacitances add to the value of C and can even become dominant if not carefully minimized.

The dark current of a vacuum photodiode comes from thermionic emission of the photocathode. This dark current is exceedingly small, on the order of femtoamperes at room temperature; it can be ignored in the presence of other noise sources. The dominant shot-noise sources for a vacuum photodiode are the photocurrent and the background radiation current. The dominant thermal noise is that from the load resistance because the load resistance is much smaller than the internal resistance of a vacuum photodiode. Ignoring the dark current from thermionic emission, the total noise of a vacuum photodiode is

$$\overline{i_n^2} = 2eB(\overline{i_s} + \overline{i_b}) + \frac{4k_B T B}{R_L}, \tag{14.57}$$

where i_s is the photocurrent, and $i_s = i_{ph}$ because a vacuum photodiode has no gain. The NEP of a vacuum photodiode is on the order of 1 fW.

The response speed of a vacuum photodiode is determined by two factors: (1) the transit time and the transit-time spread of the photoelectrons from the photocathode to the anode and (2) the RC time constant of its equivalent circuit shown in Fig. 14.6(*b*). The transit time is the time for a photoelectron to travel from the photocathode to the anode. The transit-time spread is the spread in the transit time among different photoelectrons caused primarily by the difference in the initial kinetic energies of the photoelectrons when they are emitted from the photocathode. Both the transit time and the transit-time spread can be reduced by carefully designing the geometry of the device and then by applying a large anode-to-cathode voltage, V_{ak}. For high-speed applications, the RC time constant has to be chosen not to be the limiting factor by using a sufficiently small load resistance, which is typically 50 Ω, and by eliminating all stray capacitances. The typical speed of a fast vacuum photodiode ranges from 100 ps

to 1 ns, with a corresponding 3-dB bandwidth ranging from a few hundred megahertz to about 3 GHz.

Photomultipliers

A photomultiplier tube (PMT) is basically a vacuum photodiode with a built-in high-gain, low-noise electron multiplier. A PMT consists of four major parts: (1) a photocathode for emitting photoelectrons, (2) an electron optics consisting of focusing electrodes for accelerating and focusing the photoelectrons to the first *dynode*, (3) an electron multiplier consisting of a chain of dynodes for secondary electron emission, and, finally, (4) an anode to collect the electrons for the output signal. Depending on the structure of the electron multiplier used in a PMT, there are many different photomultiplier structures, such as the circular cage, the box and grid, the venetian blind, the linear dynode chain, and the microchannel plate, to name a few. Similarly to a vacuum photodiode, a PMT can use either a reflection-mode or a transmission-mode photocathode and can have either a side-on or a head-on configuration. Figures 14.7(*a*) and (*b*) show, as examples, the configurations and structures of a side-on reflection-mode PMT with a circular-cage structure and a head-on transmission-mode PMT with a box-and-grid structure, respectively.

The electron multiplier of a PMT consists of a series of electrodes, called dynodes, as shown in Figs. 14.7(*a*) and (*b*). The dynodes are biased at successively higher voltages through a voltage-divider circuit consisting of a series of resistors, as shown in Fig. 14.8(*a*). When a PMT is used in high-current pulse operation, capacitors are placed

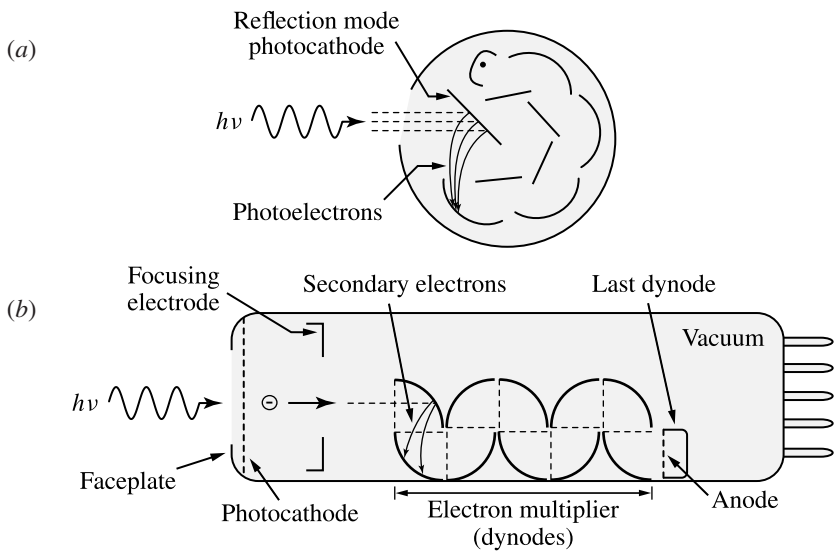


Figure 14.7 Configurations and structures of (*a*) a side-on reflection-mode PMT with a circular-cage structure and (*b*) a head-on transmission-mode PMT with a box-and-grid structure.

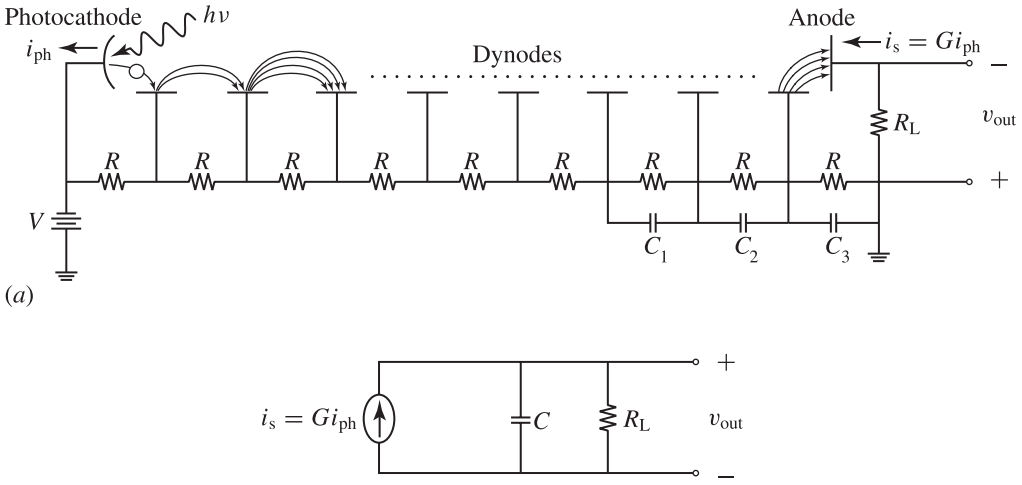


Figure 14.8 (*a*) Basic circuitry and (*b*) small-signal equivalent circuit of a photomultiplier.

in parallel to the resistors in the last two or three stages of the divider circuit. By providing a bypass when the current surges at the peak of the pulse, these capacitors help to maintain constant voltages on the last few dynodes, thus allowing the PMT to have a large linear dynamic range. Electron multiplication is accomplished by *secondary electron emission*, which is similar to photoemission except that the incident particle is an electron instead of a photon. Many photocathode materials, including the NEA semiconductors and cesiated oxides, are also used for the dynodes. A photoelectron emitted from the photocathode is accelerated by the high voltage between the photocathode and the first dynode to an energy of typically 100–200 eV. When such a high-energy electron strikes a dynode, a number of secondary electrons are emitted. This process continues through successive dynode stages.

For a PMT, the total current gain, G, as defined in (14.23) is the ratio of the output signal current at the anode to the photocurrent at the photocathode. This gain is given by the total *electron multiplication gain* through the dynode chain. If the average *electron multiplication factor* for each dynode stage is m, the total gain for a PMT with a chain of n dynodes is

$$G = \frac{i_s}{i_{ph}} = m^n. \tag{14.58}$$

The total gain can be quite significant even when the single-stage multiplication factor is modest. For example, a ten-stage dynode chain has a gain of $G \approx 10^6$ for $m = 4$ and $G \approx 10^7$ for $m = 5$. Clearly, a small variation in the multiplication factor m leads to a large change in the total gain G. Because the value of m is very sensitive to dynode voltages, both the power supply and the bias circuitry have to be kept very stable for reliable operation of a PMT. A PMT typically has 9–12 dynode stages biased

with a total voltage of 500 V to 3 kV from the anode to the photocathode, depending on the material used for the dynode and the gain desired. The typical multiplication factor ranges from $m = 3$ to 8 for a total gain ranging from $G = 10^3$ to 10^8. Because of the internal gain, the responsivity of a PMT is $\mathcal{R} = G\mathcal{R}_0$, where $\mathcal{R}_0$ is the intrinsic responsivity of the photocathode described earlier and shown in Fig. 14.5.

The small-signal equivalent circuit, including the noise sources, of a PMT is shown in Fig. 14.8(*b*). Similarly to that in the equivalent circuit of a vacuum photodiode, the capacitance C is the total equivalent capacitance from the anode to the ground, including the capacitances from the anode to all other electrodes and stray capacitances.

The major source of dark current in a PMT is the thermionic emission from the photocathode and the dynodes. Other less significant sources of dark current include leakage current, field emission, and electron emission by cosmic rays. The total amplified dark current, i_d, of a PMT is that at the anode. The anode dark current, i_da, of a PMT is contributed by the photocathode dark current, i_dk, amplified by the gain G and the dark current of all dynodes amplified by a gain less than G. If we take an effective dynode dark current, i_dd, such that the dynode contribution to the anode dark current is equivalently Gi_dd, the total dark current at the anode can be expressed as

$$i_\mathrm{d} = i_\mathrm{da} = G(i_\mathrm{dk} + i_\mathrm{dd}). \tag{14.59}$$

Like that of a vacuum photodiode, the dark current of a PMT is very small. However, the anode dark current of a PMT is detectable and thus cannot be ignored because of the high gain of a PMT. The total anode dark current i_d of a PMT is in the range between 10 pA and 10 nA, depending on the materials of the photocathode and the dynodes and on the operating temperature. Indeed, the NEP of a PMT is usually limited by the shot noise of its anode dark current. From the discussions in Section 14.1, we can then express the shot noise of a PMT as

$$\overline{i_{\mathrm{n,sh}}^2} = 2eBG^2F(\overline{i_\mathrm{ph}} + \overline{i_\mathrm{bk}} + \overline{i_\mathrm{dk}} + \overline{i_\mathrm{dd}}) = 2eBGF(\overline{i_\mathrm{s}} + \overline{i_\mathrm{b}} + \overline{i_\mathrm{d}}), \tag{14.60}$$

where $i_\mathrm{b} = Gi_\mathrm{bk}$ is the anode current due to background radiation. The excess noise factor F for an n-stage PMT is a function of the multiplication factor m. For $m > 2$,

$$F = \frac{m^{n+1} - 1}{m^n(m-1)} \approx \frac{m}{m-1}. \tag{14.61}$$

Therefore, F is on the order of unity for a PMT. Including the thermal noise, the total current noise of a PMT is

$$\overline{i_\mathrm{n}^2} = 2eBGF(\overline{i_\mathrm{s}} + \overline{i_\mathrm{b}} + \overline{i_\mathrm{d}}) + \frac{4k_\mathrm{B}TB}{R_\mathrm{L}}. \tag{14.62}$$

With this total noise, the SNR of a PMT has the form given in (14.33).

Because of its high gain and low noise, a high-gain PMT can generate an output signal with a good SNR for a single photoelectron emitted by its photocathode. Some PMTs are capable of photon counting and are among the most sensitive photodetectors available. The NEP of a PMT is typically on the order of 1 fW with a detectivity D^* on the order of 10^{16} cm $\text{Hz}^{1/2}$ W^{-1}. The NEP of a PMT operating in the photon-counting mode can be as low as 10^{-19} W. A PMT has a large linear dynamic range, typically on the order of 60–80 dB.

The response speed of a PMT is determined by the same two factors that limit the speed of a vacuum photodiode discussed earlier. Because of the dynode chain, the electron transit time from photocathode to anode in a PMT is much longer than that in a vacuum photodiode and is typically in the range of 10–100 ns. The transit-time spread, however, is much less than the transit time, typically ranging from 100 ps to about 2 ns. In terms of the impulse response, the long transit time causes a delay in the response, but the risetime of the response pulse is primarily determined by the combined effect of the transit-time spread and the RC time constant of the PMT circuit. Therefore, the risetime of a PMT is typically on the order of a few nanoseconds and can be as short as 1 or 2 ns, which is somewhat greater than the transit-time spread but is much less than the electron transit time. Consequently, a PMT is a very fast detector with a frequency bandwidth on the order of a few hundred megahertz. Combining its high speed with its large gain, a PMT is a superb photodetector that has a gain–bandwidth product unmatched by other types of photodetectors though it is not the fastest photodetector.

EXAMPLE 14.8 A PMT has a side-on configuration and nine dynode stages. Its photocathode has an effective area of 8 mm × 24 mm and an external quantum efficiency of $\eta_e = 23\%$ at $\lambda = 400$ nm. In a typical operating condition with a voltage of 1 kV applied across the anode and the photocathode, the PMT has the following operating parameters: the average electron multiplication factor per stage is $m = 6$; the anode dark current is 5 nA; the transit time is 22 ns; the transit-time spread is 1.2 ns; the impulse-response risetime is 2.2 ns; the total equivalent capacitance is $C = 6$ pF; the background radiation current is negligible. Answer the following questions for the PMT response at $\lambda = 400$ nm. (a) Find the intrinsic responsivity of the photocathode. (b) What are the gain and the responsivity of the PMT? (c) Find the NEP for the bandwidth of 1 Hz and the specific detectivity of the PMT by assuming a large load resistance. (d) What are the cutoff frequency, $f_{3\text{dB}}$, and the internal bandwidth, B, of the PMT? (e) What is the limitation on the load resistance for high-speed applications of the PMT?

Solution (a) At $\lambda = 400$ nm, the photon energy

$$h\nu = \frac{1239.8}{400} \text{ eV} = 3.1 \text{ eV}.$$

With $\eta_e = 23\%$, the intrinsic responsivity of the photocathode at $\lambda = 400$ nm is

$$\mathcal{R}_0 = \eta_e \frac{e}{h\nu} = 0.23 \times \frac{1}{3.1} \text{ A W}^{-1} = 74.2 \text{ mA W}^{-1}.$$

(b) For $n = 9$ and $m = 6$, the gain

$$G = m^n = 6^9 = 1.0 \times 10^7.$$

Therefore, the responsivity of the PMT at $\lambda = 400$ nm is

$$\mathcal{R} = G\mathcal{R}_0 = 1.0 \times 10^7 \times 74.2 \text{ mA W}^{-1} = 742 \text{ kA W}^{-1}.$$

(c) Thermal noise can be ignored for a PMT when the load resistance is sufficiently large (see Problem 14.3.4). Because the background radiation noise is also negligible, the NEP of the PMT is dark-current limited. For $m = 6$, we find from (14.61) that the excess noise factor is $F = 1.2$. Thus, with $i_d = 2$ nA, we have

$$\begin{aligned}\overline{i_n^2} &= \overline{i_{n,sh}^2} = 2eBGF\overline{i_d} \\ &= 2 \times 1.6 \times 10^{-19} \times 1.0 \times 10^7 \times 1.2 \times 2 \times 10^{-9} \times B \text{ A}^2 \text{ Hz}^{-1} \\ &= 7.68 \times 10^{-21} B \text{ A}^2 \text{ Hz}^{-1}.\end{aligned}$$

Then, the NEP for a bandwidth of 1 Hz is

$$\frac{\text{NEP}}{B^{1/2}} = \frac{\overline{i_n^2}^{1/2}}{B^{1/2}\mathcal{R}} = \frac{(7.68 \times 10^{-21})^{1/2}}{7.42 \times 10^5} \text{ W Hz}^{-1/2} = 0.118 \text{ fW Hz}^{-1/2}.$$

The active area of the photocathode is $\mathcal{A} = 8 \times 24 \text{ mm}^2 = 1.92 \times 10^{-4} \text{ m}^2$. Therefore, the specific detectivity of the PMT is

$$\begin{aligned}D^* &= \frac{(\mathcal{A}B)^{1/2}}{\text{NEP}} = \frac{(1.92 \times 10^{-4})^{1/2}}{0.118 \times 10^{-15}} \text{ m Hz}^{1/2} \text{ W}^{-1} \\ &= 1.17 \times 10^{14} \text{ m Hz}^{1/2} \text{ W}^{-1} \\ &= 1.17 \times 10^{16} \text{ cm Hz}^{1/2} \text{ W}^{-1}.\end{aligned}$$

(d) The cutoff frequency is determined by the risetime, which is $t_r = 2.2$ ns. Thus,

$$f_{3dB} = \frac{0.35}{2.2 \times 10^{-9}} \text{ Hz} = 159 \text{ MHz}.$$

We then find that

$$B = \frac{f_{3dB}}{0.886} = \frac{159}{0.886} \text{ MHz} = 180 \text{ MHz}.$$

Note that the transit-time spread of 1.2 ns contributes to a large part of the risetime of 2.2 ns. The transit time of 22 ns is 10 times the risetime, but it has no consequence on either the risetime or the bandwidth of the PMT.

(e) For high-speed applications of the PMT, it is required that the circuit RC time constant of the PMT be much smaller than the intrinsic response time of the PMT. More

precisely, the RC-time-limited 3-dB cutoff frequency of the circuit response is required to be much higher than the intrinsic cutoff frequency of the PMT. As we shall see later, the RC-time-limited 3-dB cutoff frequency

$$f_{3\text{dB}}^{\text{ckt}} = \frac{1}{2\pi R_{\text{L}} C}.$$

To make sure this PMT is not limited by the RC time, we need $f_{3\text{dB}}^{\text{ckt}} \gg 159$ MHz. With $C = 6$ pF, we find the following limitation for the load resistance:

$$R_{\text{L}} \ll \frac{1}{2\pi \times 6 \times 10^{-12} \times 159 \times 10^{6}}\ \Omega = 167\ \Omega.$$

Thus a typical choice of $R_{\text{L}} = 50\ \Omega$ will satisfy this requirement for high-speed applications of the PMT. Even when the load resistance is chosen to be so low, the PMT still operates in the quantum regime with a D^* limited by the shot noise from its dark current (see Problem 14.3.4). If speed is not a concern, a large R_{L} is usually chosen for the PMT to have a large dynamic range.

14.4 Photoconductive detectors

Photoconductive detectors are based on the phenomenon of *photoconductivity*. The conductivity of a photoconductor, which can be an insulator but is usually a semiconductor, increases with optical illumination due to photogeneration of free carriers. The conductivity of a semiconductor that has electron and hole concentrations of n and p, respectively, is

$$\sigma = e(\mu_{\text{e}} n + \mu_{\text{h}} p), \tag{14.63}$$

where e is the electronic charge and μ_{e} and μ_{h} are the electron and hole mobilities, respectively. In the absence of optical illumination, the conductivity, known as the *dark conductivity*, $\sigma_0 = e(\mu_{\text{e}} n_0 + \mu_{\text{h}} p_0)$ because the electron and hole concentrations in this situation are the equilibrium concentrations, n_0 and p_0, respectively. When a semiconductor is illuminated with light of a sufficient photon energy, carriers in excess of the equilibrium concentrations are generated. The photoconductivity is the additional conductivity contributed by these photogenerated excess carriers:

$$\Delta\sigma = \sigma - \sigma_0 = e(\mu_{\text{e}} \Delta n + \mu_{\text{h}} \Delta p), \tag{14.64}$$

where $\Delta n = n - n_0$ and $\Delta p = p - p_0$ are the photogenerated excess electron and hole concentrations, respectively.

Similarly to photoemission, photoconductivity also has a threshold photon energy, E_{th}, and a corresponding threshold wavelength, λ_{th}, that are characteristic of a given photoconductor. Together with the spectral dependence of the absorption coefficient, they determine the spectral response of a photoconductor. Depending on the processes

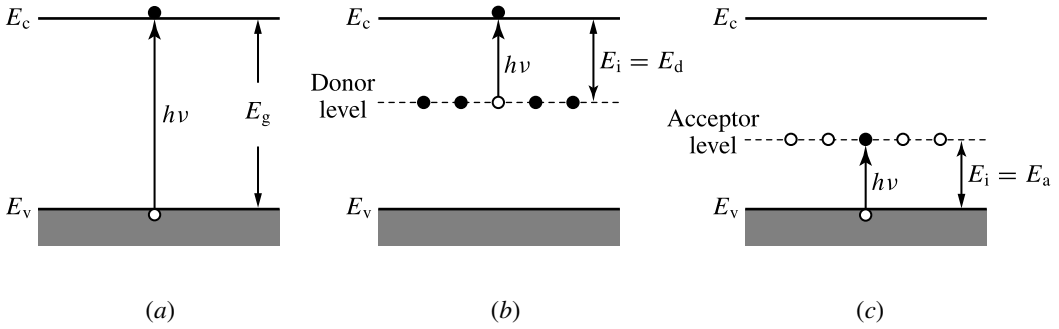


Figure 14.9 Optical transitions for (*a*) intrinsic photoconductivity, (*b*) n-type extrinsic photoconductivity, and (*c*) p-type extrinsic photoconductivity.

involved in the photogeneration of free carriers, there are two principal types of photoconductivity. The *intrinsic photoconductivity* is contributed by the excess electrons and holes that are generated by band-to-band absorption of incident photons, as shown in Fig. 14.9(*a*). The threshold photon energy of intrinsic photoconductivity is clearly the bandgap energy of the photoconductor:

$$E_{th} = E_g. \tag{14.65}$$

The *extrinsic photoconductivity* is contributed by carriers that are generated by optical transitions associated with impurity levels within the bandgap of an extrinsic semiconductor. In an n-type extrinsic photoconductor, the impurity levels have an energy $E_i = E_d$ below the conduction-band edge; electrons are excited from these donor levels to the conduction band, as shown in Fig. 14.9(*b*). In a p-type extrinsic photoconductor, the impurity levels have an energy $E_i = E_a$ above the valence-band edge; electrons are excited from the valence band to these acceptor levels, as shown in Fig. 14.9(*c*). Thus, the threshold photon energy of extrinsic photoconductivity for either n-type or p-type photoconductors is

$$E_{th} = E_i. \tag{14.66}$$

Photoconductors cover a broad spectral range from the ultraviolet to the far infrared. In particular, there are many sensitive photoconductors in the infrared region beyond 1.2 μm wavelength where no photoemissive detectors exist. Both direct-gap and indirect-gap semiconductors can be used for photoconductors. All of the semiconductors discussed in Section 12.1, including the group IV semiconductors, the III–V and II–VI compounds, and the IV–VI compounds, can be used for intrinsic photoconductors. Among them, intrinsic silicon photoconductors are the most important photoconductive detectors in the visible and near infrared spectral regions at wavelengths shorter than 1.1 μm, while intrinsic germanium photoconductors are the most important photoconductive detectors in the near infrared region at wavelengths up to 1.8 μm. In the mid infrared region between 2 and 7 μm wavelengths, one finds intrinsic

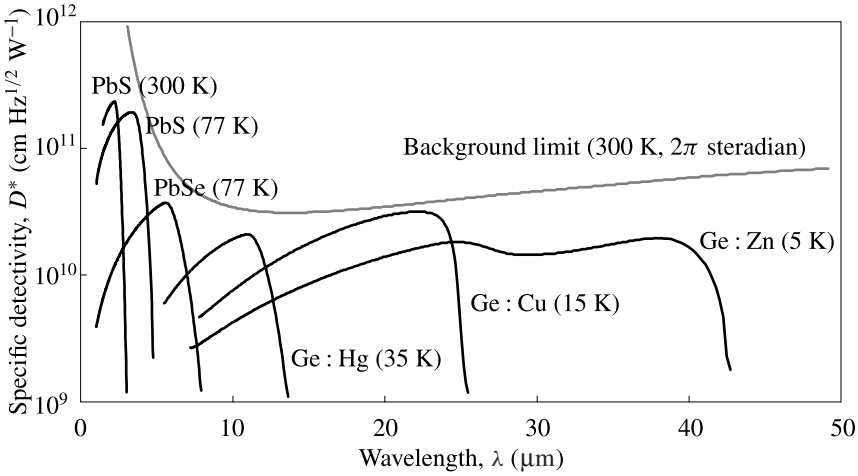


Figure 14.10 Specific detectivity, D^*, of representative photoconductive detectors as a function of optical wavelength. The gray curve shows for comparison the ideal D^* for a background-limited photoconductor of unity quantum efficiency. (Based on data from assorted sources.)

photoconductors based on InAs, InSb, PbS, and PbSe. Extrinsic photoconductors are available for these mid infrared wavelengths as well as for longer wavelengths well into the far infrared region. The most important extrinsic photoconductive detectors are p-type germanium photoconductors such as Ge : Au, Ge : Hg, Ge : Cd, Ge : Cu, and Ge : Zn. Figure 14.10 shows the specific detectivity of representative photoconductive detectors as a function of optical wavelength.

A sensitive photoconductor must have a low dark conductivity so that the photoconductivity caused by optical illumination amounts to a significant change in its total conductivity. For this reason, it is necessary to minimize the thermal equilibrium concentrations, n_0 and p_0, of free electrons and free holes in a photoconductor. According to the law of mass action given in (12.31), $n_0 p_0 = n_i^2(T)$.

In an intrinsic semiconductor, both electron and hole concentrations in the dark can be reduced by lowering the temperature because $n_0 = p_0 = n_i(T)$. Because n_i depends exponentially on $-E_g/2k_BT$, as seen in (12.29), the dark electron and hole concentrations can be significant for a semiconductor that has a small bandgap energy. Reduction of the dark free carrier concentrations by lowering temperature is particularly important for intrinsic photoconductors of small bandgap energies, such as InSb and HgCdTe. For this reason, such small-bandgap photoconductors are normally operated at the liquid nitrogen temperature of 77 K or lower.

In an extrinsic semiconductor, conductivity is predominantly contributed by the majority carriers because the majority carrier concentration is much higher than both n_i and the minority carrier concentration. It is therefore important for the functioning of an extrinsic photoconductor that most of the free majority carriers be photogenerated rather than thermally generated. This condition requires that the donors in an n-type

photoconductor and the acceptors in a p-type photoconductor not be ionized when an extrinsic photoconductor is not optically illuminated. Ideally, they should be ionized only optically when the photoconductor is illuminated. Because the value of E_i for an extrinsic photoconductor is small, it is normally necessary to operate an extrinsic photoconductor at a low temperature to reduce the dark concentration of the majority free carriers. Some extrinsic photoconductors, such as Ge : Au, are operated below 77 K. Some, such as Ge : Cu and Ge : Zn, are often operated at the liquid helium temperature of 4 K.

From these discussions, it is clear that a photoconductor of a small threshold photon energy, thus a long threshold wavelength, is required to operate at a low temperature irrespective of whether it is an intrinsic or an extrinsic type. As a rule, the operating temperature for a detector with a threshold energy E_{th} has to be $T < E_{th}/25k_B \approx 460E_{th}$ (eV). A photoconductor for the mid infrared normally requires an operating temperature of 77 K, and one for the far infrared requires an even lower operating temperature often down to 4 K.

The operation of a photoconductor requires that a voltage be applied to the device. A photoconductor has a photoconductive gain that depends on many parameters of the photoconductor and on the properties of the electrical contacts. To facilitate quantitative discussions, we consider a simple photoconductor of a length l between its electrodes, a width w, and a thickness d, as shown in Fig. 14.11. Thus, the optically illuminated area is $\mathcal{A} = lw$, but the cross-sectional area between the electrodes is wd. A voltage V is applied across the length l while the photoconductor is uniformally illuminated with an optical beam of a power P_s. The external quantum efficiency of the photoconductor, which is illuminated on surface $\mathcal{A}$, can be expressed as

$$\eta_e = \eta_{coll}\eta_t\eta_i = \eta_{coll}(1 - R)(1 - e^{-\alpha d}), \tag{14.67}$$

where η_{coll} is the collection efficiency of the photogenerated carriers, $\eta_t = 1 - R$ with R being the reflectivity of the incident surface, and $\eta_i = 1 - e^{-\alpha d}$ with α being the absorption coefficient of the photoconductor. To improve the external quantum efficiency, the electrodes have to be carefully designed to maximize the collection efficiency, and the surface reflectivity can be reduced by antireflection coating. In practice, both η_{coll} and η_t can be made close to unity. Then η_e can be made very close to 100% by increasing

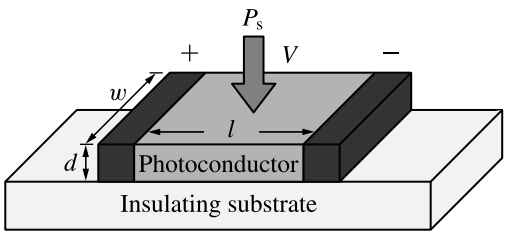


Figure 14.11 Simple geometry of a photoconductive detector.

the thickness of the photoconductor for $\eta_i = 1$ with $d \gg \alpha^{-1}$. However, the practically important performance indicator of a photodetector is not its external quantum efficiency but rather its detectivity, or its specific detectivity. The optimum thickness of a photoconductor for a maximum value of D^* is $d = 1.256\alpha^{-1}$ for $\eta_i = 71.5\%$ (see Problem 14.4.6). Because the value of D^* varies very slowly with d around this optimum thickness, there is much freedom to choose the thickness of a photoconductor in the range of $\alpha^{-1} \leq d \leq 1.5\alpha^{-1}$ for D^* to have a value that is more than 99% of its maximum value (see Problem 14.4.6).

