Handbook of Laser Technology and Applications

Volume I: Principles

Edited by

Colin E Webb

University of Oxford

and

Julian D C Jones

Heriot-Watt University

IoP

Institute of Physics Publishing
Bristol and Philadelphia

© IOP Publishing Ltd 2004

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher. Multiple copying is permitted in accordance with the terms of licences issued by the Copyright Licensing Agency under the terms of its agreement with Universities UK (UUK).

The publisher has attempted to trace the copyright holders of all the figures reproduced in this publication and apologizes to them if permission to publish in this form has not been obtained.
British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN 0 7503 0960 1 (Vol. I)
0 7503 0963 6 (Vol. II)
0 7503 0966 0 (Vol. III)
0 7503 0607 6 (3 Vol. set)

Library of Congress Cataloging-in-Publication Data are available

Development Editor: David Morris
Production Editor: Simon Laurenson
Production Control: Sarah Plenty
Cover Design: Victoria Le Billon
Marketing: Nicola Newey and Verity Cooke

Published by Institute of Physics Publishing, wholly owned by The Institute of Physics, London

Institute of Physics Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

US Office: Institute of Physics Publishing, The Public Ledger Building, Suite 929, 150 South Independence Mall West, Philadelphia, PA 19106, USA

Typeset in LaTeX 2_ε by Text 2 Text Limited, Torquay, Devon
Index by Indexing Specialists (UK) Ltd, Hove, East Sussex
Printed in the UK by MPG Books Ltd, Bodmin, Cornwall

Contents

	Editorial and Advisory Board		xii
	List of contributors		xiii
	Foreword *Charles Townes*		xxiii
	Introduction *Colin Webb*		xxv
VOLUME I: PRINCIPLES			
PART A	**PRINCIPLES**		**1**
	A	Principles *Richard Shoemaker*	3
	A1	Basic laser principles *Christopher C Davis*	5
	A2.1	Free-space laser resonators *Robert C Eckardt*	81
	A2.2	Waveguide laser resonators *Chris Hill*	115
	A3	Laser beam control *Jacky Byatt*	135
	A4	Nonlinear optics *Robert W Boyd*	161
	A5	Interference and polarization *Alan Rogers*	185
	A6	Optical waveguide theory *G Stewart*	223
	A7	Optical detection and noise *Gerald Buller and Jason Smith*	251
	A8	Introduction to numerical analysis for laser systems *George Lawrence*	281
VOLUME II: LASER DESIGN AND LASER SYSTEMS			
PART B	**LASER DESIGN, FABRICATION AND PROPERTIES**		**303**
	B1	Solid state lasers *R C Powell*	305

B1.1	Transition metal ion lasers—Cr^{3+} *Georges Boulon*	307
B1.2	Transition metal ion lasers other than Cr^{3+} *Stephen A Payne*	339
B1.3	Rare earth ion lasers—Nd^{3+} *A I Zagumennyi, V A Mikhailov and I A Shcherbakov*	353
B1.4	Lanthanide series lasers—near infrared *Norman P Barnes*	383
B1.5	Rare-earth ions—miscellaneous: Ce^{3+}, U^{3+}, divalent, etc *Gregory J Quarles*	411
B1.6	Lasers based on nonlinear effects *Fabienne Pellé*	431
B1.7	Solid state Raman lasers *Tasoltan T Basiev and Richard C Powell*	469
B1.8	Colour centre lasers *T T Basiev, P G Zverev and S B Mirov*	499
B2	Laser diodes *Ian White*	523
B2.1	Basic principles of laser diodes *N K Dutta*	525
B2.2	Spectral control in laser diodes *Markus-Christian Amann*	561
B2.3	High-speed laser diodes *Peter P Vasil'ev*	585
B2.4	High-power laser diodes and laser diode arrays *Peter Unger*	605
B2.5.1	Visible laser diodes: properties of III–V red-emitting laser diodes *Peter Blood*	619
B2.5.2	Visible laser diodes: properties of blue laser diodes *Robert Martin*	641
B2.6	Vertical-cavity surface-emitting lasers *B M A Rahman and K T V Grattan*	659
B2.7	Long wavelength laser diodes *S Anders, G Strasser and E Gornik*	691
B2.8	Semiconductor lasers and optical amplifiers for switching and signal processing *Hitoshi Kawaguchi*	707
B3	Gas lasers *Julian Jones*	749
B3.1	Carbon dioxide lasers *Denis R Hall*	751
B3.2	Excimer, F_2, N_2 and H_2 lasers *W J Witteman*	791
B3.3	Copper and gold vapour lasers *Colin Webb*	847
B3.4.1	Chemical lasers: COIL *B D Barmashenko and S Rosenwaks*	861