We first consider intrinsic photoconductivity under the conditions leading to the relations in (12.55) and (12.56) so that the photogenerated electrons and holes have the same concentration and the same lifetime: $\Delta n = \Delta p = N$ and $\tau_e = \tau_h = \tau_s$. We also assume that the contacts of the electrodes are ohmic contacts that allow electrons and holes to be freely removed from or injected into the semiconductor. With a quantum efficiency of η_e, the photogeneration rate of free carriers in the semiconductor is $\eta_e P_s / h\nu$, which is equal to $lwd \cdot N/\tau_s$ because the generation rate equals the recombination rate in the steady state. Therefore, the total number of photogenerated free carriers in the photoconductor is

$$\mathcal{N} = lwd \cdot N = \eta_e \frac{P_s}{h\nu} \tau_s. \tag{14.68}$$

Because the carriers have a lifetime of τ_s, the photocurrent resulting from the photogeneration of these carriers is

$$i_{ph} = \frac{e\mathcal{N}}{\tau_s} = \eta_e \frac{eP_s}{h\nu}, \tag{14.69}$$

which is exactly the relation given in (14.14), as expected. This is not the external signal current of the photoconductor, however. The external signal current, i_s, is generated by the applied voltage V on the photoconductance, which is $\Delta\sigma wd/l$ for the photoconductor of the geometry shown in Fig. 14.11. Therefore, we find that (see Problem 14.4.3)

$$i_s = V \frac{\Delta\sigma wd}{l} = \frac{e\mathcal{N}}{\tau_{tr}^e} + \frac{e\mathcal{N}}{\tau_{tr}^h}, \tag{14.70}$$

where

$$\tau_{tr}^e = \frac{l}{\mu_e E} = \frac{l^2}{\mu_e V} \quad \text{and} \quad \tau_{tr}^h = \frac{l}{\mu_h E} = \frac{l^2}{\mu_h V} \tag{14.71}$$

are the *transit times* of electrons and holes in the photoconductor, respectively. The transit time of an electron or hole is the time it takes for the charge carrier to cross the length l of the semiconductor at its drift velocity of $v = \mu E$ under an applied field of $E = V/l$. From these results, we find the following photoconductive gain

(see Problem 14.4.3):

$$G = \frac{i_s}{i_{ph}} = \frac{\tau_s}{\tau_{tr}^e} + \frac{\tau_s}{\tau_{tr}^h} = \frac{\tau_s}{\tau_{tr}^e}\left(1 + \frac{\mu_h}{\mu_e}\right). \tag{14.72}$$

The photoconductive gain in (14.72) is obtained for an intrinsic photoconductor with ohmic contacts on both electrodes. It is not valid if the photogenerated electrons and holes do not have the same concentration, as is the case in an extrinsic photoconductor, or if one or both contacts are not ohmic. It is also clearly not valid for all values of the applied voltage because (14.71) leads to an unphysical conclusion that the gain can be made arbitrarily large simply by increasing the voltage V to reduce the transit times. Nevertheless, the photoconductive gain can be generally expressed as

$$G = \frac{\tau}{\tau_r}, \tag{14.73}$$

where τ is a *carrier lifetime* that can take different forms in different situations and τ_r is a *relaxation time constant* that depends on the properties of the photoconductor, the contacts, and the applied voltage. The values of τ and τ_r in this relation depend on the properties and the operating condition of a photoconductor, as discussed below. In particular, when the applied voltage is large, τ_r is not simply determined by the carrier transit times given in (14.71). There is a capacitance, $C = \epsilon wd/l$, between the anode and the cathode of the photoconductor. A space-charge effect in the photoconductor appears when the number of charges, $Q = CV$, supplied by the applied voltage V on this capacitance is equal to or larger than the number of the carriers in the photoconductor. This situation takes place under the following condition (see Problem 14.4.4):

$$V \geq V_{SC} = \frac{\sigma l^2}{\mu\epsilon}, \tag{14.74}$$

where μ is a mobility that can take the form of $\mu = \mu_e + \mu_h$, $\mu = \mu_e$, or $\mu = \mu_h$ depending on the properties of the photoconductor and its electrode contacts, as discussed below. In the presence of this space-charge effect, $\tau_r = \tau_d$, where

$$\tau_d = \frac{\epsilon}{\sigma} = \frac{\epsilon}{e(\mu_e n + \mu_h p)} \tag{14.75}$$

is the *dielectric relaxation time* of the semiconductor. Clearly, the gain does not continue to increase with increasing voltage when the space-charge effect appears.

The following cases are of interest.

1. An intrinsic photoconductor in which both electrons and holes can freely move, and both the anode and the cathode have nonblocking ohmic contacts. In this case, $\tau = \tau_s$, which is the spontaneous carrier recombination lifetime defined in (12.56), and $V_{SC} = \sigma l^2/(\mu_e + \mu_h)\epsilon$. For $V < V_{SC}$, $\tau_r = \tau_{tr}^e(1 + \mu_h/\mu_e)^{-1}$. When the applied

voltage is large enough that $V > V_{SC}$, $\tau_r = \tau_d$, and the gain saturates at $G = \tau_s/\tau_d$ (see Problem 14.4.4).

2. An intrinsic or extrinsic photoconductor in which only one type of carrier can freely move, and the electrodes are nonblocking ohmic contacts for such carriers. In this case, only the free-moving carriers contribute to the photocurrent. Then τ and τ_r are, respectively, the lifetime and transit time of such carriers. The space-charge effect is determined by $V_{SC} = \sigma l^2/\mu_e\epsilon$ if only electrons can freely move but by $V_{SC} = \sigma l^2/\mu_h\epsilon$ if only holes can freely move. When the space-charge effect occurs at $V > V_{SC}$, $\tau_r = \tau_d$ also in this case.
3. An intrinsic photoconductor in which both electrons and holes can freely move and the cathode is ohmic but the anode is blocking holes. Then, $\tau = \tau_e = \tau_h = \tau_{tr}^h$ and $\tau_r = \tau_{tr}^e$. In this case, $G = 1 + \mu_e/\mu_h$.
4. An intrinsic or extrinsic photoconductor in which both electrons and holes can freely move but both the cathode and the anode have blocking nonohmic contacts. In this case, $\tau_r = \tau = \tau_{tr}^e$ and $G = 1$. With blocking contacts on both sides, the gain is unity. This is the case of the junction photodiodes discussed in the following section.

When the external quantum efficiency and the gain of a photoconductor are determined, its responsivity can be easily calculated as

$$\mathcal{R} = G\eta_e \frac{e}{h\nu}. \tag{14.76}$$

Because the gain G varies with the applied voltage V, the responsivity $\mathcal{R}$ of a photoconductor is also a function of V in addition to being a function of the optical wavelength and the device parameters.

EXAMPLE 14.9 An n-type GaAs intrinsic photoconductive detector for $\lambda = 850$ nm has the following parameters: $l = w = 100$ μm, $d = 1$ μm, $\alpha = 1 \times 10^4$ cm^{-1} $= 1 \times 10^6$ m^{-1} at 850 nm, $\eta_{coll} = 1$, and $\eta_t = 1$ for $R = 0$ with antireflection coating on the incident surface. It is lightly doped with $n_0 = 1 \times 10^{12}$ cm^{-3} $= 1 \times 10^{18}$ m^{-3} and has a lifetime of $\tau_s = 100$ μs for photogenerated carriers. Both electrons and holes can freely move in the photoconductor, and both electrodes have ohmic contacts. The device is biased at $V = 2$ V across its electrodes. GaAs has the following characteristic parameters: $\epsilon = 13.18\epsilon_0$ at DC or low frequencies, $\mu_e = 8500$ cm^2 V^{-1} s^{-1} $= 0.85$ m^2 V^{-1} s^{-1}, $\mu_h = 400$ cm^2 V^{-1} s^{-1} $= 0.04$ m^2 V^{-1} s^{-1}, and $n_i = 2.33 \times 10^{12}$ m^{-3} at 300 K (see Example 12.2). (a) Find the dark conductivity. With the given bias voltage, is the device limited by a space-charge effect at any level of input optical signal? (b) Find the external quantum efficiency. What are the gain and the responsivity of this device?

Solution (a) We have $n_0 = 1 \times 10^{18}$ m^{-3} and $p_0 = n_i^2/n_0 = 5.43 \times 10^6$ m^{-3}. We then find the following dark conductivity for the device at 300 K:

$$\sigma_0 = e(\mu_e n_0 + \mu_h p_0) = 1.6 \times 10^{-19} \times (0.85 \times 1 \times 10^{18} + 0.04 \times 5.43 \times 10^6)\ \Omega^{-1}\ \text{m}^{-1}$$
$$= 0.136\ \Omega^{-1}\ \text{m}^{-1}.$$

Because $\sigma > \sigma_0$ at any level of input optical signal, we have

$$V_{SC} > \frac{\sigma_0 l^2}{(\mu_e + \mu_h)\epsilon} = \frac{0.136 \times (100 \times 10^{-6})^2}{(0.85 + 0.04) \times 13.18 \times 8.85 \times 10^{-12}}\ \text{V} = 13.1\ \text{V}.$$

Because $V < V_{SC}$ for $V = 2$ V, the device is not limited by a space-charge effect at any level of input optical signal.

(b) We find that $\alpha d = 1$ for this device. With $\eta_{coll} = \eta_t = 1$, the external quantum efficiency

$$\eta_e = \eta_i = 1 - e^{-\alpha d} = 1 - e^{-1} = 63.2\%.$$

The electron transit time at the bias voltage of $V = 2$ V is

$$\tau_{tr}^e = \frac{l^2}{\mu_e V} = \frac{(100 \times 10^{-6})^2}{0.85 \times 2}\ \text{s} = 5.88\ \text{ns}.$$

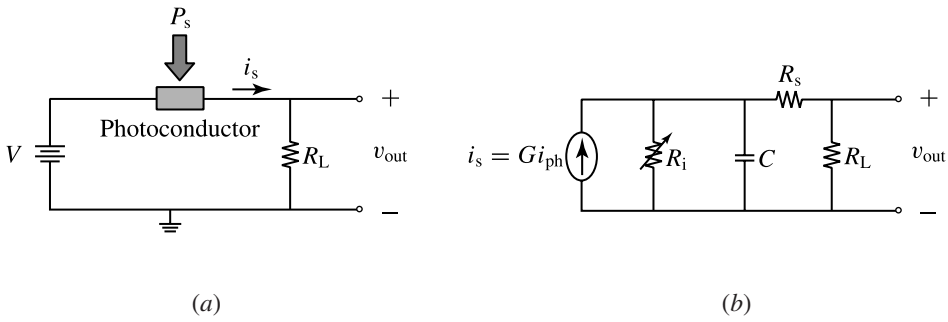
Because both electrons and holes can freely move, the gain

$$G = \frac{\tau_s}{\tau_{tr}^e}\left(1 + \frac{\mu_h}{\mu_e}\right) = \frac{100 \times 10^{-6}}{5.88 \times 10^{-9}}\left(1 + \frac{0.04}{0.85}\right) = 1.78 \times 10^4.$$

At $\lambda = 850$ nm, the responsivity

$$\mathcal{R} = G\eta_e \frac{e}{h\nu} = 1.78 \times 10^4 \times 0.632 \times \frac{850}{1239.8}\ \text{A W}^{-1} = 7.71\ \text{kA W}^{-1}.$$

It is necessary to apply a voltage or a current to a photoconductor for measuring its photoconductivity in terms of an electrical signal. Normally a bias voltage is applied and a load resistance is used to convert the signal current into an output voltage. Figure 14.12(a) shows the basic circuitry of a photoconductive detector that is biased



(a) (b)

Figure 14.12 (a) Basic circuitry and (b) small-signal equivalent circuit of a photoconductive detector.

with a voltage. Figure 14.12(*b*) shows the small-signal equivalent circuit, including the noise sources, of a photoconductive detector. The intrinsic speed of the device is not limited by its capacitance, but the speed of the output signal at the load resistance is still influenced by the device capacitance.

The shot noise in a photoconductor is associated with the statistical nature of the generation and recombination of carriers, which leads to random fluctuations in the carrier number. This shot noise is known as the *generation–recombination noise*. It has contributions from the optical signal, the background radiation, and the dark current. The dark current in this case comes from the dark conductivity of the device due to thermal excitation of free carriers. For long-wavelength infrared detectors, this dark current can be a major source of noise because of the small excitation energy. For this reason, such detectors have to be operated at low temperatures in order to minimize this noise. Because of the gain in a photoconductor, the generation–recombination noise has the form of the amplified shot noise given in (14.25):

$$\overline{i_{\text{n,GR}}^2} = 2eBGF(\overline{i_{\text{s}}} + \overline{i_{\text{b}}} + \overline{i_{\text{d}}}). \tag{14.77}$$

Because the gain G of a photoconductor is a function of the carrier lifetime τ in the form of (14.73), the excess noise factor F is determined by the statistics of the carrier lifetime and the signal frequency as

$$F = \frac{\overline{G^2}}{\overline{G}^2} = \frac{\overline{\tau^2}}{\overline{\tau}^2}. \tag{14.78}$$

For a photoconductor, in which the carrier lifetime is primarily determined by the carrier recombination process, the probability distribution of τ is characterized by the Poisson process of a continuous random variable with the consequence that $F = 2$ (see Problem 14.4.5). In addition to this shot noise, there is also thermal noise from the photoconductor resistance and the load resistance. Therefore, the total noise of a photoconductor is

$$\overline{i_{\text{n}}^2} = 4eBG(\overline{i_{\text{s}}} + \overline{i_{\text{b}}} + \overline{i_{\text{d}}}) + \frac{4k_{\text{B}}TB}{R_{\text{eq}}}, \tag{14.79}$$

where R_{eq} is an equivalent resistance seen at the output of the device. Most photoconductors are shot-noise limited by their dark current. The SNR of a photoconductor has the form given in (14.33) for a photodetector with an internal gain of G.

EXAMPLE 14.10 The photoconductive detector considered in Example 14.9 is loaded with a sufficiently large resistance such that the resistive thermal noise is negligible compared to the shot noise from its dark current at the operating temperature of 300 K. The background radiation noise is also negligible. (a) Find the dark resistance of the device. Then, find its dark current at a bias voltage of $V = 2$ V. (b) Find the NEP of the

device for a bandwidth of 1 Hz at $\lambda = 850$ nm. (c) Find the value of D^* for the device at $\lambda = 850$ nm.

Solution (a) From Example 14.9, we have $\sigma_0 = 0.136\ \Omega^{-1}\,\mathrm{m}^{-1}$. Thus, the dark resistance of the device is

$$R_0 = \frac{l}{\sigma_0 w d} = \frac{100 \times 10^{-6}}{0.136 \times 100 \times 10^{-6} \times 1 \times 10^{-6}}\ \Omega = 7.35\ \mathrm{M}\Omega.$$

The dark current at a bias voltage of $V = 2$ V is

$$i_\mathrm{d} = \frac{V}{R_0} = \frac{2}{7.35 \times 10^6}\ \mathrm{A} = 272\ \mathrm{nA}.$$

(b) With $G = 1.78 \times 10^4$ found in Example 14.9, the noise of this photoconductor is

$$\begin{aligned}\overline{i_\mathrm{n}^2} &= \overline{i_{\mathrm{n,sh}}^2} = 4eBG\overline{i_\mathrm{d}} \\ &= 4 \times 1.6 \times 10^{-19} \times 1.78 \times 10^4 \times 272 \times 10^{-9} \times B\ \mathrm{A}^2\,\mathrm{Hz}^{-1} \\ &= 3.1 \times 10^{-21} B\ \mathrm{A}^2\,\mathrm{Hz}^{-1}.\end{aligned}$$

With $\mathcal{R} = 7.71\ \mathrm{kA\ W}^{-1}$ from Example 14.9, we find the following NEP for a bandwidth of 1 Hz:

$$\frac{\mathrm{NEP}}{B^{1/2}} = \frac{(3.1 \times 10^{-21})^{1/2}}{7.71 \times 10^3}\ \mathrm{W\,Hz}^{-1/2} = 7.22\ \mathrm{fW\,Hz}^{-1/2}.$$

(c) This device has an illumination area of $\mathcal{A} = lw = (100 \times 10^{-6})^2\ \mathrm{m}^2 = 1 \times 10^{-8}\ \mathrm{m}^2$. Thus, its specific detectivity at $\lambda = 850$ nm is found to be

$$\begin{aligned}D^* &= \frac{(1 \times 10^{-8})^{1/2}}{7.22 \times 10^{-15}}\ \mathrm{m\,Hz}^{1/2}\,\mathrm{W}^{-1} = 1.39 \times 10^{10}\ \mathrm{m\,Hz}^{1/2}\,\mathrm{W}^{-1} \\ &= 1.39 \times 10^{12}\ \mathrm{cm\,Hz}^{1/2}\,\mathrm{W}^{-1}.\end{aligned}$$

The frequency response of a photoconductor that has a gain generally described by (14.73) is characterized by the following electrical power spectrum:

$$\mathcal{R}^2(f) = \frac{\mathcal{R}^2(0)}{1 + 4\pi^2 f^2 \tau^2}, \tag{14.80}$$

which has a 3-dB cutoff frequency given by

$$f_\mathrm{3dB} = \frac{1}{2\pi\tau}. \tag{14.81}$$

Therefore, a photoconductor has the following gain–bandwidth product:

$$Gf_\mathrm{3dB} = \frac{\tau}{\tau_\mathrm{r}} \cdot \frac{1}{2\pi\tau} = \frac{1}{2\pi\tau_\mathrm{r}}. \tag{14.82}$$

Figure 14.13 shows the frequency response of a typical photoconductive detector.

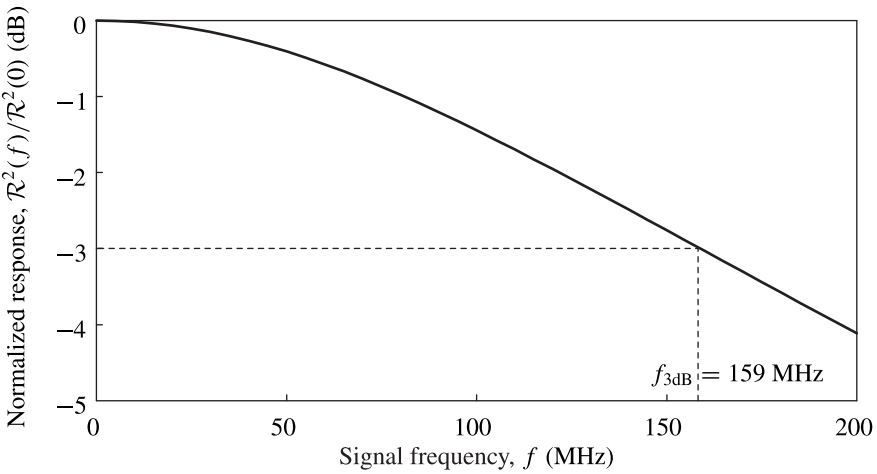


Figure 14.13 Typical frequency response, normalized to the zero-frequency response, of a photoconductive detector characterized by the electrical power spectrum as a function of signal frequency. This plot is generated with $\tau = 1$ ns.

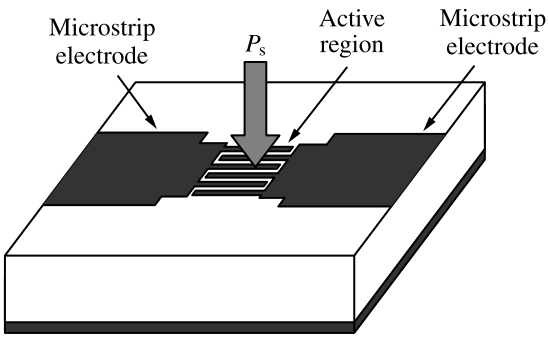


Figure 14.14 Structure of a high-speed MSM photoconductor with interdigitated electrodes.

Although the response speed of a photoconductor can be increased by reducing the carrier lifetime τ, the gain–bandwidth product is solely determined by the time constant τ_r. As discussed above, the value of τ_r can be reduced by increasing the applied voltage until $V = e\mathcal{N}/C$, when $\tau_r = \tau_d$. At this point and beyond, the gain–bandwidth product saturates at $Gf_{3dB} = 1/2\pi\tau_d$; thus, increasing the speed will reduce the gain, and vice versa. We can therefore expect a high-gain photoconductor to be very slow and a high-speed photoconductor to be insensitive. Both the gain and the speed of practical photoconductors cover a wide range, from less than 1 to over 10^5 for the gain and from less than 1 ps to over 1 ms for the response speed. Figure 14.14 shows the structure of a high-speed photoconductor. It has the *metal–semiconductor–metal* (MSM) structure with interdigitated electrodes of submicrometer features to minimize the electron transit time so that a higher gain can be obtained for a given speed. Because its speed is limited by the carrier lifetime, the device is fabricated on a high-resistivity

semi-insulating semiconductor that has a very short carrier lifetime. For the transmission of high-frequency electrical signals, the electrodes are connected to microstrip lines. A high-speed photoconductor is often used as a high-speed optoelectronic switch for switching an electronic circuit with an ultrashort optical pulse.

EXAMPLE 14.11 Find the 3-dB cutoff frequency, the gain–bandwidth product, and the NEP over the entire bandwidth for the photoconductor considered in the preceding two examples.

Solution Because $\tau = \tau_s = 100$ μs, the 3-dB cutoff frequency

$$f_{3dB} = \frac{1}{2\pi \times 100 \times 10^{-6}} \text{ Hz} = 1.59 \text{ kHz},$$

In the operating condition under consideration in the preceding examples, we have $G = 1.78 \times 10^4$. Thus, the gain–bandwidth product

$$Gf_{3dB} = 1.78 \times 10^4 \times 1.59 \text{ kHz} = 28.3 \text{ MHz}.$$

It is easily verified that $Gf_{3dB} = 1/2\pi\tau_r$ as expressed in (14.82) for $\tau_r = \tau_{tr}^e(1 + \mu_h/\mu_e)^{-1} = 5.63$ ns. With $f_{3dB} = 1.59$ kHz, we have $B = f_{3dB}/0.886 = 1.79$ kHz. Therefore, the NEP over the entire bandwidth is

$$\text{NEP} = 7.22 \times (1.79 \times 10^3)^{1/2} \text{ fW} = 305 \text{ fW}.$$

We find that this detector has a low cutoff frequency at the kilohertz level because of its large carrier lifetime of 100 μs. Because of its small bandwidth, it also has a low total NEP over its entire bandwidth. The speed of the device can be increased by reducing its carrier lifetime, but both the gain and the total NEP will suffer if other parameters of the device remain unchanged.

14.5 Junction photodiodes

Every junction diode has a photoresponse that can be utilized for optical detection. Junction photodiodes are the most commonly used photodetectors in the photonics industry. They can take many different forms, including semiconductor homojunctions, semiconductor heterojunctions, and metal–semiconductor junctions. Similarly to that of a photoconductor, the photoresponse of a photodiode results from the photogeneration of electron–hole pairs. In contrast to photoconductors, which can be of either intrinsic or extrinsic type, a photodiode is normally of intrinsic type, in which electron–hole pairs are generated through band-to-band optical absorption. Therefore, the threshold photon energy of a semiconductor photodiode is the bandgap energy of its active region:

$$E_{th} = E_g. \tag{14.83}$$

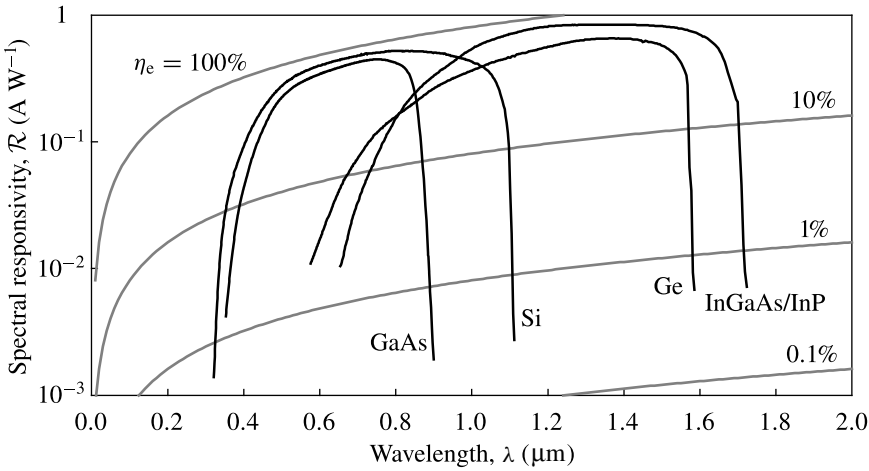


Figure 14.15 Spectral responsivity of representative photodiodes as a function of optical wavelength at 300 K. The quantum efficiency is indicated by the gray curves. (Based on data from assorted sources.)

Junction photodiodes cover a wide spectral range from ultraviolet to infrared. All of the semiconductor materials used for intrinsic photoconductors discussed in the preceding section can be used for photodiodes with similar spectral characteristics. Figure 14.15 shows the spectral responsivity of representative photodiodes as a function of optical wavelength at 300 K.

All junction photodiodes share some basic principles and characteristics. Therefore, we first consider a simple p–n homojunction photodiode for a general discussion of the common principles and characteristics. Specific characteristics of photodiodes with different structures are discussed later in this section.

The general characteristics of a semiconductor p–n homojunction in the absence of optical illumination are thoroughly discussed in Section 12.5. In a semiconductor photodiode, generation of electron–hole pairs by optical absorption can take place in any of the different regions: the depletion layer, the diffusion regions, and the homogeneous regions. In the depletion layer of a diode, the immobile space charges create an internal electric field with a polarity from the n side to the p side, resulting in an electron energy-band gradient shown in Fig. 14.16. When an electron–hole pair is generated in the depletion layer by photoexcitation, the internal field sweeps the electron to the n side and the hole to the p side, as illustrated in Fig. 14.16. This process results in a drift current that flows in the reverse direction from the cathode on the n side to the anode on the p side. If a photoexcited electron–hole pair is generated within one of the diffusion regions at the edges of the depletion layer, the minority carrier, which is the electron in the p-side diffusion region or the hole in the n-side diffusion region, can reach the depletion layer by diffusion and then be swept to the other side by the internal field, as also illustrated in Fig. 14.16. This process results in a diffusion current that also flows

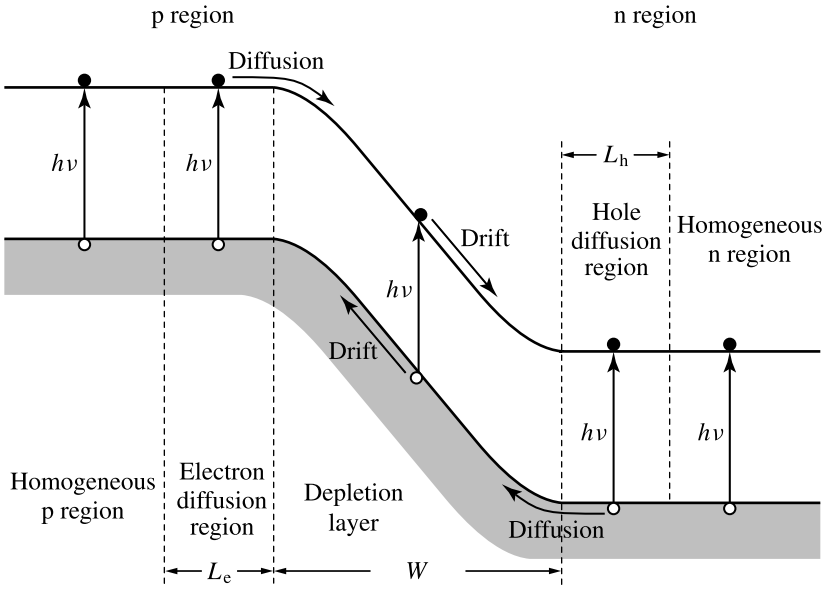


Figure 14.16 Photoexcitation and energy-band gradient of a p–n photodiode.

in the reverse direction. For an electron–hole pair generated by absorption of a photon in the p or n homogeneous region, no current is generated because there is no internal field to separate the charges and a minority carrier generated in a homogeneous region cannot diffuse to the depletion layer before recombining with a majority carrier.

Because photons absorbed in the homogeneous regions do not generate any photocurrent, the active region of a photodiode consists of only the depletion layer and the diffusion regions. For a high-performance photodiode, the diffusion current is undesirable and is minimized. Therefore, the active region mainly consists of the depletion layer where a drift photocurrent is generated. The external quantum efficiency, η_e, of a photodiode is the fraction of total incident photons absorbed in the active region that actually contribute to the photocurrent. For a vertically illuminated photodetector, in which the optical signal reaches the active region in a direction perpendicular to the junction plane, the external quantum efficiency can be expressed as

$$\eta_e = \eta_{coll}\eta_t\eta_i = \eta_{coll}(1 - R)T_h(1 - e^{-\alpha W}), \tag{14.84}$$

where η_{coll} is the collection efficiency of the photogenerated carriers, $\eta_t = (1 - R)T_h$, and $\eta_i = 1 - e^{\alpha W}$. Here, R is the reflectivity of the incident surface of the photodiode, T_h is the transmittance of the homogenous region between the incident surface and the active region, α is the absorption coefficient of the active region, and W is the width of the depletion layer that defines the active region. To improve the quantum efficiency, the surface reflectivity can be reduced by antireflection coating. Besides, the homogeneous region through which the optical signal enters must be made thin to reduce absorption of the optical signal in this region. For a p–n photodiode that has the

incident surface on the p side, the p region has to be very thin and heavily doped so that the depletion layer extends mostly into the thick and lightly doped n region. Ultimately, the quantum efficiency of a photodiode is determined by the absorption coefficient α and the depletion layer thickness W.

Clearly, there are two contributions to the photocurrent in a junction photodiode: a drift current from photogeneration in the depletion layer and a diffusion current from photogeneration in the diffusion regions. The homogeneous regions on the two ends of the diode act like blocking layers for the photogenerated carriers because carriers neither drift nor diffuse through these regions. Consequently, a junction photodiode acts like a photoconductor with two blocking contacts, which is discussed in the preceding section. It has a unity gain, $G = 1$, with the external signal current simply being equal to the photocurrent:

$$i_s = i_{ph} = \eta_e \frac{eP_s}{h\nu}. \tag{14.85}$$

This photocurrent is a reverse current that depends only on the power of the optical signal. When a bias voltage is applied to the photodiode, the total current of the photodiode is the combination of the diode current given in (12.117) and the photocurrent:

$$i(V, P_s) = I_0 \left(e^{eV/ak_BT} - 1\right) - i_s = I_0 \left(e^{eV/ak_BT} - 1\right) - \eta_e \frac{eP_s}{h\nu}, \tag{14.86}$$

which is a function of both the bias voltage V and the optical signal power P_s. Figure 14.17 shows the current–voltage characteristics of a junction photodiode at various power levels of optical illumination. The dark characteristics for $P_s = 0$ are simply those of an unilluminated diode described by (12.117). According to (14.86), the current–voltage characteristics of an illuminated photodiode shift downward from the dark characteristics by the amount of the photocurrent, which is linearly proportional to the optical power but is independent of the bias voltage.

As shown in Fig. 14.17, there are two modes of operation for a junction photodiode. The device functions in *photoconductive mode* in the third quadrant of its current–voltage characteristics, including the short-circuit condition on the vertical axis for $V = 0$. It functions in *photovoltaic mode* in the fourth quadrant, including the open-circuit condition on the horizontal axis for $i = 0$. The mode of operation is determined by the external circuitry and the bias condition.

The circuitry for the photoconductive mode, shown in Fig. 14.17(a), normally consists of a reverse bias voltage of $V = -V_r$ and a load resistance R_L. In this mode of operation, it is necessary to keep the output voltage, v_{out}, smaller than the bias voltage, V_r, so that a reverse voltage is maintained across the photodiode. This requirement can be fulfilled if the bias voltage is sufficiently large while the load resistance is smaller than the internal resistance of the photodiode in reverse bias, as illustrated with the load line in the third quadrant of Fig. 14.17. In the photoconductive mode under the

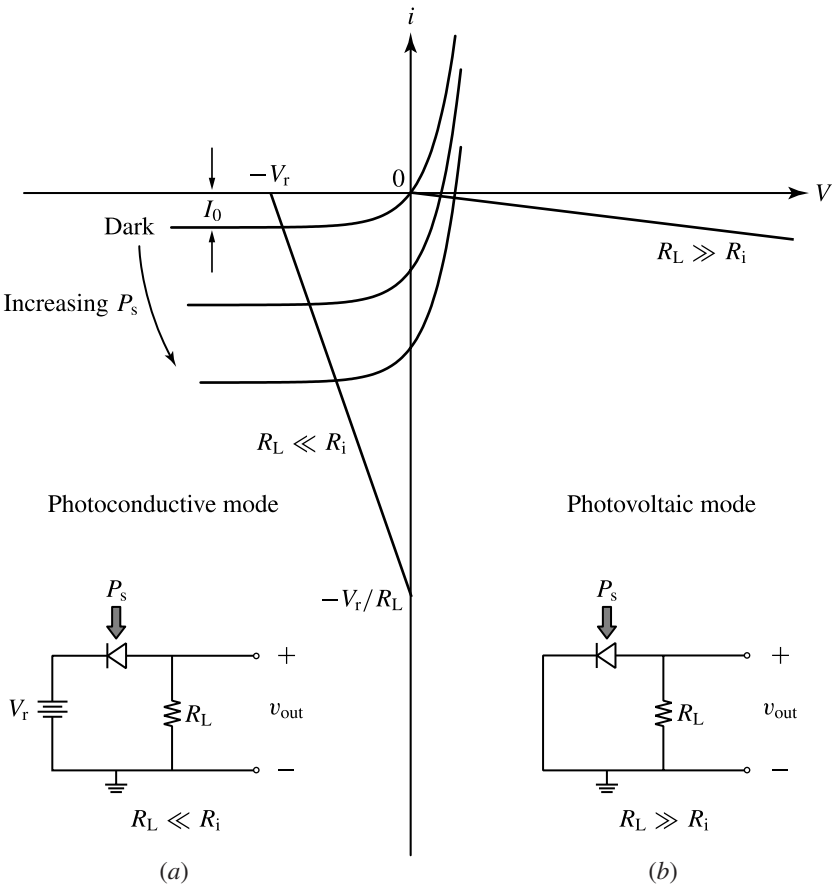


Figure 14.17 Current–voltage characteristics of a junction photodiode at various power levels of optical illumination. The basic circuitry and load line are shown for the photodiode operating in (*a*) photoconductive mode and (*b*) photovoltaic mode.

conditions that $R_L < R_i$ and $v_{out} < V_r$, a photodiode has the following linear response before it saturates:

$$v_{out} = (I_0 + i_s)R_L = \left(I_0 + \eta_e \frac{eP_s}{h\nu}\right) R_L. \tag{14.87}$$

The circuitry for the photovoltaic mode, shown in Fig. 14.17(*b*), does not require a bias voltage but requires a large load resistance. In this mode of operation, the photovoltage appears as a forward bias voltage across the photodiode. As illustrated with the load line in the fourth quadrant of Fig. 14.17, the load resistance is required to be much larger than the internal resistance of the photodiode in forward bias, $R_L \gg R_i$, so that the current i flowing through the diode and the load resistance is negligibly small. In the photovoltaic mode under this condition, the response of the photodiode is not linear

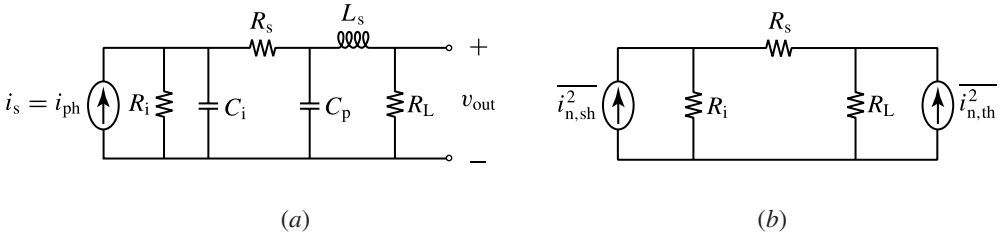


Figure 14.18 (*a*) Small-signal equivalent circuit and (*b*) noise equivalent circuit of a junction photodiode.

but is logarithmic to the optical signal:

$$v_{out} \approx \frac{ak_B T}{e} \ln\left(1 + \frac{i_s}{I_0}\right) = \frac{ak_B T}{e} \ln\left(1 + \eta_e \frac{eP_s}{h\nu I_0}\right), \tag{14.88}$$

where a is a factor of a value between 1 and 2 in the diode equation of (12.117).

In photoconductive mode, electric energy supplied by the bias voltage source is delivered to the photodiode. In photovoltaic mode, electric energy generated by the optical signal can be extracted from the photodiode to the external circuit. Solar cells are basically semiconductor junction diodes operating in photovoltaic mode for converting solar energy into electricity.

Figure 14.18(*a*) shows the small-signal equivalent circuit of a junction photodiode. A photodiode has an internal resistance R_i and an internal capacitance C_i across its junction. Both R_i and C_i depend on the size and the structure of the photodiode and vary with the voltage across the junction. In photoconductive mode under a reverse voltage, the diode has a large R_i normally on the order of 1–100 MΩ for a typical photodiode, and a small C_i dominated by the junction capacitance C_j, as discussed in Section 12.5. As the reverse voltage increases in magnitude, R_i increases but C_i decreases because the depletion-layer width increases with reverse voltage. In photovoltaic mode with a forward voltage across the junction, the diode has a large C_i dominated by the diffusion capacitance C_d, as also discussed in Section 12.5. It still has a large R_i, though smaller than that in the photodiode mode, because it operates near the open-circuit condition with a very small internal current in the fourth quadrant of the current–voltage characteristics. The series resistance R_s takes into account both resistance in the homogeneous regions of the diode and parasitic resistance from the contacts. The external parallel capacitance C_p is the parasitic capacitance from the contacts and the package. The series inductance L_s is the parasitic inductance from the wire or transmission-line connections. The values of R_s, C_p, and L_s can be minimized with careful design, processing, and packaging of the device.

The noise of a photodiode consists of both shot noise and thermal noise. Because a junction photodiode has a unity gain, its shot noise can be expressed as

$$\overline{i^2_{n,sh}} = 2eB(\overline{i_s} + \overline{i_b} + \overline{i_d}), \tag{14.89}$$

where $i_s = i_{ph}$ is the photocurrent. The thermal noise seen at the output can be expressed as

$$\overline{i_{n,th}^2} = \frac{4k_B T B}{R_{eq}}, \tag{14.90}$$

where R_{eq} is the equivalent resistance seen at the output port. From the circuit shown in Fig. 14.18(*b*), we find that

$$R_{eq} = R_L \parallel (R_i + R_s) = \frac{R_L(R_i + R_s)}{R_L + R_i + R_s}. \tag{14.91}$$

In photoconductive mode, the photodiode has a dark current of $i_d = I_0$ and a relatively small load resistance. In photovoltaic mode, the dark current can be eliminated, and the load resistance is required to be very large. Therefore, a photodiode is significantly noisier in photoconductive mode under a reverse bias than in photovoltaic mode without a bias.

High-speed photodiodes are by far the most widely used photodetectors in applications requiring high-speed or broadband photodetection. The speed of a photodiode is determined by two factors: (1) the response time of the photocurrent and (2) the time constant of its equivalent circuit shown in Fig. 14.18(*a*). Because a photodiode operating in photovoltaic mode has a large RC time constant due to the large internal diffusion capacitance in this mode of operation, only photodiodes operating in photoconductive mode are suitable for high-speed or broadband applications. For this reason, we only consider the speed and the frequency response for a photodiode operating in photoconductive mode.

For a photodiode operating in photoconductive mode under a reverse bias, the response time of the photocurrent to an optical signal is determined by two factors: (1) drift of the electrons and holes that are photogenerated in the depletion layer and (2) diffusion of the electrons and holes that are photogenerated in the diffusion regions. Drift of the carriers across the depletion layer is a fast process characterized by the transit times of the photogenerated electrons and holes across the depletion layer. In contrast, diffusion of the carriers is a slow process that is caused by optical absorption in the diffusion regions outside of the high-field depletion region. It results in a diffusion current that can last as long as the lifetime of the carriers. The consequence is a long tail in the impulse response of the photodiode, which translates into a low-frequency falloff in the frequency response of the device. For a high-speed photodiode, this diffusion mechanism has to be eliminated by reducing the photogeneration of carriers outside the depletion layer through design of the device structure. When the diffusion mechanism is eliminated, the frequency response of the photocurrent is only limited by the transit times of electrons and holes.

In general, the frequency response function that is dictated by the carrier transit time depends on the details of the electric field distribution and the photogenerated carrier distribution in the depletion layer. In a semiconductor, electrons normally have a higher

mobility, thus a smaller transit time, than holes. This difference has to be considered in the detailed analysis of the response speed of a photodiode. For a good estimate of the detector frequency response, however, the average of electron and hole transit times can be used:

$$\tau_{\mathrm{tr}} = \frac{1}{2}(\tau_{\mathrm{tr}}^{\mathrm{e}} + \tau_{\mathrm{tr}}^{\mathrm{h}}). \tag{14.92}$$

In the simple case when the process of carrier drift is dominated by a constant transit time of τ_{tr}, the temporal response of the photocurrent is ideally a rectangular function of duration τ_{tr}. Therefore, the power spectrum of the photocurrent frequency response can be approximately expressed by

$$\mathcal{R}_{\mathrm{ph}}^2(f) = \left|\frac{i_{\mathrm{ph}}(f)}{P_{\mathrm{s}}(f)}\right|^2 \approx \mathcal{R}_{\mathrm{ph}}^2(0)\left(\frac{\sin \pi f \tau_{\mathrm{tr}}}{\pi f \tau_{\mathrm{tr}}}\right)^2, \tag{14.93}$$

which has a transit-time-limited 3-dB cutoff frequency

$$f_{3\mathrm{dB}}^{\mathrm{ph}} \approx \frac{0.443}{\tau_{\mathrm{tr}}}. \tag{14.94}$$

The frequency response of the equivalent circuit shown in Fig. 14.18(*a*) is determined by (1) the internal resistance R_{i} and capacitance C_{i} of the photodiode; (2) the parasitic effects characterized by R_{s}, C_{p}, and L_{s}; and (3) the load resistance R_{L}. Clearly, the parasitic effects must be eliminated as much as possible because they can degrade the performance of a high-speed photodiode. A high-speed photodiode normally operates under the condition that $R_{\mathrm{i}} \gg R_{\mathrm{L}}, R_{\mathrm{s}}$. Therefore, when parasitic inductance is eliminated, the ultimate speed of the circuit is dictated by the RC time constant $\tau_{\mathrm{RC}} = (R_{\mathrm{L}} + R_{\mathrm{s}})(C_{\mathrm{i}} + C_{\mathrm{p}})$. Its frequency response has the following power spectrum:

$$\mathcal{R}_{\mathrm{ckt}}^2(f) \approx \frac{\mathcal{R}_{\mathrm{ckt}}^2(0)}{1 + 4\pi^2 f^2 \tau_{\mathrm{RC}}^2}, \tag{14.95}$$

which has an RC-time-limited 3-dB cutoff frequency

$$f_{3\mathrm{dB}}^{\mathrm{ckt}} \approx \frac{1}{2\pi\tau_{\mathrm{RC}}} = \frac{1}{2\pi(R_{\mathrm{L}} + R_{\mathrm{s}})(C_{\mathrm{i}} + C_{\mathrm{p}})}. \tag{14.96}$$

Combining the photocurrent response and the circuit response, the total output power spectrum of an optimized photodiode operating in photoconductive mode is

$$\mathcal{R}^2(f) = \mathcal{R}_{\mathrm{ph}}^2(f)\mathcal{R}_{\mathrm{ckt}}^2(f) = \frac{\mathcal{R}^2(0)}{1 + 4\pi^2 f^2 \tau_{\mathrm{RC}}^2}\left(\frac{\sin \pi f \tau_{\mathrm{tr}}}{\pi f \tau_{\mathrm{tr}}}\right)^2. \tag{14.97}$$

This total frequency response has a 3-dB cutoff frequency, $f_{3\mathrm{dB}}$, that can be found

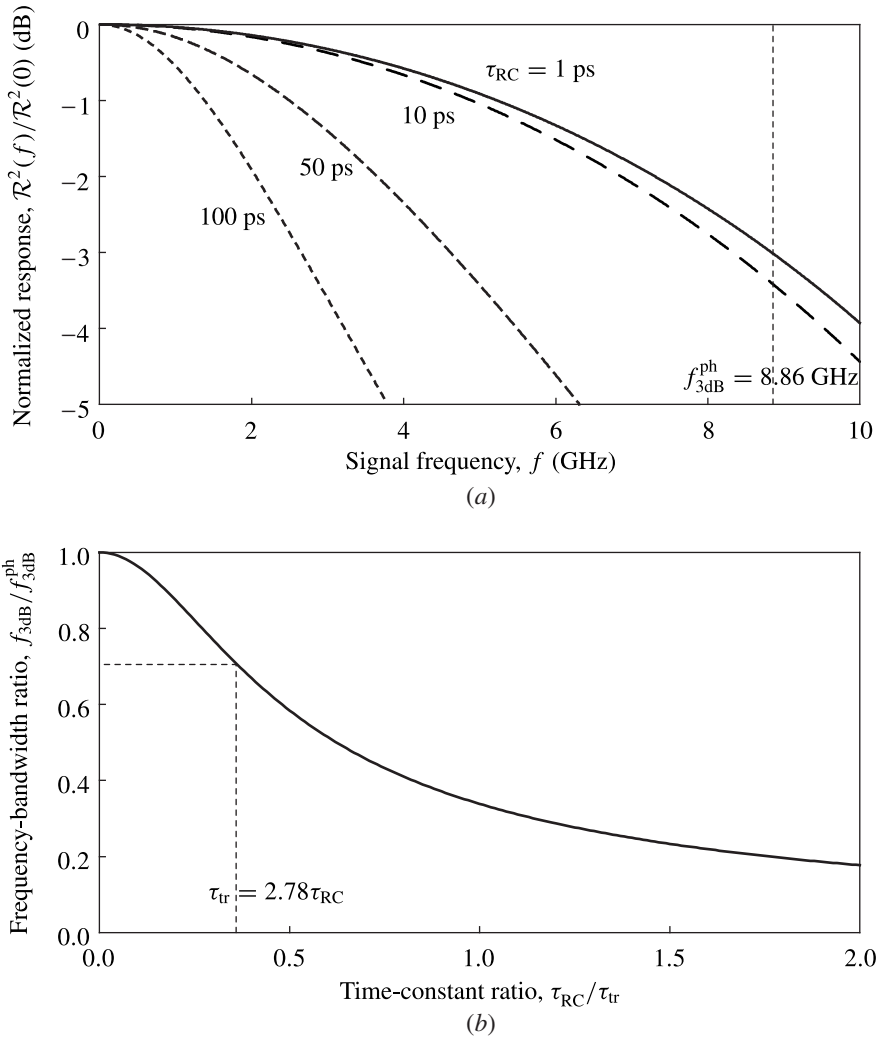


Figure 14.19 (*a*) Total frequency response, normalized to the zero-frequency response, of a photodiode for a fixed value of $\tau_{tr} = 50$ ps but for a few different values of τ_{RC}. (*b*) Dependence of the ratio f_{3dB}/f_{3dB}^{ph} on the ratio τ_{RC}/τ_{tr}.

approximately by using the following rule of the sum of squares:

$$\frac{1}{f_{3dB}^2} = \frac{1}{(f_{3dB}^{ph})^2} + \frac{1}{(f_{3dB}^{ckt})^2}. \tag{14.98}$$

By using (14.94) for f_{3dB}^{ph} and (14.96) for f_{3dB}^{ckt}, the 3-dB cutoff frequency of a photodiode including transit-time and circuit limitations can be expressed approximately as

$$f_{3dB} \approx \frac{0.443}{[\tau_{tr}^2 + (2.78\tau_{RC})^2]^{1/2}} = \frac{1}{2\pi[\tau_{RC}^2 + (0.36\tau_{tr})^2]^{1/2}}. \tag{14.99}$$

Figure 14.19 shows the total frequency response given by (14.97) for a fixed value of τ_{tr} but for a few different values of τ_{RC}. It is seen that the total frequency response is transit-time-limited when $\tau_{tr} > 2.78\tau_{RC}$, but is RC-time-limited when $\tau_{tr} < 2.78\tau_{RC}$. The characteristics given by (14.97) and shown in Fig. 14.19 represent the ultimate frequency response of a photodiode. In practice, the frequency response of a photodiode can be substantially degraded by the presence of a significant diffusion current and by parasitic effects. The optimum design of a high-speed photodiode requires (1) elimination of the diffusion current, (2) elimination of parasitic effects, and (3) equalization of the transit-time-limited bandwidth and the RC-time-limited bandwidth by making $\tau_{tr} = 2.78\tau_{RC}$.

An important consideration for a high-speed photodiode is the bandwidth–efficiency product, $\eta_e f_{3dB}$, rather than the bandwidth alone because increasing the bandwidth can often result in a reduced efficiency in many device structures. Many different approaches can be taken to optimize both the bandwidth and the efficiency for a maximum bandwidth–efficiency product. This issue is further addressed in the following discussions of various device structures.

p–i–n photodiodes

A p–i–n photodiode consists of an intrinsic region sandwiched between heavily doped p^+ and n^+ regions. Figure 14.20 shows the comparison between a p–n junction photodiode and a p–i–n photodiode. In a p–n photodiode, the depletion-layer width and the junction capacitance both vary with reverse voltage across the junction. The electric field in the depletion layer is not uniform. In a p–i–n photodiode, a reverse bias voltage applied to the device drops almost entirely across the intrinsic region because of high resistivity in the intrinsic region and low resistivities in the surrounding p^+ and n^+ regions. As a result, a p–i–n diode has the following two important characteristics: (1) the depletion layer is almost completely defined by the intrinsic region; (2) the electric field in the depletion layer is uniform across the intrinsic region. In practice, the intrinsic region does not have to be truly intrinsic but only has to be highly resistive. It can be either a highly resistive p region, called a π region, or a highly resistive n region, called a ν region.

The depletion-layer width W in a p–i–n diode does not vary significantly with bias voltage but is pretty much fixed by the thickness, d_i, of the intrinsic region so that $W \approx d_i$. The internal capacitance of a p–i–n diode can be predetermined in the design of the device through the choice of the thickness of the intrinsic region and the device area $\mathcal{A}$:

$$C_i = C_j = \frac{\epsilon \mathcal{A}}{W} \approx \frac{\epsilon \mathcal{A}}{d_i}. \tag{14.100}$$

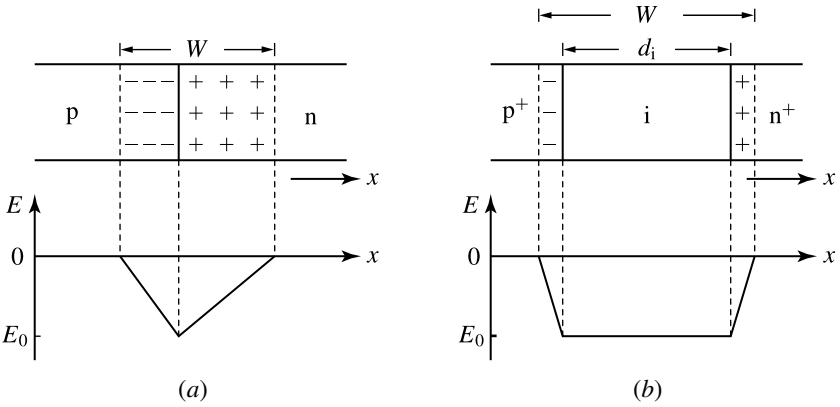


Figure 14.20 Structure and internal field distribution of (*a*) a p–n photodiode and (*b*) a p–i–n photodiode.

This capacitance is fairly independent of the bias voltage; thus it remains constant in operation.

When a reverse voltage is applied to a p–i–n diode, a uniform electric field that is linearly proportional to the reverse bias voltage exists throughout the intrinsic region:

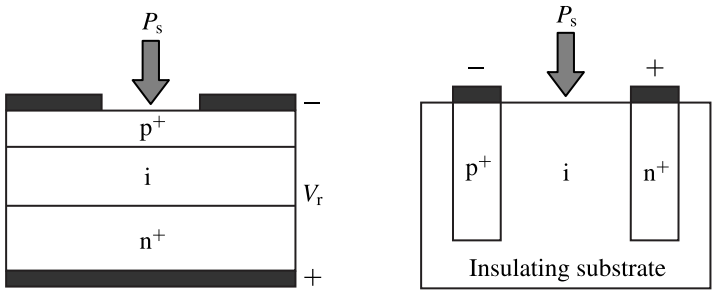
$$E \approx \frac{V_0 + V_r}{W} \approx \frac{V_r}{d_i}, \tag{14.101}$$

for $V_r \gg V_0$. Due to this uniform field, both electrons and holes have constant drift velocities across the depletion layer in a p–i–n photodiode. At low and moderate fields, the drift velocities of electrons and holes both vary linearly with the electric field strength. For a p–i–n photodiode operating in this regime with a relatively low reverse bias voltage, the average carrier transit time is given by

$$\tau_{tr} = \frac{1}{2}\left(\frac{W}{\mu_e E} + \frac{W}{\mu_h E}\right) \approx \frac{d_i^2}{2\mu V_r}, \tag{14.102}$$

where $\mu = \mu_e\mu_h/(\mu_e + \mu_h)$. Because the depletion-layer width in a p–i–n diode is dictated by the thickness of the intrinsic region, the transit time is inversely proportional to the bias voltage. Therefore, the response speed of the photodiode can be improved by increasing the reverse bias voltage. At high fields, however, both electron and hole drift velocities reach their respective saturation velocities: $v_e \approx v_e^{sat}$ and $v_h \approx v_h^{sat}$, which vary little with bias voltage. For most semiconductors, this occurs at a field strength above 100 MV m^{-1} for a saturation velocity on the order of 10^5 m s^{-1}. For a p–i–n photodiode operating in this regime with a sufficiently large reverse bias voltage, electrons and holes have a constant average transit time across the depletion layer:

$$\tau_{tr} = \frac{W}{v_{sat}} \approx \frac{d_i}{v_{sat}}, \tag{14.103}$$



(*a*) (*b*)

Figure 14.21 Schematic cross-sectional structures of (*a*) a vertical p–i–n photodiode and (*b*) a lateral p–i–n photodiode.

where $1/v_{sat} = (1/v_e^{sat} + 1/v_h^{sat})/2$. So long as the reverse bias voltage is large enough to keep electrons and holes drifting at their respective saturation velocities, τ_{tr} is independent of the bias voltage and can thus be predetermined by the thickness of the intrinsic region through the design of the device.

Compared to a p–n photodiode, in which the depletion-layer width varies with bias voltage, a p–i–n photodiode has a number of advantages because its depletion-layer width is determined by the thickness of the intrinsic region and is independent of the bias voltage. Both the quantum efficiency and the frequency response of a p–i–n photodiode can be optimized by the geometric design of the device, whereas those of a p–n photodiode depend strongly on the bias voltage. From the above discussions, it is clear that the transit time, the RC time constant, and the internal quantum efficiency of a vertically illuminated p–i–n photodiode, shown in Fig. 14.21(*a*), all depend on the thickness d_i of the intrinsic region: $\tau_{tr} \propto d_i$, $C_i \propto d_i^{-1}$, and $\eta_i = 1 - e^{-\alpha d_i}$. For a high quantum efficiency, the thickness d_i of the intrinsic region can be chosen to be larger than the absorption length: $d_i > 1/\alpha$. To optimize the speed of a p–i–n photodiode, both the thickness of the intrinsic region and the area of the device have to be properly chosen. To reduce the diffusion current, d_i can be chosen to be larger than the electron diffusion length in the p^+ region and the hole diffusion length in the n^+ region: $d_i \gg L_e, L_h$. A large d_i reduces the RC time constant of the device by reducing C_i, but it increases the transit time τ_{tr}. Because the electric field is relatively constant throughout the active region of a p–i–n photodiode, the transit time can be optimized with a chosen d_i. Because C_i can be reduced by reducing the device area, a p–i–n photodiode normally has an intrinsic region that has a thickness chosen to optimize the quantum efficiency and the transit time. For a high-speed p–i–n photodiode, the device area is made small enough that the RC time constant is not a limiting factor of its frequency response.

One major limitation of p–i–n photodiodes that are made of indirect-gap semiconductors, such as Si and Ge, is the small absorption coefficients of these semiconductors in the spectral regions where only indirect absorption takes place in such semiconductors.

For example, at $\lambda = 850$ nm, the absorption coefficient at 300 K is only about 7×10^4 m^{-1} for Si but is about 1×10^6 m^{-1} for GaAs though 850 nm is farther away from the bandgap wavelength of 1.11 μm for Si than from that of 871 nm for GaAs. This results in a low quantum efficiency, thus a small responsivity, for a Si or Ge p–i–n photodiode of even just a moderate speed because of the conflicting requirements on the thickness d_i for reducing τ_{tr} and increasing η_i in a *vertical p–i–n* photodiode shown in Fig. 14.21(*a*).