	B3.4.2	Chemical lasers: HF/DF *Lee H Sentman*	881
	B3.5	Argon and krypton ion lasers *Malcolm H Dunn and Tony Gutierrez*	893
	B3.6	Helium–neon lasers *Alan D White and Lisa Tsufura*	909
	B3.7	Helium–cadmium laser *William T Silfvast*	921
	B3.8	Optically pumped mid IR lasers: NH_3, C_2H_2 *Mary S Tobin*	929
	B3.9	Far-IR lasers: HCN, H_2O *Wilhelm Prettl*	951
	B4	Fibre and waveguide lasers *R C Powell*	961
	B4.1	Fibre lasers *David Hanna*	963
	B4.2	High power fiber lasers *Andreas Tünnermann and Holger Zellmer*	977
	B4.3	Cascaded Raman fibre lasers *Clifford Headley*	989
	B4.4	Soliton lasers *J R Taylor*	1007
	B4.5	Erbium and other doped fibre amplifiers *Kevin Cordina*	1025
	B4.6	High-power waveguide lasers *D P Shepherd*	1045
	B5	Other lasers *Colin Webb*	1063
	B5.1	Free electron lasers and synchrotron light sources *P G O'Shea and J B Murphy*	1065
	B5.2	X-ray lasers *Jorge J Rocca*	1087
	B5.3	Liquid lasers *David H Titterton*	1115
	B5.4	Solid-state dye lasers *David H Titterton*	1143
PART C	**LASER SYSTEM DESIGN**		**1163**
	C1	Optical components *Julian Jones*	1165
	C1.1	Optical components *Leo H J F Beckmann*	1167
	C1.2	Optical control elements *Alan Greenaway*	1183
	C1.3	Adaptive optics and phase conjugate reflectors *Michael J Damzen and Carl Paterson*	1193

C1.4	Opto-mechanical parts *Frank Luecke*	1203
C1.5.1	Power conditioning: supplies for driving semiconductor laser diodes *Ralph Savioli*	1211
C1.5.2	Power conditioning: supplies for driving gas discharges (gas and solid state lasers) *I Smilanski*	1217
C1.5.3	Power conditioning: supplies for driving flash tubes and arclamps for solid state lasers *Mark Greenwood and D W Miller*	1237
C2	Optical pulse generation *Clive Ireland*	1247
C2.1	Quasi-cw and modulated beams *K Washio*	1249
C2.2	Short pulses *Andreas Ostendorf*	1257
C2.3	Ultrashort pulses *Derryck T Reid*	1273
C3	Frequency conversion and filtering *Terence A King*	1313
C3.1	Harmonic generation—materials and methods *David J Binks*	1315
C3.2	Optical parametric devices *M Ebrahimzadeh*	1347
C3.3	Laser stabilization for precision measurements *G P Barwood and P Gill*	1393
C4	Beam delivery *Julian Jones*	1415
C4.1	Basic principles *D P Hand*	1417
C4.2	Free-space optics *Leo H J F Beckmann*	1425
C4.3	Fibre optic beam delivery *D P Hand*	1461
C4.4	Positioning and scanning systems *Jürgen Koch*	1475
C5	Laser beam measurement *Julian Jones*	1499
C5.1	Beam propagation *B A Ward*	1501
C5.2	Detectors *Kenny Weir*	1509
C5.3	Laser energy and power measurement *Robert K Tyson*	1523
C5.4	Irradiance and phase distribution measurement *B Schäfer*	1527
C6	Laser safety *Colin Webb*	1535