One solution to this problem is provided by the *lateral p–i–n* geometry shown in Fig. 14.21(*b*). In a lateral p–i–n, both τ_{tr} and C_i still depend on d_i in the same manner as in a vertical p–i–n, but the internal quantum efficiency is not a function of d_i but is a function of the trench depth d as $\eta_i = 1 - e^{-\alpha d}$. Thus, f_{3dB} and η_i can be independently optimized by properly choosing a value of d_i to optimize τ_{tr} and C_i for a large f_{3dB} while making a deep enough trench for a high value of η_i. One additional advantage of a lateral p–i–n photodiode is that the incident optical signal does not have to pass through the homogeneous p^+ or n^+ region before it reaches the active intrinsic region, thus improving the external quantum efficiency. This feature is significant for a homojunction p–i–n used for optical detection at short optical wavelengths, such as a Si p–i–n for blue or ultraviolet wavelengths, where the absorption coefficient is very high and the optical penetration depth is very small.

EXAMPLE 14.12 A vertically illuminated InGaAs/InP p–i–n photodiode for $\lambda = 1.3$ μm consists of a lightly doped n^--InGaAs layer of a thickness d_i between a thin p^+-InGaAs top layer and an n^+-InP substrate. The device is reverse-biased at a sufficiently high bias voltage for both electrons and holes to reach their respective saturation velocities of $v_e^{sat} = 6.5 \times 10^4$ m s^{-1} and $v_h^{sat} = 4.8 \times 10^4$ m s^{-1}. The absorption coefficient of InGaAs at 1.3 μm is $\alpha = 1.16 \times 10^6$ m^{-1} $= 1.16$ μm^{-1}. The dielectric susceptibility of InGaAs at DC and low frequencies is $\epsilon = 14.1\epsilon_0$. Take $R = R_L + R_s = 50\ \Omega$, $C_p = 0$, and $L_s = 0$ for this device. This device can be designed to be either front or back illuminated and can be antireflection coated to have a high η_t; meanwhile, its structure can be optimized to have a high η_{coll}. In any event, its bandwidth–efficiency product is limited to $\eta_i f_{3dB}$ because $\eta_i \geq \eta_e$. The device is made to have a circular active area of a diameter $2r$. Plot its 3-dB cutoff frequency, f_{3dB}, and the upper limit of its bandwidth–efficiency product, $\eta_i f_{3dB}$, as a function of the intrinsic layer thickness d_i in the range of $0 < d_i < 3$ μm for the four different diameters of $2r = 10$, 20, 40, and 80 μm.

Solution The average transit time can be calculated using (14.103) with the following average saturation velocity for electrons and holes:

$$v_{sat} = \left[\frac{1}{2}\left(\frac{1}{v_e^{sat}} + \frac{1}{v_h^{sat}}\right)\right]^{-1} = \left[\frac{1}{2}\left(\frac{1}{6.5 \times 10^4} + \frac{1}{4.8 \times 10^4}\right)\right]^{-1} \text{ m s}^{-1}$$
$$= 5.52 \times 10^4 \text{ m s}^{-1}.$$

The active area is $\mathcal{A} = \pi r^2$. The internal capacitance of the photodiode is $C_i = \epsilon \mathcal{A}/d_i = \epsilon \pi r^2/d_i$. Thus, the RC time constant

$$\tau_{RC} = RC_i = R\frac{\epsilon \pi r^2}{d_i},$$

with $R = 50\ \Omega$ and $\epsilon = 14.1\epsilon_0$. From (14.99), we then have

$$f_{3dB} \approx \frac{0.443}{\left[\tau_{tr}^2 + (2.78\tau_{RC})^2\right]^{1/2}} = \frac{0.443}{\left\{(d_i/v_{sat})^2 + \left[2.78R(\epsilon \pi r^2/d_i)\right]^2\right\}^{1/2}}.$$

The values of f_{3dB} in the range of $0 < d_i < 3$ μm are calculated using this relation for $2r = 10, 20, 40$, and 80 μm. Then the bandwidth–efficiency product is calculated using

$$\eta_i f_{3dB} = (1 - e^{-\alpha d_i}) f_{3dB}.$$

The values of both f_{3dB} and $\eta_i f_{3dB}$ are plotted as a function of d_i in Fig. 14.22. From the data shown in this figure, we see that for a given device diameter there is an optimum intrinsic layer thickness of d_{opt} for a maximum value of f_{3dB} and a different optimum intrinsic layer thickness of d'_{opt} for a maximum value of $\eta_i f_{3dB}$. We also find that $d'_{opt} > d_{opt}$. The cutoff frequency is primarily limited by τ_{RC} if $d_i < d_{opt}$, whereas it is primarily limited by τ_{tr} if $d_i > d_{opt}$. For a given device diameter, there is one possible choice of d_i on either side of d_{opt} for a sufficiently large value of f_{3dB}. For a desired f_{3dB}, the choice of $d_i > d_{opt}$ has a larger bandwidth–efficiency product than that of $d_i < d_{opt}$.

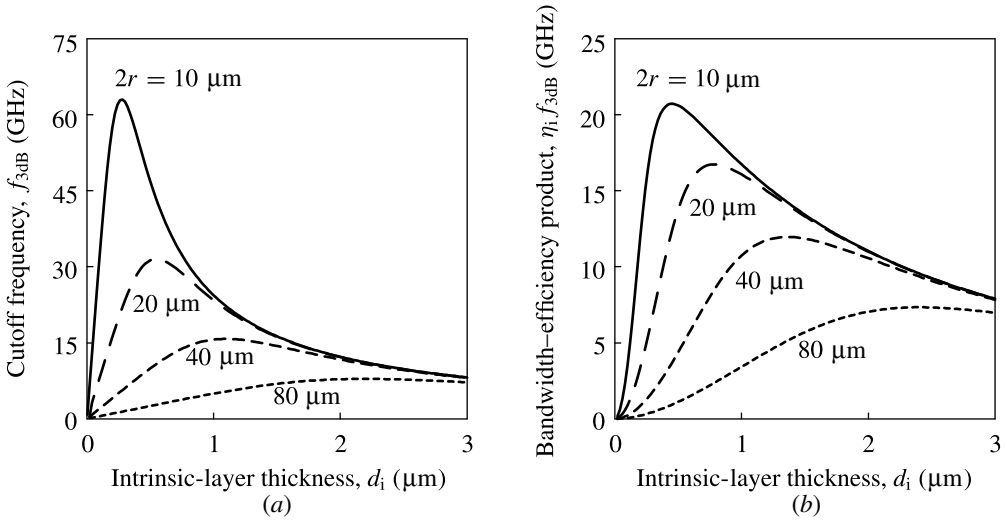


Figure 14.22 (*a*) Cutoff frequency, f_{3dB}, and (*b*) bandwidth–efficiency product, $\eta_i f_{3dB}$, of an InGaAs/InP p–i–n photodiode for 1.3 μm wavelength as a function of intrinsic layer thickness d_i for four different device diameters of $2r = 10, 20, 40$, and 80 μm.

Heterojunction photodiodes

Heterojunction structures offer additional flexibility in optimizing the performance of a photodiode. In a heterojunction photodiode, the active region normally has a bandgap that is smaller than one or both of the homogeneous regions. A large-gap homogeneous region, which can be either the top p^+ region or the substrate n region, serves as a window for the optical signal to enter. The small bandgap of the active region determines the threshold wavelength, λ_{th}, of the detector on the long-wavelength side, while the large bandgap of the homogeneous window region sets a cutoff wavelength, λ_c, on the short-wavelength side. For an optical signal that has a wavelength λ_s in the range of $\lambda_{th} > \lambda_s > \lambda_c$, the quantum efficiency and the responsivity can be optimized. A limiting factor for the speed of a heterojunction photodiode is the trapping of electrons at the conduction-band discontinuity and that of holes at the valence-band discontinuity. For high-speed applications, this limitation has to be removed by reducing the barrier height through compositional grading at the interface of the heterojunction. Many III–V p–i–n photodiodes have heterojunction structures, which can be either symmetric with a small-bandgap active intrinsic region sandwiched between large-bandgap p^+ and n^+ regions, such as p^+-AlGaAs/GaAs/n^+-AlGaAs and p^+-InP/InGaAs/n^+-InP, or asymmetric with a large-bandgap p^+ or n^+ region on only one side, such as p^+-AlGaAs/GaAs/n^+-GaAs or p^+-InGaAs/InGaAs/n^+-InP. Figure 14.23 shows some structures of heterojunction photodiodes.

Sophisticated heterojunction structures such as quantum wells and strained quantum wells, as well as quantum wires and quantum dots, are also used for the active region

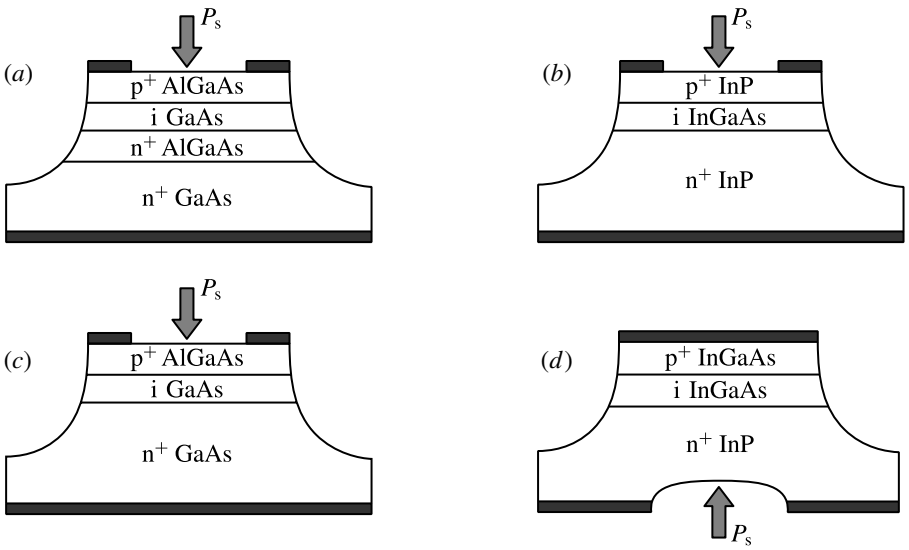


Figure 14.23 Structures of heterojunction photodiodes.

of photodiodes. Such quantum structures have the advantage of high peak absorption coefficients, which lead to an improved quantum efficiency for a given thickness of the active region. They are often used for improving the bandwidth–efficiency products of high-speed photodetectors.

Schottky photodiodes

The property of the interface between a metal and a semiconductor depends on the work functions of the metal and the semiconductor, $e\phi_m$ and $e\phi_s$, respectively, and the type of semiconductor. The metal–semiconductor junction is an ohmic contact without a potential barrier if $\phi_s > \phi_m$ in the case of an n-type semiconductor or $\phi_s < \phi_m$ in the case of a p-type semiconductor. A *Schottky barrier* of a height $E_b = e(\phi_m - \chi)$ for electrons to flow from the metal to the semiconductor exists at the metal–semiconductor junction if $\phi_s < \phi_m$ in the case of an n-type semiconductor, as shown in Fig. 14.24(a). A Schottky barrier of a height $E_b = E_g - e(\phi_m - \chi)$ for holes to flow from the metal to the semiconductor exists at the metal–semiconductor junction if $\phi_s > \phi_m$ in the case of a p-type semiconductor, as shown in Fig. 14.24(b).

The general characteristics of a Schottky junction are similar to those of a p–n junction. The characteristics of a Schottky junction formed between a metal and an n-type semiconductor can be approximated by those of a p^+–n junction with a built-in potential of $V_0 = \phi_m - \phi_s$, as shown in Fig. 14.24(a). Similarly, a Schottky junction between a

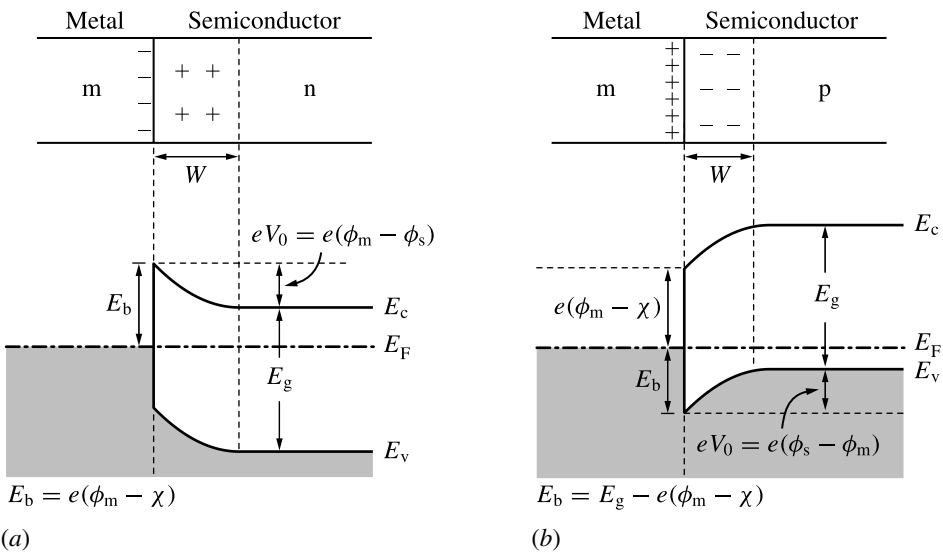


Figure 14.24 Schottky junctions at (a) the interface of a metal and an n-type semiconductor with $\phi_s < \phi_m$ and (b) the interface of a metal and a p-type semiconductor with $\phi_s > \phi_m$.

metal and a p-type semiconductor can be considered as an n^+–p junction with a built-in potential of $V_0 = \phi_s - \phi_m$, as shown in Fig. 14.24(*b*). Therefore, the depletion-layer width W of a Schottky junction and its dependence on bias voltage can be found by using (12.100) and by taking $N_a \gg N_d$ in the case of an n-type semiconductor or $N_d \gg N_a$ in the case of a p-type semiconductor. The junction capacitance simply has the same form as that of a p–n junction given in (12.119).

It is also possible for a Schottky diode to function like a p–i–n diode by inserting a lightly doped n^--semiconductor layer between a metal and a heavily doped n^+-semiconductor region. In such a structure, the metal functions as a p^+-homogeneous region, and the n^- layer functions as the intrinsic region in a p–i–n diode. The depletion layer, which exists almost entirely in the n^- region, broadens as the reverse bias voltage increases until it reaches the metal at a voltage known as the *punchthrough voltage*. When the reverse bias voltage is larger than the punchthrough voltage, the depletion-layer width of such a Schottky diode becomes independent of the voltage and is simply defined by the thickness of the n^- layer.

The characteristics and the equivalent circuit of a Schottky photodiode are similar to those of a semiconductor junction photodiode discussed above. A Schottky photodiode can also operate in either photoconductive or photovoltaic mode, but it normally operates in photoconductive mode in most of its application for the same reasons as discussed above for other junction photodiodes. A Schottky photodiode operating in photoconductive mode can have a very high speed, particularly when an n-type semiconductor is used. Because the optical signal is absorbed in a thin layer at the junction interface, only the majority carriers, which are electrons in the case of an n-type semiconductor, have to drift across the active region. A well-designed Schottky photodiode can reach an intrinsic frequency bandwidth as high as 100 GHz.

The spectral response of a Schottky photodiode depends on whether an optical signal is absorbed by the semiconductor or by the metal. If the optical signal is absorbed by the semiconductor, the spectral characteristic of a Schottky photodiode is the same as that of a semiconductor junction photodiode with a threshold photon energy defined by the bandgap of the absorbing semiconductor: $h\nu > E_{th} = E_g$. This process takes place when the Schottky photodiode has a thin, semi-transparent metallic layer to allow the optical signal to enter with little attenuation before it reaches the depletion layer. This is the normal mode of operation for a high-efficiency, high-speed Schottky photodiode. Absorption of a photon by the metal at the junction interface can also produce a photoresponse if the photon has sufficient energy to excite an electron over the Schottky barrier. For a Schottky photodiode to operate in this mode, the metallic layer has to be thick and absorbing, but the absorption has to take place at the junction interface. The spectral response range in this mode of operation is then $E_b < h\nu < E_g$ for the optical signal to enter from the semiconductor side without being absorbed by the semiconductor. A Schottky photodiode operating in this mode is useful

as an infrared detector, but its efficiency is low because a metal does not absorb light efficiently.

EXAMPLE 14.13 An InGaAs/InP Schottky photodiode has a structure similar to that of the InGaAs/InP p–i–n photodiode considered in Example 14.12, but it has a metallic layer in place of the p^+ layer of the p–i–n photodiode. The thickness of the n^- layer is $d_i = 1\ \mu m$. The diameter of the device is $2r = 12\ \mu m$. It is back illuminated through the InP substrate. The device is biased above the punchthrough voltage, and the electrons have reached their saturation velocity. (a) What is the spectral response range of this photodiode at 300 K? (b) Find the 3-dB cutoff frequency of this photodiode if $R = R_L + R_s = 50\ \Omega$ and $C_p = 0$.

Solution (a) The spectral response range of this back-illuminated photodiode is limited at the short-wavelength end by a cutoff wavelength λ_c determined by the bandgap of the InP window layer because an optical signal has to pass through the InP substrate to reach the InGaAs active layer. It is limited at the long-wavelength end by the threshold wavelength λ_{th} determined by the bandgap of InGaAs that is lattice matched to InP. From the discussions following (12.9), we find that the absorption edge of InP is at 919 nm and that of InGaAs is at 1.65 μm. Therefore, the spectral response range of this Schottky photodiode at 300 K is from $\lambda_c = 919$ nm to $\lambda_{th} = 1.65\ \mu m$.

(b) In a Schottky photodiode, only the majority carriers, which in this case are electrons, have to drift across the active region. Thus, the transit time is simply that of the electrons. From Example 14.12, we have $v_e^{sat} = 6.5 \times 10^4\ m\ s^{-1}$. With $d_i = 1\ \mu m$, we find that

$$\tau_{tr} = \frac{d_i}{v_e^{sat}} = \frac{1 \times 10^{-6}}{6.5 \times 10^4}\ s = 15.4\ ps.$$

With $\epsilon = 14.1\epsilon_0$ from Example 14.12, we find that the internal capacitance of the device for $d_i = 1\ \mu m$ and $2r = 12\ \mu m$ is

$$C_i = \frac{\epsilon \pi r^2}{d_i} = \frac{14.1 \times 8.85 \times 10^{-12} \times \pi \times (12 \times 10^{-6}/2)^2}{1 \times 10^{-6}}\ F = 14.1\ fF.$$

With $R = R_L + R_s = 50\ \Omega$, the RC time constant

$$\tau_{RC} = RC_i = 50 \times 14.1 \times 10^{-15}\ s = 705\ fs.$$

Therefore, the 3-dB cutoff frequency of this photodiode is

$$f_{3dB} = \frac{0.443}{\left[(15.4 \times 10^{-12})^2 + (2.78 \times 705 \times 10^{-15})^2\right]^{1/2}}\ Hz = 28.5\ GHz.$$

Because $\tau_{tr} \gg \tau_{RC}$ for this device, f_{3dB} is almost entirely determined by the electron transit time.

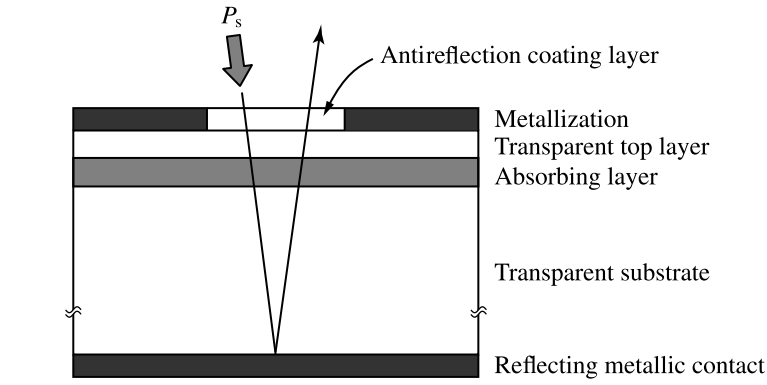
Photodiodes with multipass structures

A high-speed photodiode requires a thin depletion layer for a short transit time, but, according to (14.84), the quantum efficiency of the photodiode decreases as the depletion-layer width W is reduced. Therefore, there is a trade-off between its frequency bandwidth and quantum efficiency. To optimize both the bandwidth and the efficiency of a high-speed photodiode, a large bandwidth–efficiency product $\eta_e f_{3dB}$ is desired. From (14.84), it can be seen that the external quantum efficiency of a photodiode can be increased without changing the depletion-layer width by (1) antireflection coating the incident surface to make $R = 0$ and (2) using a heterostructure with a nonabsorbing large-bandgap homogeneous region for $T_h = 1$. Many different device structures have been developed to increase the bandwidth–efficiency product further beyond that obtained with these two simple steps. They can be divided into three basic categories: (1) *vertically illuminated photodetectors* with multiple optical passes through the active region, which are discussed here; (2) *laterally illuminated photodetectors* such as the lateral p–i–n photodetectors, which are discussed earlier; and (3) *guided-wave photodetectors*, which are discussed in Section 14.7.

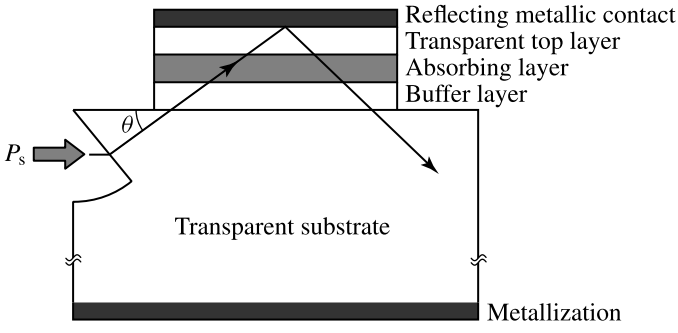
Figure 14.25 shows three approaches to increasing the bandwidth–efficiency product of a photodiode by increasing its quantum efficiency without increasing the thickness of its active region. The simple *double-pass* structure, shown in Fig. 14.25(*a*), directs the optical signal to pass through the active region twice with a back reflector of a reflectivity R_b, which can be simply the substrate electrode if the substrate is transparent. With this structure, the quantum efficiency can be improved by a factor close to $1 + R_b$ if the absorbing active region has a thickness of $W < \alpha^{-1}$.

To increase the quantum efficiency further, the effective optical path length in the active region can be increased without increasing the physical thickness of the active region by using the *refracting-facet* structure shown in Fig. 14.25(*b*). In this structure, the top electrode reflects the optical signal for a second pass through the active region to keep the advantage of a double-pass structure, but the optical signal passes through the active region at an angle θ for a total effective path length of $2W/\sin\theta$. Therefore, the quantum efficiency is further increased over that of the simple double-pass structure shown in Fig. 14.25(*a*). A bandwidth–efficiency product around 40 GHz has been obtained for refracting-facet photodiodes.

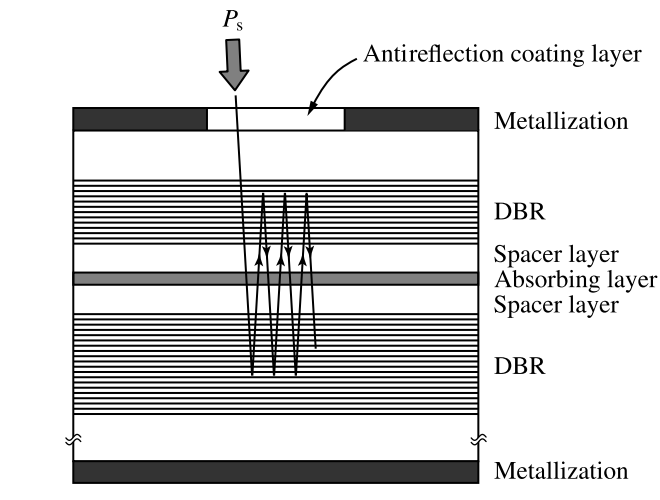
To push the quantum efficiency close to unity in a high-speed photodiode with a very thin active region, a *resonant-cavity-enhanced* structure shown in Fig. 14.25(*c*) can be used. This structure consists of both a front and a back reflector to form a resonant cavity. It functions in a manner similar to that of a VCSEL by forming a standing wave with its high-intensity crest located at the thin absorbing active region. By using DBR reflectors of a reflectivity greater than 99% for a high-Q cavity, a quantum efficiency higher than 90% can be achieved with this scheme. A bandwidth–efficiency product around 20 GHz has been obtained for resonant-cavity-enhanced p–i–n and



(*a*)



(*b*)



(*c*)

Figure 14.25 Photodiodes with multiple optical passes to increase quantum efficiency: (*a*) double-pass photodiode, (*b*) refracting-facet photodiode, and (*c*) resonant-cavity-enhanced photodiode.

Schottky photodiodes. The resonant-cavity-enhanced structure is highly wavelength selective because of its resonance nature. This wavelength selectivity is a disadvantage for general applications because of its narrow optical bandwidth, but it is a useful feature for applications in wavelength-selective detection systems such as wavelength-division multiplexing systems.

14.6 Avalanche photodiodes

The avalanche photodiode (APD) is the solid-state counterpart of the PMT. An APD versus an ordinary junction photodiode is similar to a PMT versus a vacuum photodiode. However, the high-gain and low-noise characteristics of PMTs are difficult for conventional APDs to match. An internal gain is built into an APD to multiply the photogenerated electrons and holes. The physical process responsible for the internal gain in an APD is *avalanche multiplication* of charge carriers through *impact ionization*, as illustrated in Fig. 14.26. In the impact ionization process, an electron or hole of a sufficiently high kinetic energy can create a secondary electron–hole pair by transferring its kinetic energy to the excitation of the secondary carriers through collision with the lattice. In the presence of a high electric field, the newly generated electron and hole can be accelerated to gain sufficient kinetic energies for impact ionization to generate more electron–hole pairs. A cascade of these events leads to avalanche multiplication of the photogenerated carriers. This process does not take place in an ordinary photodiode.

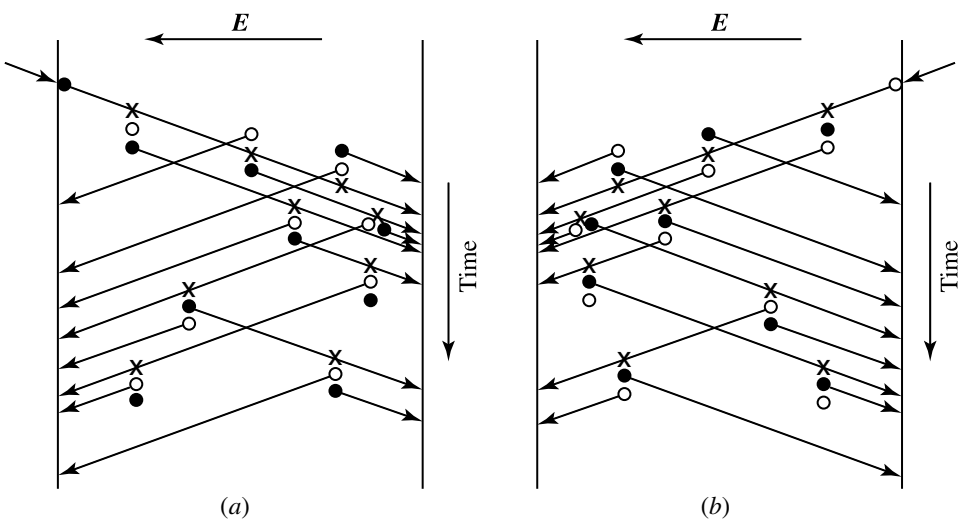


Figure 14.26 Avalanche multiplication of electrons and holes through impact ionization in a semiconductor in the presence of a high electric field with (*a*) electron injection in the case of $k < 1$ and (*b*) hole injection in the case of $k > 1$.

The spectral response of an APD is similar to that of an ordinary photodiode with a threshold photon energy of $E_{th} = E_g$ determined by the bandgap of the absorption region where electron–hole pairs are photogenerated. The threshold kinetic energies for an electron or hole to initiate impact ionization in a semiconductor of a bandgap E_g fall between E_g and $2E_g$, depending on the effective electron and hole masses and the details of the band structure. These threshold energies are much higher than the kinetic energies of electrons and holes at their respective saturation velocities. Therefore, no avalanche multiplication takes place in an ordinary photodiode even when the photogenerated electrons and holes in the device are accelerated to reach their respective saturation velocities, such as in a high-speed p–i–n photodiode. In an APD, the average drift velocities of electrons and holes remain at the saturation velocities, but high-energy carriers at the tail of the energy distribution can have kinetic energies higher than the threshold energies for impact ionization.

The impact ionization process is characterized quantitatively by the *ionization coefficients*, α_e for electrons and α_h for holes (quoted per meter, but also often quoted per centimeter). The ionization coefficient for an electron or hole represents the probability for an electron or hole that travels a unit distance to create an electron–hole pair through impact ionization. Both α_e and α_h are characteristics of a semiconductor and are strong functions of both electric field strength and temperature. They increase rapidly with an increasing electric field strength but decrease with increasing temperature. Their ratio, known as the *ionization ratio*, is defined as

$$k = \frac{\alpha_h}{\alpha_e}. \tag{14.104}$$

The ionization ratio is a function of field strength and temperature. It also varies among different semiconductors. When $k < 1$, impact ionization by electrons dominates. When $k > 1$, impact ionization by holes dominates. For Si, $k < 1$, and the value of k can be as small as 0.01, depending on the field strength. Therefore, impact ionization in Si is completely dominated by electrons. For Ge and InP, $k > 1$, but the value of k is not large. For GaAs, $k \approx 1$. As we shall see below, to maximize the avalanche gain and minimize the excess noise, an ideal APD must have only electrons initiating impact ionization, thus $k \ll 1$, or only holes initiating impact ionization, thus $k \gg 1$. A k value close to unity is not desirable because it limits the avalanche gain due to a large excess noise.

The total current gain, $G = i_s/i_{ph}$ as defined in (14.23), of an APD is the *avalanche multiplication factor* of photogenerated carriers. It depends on the thickness and the structure of the avalanche region in the APD, as well as on the reverse voltage applied to the APD. For an APD that has a uniform field across its avalanche multiplication region of thickness d_m, the field-dependent parameters α_e, α_h, and k have spatially independent, constant values over the thickness d_m. In this ideal situation, the avalanche multiplication gain for electron or hole injection into the avalanche region can be

expressed as

$$G = \frac{1-k}{e^{-(1-k)\alpha_e d_m} - k} = \frac{1-1/k}{e^{-(1-1/k)\alpha_h d_m} - 1/k}. \tag{14.105}$$

When $k = 1$,

$$G = \frac{1}{1-\alpha_e d_m} = \frac{1}{1-\alpha_h d_m}. \tag{14.106}$$

We see from the above two relations that the multiplication gain G increases nonlinearly with an increase in the value of $\alpha_e d_m$, with a corresponding increase in that of $\alpha_h d_m$, for any given value of k. At a certain value of $\alpha_e d_m$ and its corresponding value of $\alpha_h d_m$ for a given k, however, G increases quickly to approach infinity (see Problem 14.6.2). The consequence is an instability leading to avalanche breakdown.

In practice, the gain of an APD is often expressed empirically as

$$G = \frac{1}{1-(V_r/V_{br})^n}, \tag{14.107}$$

where V_r is the reverse voltage on the APD, V_{br} is the avalanche breakdown voltage, and n is an empirically fitted parameter typically in the range of 3–6. The values of V_{br} and n depend strongly on the device structure and operating temperature. The gain of an APD is very sensitive to both reverse bias voltage and temperature. Voltage and temperature stabilization is often required for the operation of an APD at a constant gain. In normal operation, an APD is biased at a fixed voltage below, but close to, the breakdown voltage. Typical gains range from 10 to 20 for Ge and InGaAs APDs, and from 50 to 200 for Si APDs. Because of the internal gain, the responsivity of an APD is $\mathcal{R} = G\mathcal{R}_0$, where $\mathcal{R}_0$ is the intrinsic responsivity of an equivalent photodiode without an internal gain.

EXAMPLE 14.14 A superlattice InGaAs/InP APD, which is described in further detail in Example 14.16, has an avalanche multiplication region that consists of an InAlGaAs/InAlAs superlattice layer of $d_m = 231$ nm. It has an ionization ratio of $k = 0.25$. When an average electric field of $E_m = 63$ MV m^{-1} is established by a reverse bias voltage in this avalanche multiplication layer, the electron ionization coefficient is $\alpha_e = 6.5 \times 10^6$ m^{-1}. (a) Find the avalanche multiplication gain in this condition. (b) If the device has a breakdown voltage of $V_{br} = 20$ V, what is the reverse bias voltage?

Solution (a) With the given parameters, we have $\alpha_e d_m = 6.5 \times 10^6 \times 231 \times 10^{-9} = 1.5$. The multiplication gain

$$G = \frac{1-k}{e^{-(1-k)\alpha_e d_m} - k} = \frac{1-0.25}{e^{-(1-0.25)\times 1.5} - 0.25} = 10.$$

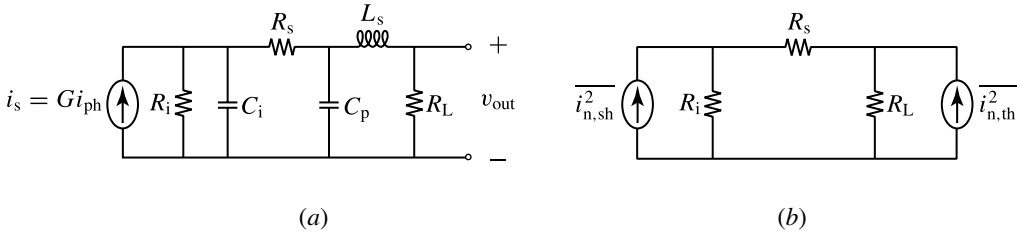


Figure 14.27 (*a*) Small-signal equivalent circuit and (*b*) noise equivalent circuit of an APD.

(b) We can estimate the reverse bias voltage by using (14.107). With $V_{br} = 20$ V and $G = 10$, we have

$$V_r = \left(1 - \frac{1}{G}\right)^{1/n} V_{br} = 0.9^{1/n} \times 20 \text{ V}.$$

Because we do not have the information on the parameter n, we can only calculate the limits of the reverse bias voltage to be $19.31 \text{ V} \leq V_r \leq 19.65 \text{ V}$ by assuming that $3 \leq n \leq 6$. Thus, the bias voltage is below, but very close to, the breakdown voltage. Because V_r is very close to V_{br}, the multiplication gain is very sensitive to the reverse bias voltage. For example, if we take $n = 3$ but use $V_r = 19.65$ V, which is obtained for $n = 6$, we find a gain of $G = 19.4$ instead of 10. This example shows that stabilization of both voltage and temperature is very important for an APD to function at a constant gain as both V_{br} and n vary sensitively with temperature.

The small-signal equivalent circuit of an APD is shown in Fig. 14.27(*a*). It is similar to that of an ordinary junction photodiode, except that the avalanche multiplication gain is included in the signal current $i_s = Gi_{ph}$ for an APD. Figure 14.27(*b*) shows the noise equivalent circuit of an APD.

The shot noise of an APD has the form given in (14.25) for a photodetector that has an internal gain. All APDs generate excess noise because of the statistical nature of the avalanche multiplication process. The excess noise factor F for an APD is a function of the avalanche multiplication gain G and the ionization ratio k. For conventional APDs, the excess noise factor for avalanche multiplication initiated by electrons can be expressed as

$$F = F_e = kG + (1 - k)\left(2 - \frac{1}{G}\right), \tag{14.108}$$

and that for avalanche multiplication initiated by holes can be expressed as

$$F = F_h = \frac{G}{k} + \left(1 - \frac{1}{k}\right)\left(2 - \frac{1}{G}\right). \tag{14.109}$$

The excess noise of an APD is minimized if $k < 1$ when only electrons contribute to avalanche multiplication, or if $k > 1$ when only holes contribute to avalanche multiplication. The theoretical minimum of the excess noise factor for an APD is $F = 2 - 1/G$ for $k = 0$ in (14.108) or $k = \infty$ in (14.109). From (14.108) and (14.109), we see that it is important to have the correct type of carriers injected into the avalanche regions in order to minimize the excess noise because injection of the wrong type of carriers will lead to a very large value of F. For this reason, an avalanche region consisting of a material with $k < 1$ is placed on the n^+ side opposite to an absorption layer on the p^+ side so that electrons are injected into the avalanche region in reverse bias, whereas an avalanche region consisting of a material with $k > 1$ is placed on the p^+ side opposite to an absorption layer on the n^+ side so that holes are injected into the avalanche region in reverse bias. This point can be clearly seen in the two structures shown later in Fig. 14.29.

In practice, the excess noise factor is often expressed with the following empirically fitted formula:

$$F = G^x, \tag{14.110}$$

where x is a parameter typically in the range of 0.2–1 obtained from fitting experimental data. Including the thermal noise, which does not get amplified, the total current noise of an APD is

$$\overline{i_n^2} = 2eBGF(\overline{i_s} + \overline{i_b} + \overline{i_d}) + \frac{4k_B T B}{R_L}, \tag{14.111}$$

which has the same form as that of the PMT. The SNR of an APD has the form given in (14.33) for a photodetector that has an internal gain.

The excess noise degrades the SNR of an APD when compared with an ordinary photodiode of the same quantum efficiency. Therefore, the use of an APD instead of an ordinary photodiode such as a p–i–n photodiode makes sense only when amplifiers are needed in the use of an ordinary photodiode for the detection of low-power optical signals. Because of the noise from the amplifiers, an APD can have a better SNR than a photodiode–amplifier combination to justify the use of the APD. This situation occurs when detecting high-frequency signals at very low power levels because the amplifier noise dominates the detector noise at high frequencies. The NEP for Si APDs can be as low as 1 pW.

EXAMPLE 14.15 Find the excess noise factor for the APD considered in Example 14.14 if electrons are injected into its avalanche region to initiate the avalanche multiplication process. What is its excess noise factor if holes are injected instead?

Solution We have $k = 0.25$ and $G = 10$ from Example 14.15. If electrons are injected, the excess noise factor is found using (14.108) to be

$$F = F_e = 0.25 \times 10 + (1 - 0.25) \times \left(2 - \frac{1}{10}\right) = 3.925.$$

If holes are injected instead, we have to use (14.109) to find that

$$F = F_h = \frac{10}{0.25} + \left(1 - \frac{1}{0.25}\right) \times \left(2 - \frac{1}{10}\right) = 34.3.$$

We see that F_h is about nine times F_e. Clearly, the avalanche multiplication process in this device of $k < 1$ has to be initiated by electrons, not by holes, in order to minimize the excess noise. If holes are injected instead, the excess noise factor would be enhanced by as much as nine times, resulting in a significant increase in the APD noise.

Like any photodiode, the response time of an APD is determined by both the response time of its signal current and the time constant of its equivalent circuit. The speed of an APD is determined by four factors: (1) the transit time τ_{tr} through the absorption layer of a thickness d_a, (2) the diffusion time in the diffusion regions, (3) the avalanche buildup time τ_{av} in the avalanche multiplication layer of a thickness d_m, and (4) the circuit response time limited by the RC time constant τ_{RC}. The avalanche buildup time is unique to APDs. The other three factors are common to all photodiodes, but the transit time in an APD is different from that in an ordinary photodiode. The absorption layer of an APD is equivalent to the intrinsic region of a p–i–n photodiode. It is either intrinsic or very lightly doped and is depleted to maintain a sufficiently high field in this region for a short carrier transit time. The avalanche multiplication layer requires an even higher field. Thus, it is also either intrinsic or very lightly doped. Besides, it is much thinner than the absorption layer: $d_m \ll d_a$. Because the field strengths in these two regions are different, and their material compositions can also be different in heterostructure APDs, the carrier velocities in these two regions can be different even when they are all close to or at their respective saturation values.

A detailed analysis of the time response of an APD is very complicated because it has to take into account the spatial variations of the field strength and the carrier distribution in each region, as well as the spatial variations in α_e, α_h, and k in the avalanche region. However, by taking these parameters to be constants of their respectively spatially averaged values, a simplified analysis yields results that are very good approximations to accurate values.

In an APD where electron multiplication dominates the avalanche process, an electron generated on the p^+ side of the absorption layer can generate a secondary electron–hole pair in the avalanche region located on the n^+ side of the absorption layer after taking a time of τ_{tr}^e to drift through the absorption layer. The secondary hole then takes a time

of τ_{tr}^{h} to drift back to the p^+ side where it is collected. Because the drift of a secondary hole follows the drift of its primary electron, the transit time in an APD is twice as long as that in an ordinary photodiode of the same intrinsic absorption-layer thickness:

$$\tau_{tr} = \tau_{tr}^{e} + \tau_{tr}^{h} = \frac{d_a}{v_e^a} + \frac{d_a}{v_h^a}, \tag{14.112}$$

where v_e^a and v_h^a are, respectively, the electron and hole drift velocities in the absorption region. The same transit time is obtained in an APD where hole multiplication dominates the avalanche process.

In the avalanche region, the multiplication process is not instantaneous but takes time to build up. The avalanche buildup time is a function of the gain, the ionization ratio, and the thickness of the avalanche region. For an avalanche process initiated by electrons with $k < 1$, the avalanche buildup time can be approximated as

$$\tau_{av} \approx Gk\frac{d_m}{v_e^m} + \frac{d_m}{v_h^m}, \tag{14.113}$$

where v_e^m and v_h^m are, respectively, the electron and hole drift velocities in the avalanche multiplication region. For an avalanche process initiated by holes with $k > 1$,

$$\tau_{av} \approx \frac{G}{k}\frac{d_m}{v_h^m} + \frac{d_m}{v_e^m}. \tag{14.114}$$

When the diffusion time of carriers in the diffusion regions is minimized, the intrinsic time constant for the signal current in an APD is the sum of the transit time and the avalanche buildup time:

$$\tau = \tau_{tr} + \tau_{av}. \tag{14.115}$$

The signal-current frequency response of an APD has the form of (14.93) but with the time constant τ given in (14.115):

$$\mathcal{R}_s^2(f) = \left|\frac{i_s(f)}{P_s(f)}\right|^2 \approx \mathcal{R}_s^2(0)\left(\frac{\sin \pi f\tau}{\pi f\tau}\right)^2, \tag{14.116}$$

which has a 3-dB cutoff frequency

$$f_{3dB}^{s} \approx \frac{0.443}{\tau}. \tag{14.117}$$

The circuit response of an APD is similar to that of an ordinary photodiode, with a 3-dB cutoff frequency

$$f_{3dB}^{ckt} \approx \frac{1}{2\pi\tau_{RC}}, \tag{14.118}$$

where τ_{RC} is the RC time constant of the APD equivalent circuit. Therefore, the total

frequency response of an APD can be expressed as

$$\mathcal{R}^2(f) = \mathcal{R}_s^2(f)\mathcal{R}_{ckt}^2(f) = \frac{\mathcal{R}^2(0)}{1 + 4\pi^2 f^2 \tau_{RC}^2} \left(\frac{\sin \pi f \tau}{\pi f \tau} \right)^2. \tag{14.119}$$

The 3-dB cutoff frequency of an APD can be approximated by a relation similar to that given in (14.99):

$$f_{3dB} \approx \frac{0.443}{[\tau^2 + (2.78\tau_{RC})^2]^{1/2}} = \frac{1}{2\pi[\tau_{RC}^2 + (0.36\tau)^2]^{1/2}}. \tag{14.120}$$

An important figure of merit for an APD is the gain–bandwidth product Gf_{3dB}.

There are two modes of operation for an APD. In the normal mode of operation discussed above, the bias voltage is set at a fixed value just below the breakdown voltage. As can be seen from (14.107), the device has a fixed gain at a given operating temperature for $V_r < V_{br}$. In the photon-counting mode of operation, the reverse bias voltage is set above the breakdown voltage. In this situation, a single photon can trigger a constant flow of photocurrent because $G \to \infty$ for $V_r > V_{br}$, according to (14.107). The operation of an APD in this mode is controlled by an external circuit to quench the breakdown current by reducing the voltage on the APD to below the breakdown voltage after a photon triggers the breakdown. The APD is then ready to respond to the next incoming photon. In this mode of operation, an APD is capable of counting single photons, like a PMT. Its response speed, or time resolution, in counting successive photons is determined by the speed of the external circuit. With a passive current-quenching circuit that consists of current-limiting resistors, the time resolution is on the order of a few nanoseconds, limited by the RC time constant of the circuit. With an active current-quenching circuit consisting of a current-switching transistor, the time resolution can be as high as 20 ps, limited by the switching speed of the transistor in the circuit.

There are many different structures developed for APDs. In principle, a p–n or p–i–n diode biased near its breakdown voltage can have an avalanche multiplication gain, thus functioning as an APD. In practice, however, the structure of an APD is designed to optimize both the quantum efficiency and the avalanche multiplication gain of the device. To maximize quantum efficiency, the absorption region for photogeneration of carriers has to be relatively thick. To optimize avalanche multiplication, two conditions are required: (1) the avalanche region has to be relatively thin in order to support a very high field without local breakdown, and (2) it is best to have a single type of carrier injected into the avalanche region rather than have both electrons and holes photogenerated throughout the region. An ordinary p–n or p–i–n structure is not ideal for an APD because both photogeneration and avalanche multiplication of carriers take place in its depletion layer. Some Ge APDs have n^+–p, n^+–n–p, or p^+–n structures, which are acceptable but not optimum.

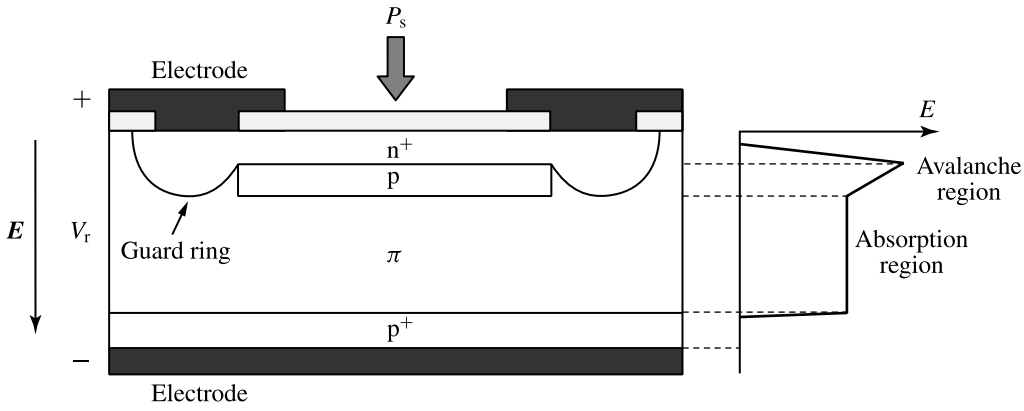


Figure 14.28 Structure and field distribution of a reach-through Si APD, for which $k < 1$.

Separate absorption and multiplication APD

A concept for optimizing both photogeneration and avalanche multiplication in an APD is to use a *separate absorption and multiplication* (SAM) structure, which has separate regions for the two functions. In such a structure, photogeneration takes place in a relatively thick region of a moderately high field to reduce the carrier transit time, whereas the ionizing carriers are injected into a thin region of a very high field for avalanche multiplication. Figure 14.28 shows the structure and the field distribution in reverse bias of a Si SAM APD consisting of p^+–π–p–n^+ layers. This structure is called the *reach-through* structure because the depletion layer under a large reverse bias voltage in the operating condition of the device reaches through the π and p regions from the p^+ region to the n^+ region. For optimum performance of a Si APD, electron injection into the avalanche region is required because $k \ll 1$ in Si. In the reach-through structure shown in Fig. 14.28, photons are absorbed to generate electron–hole pairs mainly in the thick π region. The photogenerated electrons, which are minority carriers in the π region, are accelerated and injected into the thin p–n^+ junction where avalanche multiplication takes place in the presence of a high electric field. The photogenerated holes in the π region are collected in the p^+ region without multiplication because of the low field in that region. To reduce the noise caused by the leakage current at the edges of the p–n^+ junction and to avoid local breakdown at these edges, a guard ring around the edges is often incorporated into a reach-through Si APD, as also shown in Fig. 14.28.

Figure 14.29(*a*) shows the structure and the field distribution in reverse bias of a heterojunction InGaAs/InP SAM APD. Because $k > 1$ in InP, hole injection, rather than electron injection, into the avalanche region for multiplication is desired in this device. Therefore, it is the n^--InP layer that is placed on the p^+ side next to the p^+-InP layer. The absorption region in this P^+–N^-–ν–n^+ heterostructure is the InGaAs ν region, which has a smaller bandgap than the InP layers. Holes that are photogenerated in this

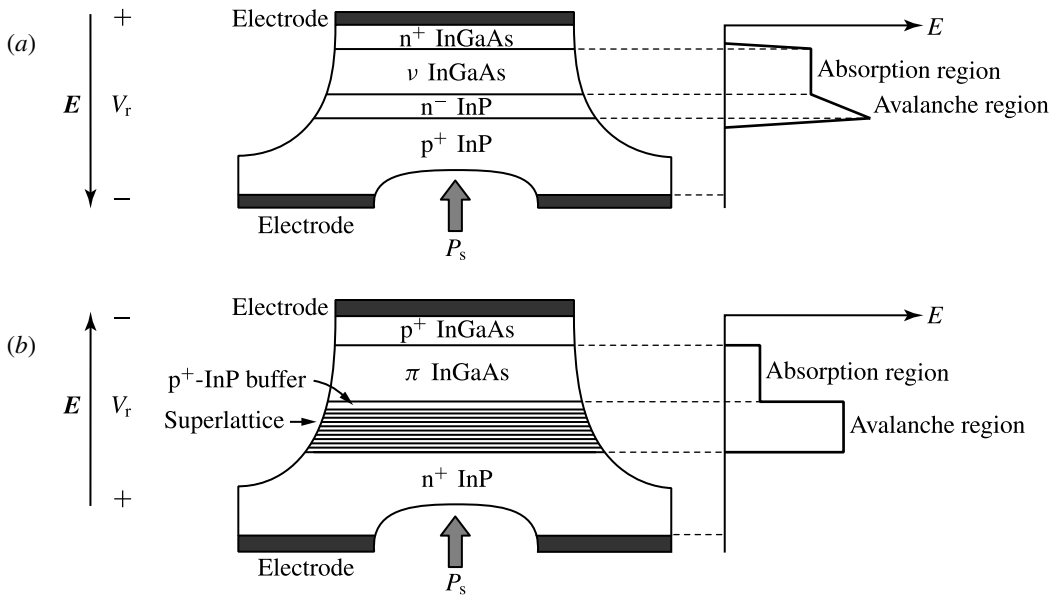


Figure 14.29 Structures and field distributions of (*a*) a heterojunction InGaAs/InP SAM APD with $k > 1$ in the InP avalanche multiplication region and (*b*) a superlattice InGaAs/InP SAM APD with $k < 1$ in the superlattice avalanche multiplication region.

region are injected into the avalanche region at the InP p^+–n^- junction for avalanche multiplication. Photogenerated electrons are collected in the InGaAs n^+ region without multiplication.

Figure 14.29(*b*) shows the structure and the field distribution in reverse bias of a superlattice InGaAs/InP SAM APD. In this structure, the avalanche region consists of either an InGaAsP/InAlAs superlattice or an InAlGaAs/InAlAs superlattice that is lattice matched to InP. Because these superlattice materials have $k < 1$, electron injection, rather than hole injection, is desired. Consequently, this superlattice multiplication layer is placed on the n^+ side of the structure, and the InGaAs absorption layer is now on the p^+ side. There is also a thin p^+-InP buffer layer in this structure. This heavily doped buffer layer allows a sharp transition from a very high field strength in the avalanche region to a lower field in the absorption region so that relatively constant, but very different, field strengths can be maintained in both regions. Its purpose is to suppress undesirable tunneling dark current generation and avalanche multiplication in the absorption layer.

EXAMPLE 14.16 A superlattice InGaAs/InP SAM APD designed for optical detection in the infrared spectral range covering 1.3 and 1.55 μm wavelengths has the structure shown in Fig. 14.29(*b*). It consists of a nearly intrinsic π-InGaAs absorption layer of $d_a = 1$ μm, an undoped InAlGaAs/InAlAs superlattice multiplication layer of $d_m = 231$ nm, and a heavily doped p^+-InP buffer layer of a very small thickness

of 30–50 nm between these two layers. The absorption coefficients of the InGaAs absorption layer at 1.3 and 1.55 μm wavelengths are $\alpha = 1.2 \times 10^6$ m^{-1} and $\alpha = 6.6 \times 10^5$ m^{-1}, respectively. In the normal operating condition of the APD, the electron and hole drift velocities are $v_e^a = 8 \times 10^4$ m s^{-1} and $v_h^a = 6 \times 10^4$ m s^{-1} in the InGaAs absorption layer and $v_e^m = 4.2 \times 10^4$ m s^{-1} and $v_h^m = 3.2 \times 10^4$ m s^{-1} in the InAlGaAs/InAlAs superlattice multiplication layer. The impact ionization ratio is $k = 0.25$. The active area of this APD has a diameter of $2r = 40$ μm. It has a total capacitance, including its internal capacitance and parasitic capacitance, of $C = 300$ fF and a parasitic series resistance of $R_s = 10\ \Omega$. Find the 3-dB cutoff frequency and the gain–bandwidth product of this APD when it operates at a multiplication gain of $G = 10$ with a load resistance of $R_L = 50\ \Omega$.

Solution With $d_a = 1$ μm, the transit time in the absorption layer is

$$\tau_{tr} = \frac{d_a}{v_e^a} + \frac{d_a}{v_h^a} = \left(\frac{1 \times 10^{-6}}{8 \times 10^4} + \frac{1 \times 10^{-6}}{6 \times 10^4}\right) \text{ s} = 29 \text{ ps}.$$

The avalanche multiplication in this APD is initiated by electrons. With $d_m = 231$ nm, $k = 0.25$, and $G = 10$, the avalanche buildup time in the multiplication layer is

$$\tau_{av} \approx Gk\frac{d_m}{v_e^m} + \frac{d_m}{v_h^m} = \left(10 \times 0.25 \times \frac{231 \times 10^{-9}}{4.2 \times 10^4} + \frac{231 \times 10^{-9}}{3.2 \times 10^4}\right) \text{ s} = 21 \text{ ps}.$$

We find that τ_{tr} is comparable to but somewhat larger than τ_{av} for this APD in the given operating condition. Thus, the intrinsic time constant

$$\tau = \tau_{tr} + \tau_{av} = 50 \text{ ps}.$$

The RC time constant

$$\tau_{RC} = (R_s + R_L)C = (10 + 50) \times 300 \times 10^{-15} \text{ s} = 18 \text{ ps}.$$

We find that $2.78\tau_{RC} = 50$ ps, which is the same as τ. Thus the bandwidth of this APD in the given operating condition is equally determined by both its intrinsic time constant and its RC time constant. We have

$$f_{3dB} = \frac{0.443}{[\tau^2 + (2.78\tau_{RC})^2]^{1/2}} = \frac{0.443}{[50^2 + 50^2]^{1/2} \times 10^{-12}} \text{ Hz} = 6.26 \text{ GHz}.$$

Therefore, with $G = 10$, the gain–bandwidth product

$$Gf_{3dB} = 62.6 \text{ GHz}.$$

Graded-gap staircase APD

Sophisticated heterostructures, including those using quantum wells and graded-gap layers, have been developed to improve the performance characteristics of APDs. Figures 14.30(*a*) and (*b*) show, respectively, the unbiased and biased band diagrams of a

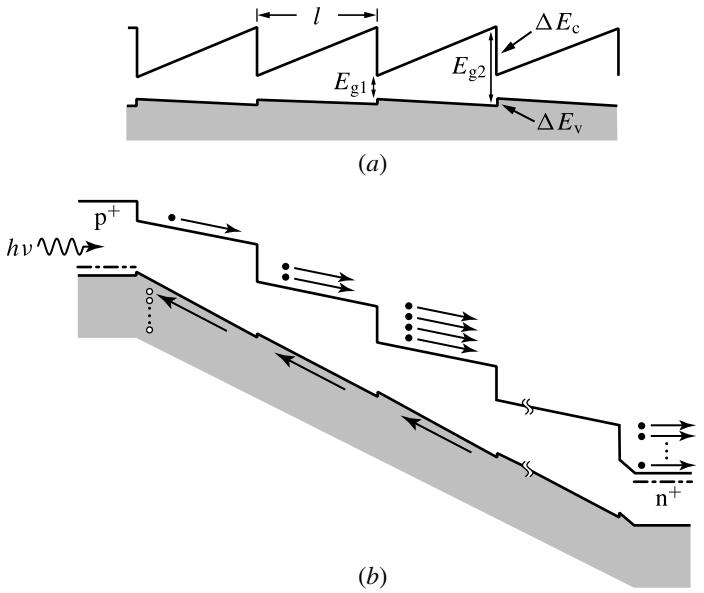


Figure 14.30 Band diagrams of a graded-gap staircase APD (*a*) in thermal equilibrium with no bias and (*b*) in reverse bias with a proper bias voltage.

staircase APD, which consists of multiple nearly intrinsic, or lightly doped, graded-gap layers between the p^+ and n^+ regions. This structure is the solid-state equivalent of the PMT, with each graded-gap layer functioning as an electron multiplication stage equivalent to a dynode in a PMT. The graded gap in this structure is made by varying the composition of a semiconductor material, such as the composition x in $Al_xGa_{1-x}As$. The bandgap in each layer increases linearly from a small value of E_{g1} to a large value of E_{g2} with an abrupt drop back to E_{g1} at the end of the layer. For typical III–V semiconductors, most of the heterostructure bandgap difference occurs in the conduction band. In reverse bias, the voltage applied to the device drops almost entirely across the nearly intrinsic multilayer graded-gap region. At the operating bias voltage of the device, the energy band has a pattern like that shown in Fig. 14.30(*b*). Photogenerated electrons in the p^+ region are injected into successive stages of alternating low-field graded-gap regions and high-field conduction-band steps. The energy drop at each conduction-band step is larger than the threshold impact-ionization energy for electrons. The electrons drift through a low-field region without multiplication, but they impact ionize when passing through an abrupt conduction-band step. Holes do not contribute to avalanche multiplication but are quickly swept away because the moderately high field in the valence band is not large enough to cause impact ionization by holes. The value of the ionization ratio k for this structure is thus substantially reduced in comparison to a conventional structure of the same material.

The excess noise factor of a staircase APD is also much reduced due to the fact that impact ionization in this device is localized at the conduction-band steps. At each

potential step, each electron acquires only enough energy to generate one secondary electron–hole pair. The only excess noise comes from the probability that an electron, though having enough energy, may or may not impact ionize at a given potential step. As a result, a staircase APD typically has a small excess noise factor close to unity similar to that of a PMT. Because this device has a very small k value and a small excess noise factor, it has improved performance characteristics in terms of optimized gain, reduced noise, and increased speed.

14.7 Guided-wave photodetectors

The photodetectors discussed in the preceding sections are vertically illuminated. In a vertically illuminated photodetector (VIPD), the optical signal propagates in a direction perpendicular to the junction interfaces of the device. This situation leads to a trade-off between the carrier transit time and the quantum efficiency, resulting in a limitation on the bandwidth–efficiency product of the device. Another limitation of a high-speed VIPD arises from the trade-off between its bandwidth and its saturation power. A large bandwidth for a VIPD requires a small absorption volume, which results in a high carrier concentration at a given power level for the optical signal. The space-charge effect in the active region caused by the high carrier concentration sets a limit on the saturation power of the photodetector. Although the bandwidth–efficiency product of a VIPD can be improved by using a multipass structure as discussed in Section 14.5, the saturation power is not increased by such a strategy but can only be improved by increasing the effective absorption volume. Guided-wave photodetectors are developed to overcome these limitations. A well-designed guided-wave photodetector can have both a large bandwidth–efficiency product and a high saturation power.

Most of the photodetectors, including the MSM photodetectors, the p–i–n photodiodes, the Schottky photodiodes, and the APDs, that are discussed in preceding sections can be made in guided-wave device form. In a guided-wave photodetector, the guided optical signal propagates in a direction that is parallel to the junction interfaces and is perpendicular to the drift of the photogenerated carriers. This geometry decouples the absorption length of the optical signal from the drift length of the photogenerated carriers. The optical signal is absorbed along the length l of the active region, while the carriers drift across the thickness d of the active region. Thus, the quantum efficiency of a guided-wave photodiode is not that given by (14.84) but can be expressed as

$$\eta_e = \eta_{coll}\eta_t\eta_i = \eta_{coll}(1 - R)\eta_c(1 - e^{-\alpha_{eff}l}), \tag{14.121}$$

where η_{coll} is the collection efficiency of the photogenerated carriers, $\eta_t = (1 - R)\eta_c$, and $\eta_i = 1 - e^{\alpha_{eff}l}$. Here R is the reflectivity at the incident surface of the waveguide, η_c is the coupling efficiency of the optical signal into the waveguide, α_{eff} is the effective

absorption coefficient of the active region, and l is the length of the active region measured along the waveguide direction. The effective absorption coefficient α_{eff} has to be calculated according to the device structure in order to take into account the fact that only a fraction of the guided optical wave overlaps with the absorbing active region of a thickness d. In the case when the entire core of the waveguide is the active region, $\alpha_{\text{eff}} = \Gamma\alpha$, where Γ is the confinement factor of the waveguide and α is the absorption coefficient of the material in the active region. In other device structures, $\alpha_{\text{eff}} \neq \Gamma\alpha$ because the active region might be located outside the waveguide core or it might occupy only a fraction of the core. Because the signal absorption length is no longer tied to the thickness of the active region, a guided-wave photodetector can have both a very thin active region for a very short transit time and a long absorption length for a high quantum efficiency and a high saturation power.

The major advantage of a guided-wave photodetector over a VIPD is that its carrier transit time can be independently reduced without sacrificing its quantum efficiency and saturation power. The transit-time-limited bandwidth of a guided-wave photodetector is easily made larger than its RC-time-limited bandwidth by using a sufficiently thin active region. As a result, the primary bandwidth limitation is not the carrier transit time but is the RC time constant of the device. Besides this major advantage, a guided-wave photodetector with a thin active region has a few additional advantages. A thin active region allows the device to operate at a low bias voltage, resulting in a small dark current and reduced noise contributed by the dark current. The waveguide geometry is also compatible with other guided-wave photonic devices and components, making it easy to incorporate photodetectors into an integrated photonic circuit with reduced input and output coupling losses.

From the standpoint of considering the RC-time-limited bandwidth, guided-wave photodetectors can be classified into two major categories: (1) lumped-circuit devices, commonly called *waveguide photodetectors*; and (2) distributed-circuit devices, commonly called *traveling-wave photodetectors*.

Waveguide photodetectors

A waveguide photodetector (WGPD) differs from a VIPD mainly in its optical waveguide structure. Being a lumped-circuit device, its electrical structure and equivalent circuit are similar to those of a VIPD. The major advantage of a WGPD over a VIPD is that it can maintain a high quantum efficiency for a high cutoff frequency, thus a large bandwidth–efficiency product. The saturation power of a WGPD is comparable to that of a VIPD.

A WGPD is formed by integrating the active region of a photodetector with an optical waveguide. There are two basic integration schemes: the *butt-coupling* configuration and the *evanescent-coupling* configuration, which are illustrated in Fig. 14.31. The optical structure of a WGPD belongs to one or other of these two integration schemes though

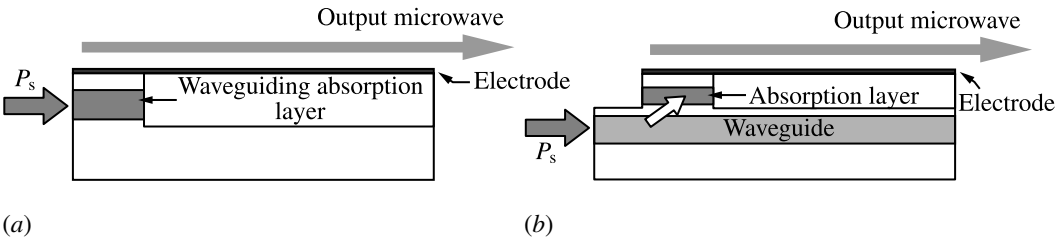


Figure 14.31 Schematic structures of waveguide photodetectors in (*a*) a butt-coupling configuration and (*b*) an evanescent-coupling configuration. (Adapted from Kato, K., "Ultrawide-band/high-frequency photodetectors," *IEEE Transactions on Microwave Theory and Techniques* **47**(7): 1265–1281, July 1999.)

its details and sophistication may vary from one device to another. In the butt-coupling scheme, also called the *end-coupling* or *end-firing* scheme, shown in Fig. 14.31(*a*), the active photoabsorption region of the photodetector is located in the waveguide core or is directly aligned with the core of a feeding waveguide. In this coupling scheme, the coupling efficiency from the feeding waveguide, if one is used, to the active region of the photodetector can be as high as 100% in principle but the coupling efficiency, η_c in (14.121), from free space to the waveguide core can be small if the waveguide core is defined by the thickness of a thin active region. One approach to improving the coupling efficiency η_c, but at the expense of reducing the effective absorption coefficient α_{eff} to a fraction of $\Gamma\alpha$, is to use a large-core waveguide with the thin active region occupying only a fraction of the waveguide core. In the evanescent-coupling scheme, shown in Fig. 14.31(*b*), the entire waveguide is nonabsorbing at the optical signal wavelength because the active photoabsorption region of the photodetector is located outside the waveguide core, typically on top of the waveguide. In this scheme, a large-core waveguide is normally used to maximize the optical coupling efficiency η_c to the waveguide. The effective absorption coefficient α_{eff} is also only a fraction of $\Gamma\alpha$ because of a small overlap factor between the optical field and the active region in this evanescent-coupling configuration. The overall quantum efficiency of a WGPD can be improved by using a large-core waveguide in either scheme to maximize η_c because the reduction in α_{eff} can be compensated by increasing the length l of the active region.

Because a WGPD is a lumped-circuit device, its equivalent circuit and RC-time-limited bandwidth have the same form as those of a corresponding VIPD discussed in preceding sections. Although the transit-time-limited bandwidth is independent of the quantum efficiency and can be made large enough that it is not a limiting factor, there is still a trade-off in maximizing both the RC-time-limited bandwidth and the quantum efficiency. As the active region is made thin to shorten the carrier transit time and is made long to increase the quantum efficiency, the junction capacitance increases. Any parasitic capacitance that exists tends to increase also. A high-speed photodetector normally has a fixed load resistance of $R_L = 50\ \Omega$. Therefore, increasing the transit-time-limited bandwidth and the quantum efficiency by making the active region thin and long results

in a reduction in the RC-time-limited bandwidth. To reduce the capacitance, the area of the active region can be reduced to a minimum required by a given quantum efficiency. However, a small area for the active region leads to a large series resistance R_s. When R_s becomes larger than R_L, there is a trade-off between the capacitance and the resistance in maximizing the RC-time-limited bandwidth. Consequently, the bandwidth–efficiency product of a WGPD is typically in the range of 20–40 GHz, which is comparable to that of a multipass VIPD, though a bandwidth–efficiency product larger than 50 GHz with a bandwidth larger than 100 GHz is possible for a well-designed WGPD.

The saturation power of a WGPD is similar to that of a VIPD of a comparable absorption volume. In a WGPD, there is a trade-off between the saturation power and the bandwidth. The saturation power can be increased by increasing the thickness and the length of the active region while reducing α_{eff} so that the absorption of the optical signal is distributed over a large volume, but the bandwidth will be reduced by such an action. A WGPD has no advantage in the saturation power if its absorption volume is limited by the consideration of reducing both the carrier transit time and the device capacitance for a large bandwidth.

EXAMPLE 14.17 High-speed InP/InGaAsP/InGaAs/InGaAsP/InP p–i–n VIPD and WGPD of the same device parameters are compared in this example. Both are used for the detection of optical signals at $\lambda = 1.55$ μm. The intrinsic InGaAs active region has a thickness of $d_i = 0.2$ μm, which is sandwiched between two InGaAsP layers that form a double-core waveguide. The absorption coefficient for InGaAs at $\lambda = 1.55$ μm is $\alpha = 6.6 \times 10^5$ m^{-1}. The electron and hole saturation velocities are $v_e^{\text{sat}} = 8 \times 10^4$ m s^{-1} and $v_h^{\text{sat}} = 6 \times 10^4$ m s^{-1}, respectively. The device area $\mathcal{A} = 50$ μm^2, which can take any shape for the VIPD but is formed by a stripe of $w = 2$ μm and $l = 25$ μm for the WGPD. With these dimensions, both devices have a capacitance $C = 30$ fF, a series resistance from the contacts and the materials of $R_s = 40$ Ω. The load resistance is $R_L = 50$ Ω. For the WGPD, the confinement factor of the active region is $\Gamma = 15\%$, and the optical coupling efficiency $\eta_c = 70\%$. Assume that both devices have $\eta_{\text{coll}} = \eta_t = 1$. (a) Find the 3-dB cutoff frequencies of both devices. (b) Find the bandwidth–efficiency product of the VIPD in a single-pass configuration for the optical signal through the active region. (c) Find the bandwidth–efficiency product of the VIPD in a double-pass configuration with 100% back reflection. (d) Find the bandwidth–efficiency product of the WGPD.

Solution (a) The VIPD and the WGPD have the same 3-dB cutoff frequency because they have identical dimensions and device parameters. We first find that

$$v_{\text{sat}} = \left[\frac{1}{2}\left(\frac{1}{v_e^{\text{sat}}} + \frac{1}{v_h^{\text{sat}}}\right)\right]^{-1} = \left[\frac{1}{2}\left(\frac{1}{8 \times 10^4} + \frac{1}{6 \times 10^4}\right)\right]^{-1} \text{ m s}^{-1}$$
$$= 6.86 \times 10^4 \text{ m s}^{-1}.$$

With $d_i = 0.2$ μm, we find that

$$\tau_{tr} = \frac{d_i}{v_{sat}} = \frac{0.2 \times 10^{-6}}{6.86 \times 10^4} \text{ s} = 2.9 \text{ ps}.$$

With $C = 30$ fF, $R_s = 40$ Ω, and $R_L = 50$ Ω, we have

$$\tau_{RC} = (R_s + R_L)C = 90 \times 30 \times 10^{-15} \text{ s} = 2.7 \text{ ps}.$$

Therefore, the 3-dB cutoff frequency

$$f_{3dB} = \frac{0.443}{[(2.9 \times 10^{-12})^2 + (2.78 \times 2.7 \times 10^{-12})^2]^{1/2}} \text{ Hz} = 55 \text{ GHz}.$$

(b) We have $\alpha d_i = 6.6 \times 10^5 \times 0.2 \times 10^{-6} = 0.132$. For the single-pass VIPD with $\eta_{coll} = \eta_t = 1$, we have

$$\eta_e = \eta_i = 1 - e^{-\alpha d_i} = 1 - e^{-0.132} = 12.4\%.$$

Therefore, its bandwidth–efficiency product

$$\eta_e f_{3dB} = 0.124 \times 55 \text{ GHz} = 6.8 \text{ GHz}.$$

(c) For the double-pass VIPD with $\eta_{coll} = \eta_t = 1$ and 100% back reflection, we have

$$\eta_e = \eta_i = 1 - e^{-2\alpha d_i} = 1 - e^{-2\times 0.132} = 23.2\%.$$

Therefore, its bandwidth–efficiency product

$$\eta_e f_{3dB} = 0.232 \times 55 \text{ GHz} = 12.8 \text{ GHz}.$$

(d) For the WGPD, we find that $\alpha_{eff} = \Gamma\alpha = 0.15 \times 6.6 \times 10^5 \text{ m}^{-1} = 9.9 \times 10^4 \text{ m}^{-1}$ and $\alpha_{eff} l = 9.9 \times 10^4 \times 25 \times 10^{-6} = 2.475$. Thus, with $\eta_{coll} = \eta_t = 1$ and $\eta_c = 70\%$, we have

$$\eta_e = \eta_c \eta_i = \eta_c(1 - e^{-\alpha_{eff} l}) = 0.7 \times (1 - e^{-2.475}) = 64.1\%.$$

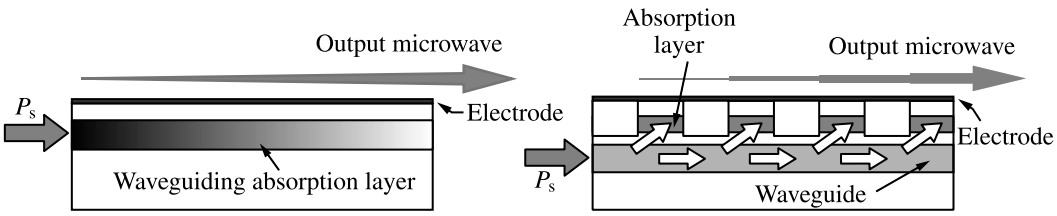
Therefore, the bandwidth–efficiency product of the WGPD is

$$\eta_e f_{3dB} = 0.641 \times 55 \text{ GHz} = 35.3 \text{ GHz}.$$

We find that the bandwidth–efficiency product of the WGPD is 2.75 times that of the double-pass VIPD and is more than five times that of the single-pass VIPD though all of them have the same 3-dB cutoff frequency.

Traveling-wave photodetectors

A traveling-wave photodetector (TWPD) differs from a WGPD mainly in its distributed electrical structure. Its optical structure is similar to that of a WGPD. A TWPD



(*a*) (*b*)

Figure 14.32 Schematic structures of traveling-wave photodetectors in (*a*) a distributed configuration and (*b*) a periodic configuration. (Adapted from Kato, K., "Ultrawide-band/high-frequency photodetectors," *IEEE Transactions on Microwave Theory and Techniques* **47**(7): 1265–1281, July 1999.)

can have a larger bandwidth–efficiency product and a higher saturation power than a WGPD.

A TWPD is formed by integrating an electrical transmission line with one or more guided-wave photodetectors. There are two different basic configurations: the distributed configuration and the periodic configuration, both of which are illustrated in Fig. 14.32. A distributed TWPD, shown in Fig. 14.32(*a*), is a traveling-wave version of a WGPD. It consists of a transmission line built on a fully distributed WGPD. A periodic TWPD, shown in Fig. 14.32(*b*), consists of a set of photodetectors that are periodically located along a transparent optical waveguide and are serially connected by a transmission line. A distributed TWPD is often simply called a TWPD. A periodic TWPD is also called a *velocity-matched distributed photodetector* (VMDP) if it is designed so that the optical wave and the microwave are velocity matched. Both the butt-coupling and evanescent-coupling schemes discussed above for WGPDs can be used for distributed and periodic TWPDs.

For a traveling-wave device, the bandwidth limitation due to the RC time constant of a lumped circuit is replaced by a bandwidth limitation due to the velocity mismatch between the optical wave propagating in the optical waveguide and the microwave propagating in the transmission line. For a traveling-wave electro-optic modulator discussed in Section 6.5, the bandwidth is determined by the mismatch between the *phase* velocities of the optical wave and microwave, v_p^o and v_p^m, respectively. Because electro-optic modulation acts upon the phase of the optical wave, it is necessary to synchronize the wavefronts of the optical wave and microwave by matching their phase velocities. For a TWPD, however, velocity matching is considered between the *group* velocity of the optical wave, v_g^o, and the *phase* velocity of the microwave, v_p^m, rather than between the phase velocities of the two waves. In a TWPD, the microwave electrical signal is generated by absorption of the optical signal energy that propagates at the optical group velocity, but propagation of the microwave is determined by its phase velocity because the electrical signals generated along the transmission line add coherently. The group

velocity of a guided optical wave of frequency ω and propagation constant β is

$$v_{\mathrm{g}}^{\mathrm{o}} = \frac{\mathrm{d}\omega}{\mathrm{d}\beta} = \frac{c}{N_\beta}, \tag{14.122}$$

where N_β is the effective group index of the guided mode. The phase velocity of a microwave electrical signal propagating in a transmission line is

$$v_{\mathrm{p}}^{\mathrm{m}} = \frac{1}{\sqrt{LC}} = \frac{1}{ZC}, \tag{14.123}$$

where L and C are, respectively, the inductance and capacitance per unit length of the transmission line and $Z = \sqrt{L/C}$ is the characteristic impedance of the transmission line. The velocity-mismatch-limited bandwidth, $f_{3\mathrm{dB}}^{\mathrm{VM}}$, of a TWPD is characterized by a time constant τ_{VM}, which measures the temporal walk-off between the optical wave and the microwave at the output of the TWPD:

$$f_{3\mathrm{dB}}^{\mathrm{VM}} = \frac{1}{2\pi\,\tau_{\mathrm{VM}}}. \tag{14.124}$$

Replacing τ_{RC} in (14.99) with τ_{VM}, the 3-dB cutoff frequency of a TWPD including transit-time and velocity-mismatch limitations can be approximated as

$$f_{3\mathrm{dB}} \approx \frac{0.443}{[\tau_{\mathrm{tr}}^2 + (2.78\tau_{\mathrm{VM}})^2]^{1/2}} = \frac{1}{2\pi[\tau_{\mathrm{VM}}^2 + (0.36\tau_{\mathrm{tr}})^2]^{1/2}}. \tag{14.125}$$

As discussed below, the form of the velocity-mismatch time constant τ_{VM} depends on the structure of a TWPD.

In a TWPD, microwaves that propagate in both forward and backward directions in the transmission line are generated by absorption of the optical signal along the waveguide. The velocity mismatch for the forward-propagating microwave is $v_{\mathrm{g}}^{\mathrm{o}} - v_{\mathrm{p}}^{\mathrm{m}}$, but that for the backward-propagating microwave is $v_{\mathrm{g}}^{\mathrm{o}} + v_{\mathrm{p}}^{\mathrm{m}}$. It is clearly not possible to velocity match the optical signal to both forward- and backward-propagating microwaves. The bandwidth and the efficiency of a TWPD depend on whether or not the backward-propagating microwave is allowed to contribute to the electrical output. At the optical input end, the termination of the transmission line can be either (1) connected with a matching impedance to eliminate the reflection of the backward-propagating microwave or (2) left open to allow total reflection of the backward-propagating microwave. The efficiency of a TWPD with an impedance-matched input electrical termination is half that of the same TWPD with an open input electrical termination.

The velocity-mismatch time constant, τ_{VM}, of a TWPD is a function of the velocity mismatch and the effective length, l_{eff}, of the device. Because forward- and backward-propagating microwaves have different velocity mismatches, τ_{VM} has different forms

for TWPDs of different input terminations. For a TWPD with a matched input electrical termination,

$$\tau_{\mathrm{VM}} = \left| \frac{l_{\mathrm{eff}}}{v_{\mathrm{g}}^{\mathrm{o}}} - \frac{l_{\mathrm{eff}}}{v_{\mathrm{p}}^{\mathrm{m}}} \right| = l_{\mathrm{eff}} \left| \frac{v_{\mathrm{g}}^{\mathrm{o}} - v_{\mathrm{p}}^{\mathrm{m}}}{v_{\mathrm{g}}^{\mathrm{o}} v_{\mathrm{p}}^{\mathrm{m}}} \right|. \tag{14.126}$$

For a TWPD with an open input electrical termination,

$$\tau_{\mathrm{VM}} \approx \frac{3}{2} \frac{l_{\mathrm{eff}}}{v_{\mathrm{p}}^{\mathrm{m}}}. \tag{14.127}$$

The velocity-mismatch time constant for a TWPD with an open input electrical termination is independent of the optical group velocity because the mismatches on the forward- and backward-propagating microwaves have opposite effects. The effective length, l_{eff}, of a TWPD depends on the physical length l of the optical waveguide, the distribution of the photoabsorption region in the waveguide, and the effective absorption coefficient α_{eff} of the device. It is the lesser of the physical length and the propagation distance of the optical signal. For a distributed TWPD,

$$l_{\mathrm{eff}} = \begin{cases} l, & \text{if} \quad l < \dfrac{1}{\alpha_{\mathrm{eff}}}, \\ \dfrac{1}{\alpha_{\mathrm{eff}}}, & \text{if} \quad l > \dfrac{1}{\alpha_{\mathrm{eff}}}. \end{cases} \tag{14.128}$$

For a periodic TWPD, the optical waveguide of a total physical length l is only periodically loaded with photodetectors. If the length of each period is l_{p} and the length of the photoabsorption region of each photodetector is l_{d}, then

$$l_{\mathrm{eff}} = \begin{cases} l, & \text{if} \quad l < \dfrac{l_{\mathrm{p}}}{\alpha_{\mathrm{eff}} l_{\mathrm{d}}}, \\ \dfrac{l_{\mathrm{p}}}{\alpha_{\mathrm{eff}} l_{\mathrm{d}}}, & \text{if} \quad l > \dfrac{l_{\mathrm{p}}}{\alpha_{\mathrm{eff}} l_{\mathrm{d}}}. \end{cases} \tag{14.129}$$

Because l_{p} can be much larger than l_{d} in a periodic TWPD, both the physical length and the effective length of a periodic TWPD can be much larger than those of a distributed TWPD.

It can be seen from (14.126) and (14.127) that the velocity-mismatch-limited bandwidth of a TWPD with a matched termination can be unlimitedly improved by velocity matching for $v_{\mathrm{p}}^{\mathrm{m}}/v_{\mathrm{g}}^{\mathrm{o}} = 1$, whereas that of a TWPD with an open termination can be improved by simply increasing the phase velocity $v_{\mathrm{p}}^{\mathrm{m}}$ of the microwave. In principle, a TWPD with a matched termination can be velocity matched to have an arbitrarily large velocity-mismatch-limited bandwidth so that its bandwidth is purely transit-time limited. In reality, however, $v_{\mathrm{p}}^{\mathrm{m}} < v_{\mathrm{g}}^{\mathrm{o}}$ for a transmission line on an optical waveguide loaded with a photodetector, and $v_{\mathrm{p}}^{\mathrm{m}} > v_{\mathrm{g}}^{\mathrm{o}}$ for a transmission line on an unloaded optical waveguide. For a distributed TWPD, the microwave phase velocity is always lower than

the optical group velocity with a ratio v_p^m/v_g^o typically falling in the range of 30–80%. Therefore, perfect velocity matching is not possible in a distributed TWPD but is only possible in a properly designed periodic TWPD.

The major advantage of a TWPD over a WGPD is that its physical length can be made larger than its effective length without degrading its bandwidth because τ_{VM} depends only on l_{eff}. The bandwidth of a TWPD becomes independent of its physical length when it is longer than the absorption length, whereas that of a WGPD continues to decrease as its length increases. To maximize the efficiency, a distributed TWPD is normally made long enough that $l \gg l_{eff}$. This flexibility allows a distributed TWPD to have an efficiency that is about 1.3 times the efficiency of a comparable WGPD designed for the same bandwidth, thus a 30% advantage in the bandwidth–efficiency product for the TWPD. For the same reason, a distributed TWPD can have a higher saturation power than a comparable WGPD, particularly when they are both designed for high-speed operation. On the other hand, when a TWPD and a WGPD are designed to have the same efficiency, the TWPD will have a larger bandwidth, thus a larger bandwidth–efficiency product, than the WGPD.

EXAMPLE 14.18 A distributed TWPD based on the WGPD described in Example 14.17 is made by using a properly designed transmission line for its electrodes. The waveguide has an optical group velocity of $v_g^o = 8.9 \times 10^7$ m s^{-1} at the 1.55 μm signal wavelength. The microwave phase velocity of the transmission line is $v_p^m = 2.9 \times 10^7$ m s^{-1}. Find the cutoff frequency and the bandwidth–efficiency product of the TWPD if (a) it has a matched termination and (b) it has an open termination.

Solution From Example 14.17, we have $\alpha_{eff} = 9.9 \times 10^4$ m^{-1}. Because $l = 25$ μm $> \alpha_{eff}^{-1} = 10.1$ μm, we have $l_{eff} = 10.1$ μm for this distributed TWPD. From Example 14.17, we also have $\tau_{tr} = 2.9$ ps.

(a) With a matched termination, we have

$$\tau_{VM} = \left| \frac{l_{eff}}{v_g^o} - \frac{l_{eff}}{v_p^m} \right| = \left| \frac{10.1 \times 10^{-6}}{8.9 \times 10^7} - \frac{10.1 \times 10^{-6}}{2.9 \times 10^7} \right| \text{ s} = 235 \text{ fs}.$$

Therefore, the 3-dB cutoff frequency

$$f_{3dB} = \frac{0.443}{[(2.9 \times 10^{-12})^2 + (2.78 \times 235 \times 10^{-15})^2]^{1/2}} \text{ Hz} = 149 \text{ GHz}.$$

For the TWPD with a matched termination, the efficiency is only half of the 64.1% efficiency found in Example 14.17. Thus, the bandwidth–efficiency product

$$\eta_e f_{3dB} = \frac{1}{2} \times 0.641 \times 149 \text{ GHz} = 47.8 \text{ GHz}.$$

(b) With an open termination, we have

$$\tau_{VM} = \frac{3l_{eff}}{2v_p^m} = \frac{3 \times 10.1 \times 10^{-6}}{2 \times 2.9 \times 10^7} \text{ s} = 522 \text{ fs}.$$

Therefore, the 3-dB cutoff frequency

$$f_{3dB} = \frac{0.443}{[(2.9 \times 10^{-12})^2 + (2.78 \times 522 \times 10^{-15})^2]^{1/2}} \text{ Hz} = 136.6 \text{ GHz}.$$

For the TWPD with an open termination, the efficiency is just the 64.1% efficiency found in Example 14.17. Thus, the bandwidth–efficiency product

$$\eta_e f_{3dB} = 0.641 \times 136.6 \text{ GHz} = 87.6 \text{ GHz}.$$

Because perfect velocity matching is not achieved in the TWPD with a matched termination, the velocity-mismatch-limited bandwidth of the TWPD with a matched termination is only about twice that of the TWPD with an open termination in this example. However, the cutoff frequencies for both cases are limited by the transit time because $\tau_{tr} \gg \tau_{VM}$ in both cases. As a result, the TWPD with a matched termination only has a slightly larger bandwidth than the TWPD with an open termination. The one with an open termination then has a larger bandwidth–efficiency product because it has twice the efficiency of the one with a matched termination. Compared to the WGPD, which is further limited by the RC time constant, however, the TWPD with either type of termination has a larger bandwidth and a larger bandwidth–efficiency product.

In a periodic TWPD, the transmission line runs along the optical waveguide with periodically alternating loaded regions, where $v_p^m < v_g^o$, and unloaded regions, where $v_p^m > v_g^o$. By properly designing the size and spacing of the periodically distributed photodetectors, it is then possible to achieve close velocity matching in a periodic TWPD. A velocity-matched periodic TWPD is known as a VMDP. The unique advantage of a VMDP is that the transmission line, the optical waveguide, and the individual photodetectors can be independently optimized. The transmission line is optimized for its impedance and for velocity matching. Because close velocity matching is possible, a VMDP always has a matched input electrical termination to realize the benefit of velocity matching. Therefore, its efficiency is limited to a maximum of 50%. The optical waveguide is optimized with the characteristics of large-core, low-coupling loss, and single-mode operation for the VMDP to have a high efficiency and a high saturation power. With close velocity matching, the bandwidth of a VMDP is no longer velocity-mismatch limited but is essentially that of the individual photodetectors. The individual photodetectors in a VMDP are optimized for a large bandwidth. Each individual photodetector is kept below its saturation current. Because velocity matching

permits a large device length without degrading the bandwidth, a high saturation power and a high efficiency can both be achieved with a long device without sacrificing the bandwidth.

EXAMPLE 14.19 What changes have to be made to the structure and dimensions of the TWPD considered in Example 14.18 in order to make it into a VMDP? What are the bandwidth and bandwidth–efficiency product that can be obtained for this VMDP instead?

Solution To make a VMDP, it is necessary to break the 25-μm length of the TWPD into individually optimized photodetectors that are then properly spaced to achieve velocity matching. Thus, if we keep the sum of the lengths of these individual photodetectors to be 25 μm, the entire length of the VMDP will be much longer. If perfect, or nearly perfect, velocity matching is accomplished, we have $\tau_{VM} \approx 0$. Then, the bandwidth of the device is purely transit-time limited:

$$f_{3dB} = \frac{0.443}{2.9 \times 10^{-12}} \text{ Hz} = 152.8 \text{ GHz}.$$

Because a VMDP is required to have a matched termination, its efficiency is only half of the 64.1% efficiency found for the WGPD. Thus, its bandwidth–efficiency product

$$\eta_e f_{3dB} = \frac{1}{2} \times 0.641 \times 152.8 \text{ GHz} = 49 \text{ GHz}.$$

We find that f_{3dB} of this VMDP is only slightly larger than that of the TWPD with a matched termination considered in Example 14.18. As a consequence, its bandwidth–efficiency product is also only slightly larger than that of the TWPD with a matched termination but is smaller than that of the TWPD with an open termination. The reason for this insignificant improvement by velocity matching is that $\tau_{VM} \ll \tau_{tr}$ for both cases of TWPD considered in Example 14.18. Thus, the 3-dB cutoff frequencies of the two cases of TWPD are already close to the transit-time limit. A VMDP can realize a significant bandwidth increase over a TWPD only when the bandwidth of the TWPD is limited by velocity mismatch rather than by the transit time. See Problem 14.7.5 for such an example.

PROBLEMS

14.0.1 Discuss the differences between photon detectors and thermal detectors in their operation principles, characteristics, and applications. How are photon detectors further classified?

14.1.1 What are the major sources of noise for a photodetector? Describe their physical origins and characteristics.

14.1.2 Answer the following questions regarding the general characteristics of noise.
   a. How is the magnitude of the noise in a given signal quantified? How does it relate to the magnitude of the signal?
   b. Why is the magnitude of noise not measured by its average value but by its root mean square value?
   c. What is the magnitude of the total noise from multiple independent noise sources in terms of the magnitudes of its contributing noise sources?

14.1.3 Show that if a variable, such as the photon number $\mathcal{S}$ or the charge number $\mathcal{N}$, is characterized by the Poisson statistics as in (14.15) for $\mathcal{S}$ or (14.17) for $\mathcal{N}$, then its variance is equal to its mean value as in (14.16) for $\mathcal{S}$ or (14.18) for $\mathcal{N}$. Use this relation to verify the relation in (14.21) for a signal current i_s generated by charge carriers that have the Poisson statistics.

14.1.4 Find the SNR of an optical signal for the photon number $\mathcal{S}$ in a time interval T by using (14.9) with $s = \mathcal{S}$. Find the SNR of a photoelectric signal for the number $\mathcal{N}$ of the carriers generated by the photoelectric effect in a time interval T by using (14.9) with $s = \mathcal{N}$. Compare the SNR of the photoelectric signal to the SNR of the optical signal for a quantum efficiency of η_e, which has a value between 0 and 1.

14.1.5 Show that a photodetector with no internal gain operates in the quantum regime if all of its currents combined satisfy the following condition:

$$\overline{i_{ph}} + \overline{i_b} + \overline{i_d} > \frac{2k_B T}{eR} = \frac{T}{300}\frac{51.8\text{ mV}}{R}, \tag{14.130}$$

where T is the temperature in kelvins. At 300 K, what is the minimum photocurrent for a photodetector with a 50 Ω resistance to operate in the quantum regime? Discuss the implications of this relation on how the detectivity of a photodetector can be improved. Find a similar relation for a photodetector with an internal gain.

14.1.6 A photodetector generates a signal photocurrent of 1 mA. Its dark current and background radiation current are both negligibly small compared to this signal current. Find its shot noise, thermal noise, and signal-to-noise ratio if (a) it has a bandwidth $B = 1$ GHz with a load resistance of $R = 50$ Ω and (b) it has a bandwidth $B = 10$ MHz with a load resistance of $R = 1$ kΩ. In each case, find out whether the detector is operating in the quantum or thermal regime. Compare the SNR for the two cases and find the reasons for the difference in the SNR.

14.2.1 Plot the maximum possible intrinsic responsivity $\mathcal{R}_0$ for a photodetector as a function of optical wavelength in the ideal situation of a unity external quantum efficiency of $\eta_e = 1$. What are its values at the following wavelengths: 200, 400, 550, and 850 nm, and 1, 1.3, 1.55, 5, and 10 μm?

14.2.2 An InGaAs photodetector for $\lambda = 1.3$ μm has a responsivity of $\mathcal{R} = 0.8$ A W^{-1}, a specific detectivity $D^* = 7 \times 10^{10}$ cm Hz$^{1/2}$ W^{-1}, a bandwidth of $B = 2.5$ GHz, and a dynamic range DR = 60 dB. It has a circular active area of 80 μm in diameter. The total resistance including the load is $R = 50\,\Omega$. The dark current and the background radiation current of the detector are not known.

a. What is the NEP of this photodetector? What is the photocurrent at this power level?
b. What is the saturation optical signal power for this photodetector? What is the photocurrent at this power level?
c. What is the SNR at the saturation power level of this photodetector? Does the photodetector operate in the quantum or thermal regime at this optical power level?
d. What is the risetime of the detector response to an impulse signal?

14.2.3 The bandwidth B, the cutoff frequency $f_{3\text{dB}}$, and the response risetime t_r of a photodetector are separately defined but are related to one another. Use their definitions given in the text to show that (a) $t_\text{r} = 0.35/f_{3\text{dB}}$ as given in (14.50) and (b) $f_{3\text{dB}} = 0.886B$ as given in (14.52).

14.2.4 What are the fastest-rising optical signals that can be detected with a sufficient temporal resolution by photodetectors of the following 3-dB cutoff frequencies: (a) 1 GHz, (b) 2.5 GHz, (c) 10 GHz, and (d) 50 GHz? What are the shortest rectangular optical pulses that can be detected using such photodetectors?

14.3.1 How is the threshold photon energy determined for photoelectric emission from the following materials: (a) a metal, (b) a nondegenerate semiconductor, (c) an n-type degenerate semiconductor, (d) a p-type degenerate semiconductor, and (e) an NEA semiconductor?

14.3.2 Why are elemental metals and ordinary semiconductors not useful for photocathodes? What are the practically useful materials for photocathodes?

14.3.3 What are the factors that determine the speed of a photocathode or a PMT? What can be done to increase the response speed?

14.3.4 A PMT is usually limited by shot noise generated by its dark current, whereas a vacuum photodiode is normally limited by thermal noise. In this problem, we consider the PMT and its photocathode discussed in Example 14.8. Ignore the background radiation noise in answering the following questions.

a. Show that even when the load resistance is chosen to be as low as $R_\text{L} = 50\,\Omega$ for high-speed applications, the PMT is still shot-noise limited by its dark current. Thus, it always operates in the quantum regime.
b. Assume that all of the dark current of the PMT originates from its photocathode but is amplified through the dynode chain to reach its specified level at the anode. This assumption grossly overestimates the dark current

of the photocathode. Now consider a vacuum photodiode that is made of this photocathode and is biased with the same voltage, about 100 V, on the photocathode of the PMT. Show that this vacuum photodiode is thermal-noise limited for any practical load resistance. Therefore, it always operates in the thermal regime.

14.3.5 The photocathode of the PMT described in Example 14.8 has a much reduced external quantum efficiency of $\eta_e = 0.51\%$ at 850 nm wavelength. Answer the questions in Example 14.8 for this PMT when it is used for optical detection at 850 nm wavelength.

14.4.1 Describe the different types of photoconductors, their threshold photon energies, and their spectral coverage ranges.

14.4.2 Many photoconductive detectors are required to operate at low temperatures, but many others are normally used at room temperature. (a) Which kinds of photoconductive detectors are required to operate at low temperatures? (b) Why are they required to operate at low temperatures? (c) How is the required operating temperature of a photoconductive detector determined?

14.4.3 Show that the gain of an intrinsic photoconductor with ohmic contacts and free-moving electrons and holes is that given in (14.72) by verifying that the signal current is that given in (14.70) with the transit times of electrons and holes given in (14.71).

14.4.4 When the voltage applied to an intrinsic photoconductor is such that $V > V_{SC}$, the photogenerated carriers are screened by a space-charge effect so that they only see a voltage of $V_{SC} = e\mathcal{N}/C$ even as the applied voltage continues to increase.

a. Show that when the space-charge effect appears, $\tau_{tr}^{e}(1 + \mu_h/\mu_e)^{-1} = \tau_d$ if both electrons and holes can freely move, $\tau_{tr}^{e} = \tau_d$ if only electrons can freely move, and $\tau_{tr}^{h} = \tau_d$ if only holes can freely move.

b. Under the condition when the photogenerated carrier density is so high that $N \gg n_0$, p_0, show that the space-charge effect appears when $CV > e\mathcal{N}$.

14.4.5 For a photoconductor, in which the carrier lifetime is primarily determined by the carrier recombination process, the probability distribution function of τ is

$$p(\tau) = \frac{1}{\overline{\tau}} e^{-\tau/\overline{\tau}}, \tag{14.131}$$

which characterizes the Poisson process of a continuous random variable τ. By using this probability distribution function to calculate F defined in (14.78), show that $F = 2$.

14.4.6 The NEP and the detectivity of a photoconductor with a geometry as shown in Fig. 14.11 and a bias circuit as shown in Fig. 14.12(a) are usually limited by the shot noise from its dark current when the device is properly biased and is

loaded with a sufficiently large load resistance to maximize its output signal. Under this condition, the NEP can be minimized while D^* is maximized by choosing an optimum thickness d for the device.

a. Show that the resistive thermal noise is negligible compared to the shot noise from the dark current as long as the voltage applied across the electrodes of the photoconductor satisfies the following condition:

$$V \gg \frac{1}{G}\frac{R_0}{R_{\text{eq}}}\frac{k_{\text{B}}T}{e}, \tag{14.132}$$

where G is the gain, R_0 is the dark resistance of the photoconductor, and R_{eq} is the equivalent resistance of the device including its load. Based on this relation, discuss why the NEP and the detectivity of a photoconductor are usually limited by the shot noise from its dark current.

b. Show that the specific detectivity of a shot-noise-limited photoconductor can be expressed as

$$D^* = \frac{1 - \mathrm{e}^{-\alpha d}}{d^{1/2}}\frac{\eta_{\text{coll}}\eta_{\text{t}}}{h\nu}\left(\frac{eGl^2}{4\sigma_0 V}\right)^{1/2}. \tag{14.133}$$

c. From the relation in (14.133), find the optimum value of the thickness d for a fixed value of α to maximize D^* of the device. What are the value of η_{i} and the value of this D^*_{max} when d is chosen to be its optimum value? Show that $D^* > 99\%D^*_{\text{max}}$ for $\alpha^{-1} < d < 1.5\alpha^{-1}$ and $D^* > 90\%D^*_{\text{max}}$ for $0.6\alpha^{-1} < d < 2.6\alpha^{-1}$.

14.4.7 Find the gain, the responsivity, the NEP for a bandwidth of 1 Hz, and the value of D^* at $\lambda = 850$ nm for the photoconductive detector considered in Examples 14.9 and 14.10 if its cathode is ohmic but its anode has a nonohmic contact that blocks holes.

14.4.8 Show that the space-charge effect can appear if the length of the photoconductor considered in Example 14.9 is reduced to $l = 10$ μm while all other parameters remain unchanged. Find the optical signal powers for which the device is limited by the space-charge effect and those for which it is free of the space-charge effect.

14.4.9 Find the gain, the responsivity, the NEP for a bandwidth of 1 Hz, and the value of D^* at $\lambda = 850$ nm for the photoconductive detector considered in Problem 14.4.8.

14.5.1 How is the active region of a junction photodiode defined? Explain why a photodiode has a unity gain.

14.5.2 Compare the photoconductive and photovoltaic modes of operation of a junction photodiode in terms of (a) the requirement on the bias voltage, (b) the

relative magnitude of the load resistance as compared to the internal resistance of the photodiode, (c) the response as a function of the optical signal power, (d) the noise characteristics, and (e) the response speed and the bandwidth.

14.5.3 Show that a photodiode has the linear response given in (14.87) when it is operated in photoconductive mode, and the logarithmic response given in (14.88) when it is operated in photovoltaic mode. For each case, discuss the conditions under which the given response is valid.

14.5.4 What are the major factors that determine the response speed of a photodiode? What has to be done to increase the response speed?

14.5.5 Compare the advantages and disadvantages between a p–n junction photodiode and a p–i–n junction photodiode.

14.5.6 Compare the advantages and disadvantages between a homojunction photodiode and a heterojunction photodiode.

14.5.7 Answer the following questions regarding Schottky photodiodes.

a. How is a Schottky junction formed?
b. Is a Schottky photodiode a photoconductive or a photovoltaic device?
c. Is a p-type or as n-type semiconductor preferred for a high-speed Schottky photodiode? Explain.
d. How is the spectral response range of a Schottky photodiode determined?

14.5.8 What can be done to maximize the bandwidth–efficiency product of a photodiode? What are the practical device structures that are devised for this purpose?

14.5.9 In this problem, we compare the performances of Si and GaAs p–i–n photodiodes that have the same physical structures for optical detection at 850 nm wavelength. Both have the same i-region thickness $d_i = 3$ μm and the same active-area diameter $2r = 40$ μm. Both are reverse biased at 3 V for a field of 1 MV m^{-1} in the intrinsic region. At 300 K under these conditions, Si has the following parameters: $\alpha = 7 \times 10^4$ m^{-1}, $v_e = 8 \times 10^4$ m s^{-1}, $v_h = 3.2 \times 10^4$ m s^{-1}, and $\epsilon = 11.8\epsilon_0$, while GaAs has the following parameters: $\alpha = 1 \times 10^6$ m^{-1}, $v_e = 1.2 \times 10^5$ m s^{-1}, $v_h = 1.7 \times 10^4$ m s^{-1}, and $\epsilon = 13.18\epsilon_0$. Take $R_L + R_s = 50\ \Omega$ and $C_p = 0$ for both devices. Find the 3-dB cutoff frequency, f_{3dB}, and the internal bandwidth–efficiency product, $\eta_i f_{3dB}$, for both devices. Compare the performances of these two devices.

14.5.10 What is the expected 3-dB cutoff frequency of the InGaAs/InP Schottky photodiode considered in Example 14.13 if it uses p-type semiconductors with a p$^-$-InGaAs layer and a p$^+$-InP substrate?

14.5.11 An InGaAs/InP p–i–n photodiode as considered in Example 14.12 has a diameter of $2r = 20$ μm. It is desired that it has a 3-dB cutoff frequency of $f_{3dB} = 20$ GHz.

a. There are two possible choices of the i-region thickness d_i for this device to have the desired f_{3dB}. Find these two possible values of d_i. Then, compare the two choices in terms of their quantum efficiencies, bandwidth–efficiency products, and responsivities.
b. A double-pass structure with a 100% back reflector is adopted to increase the efficiency, thus the responsivity and the bandwidth–efficiency product, of the device. Compare the two choices of d_i in this double-pass structure in terms of their quantum efficiencies, bandwidth–efficiency products, and responsivities.

14.6.1 Compare the effects of a bias voltage on a photoconductor, a junction photodiode, and an APD.

14.6.2 The gain of an APD results from avalanche multiplication, which can be initiated by electrons or holes. Consider an APD where this process takes place in an avalanche multiplication region of a thickness d_m with finite, nonzero ionization coefficients of α_e and α_h for electrons and holes, respectively, so that the ionization ratio, $k = \alpha_h/\alpha_e$, has a finite, nonzero value.
a. What are the factors that determine the avalanche multiplication factor?
b. Show that for $k \neq 1$, avalanche breakdown occurs at

$$\alpha_e d_m = \frac{\ln k}{k-1}, \tag{14.134}$$

and that for $k = 1$ it occurs at $\alpha_e d_m = \alpha_d d_m = 1$.
c. For a given value of the product $\alpha_e d_m$, show that the largest gain is obtained when the impact ionization ratio is $k = 1$.
d. If, according to (c), the largest gain for a given value of $\alpha_e d_m$ or a given value of $\alpha_h d_m$ is obtained when $k = 1$, why is a material with a value of k close to unity not a good choice for an APD?

14.6.3 What are the major factors that determine the response speed of an APD? How are they different from those that determine the speed of a junction photodiode, which has no gain?

14.6.4 What are the two different modes of operation of an APD? What are their differences in terms of their operating conditions, purposes, and characteristics?

14.6.5 The surface of the APD described in Example 14.16 is antireflection coated for the optical signal to enter without a reflection loss. Absorption of the optical signal takes place only in the InGaAs absorption layer because all other layers have bandgaps larger than the photon energies of interest at 1.3 and 1.55 μm signal wavelengths. The APD operates in the condition described in Example 14.16 with a multiplication gain of $G = 10$. The optical signal makes only a single pass through the device. Answer each of the following questions for both 1.3 and 1.55 μm signal wavelengths.

a. Find the external quantum efficiency and the responsivity of this APD.
b. At $G = 10$, this APD has a dark current of $i_d = 150$ nA and a negligible background radiation current. Find its shot-noise-limited NEP that includes the excess noise and its total NEP that includes the resistive thermal noise, both for a bandwidth of 1 Hz.
c. What is the shot-noise-limited specific detectivity D^* of this ADP? What is the value of D^* if all noise sources are considered?

14.6.6 Find the 3-dB cutoff frequency and the gain–bandwidth product for the ADP described in Example 14.16 when it operates at a gain $G = 20$. The ionization ratio remains at $k = 0.25$ in this operating condition.

14.7.1 Consider a VIPD versus a comparable GWPD.
a. What are the major considerations in using a guided-wave configuration to replace a vertically illuminated configuration?
b. What is the major advantage of a GWPD versus a VIPD?
c. What are the additional advantages?

14.7.2 Compare the differences and the relative advantages of different types of guided-wave photodiodes.

14.7.3 What are the factors that determine the response speeds, thus the bandwidths, of WGPDs and TWPDs, respectively? Explain why a TWPD can have a larger bandwidth–efficiency product than a WGPD if both are designed to have either the same bandwidth or the same efficiency.

14.7.4 Both the bandwidth and the efficiency of a p–i–n photodiode can sometimes be increased but can be reduced in other situations if the thickness, d_i, of its intrinsic active region is increased. In this problem, we increase the value of d_i for the devices considered in Examples 14.17–14.19 to have $d_i = 0.3$ μm. Except those parameters that vary with d_i, all other parameters given in Examples 14.17–14.19 remain unchanged. The parameters that change with d_i have the following new values: $\Gamma = 22.5\%$, $C = 20$ fF for the lumped-circuit devices, and $v_p^m = 4.35 \times 10^7$ m s^{-1} for the distributed TWPD. With this change in d_i, find the bandwidths and the bandwidth–efficiency products for the devices considered in Examples 14.17–14.19: (a) the VIPD with either single-pass or double-pass configuration, (b) the WGPD, (c) the TWPD with either matched or open termination, and (d) the VMDP with perfect velocity matching. For each case considered, indicate whether the bandwidth and the bandwidth–efficiency product are increased or reduced by this change in d_i.

14.7.5 The carrier transit time of a p–i–n photodiode can be reduced by reducing the thickness, d_i, of the intrinsic active region. This reduction in τ_{tr} can increase the bandwidth of a transit-time-limited device, but it can sometimes reduce the bandwidth or the bandwidth–efficiency product of a device in other situations. In this problem, we reduce the value of d_i for the devices considered in

Examples 14.17–14.19 to have $d_i = 0.1$ μm. Except those parameters that vary with d_i, all other parameters given in Examples 14.17–14.19 remain unchanged. The parameters that change with d_i have the following new values: $\Gamma = 7.5\%$, $C = 60$ fF for the lumped-circuit devices, and $v_p^m = 1.45 \times 10^7$ m s^{-1} for the distributed TWPD. With this change in d_i, find the bandwidths and the bandwidth–efficiency products for the devices considered in Examples 14.17–14.19: (a) the VIPD with either single-pass or double-pass configuration, (b) the WGPD, (c) the TWPD with either matched or open termination, and (d) the VMDP with perfect velocity matching. For each case considered, indicate whether the bandwidth and the bandwidth–efficiency product are increased or reduced by this change in d_i.

SELECT BIBLIOGRAPHY

Bhattacharya, P., *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall, 1997.

Bube, R. H., *Photoconductivity of Solids*. New York: Wiley, 1960.

Chuang, S. L., *Physics of Optoelectronic Devices*. New York: Wiley, 1995.

Davis, C. C., *Lasers and Electro-Optics: Fundamentals and Engineering*. Cambridge: Cambridge University Press, 1996.

Donati, S., *Photodetectors: Devices, Circuits, and Applications*. Upper Saddle River, NJ: Prentice-Hall, 2000.

Ebeling, K. J., *Integrated Optoelectronics: Waveguide Optics, Photonics, Semiconductors*. Berlin: Springer-Verlag, 1993.

Gowar, J., *Optical Communication Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Haus, H. A., *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

Iizuka, K., *Elements of Photonics for Fiber and Integrated Optics*, Vol. II. New York: Wiley, 2002.

Kasap, S. O., *Optoelectronics and Photonics: Principles and Practices*. Upper Saddle River, NJ: Prentice-Hall, 2001.

Nalwa, H. S., ed., *Photodetectors and Fiber Optics*. San Diego, CA: Academic Press, 2001.

Pollock, C. R., *Fundamentals of Optoelectronics*. Chicago, IL: Irwin, 1995.

Powers, J., *An Introduction to Fiber Optic Systems*. 2nd edn. Chicago, IL: Irwin, 1997.

Rosencher, E. and Vinter, B., *Optoelectronics*. Cambridge: Cambridge University Press, 2002.

Saleh, B. E. A. and Teich, M. C., *Fundamentals of Photonics*. New York: Wiley, 1991.

Verdeyen, J. T., *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Willardson, R. K. and Beer, A. C., eds., *Semiconductors and Semimetals*, Vol. 22, W. T. Tsang, ed., *Lightwave Communications Technology, Part D, Photodetectors*. New York: Academic Press, 1985.

Wilson, J. and Hawkes, J. F. B., *Optoelectronics: An Introduction*, 3rd edn. London: Prentice-Hall Europe, 1998.

Yariv, A., *Optical Electronics in Modern Communications*, 5th edn. Oxford: Oxford University Press, 1997.

Yeh, C., *Applied Photonics*. San Diego, CA: Academic Press, 1994.

ADVANCED READING LIST

Alping, A., "Waveguide pin photodetectors: theoretical analysis and design criteria," *IEE Proceedings*, *Part J Optoelectronics* **136**(3): 177–182, June 1989.

Brennan, K. F., Haralson, II, J., Parks, J. W., Jr., and Salem, A., "Review of reliability issues of metal–semiconductor–metal and avalanche photodiode photonic detectors," *Microelectronics Reliability* **39**(12): 1873–1883, Dec. 1999.

Bowers, J. E. and Burrus, C. A., Jr., "Ultrawide-band long-wavelength p–i–n photodetectors," *Journal of Lightwave Technology* **LT-5**(10): 1339–1350, Oct. 1987.

Giboney, K. S., Rodwell, M. J., and Bowers, J. E., "Traveling-wave photodetector theory," *IEEE Transactions on Microwave Theory and Techniques* **45**(8): 1310–1319, Aug. 1997.

Kagawa, T., Kawamura, Y., and Iwamura, H., "InGaAsP–InAlAs superlattice avalanche photodiode," *IEEE Journal of Quantum Electronics* **28**(6): 1419–1423, June 1992.

Kato, K., "Ultrawide-band/high-frequency photodetectors," *IEEE Transactions on Microwave Theory and Techniques* **47**(7): 1265–1281, July 1999.

Kumar, A. and Bose, D. N., "Compound semiconductor photodetectors: a review," *IETE Journal of Research* **43**(2–3): 257–265, Mar.–June 1997.

Levine, B. F., "Quantum-well infrared photodetectors," *Journal of Applied Physics* **74**(8): R1–R81, Oct. 1993.

Lin, L. Y., Wu, M. C., Itoh, T., Vang, T. A., Muller, R. E., Sivco, D. L., and Cho, A. Y., "High-power high-speed photodetectors: design, analysis, and experimental demonstration," *IEEE Transactions on Microwave Theory and Techniques* **45**(8): 1320–1331, Aug. 1997.

McIntyre, R. J., "A new look at impact ionization: Part I. A theory of gain, noise, breakdown probability, and frequency response," *IEEE Transactions on Electron Devices* **46**(8): 1623–1631, Aug. 1999.

Pan, J. L. and Fonstad, C. G., "Theory, fabrication and characterization of quantum well infrared photodetectors," *Materials Science and Engineering R*: *Reports*, **R28**(3–4): 65–147, July 2000.

Razeghi, M. and Rogalski, A., "Semiconductor ultraviolet detectors," *Journal of Applied Physics* **79**(10): 7433–7473, May 1996.

Watanabe, I., Tsuji, M., Hayashi, M., Makita, K., and Taguchi, K., "Design and performance of InAlGaAs/InAlAs superlattice avalanche photodiodes," *Journal of Lightwave Technology* **15**(6): 1012–1019, June 1997.

Yuan, P., Anselm, A., Hu, C., Nie, H., Lenox, C., Holmes, A. L., Streetman, B. G., Campbell, J. C., and McIntyre, R. J., "A new look at impact ionization: Part II. Gain and noise in short avalanche photodiodes," *IEEE Transactions on Electron Devices* **46**(8): 1632–1639, Aug. 1999.

Appendix A
Symbols and notations

A.1 Fields

Field vectors and their scalar magnitudes are represented using a consistent system of symbols and fonts. All vectors are represented in bold-face fonts with the exceptions of unit vectors; whereas all scalar quantities are represents in nonbold fonts. This system is illustrated in the following using the electric field as an example.

Real fields

All real fields are defined in the real space and time domain only. All real field vectors are represented in a slanted bold capital Roman font, such as

$$\boldsymbol{E}(\mathbf{r}, t), \tag{A.1}$$

for the real electric field vector. Other real field vectors are $\boldsymbol{H}(\mathbf{r}, t)$, $\boldsymbol{D}(\mathbf{r}, t)$, $\boldsymbol{B}(\mathbf{r}, t)$, $\boldsymbol{P}(\mathbf{r}, t)$, $\boldsymbol{M}(\mathbf{r}, t)$, $\boldsymbol{J}(\mathbf{r}, t)$, and $\boldsymbol{S}(\mathbf{r}, t)$. Except for current density, all real fields are always represented in vector form without separate symbols defined for their scalar magnitudes. The scalar magnitude of $\boldsymbol{J}$ is represented as J.

Complex fields

All complex field vectors are represented in a nonslanted bold capital Roman font. All complex field vectors in the real space and time domain are defined in relation to their respective real field vectors, such as $\mathbf{E}(\mathbf{r}, t)$ defined in (1.39) for the complex electric field vector:

$$\boldsymbol{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \mathbf{E}^*(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) + \text{complex conjugate}. \tag{A.2}$$

Other complex field vectors defined in a similar manner are $\mathbf{H}(\mathbf{r}, t)$, $\mathbf{D}(\mathbf{r}, t)$, $\mathbf{B}(\mathbf{r}, t)$, $\mathbf{P}(\mathbf{r}, t)$, and $\mathbf{M}(\mathbf{r}, t)$. No complex vector is used for current density. The complex Poynting vector is defined differently as given in (1.50):

$$\overline{\boldsymbol{S}} = \mathbf{S} + \mathbf{S}^*. \tag{A.3}$$

The scalar magnitude of a complex field vector is represented in a nonbold mathematic capital Roman font, such as E for the magnitude of $\mathbf{E}$:

$$\mathbf{E} = E\hat{e}, \tag{A.4}$$

where $\hat{e}$ is the unit vector of $\mathbf{E}$. Other scalar field magnitudes represented in a similar manner are H, D, B, P, and M. No complex current density vector is used, and the scalar J represents the magnitude of the real current density vector $\boldsymbol{J}$. No scalar magnitude of the complex Poynting vector $\mathbf{S}$ is used.

Complex field amplitudes

The slowly varying amplitude vector of a complex field vector is represented in a bold capital script font, such as $\boldsymbol{\mathcal{E}}$ for the slowly varying amplitude of $\mathbf{E}$. It is defined as the slow variation of the field envelope on its carrier frequency through the following relation:

$$\mathbf{E}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}(\mathbf{r}, t) \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r} - \mathrm{i}\omega t), \tag{A.5}$$

as expressed in (1.47) for the electric field. Other slowly varying field amplitude vectors defined in a similar manner are $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\mathcal{P}}$, and $\boldsymbol{\mathcal{M}}$, but not all of them are used in the text. No slowly varying field amplitudes are defined for current density and the Poynting vector.

The scalar magnitude of a slowly varying field amplitude vector is represented in a nonbold capital script font, such as $\mathcal{E}$ for the magnitude of $\boldsymbol{\mathcal{E}}$:

$$\boldsymbol{\mathcal{E}} = \mathcal{E}\hat{e}. \tag{A.6}$$

Other scalar magnitudes of slowly varying field amplitudes represented in a similar manner are $\mathcal{H}$, $\mathcal{D}$, $\mathcal{B}$, $\mathcal{P}$, and $\mathcal{M}$, but not all of them are used in the text.

Mode fields

Complex mode field vectors are represented as $\mathbf{E}_\nu(\mathbf{r}, t)$ and $\mathbf{H}_\nu(\mathbf{r}, t)$ with their scalar magnitudes represented as $E_\nu(\mathbf{r}, t)$ and $H_\nu(\mathbf{r}, t)$, respectively, where the subscript index ν represents a compound mode index such as m or mn for waveguide modes, or mnq for Gaussian modes. The vectorial field profiles of a waveguide mode characterize the transverse spatial distributions of the mode fields. They are a function of transverse spatial coordinates only. These vectorial waveguide mode field profiles are represented as $\boldsymbol{\mathcal{E}}_\nu(x, y)$ and $\boldsymbol{\mathcal{H}}_\nu(x, y)$, or $\boldsymbol{\mathcal{E}}_\nu(\phi, r)$ and $\boldsymbol{\mathcal{H}}_\nu(\phi, r)$, with their scalar magnitudes represented as $\mathcal{E}_\nu(x, y)$ and $\mathcal{H}_\nu(x, y)$, or $\mathcal{E}_\nu(\phi, r)$ and $\mathcal{H}_\nu(\phi, r)$. Normalized vectorial mode field patterns, defined in (2.41), are represented as $\hat{\boldsymbol{\mathcal{E}}}_\nu(x, y)$ and $\hat{\boldsymbol{\mathcal{H}}}_\nu(x, y)$, or $\hat{\boldsymbol{\mathcal{E}}}_\nu(\phi, r)$ and $\hat{\boldsymbol{\mathcal{H}}}_\nu(\phi, r)$. Gaussian modes are represented using similar symbols, but they are a function of x, y, and z, as seen in (1.138).

A.2 Vectors and tensors

All vectors are represented in bold face, with the exceptions of unit vectors, and their magnitudes are represented with corresponding symbols in nonbold fonts. A vector is also represented in the form of a 3×1 column matrix. Besides the field vectors and their magnitudes described in the preceding section, we have

$$\mathbf{k}, k;\ \mathbf{K}, K;\ \mathbf{r}, r;\ \mathbf{u}, u;\ \Delta\mathbf{k}, \Delta k.$$

All tensors and transformation matrices are represented in bold face or in terms of their elements with subscript indices. Second-rank tensors and transformation matrices are also represented in the

form of 3×3 square matrices. The tensors used include:

$$
\begin{array}{lllll}
[c_{ijkl}], & \mathbf{d} = [d_{ijk}], & [f_{ijk}], & [p_{ijkl}], & [r_{ijk}], \\
\mathbf{R} = [R_{ij}], & [s_{ijkl}], & \mathbf{S} = [S_{ij}], & \boldsymbol{\epsilon} = [\epsilon_{ij}], & \boldsymbol{\eta} = [\eta_{ij}], \\
\boldsymbol{\mu} = [\mu_{ij}], & \boldsymbol{\chi} = [\chi_{ij}], & \boldsymbol{\chi}^{(2)} = [\chi^{(2)}_{ijk}], & \boldsymbol{\chi}^{(3)} = [\chi^{(3)}_{ijkl}], & \boldsymbol{\chi}_{\mathrm{m}} = [(\chi_{\mathrm{m}})_{ij}].
\end{array}
$$

The transformation matrices used in the text include:

$\mathbf{F}(z; z_0)$, $\mathbf{R}(z; 0, l)$, $\mathbf{S}(z; z_0)$, $\mathbf{T}$, $\tilde{\mathbf{T}}$.

A.3 Fourier-transform pairs

The same symbol is used for a quantity in real space and its counterpart in momentum space, or one in the time domain and its counterpart in the frequency domain. The difference is indicated by expressing a quantity as a function of $\mathbf{r}$ or $\mathbf{k}$, or as a function of t or ω. Note that the unit of a quantity is multiplied by a length unit of a meter each time one of the three spatial dimensions is transformed to momentum space, and is multiplied by a time unit of a second when the quantity is transformed from the time domain to the frequency domain. For example, the electric field $E(\mathbf{r}, t)$ in the real space and time domain has units of volts per meter (V m^{-1}), but $E(\mathbf{k}, t)$ has units of volt-square-meters (V m^2), $E(\mathbf{r}, \omega)$ has units of volt-seconds per meter (V s m^{-1}), and $E(\mathbf{k}, \omega)$ has units of volt-second-square-meters (V s m^2).

A.4 Special notations

A few special notations are used to label symbols for special meanings.

Unit vectors and normalized quantities

Unit vectors are denoted with a hat on top of a symbol. The following unit vectors appear in the text:

$\hat{e}$, $\hat{k}$, $\hat{n}$, $\hat{r}$, $\hat{u}$, $\hat{x}$, $\hat{y}$, $\hat{z}$, $\hat{X}$, $\hat{Y}$, $\hat{Z}$.

Normalized quantities are also denoted with a hat on top of a symbol. The following normalized mode field profiles appear in both vector and scalar forms:

$\hat{\boldsymbol{\mathcal{E}}}_\nu$, $\hat{\mathcal{E}}_\nu$, $\hat{\boldsymbol{\mathcal{H}}}_\nu$, $\hat{\mathcal{H}}_\nu$.

Other normalized quantities that appear in the text include:

$\hat{f}$, $\hat{g}(\nu)$, $\hat{P}_{\mathrm{sp}}$, $\hat{T}_{\mathrm{c}}$, $\hat{\eta}_{\mathrm{SH}}$.

Modified quantities

A quantity that is modified from the original quantity in some manner is denoted with a tilde on top of a symbol. Modified quantities that appear in the text include:

$\tilde{A}$, $\tilde{B}$, $\tilde{g}_{\mathrm{B}}$, $\tilde{g}_{\mathrm{R}}$, $\tilde{\mathbf{T}}$, $\Delta\tilde{\epsilon}$, $\Delta\tilde{\epsilon}$, $\tilde{\kappa}$.

Average values

The spatial average, temporal average, weighted average, or mean value of a quantity is denoted with a bar on top of a symbol, such as:

$\overline{i}$, $\overline{i^2}$, $\overline{k}$, $\overline{\mathcal{N}}$, $\overline{\mathcal{N}^2}$, $\overline{P}$, $\overline{s}$, $\overline{s^2}$, $\overline{\boldsymbol{S}}$, $\overline{\mathcal{S}}$, $\overline{\mathcal{S}^2}$, $\overline{v^2}$, $\overline{W}_{\rm p}$, $\overline{\alpha}$, $\overline{\beta}$.

A.5 Subscripts and superscripts

Various fonts and notations are used for subscripts and superscripts. They include numerals, the mathematical font, the Greek font, coordinate symbols, and the Roman font. Bare numerals and mathematic and Greek letters that represent indices or variables are used only for subscripts. Roman letters and some special notations that have literal meanings can appear either as subscripts or as superscripts.

Numerals

Bare numerals are used only for subscripts. The following four numbers have special meanings in a proper context:

0 background value (L_0), base value (α_0, m_0), free-space value (ϵ_0, μ_0), center value (f_0, ν_0), unsaturated value (g_0), equilibrium value (n_0, p_0), beam waist (w_0), or static field ($\boldsymbol{E}_0$, $\boldsymbol{H}_0$);

1 parameters for waveguide core (n_1, N_1, D_1, k_1, h_1) or parameters for the lower laser level $|1\rangle$ (E_1, N_1, R_1);

2 parameters for waveguide substrate or fiber cladding (n_2, N_2, D_2, k_2, γ_2) or parameters for the upper laser level $|2\rangle$ (E_2, N_2, R_2);

3 parameters for waveguide cover layer (n_3, N_3, D_3, k_3, γ_3) or parameters for the energy level $|3\rangle$.

Note that the same symbol can have different meanings in different contexts. For example, n_2 in nonlinear optics also represents the coefficient of intensity-dependent index change defined in (9.49).

The numbers 1, 2, and 3 are also used as subscripts to represent the orthogonal coordinates of a general three-dimensional spatial coordinate system. The numbers 1 through 6 are also used as subscripts representing double indices to label tensor elements under the following index contraction rule:

xx	yy	zz	yz, zy	zx, xz	xy, yx
1	2	3	4	5	6

which is also defined in (1.115).

A numeral in the superscript is always placed in parentheses so that it is never confused with an exponent. It represents a perturbation order or the order of an interaction process. For example, $\chi^{(1)}$ is a linear susceptibility, $\chi^{(2)}$ is a second-order nonlinear susceptibility, $\chi^{(3)}$ is a third-order nonlinear susceptibility, and so forth.

Mathematic and Greek subscripts

Mathematic and Greek fonts are used only for subscripts. They represent variable indices with the following well-defined meanings:

a, b, c	general indices or general mode indices;
i, j, k, l	integers or coordinate indices;
m, n, p, q	integers or frequency component indices;
m, n	transverse mode indices, each labeling a spatial dimension;
q	longitudinal mode index or diffraction order;
α, β	contracted indices representing double coordinate indices;
μ, ν, ξ, ζ	compound transverse mode indices, each representing a mode.

Some Greek subscripts do not represent indices or variables but express literal meanings. They include:

$\beta,\ \lambda,\ \lambda/2,\ \lambda/4,\ \pi,\ \pi/2,\ \chi,\ \omega,\ 2\omega.$

Coordinate labels

General orthogonal spatial coordinates are labeled as 1, 2, and 3. Specific coordinates include the rectangular coordinates (x, y, z), the cylindrical coordinates (r, ϕ, z), and the spherical coordinates (r, θ, ϕ). One set of special rectangular coordinates (X, Y, Z) is used for the new principal axes $\hat{X}$, $\hat{Y}$, and $\hat{Z}$ of a crystal transformed under the Pockels effect, as described in (6.6) and (6.9). Two orthogonal unit vectors, $\hat{e}_+$ and $\hat{e}_-$ defined in (1.72) and (1.75), are used for left- and right-circular polarizations, respectively. Two special symbols are also used to represent directions: $\perp$ for perpendicular and $\parallel$ for parallel.

Coordinate labels generally appear as subscripts with commonly accepted meanings, with one exception. This exception takes place when labeling a propagation constant k and the corresponding wavevector $\mathbf{k}$ of an optical field that has a particular normal mode polarization. Because k_x conventionally represents the x component of the $\mathbf{k}$ vector, meaning $k_x = \mathbf{k} \cdot \hat{x}$, the propagation constant of an x-polarized optical field that can propagate in any direction perpendicular to $\hat{x}$ is represented as k^x in order to avoid confusion. To be consistent, the corresponding wavevector is labeled as $\mathbf{k}^x$. Thus, $k^x = n_x\omega/c \neq k_x$, and $\mathbf{k}^x = k^x\hat{k}$ where $\hat{k} \perp \hat{x}$. Such superscript coordinate labeling for k and $\mathbf{k}$ applies only to the following:

$k^x,\ k^y,\ k^z,\ k^X,\ k^Y,\ k^Z,\ k^+,\ k^-,$

$\mathbf{k}^x,\ \mathbf{k}^y,\ \mathbf{k}^z,\ \mathbf{k}^X,\ \mathbf{k}^Y,\ \mathbf{k}^Z,\ \mathbf{k}^+,\ \mathbf{k}^-.$

Roman labels

All superscript and subscript labels in Roman font have literal meaning. A given Roman label can appear either as a subscript or as a superscript, depending on the convenience of the situation, with exactly the same meaning. Among all subscript and superscript labels, only Roman labels have such flexibility. With only a few exceptions for avoiding confusion, the conventional rules for abbreviations are largely followed: (1) abbreviations for common words are in lower case, with the exceptions of E for TE, M for TM, L for longitudinal, linear, or load, T for transverse, and Q for quasi-phase matching; (2) abbreviations for proper nouns are in upper case; and (3) acronyms are all in upper case, with the exception of hh for heavy hole and lh for light hole. Two Roman numbers are also used: I for type I and II for type II. The Roman labels used in this book are listed below.

Label	Meaning	Label	Meaning
a	absorption, acceptance, acceptor, acoustic, anode, aperture	m	magnetization, microwave, modulation, multiplication
AS	anti-Stokes	M	TM mode
att	attenuation	max	maximum
av	avalanche	MC	mode conversion
b	background, backward, bias	min	minimum
B	Boltzmann, Bragg, Brewster, Brillouin	n	n type, noise
br	breakdown	NL	nonlinear
c	carrier, cavity, center, characteristic, circular, coercive, conduction band, conversion, coupling, critical, cutoff	nonrad	nonradiative
ckt	circuit	nonrec	nonreciprocal
coh	coherence	o	optical, ordinary
coll	collection	opt	optimum
comp	compensation	out	output
d	dark, data, detector, dielectric, diffracted, diffusion, donor	p	p-polarized (TM), p type, parallel, period, phase, polarization, prism, pump
D	Doppler	ph	photo, photon
def	deflection	pk	peak
e	electrical, electrode, electron, emission, external, extraordinary	PM	phase matched
E	TE mode	ps	pulse
eff	effective	q	quantum
eq	equivalent	Q	quasi-phase matching
esc	escape	QW	quantum well
f	fall, forward	r	radiation, reduced, recombination, reflection, relaxation, reversed, rise
F	Faraday, Fermi	R	Raman, Rayleigh
g	bandgap, gain, gap, grating, group	rad	radiative
GR	generation–recombination	RC	RC time
h	hole, homogeneous	rec	reciprocal
hh	heavy hole	res	resonance
i	incidence, initial, internal, intrinsic	RT	round trip
in	input	s	s-polarized (TE), saturation, series, shield, signal, slope, spontaneous, source, switching
inh	inhomogeneous	S	Stokes
inj	injection	sat	saturation
j	junction	SB	stop band
k	cathode	SC	space charge
K	Kerr	sh	shot
l	luminous	SH	second harmonic
L	linear, load, longitudinal	sp	spontaneous
lh	light hole	SR	Shockley–Reed
		ST	Shawlow–Townes

Label	Meaning	Label	Meaning
t	extraction, tangential, total, transducer, transmitted	th	thermal, threshold
		v	valence band
T	transverse	vac	vacuum
tr	transit, transparency	VM	velocity mismatch

Appendix B
Table of prerequisites

Section	Prerequisite sections	Section	Prerequisite sections
1 General background			
1.1	none	**1.7**	none
1.2	1.1	**1.8**	none
1.3	1.2	**1.9**	none
1.4	none	**1.10**	1.1–1.3
1.5	1.2, 1.3	**1.11**	none
1.6	1.2–1.5		
2 Optical waveguides			
2.1	1.8	**2.5**	2.1–2.4
2.2	1.2, 2.1	**2.6**	2.5
2.3	1.5, 2.2	**2.7**	2.1–2.5
2.4	1.1, 2.2	**2.8**	2.1–2.7
3 Optical fibers			
3.1	2.1–2.5	**3.4**	none
3.2	3.1	**3.5**	1.9, 1.10, 3.3
3.3	2.7, 3.1, 3.2		
4 Coupling of waves and modes			
4.1	1.5, 1.6	**4.3**	4.2
4.2	2.2, 2.4		
5 Optical couplers			
5.1	2.5, 4.2, 4.3	**5.3**	2.1
5.2	2.5, 4.2, 4.3		
6 Electro-optic devices			
6.1	1.1, 1.6	**6.4**	2.5, 4.2, 4.3, 5.1, 5.2, 6.3
6.2	1.6, 6.1	**6.5**	6.4
6.3	1.6, 6.2		
7 Magneto-optic devices			
7.1	1.1, 1.6, 6.1	**7.5**	7.2
7.2	1.4, 7.1	**7.6**	7.3
7.3	1.4, 7.1	**7.7**	2.5, 4.2, 4.3, 5.1, 5.2, 6.4, 7.1, 7.4
7.4	1.6, 7.2		

Section	Prerequisite sections	Section	Prerequisite sections
8 Acousto-optic devices			
8.1	none	**8.5**	8.3, 8.4
8.2	1.6, 6.1, 6.2, 8.1	**8.6**	8.3
8.3	4.1, 4.3, 8.2	**8.7**	8.3–8.6
8.4	1.7, 8.3		
9 Nonlinear optical devices			
9.1	1.1–1.3, 4.1	**9.7**	6.3, 9.3
9.2	1.3, 1.6, 6.1, 6.2, 9.1	**9.8**	9.7
9.3	1.3, 9.1, 9.2	**9.9**	1.10, 8.3, 9.3
9.4	4.1, 9.1–9.3	**9.10**	4.1, 4.2
9.5	4.3, 9.3	**9.11**	9.4–9.6, 9.10
9.6	1.6, 5.1, 9.4	**9.12**	4.2, 4.3, 9.7, 9.10
10 Laser amplifiers			
10.1	1.10	**10.4**	10.3
10.2	1.5, 1.10, 10.1	**10.5**	10.4
10.3	10.2		
11 Laser oscillators			
11.1	1.7	**11.4**	11.3
11.2	11.1	**11.5**	10.5, 11.1–11.4
11.3	10.3, 11.2		
12 Semiconductor basics			
12.1	none	**12.4**	none
12.2	none	**12.5**	12.2, 12.4
12.3	none		
13 Semiconductor lasers and light-emitting diodes			
13.1	12.3	**13.6**	13.5
13.2	10.1, 12.2	**13.7**	12.1, 12.5, 13.1, 13.5, 13.6
13.3	10.2, 10.3, 12.2, 12.3, 13.2	**13.8**	10.4, 10.5
13.4	12.3, 13.3	**13.9**	5.1, 5.3, 11.1, 11.2, 12.1, 13.5–13.7
13.5	12.2, 12.5, 13.3	**13.10**	11.2, 11.3, 12.5, 13.3, 13.7, 13.9
14 Photodetectors			
14.1	none	**14.5**	12.5, 14.1, 14.2
14.2	none	**14.6**	12.5, 14.1, 14.2, 14.5
14.3	14.1, 14.2	**14.7**	6.5, 14.5, 14.6
14.4	12.2–12.4, 14.1, 14.2		

Appendix C
SI metric system

Table C.1 *SI base units*

Quantity	Name	Symbol
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

Table C.2 *SI derived units*

Quantity	Name	Symbol	Equivalent
Plane angle	radian	rad	$\mathrm{m\ m^{-1}} = 1$
Solid angle	steradian	sr	$\mathrm{m^2\ m^{-2}} = 1$
Frequency	hertz	Hz	$\mathrm{s^{-1}}$
Force	newton	N	$\mathrm{kg\ m\ s^{-2}}$
Pressure	pascal	Pa	$\mathrm{N\ m^{-2}}$
Energy	joule	J	$\mathrm{kg\ m^2\ s^{-2}}$
Power	watt	W	$\mathrm{J\ s^{-1}}$
Electric charge	coulomb	C	A s
Electric potential	volt	V	$\mathrm{J\ C^{-1}}$, $\mathrm{W\ A^{-1}}$
Magnetic flux	weber	Wb	V s
Magnetic flux intensity	tesla	T	$\mathrm{Wb\ m^{-2}}$
Resistance	ohm	Ω	$\mathrm{V\ A^{-1}}$
Conductance	siemens	S	$\mathrm{A\ V^{-1}}$, Ω^{-1}
Capacitance	farad	F	$\mathrm{C\ V^{-1}}$
Inductance	henry	H	$\mathrm{Wb\ A^{-1}}$
Luminous flux	lumen	lm	cd sr
Illuminance	lux	lx	$\mathrm{lm\ m^{-2}}$

Source: Nelson, R. A., "Guide for metric practice," *Physics Today* BG15–BG16, August, 2002.

Table C.3 *Metric prefixes*

Name	Symbol	Factor	Name	Symbol	Factor
Exa	E	10^{18}	Deci	d	10^{-1}
Peta	P	10^{15}	Centi	c	10^{-2}
Tera	T	10^{12}	Milli	m	10^{-3}
Giga	G	10^{9}	Micro	μ	10^{-6}
Mega	M	10^{6}	Nano	n	10^{-9}
Kilo	k	10^{3}	Pico	p	10^{-12}
Hecto	h	10^{2}	Femto	f	10^{-15}
Deca	da	10	Atto	a	10^{-18}
Unit		1			

Appendix D
Fundamental physical constants

Table D.1 *Physical constants*

Quantity	Symbol	Value	Unit
Speed of light in free space	c	$2.997\,924\,58 \times 10^{8}$	$\mathrm{m\ s^{-1}}$
Magnetic permeability of free space	μ_0	$4\pi \times 10^{-7}$	$\mathrm{H\ m^{-1}}$
		$1.256\,637\,061\,4 \times 10^{-6}$	$\mathrm{H\ m^{-1}}$
Electric permittivity of free space $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817 \times 10^{-12}$	$\mathrm{F\ m^{-1}}$
Impedance of free space $(\mu_0/\epsilon_0)^{1/2}$	Z_0	376.730 313 461	Ω
Planck constant	h	$6.626\,068\,765\,2 \times 10^{-34}$	J s
		$4.135\,667\,271\,6 \times 10^{-15}$	eV s
Planck constant $h/2\pi$	$\hbar$	$1.054\,571\,596\,8 \times 10^{-34}$	J s
		$6.582\,118\,892\,6 \times 10^{-16}$	eV s
Elementary charge	e	$1.602\,176\,462\,6 \times 10^{-19}$	C
Electron rest mass	m_0	$9.109\,381\,887\,2 \times 10^{-31}$	kg
Proton rest mass	m_p	$1.672\,621\,581\,3 \times 10^{-27}$	kg
Atomic mass unit	m_u	$1.660\,538\,731\,3 \times 10^{-27}$	kg
Boltzmann constant	k_B	$1.380\,650\,324 \times 10^{-23}$	$\mathrm{J\ K^{-1}}$
		$8.617\,342\,15 \times 10^{-5}$	$\mathrm{eV\ K^{-1}}$
Thermal energy at $T = 300$ K	$k_\mathrm{B}T$	$2.585\,202\,645 \times 10^{-2}$	eV
Photon constant $hc = \lambda h\nu$	hc	$1.239\,841\,86 \times 10^{-6}$	eV m

Source: Mohr, P. J., and Taylor, B. N., "The fundamental physical constants," *Physics Today* BG6–BG13, August, 2002.

Appendix E
Fourier-transform relations

According to the discussions in Chapter 1, we define the Fourier transform between the time domain and the frequency domain in terms of angular frequency as follows:

$$E(\omega) = \mathcal{F}\{E(t)\} = \int_{-\infty}^{\infty} E(t)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t \tag{E.1}$$

and

$$E(t) = \mathcal{F}^{-1}\{E(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} E(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}\omega. \tag{E.2}$$

In terms of the real frequency $\nu = \omega/2\pi$, we have

$$E(\nu) = \int_{-\infty}^{\infty} E(t)\mathrm{e}^{\mathrm{i}2\pi\nu t}\mathrm{d}t \tag{E.3}$$

and

$$E(t) = \int_{-\infty}^{\infty} E(\nu)\mathrm{e}^{-\mathrm{i}2\pi\nu t}\mathrm{d}\nu. \tag{E.4}$$

The Fourier-transform relations for common functions encountered in the description of various waveforms are listed in Table E.1. In this table, the Heaviside function $H(x)$ is defined as

$$H(x) = \begin{cases} 1, & \text{if} \quad x > 0, \\ 0, & \text{if} \quad x < 0; \end{cases} \tag{E.5}$$

the rectangular function $\Pi(x)$ is defined as

$$\Pi(x) = \begin{cases} 1, & \text{if} \quad |x| < 1/2, \\ 0, & \text{if} \quad |x| > 1/2; \end{cases} \tag{E.6}$$

Table E.1 *Fourier-transform relations*

Function form	$E(t)$	$E(\omega)$	Function form
Gaussian	e^{-t^2/τ^2}	$\sqrt{\pi}\tau\mathrm{e}^{-\omega^2\tau^2/4}$	Gaussian
sech	$\operatorname{sech}\dfrac{t}{\tau}$	$\pi\tau\operatorname{sech}\dfrac{\pi\omega\tau}{2}$	sech
Infinite impulse sequence	$\displaystyle\sum_m \delta\left(\frac{t}{\tau}-m\right)$	$\displaystyle\tau\sum_n \delta\left(\frac{\omega\tau}{2\pi}-n\right)$	infinite impulse sequence
Complex exponential	$\mathrm{e}^{-\mathrm{i}\omega_0 t}$	$2\pi\delta(\omega-\omega_0)$	delta
Double-sided exponential	$\mathrm{e}^{-\lvert t/\tau\rvert}$	$\dfrac{2\tau}{1+\omega^2\tau^2}$	Lorentzian
Single-sided exponential	$\mathrm{e}^{-t/\tau}H(t)$	$\dfrac{\tau}{1+\mathrm{i}\omega\tau}$	complex Lorentzian
Rectangular	$\Pi\left(\dfrac{t}{\tau}\right)$	$\tau\dfrac{\sin(\omega\tau/2)}{\omega\tau/2}$	sinc
Triangular	$\Lambda\left(\dfrac{t}{\tau}\right)$	$\tau\dfrac{\sin^2(\omega\tau/2)}{(\omega\tau/2)^2}$	sinc^2
Convolution	$f(t)*g(t)$	$f(\omega)g(\omega)$	product
Product	$f(t)g(t)$	$\dfrac{1}{2\pi}f(\omega)*g(\omega)$	convolution
Complex conjugate	$f^*(t)$	$[f(-\omega)]^*$	

and the triangular function $\Lambda(x)$ is defined as

$$\Lambda(x)=\begin{cases}1-|x|, & \text{if} \quad |x|\le 1,\\ 0, & \text{if} \quad |x|>1.\end{cases} \tag{E.7}$$

The convolution integral is defined as

$$f(x)*g(x)=\int_{-\infty}^{\infty} f(x')g(x-x')\mathrm{d}x'. \tag{E.8}$$

Using the Fourier-transform relation between $f(t)*g(t)$ and $f(\omega)g(\omega)$ and that between $f^*(t)$ and $f^*(-\omega)$ shown in Table E.1, the following useful relations can be obtained:

$$\text{Correlation Theorem}\quad \int_{-\infty}^{\infty} f^*(t)g(t+\tau)\mathrm{d}t=\frac{1}{2\pi}\int_{-\infty}^{\infty} f^*(\omega)g(\omega)\mathrm{e}^{-\mathrm{i}\omega\tau}\mathrm{d}\omega, \tag{E.9}$$

$$\text{Autocorrelation Theorem}\quad \int_{-\infty}^{\infty} f^*(t)f(t+\tau)\mathrm{d}t=\frac{1}{2\pi}\int_{-\infty}^{\infty} |f(\omega)|^2\mathrm{e}^{-\mathrm{i}\omega\tau}\mathrm{d}\omega, \tag{E.10}$$

Power Theorem
$$\int_{-\infty}^{\infty} f^*(t)g(t)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} f^*(\omega)g(\omega)\mathrm{d}\omega, \tag{E.11}$$

Parseval's Theorem
$$\int_{-\infty}^{\infty} |f(t)|^2\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} |f(\omega)|^2\mathrm{d}\omega. \tag{E.12}$$

Using (E.3) and (E.4), Parseval's theorem can also be written as

$$\int_{-\infty}^{\infty} |E(t)|^2\mathrm{d}t = \int_{-\infty}^{\infty} |E(\nu)|^2\mathrm{d}\nu = \frac{1}{2\pi}\int_{-\infty}^{\infty} |E(\omega)|^2\mathrm{d}\omega. \tag{E.13}$$

Index