C6.1 Laser safety 1537
J Michael Green and Karl Schulmeister

VOLUME III: APPLICATIONS

PART D APPLICATIONS: CASE STUDIES 1557

D1 Materials processing 1559
Clive Ireland
D1.1 Welding 1561
H Hügel and C Schinzel
D1.2 Cutting 1587
John Powell and Claes Magnusson
D1.3 Laser marking 1613
Terry J McKee
D1.4 Drilling 1633
S Williams
D1.5 Photolithography 1653
Shinji Okazaki
D1.6 Laser micromachining 1661
Malcolm Gower
D1.7 Rapid manufacturing 1693
Gary K Lewis
D1.8 Pulsed laser deposition of thin films 1705
Ian Boyd and D B Chrisey
D2 Optical measurement techniques 1721
Julian Jones
D2.1 Fundamental length metrology 1723
J Flügge, F Riehle and H Kunzmann
D2.2 Laser velocimetry 1749
C Tropea
D2.3 Laser vibrometers 1779
Neil A Halliwell
D2.4 Electronic speckle pattern interferometry (ESPI) 1805
Dave Towers and Clive Buckberry
D2.5 Optical fibre hydrophones 1839
Geoffrey A Cranch and Philip J Nash
D2.6 Optical fibre Bragg grating sensors for strain measurement 1881
David A Jackson and David J Webb
D2.7 High-speed imaging 1919
Adam Whybrew
D2.8 Particle sizing 1931
Nils Damaschke, Maurice Wedd, Adam Whybrew and Damien Blondel
D3 Medical 1951
Terence A King and Brian C Wilson
D3.1 Light–tissue interactions 1955
Steven Jacques and Michael Patterson

D3.2	Therapeutic applications: introduction *Reginald Birngruber*	1995
D3.2.1	Therapeutic applications: ophthalmology *Reginald Birngruber*	1999
D3.2.2	Therapeutic applications: refractive surgery *Giovanni Cennamo and Raimondo Forte*	2009
D3.2.3	Therapeutic applications: photodynamic therapy *Brian C Wilson and Stephen G Bown*	2019
D3.2.4	Therapeutic applications: thermal treatment of tumours *Stephen G Bown*	2037
D3.2.5	Therapeutic applications: dermatology—selective photothermolysis *Sean Lanigan*	2045
D3.2.6	Therapeutic applications: lasers in vascular surgery *Mahesh Pai*	2055
D3.2.7	Therapeutic applications: hardtissue/dentistry *Raimund Hibst*	2065
D3.2.8	Therapeutic applications: free-electron laser *E Duco Jansen, Michael Copeland, Glenn S Edwards, William Gabella, Karen Joos, Mark A Mackanos, Jin H Shen and Stephen R Uhlhorn*	2075
D3.3	Medical diagnostics *Brian C Wilson*	2087
D3.4	Laser applications in biology and biotechnology *Sebastian Wachsmann-Hogiu, Alexander J Annala and Daniel L Farkas*	2123
D3.5	Biomedical laser safety *Harry Moseley and Bill Davies*	2155
D4	Communications *John Marsh*	2181
D4.1	The basic point-to-point communications system *John Gowar*	2183
D4.2	High-capacity optical transmission systems *Paul Urquhart*	2231
D4.3	Local area networks *J Lehman and K L Johnson*	2289
D4.4	Fibre-to-the-chip: development of vertical cavity surface emitting laser arrays designed for integration with VLSI circuits *A V Krishnamoorthy, L M F Chirovsky, K W Goosen, J Lopata and W S Hobson*	2321
D4.5	Optical satellite communications *A Coello-Vera and M Maignan*	2345
D4.6	Smart pixel technologies and optical interconnects *Marc P Y Desmulliez and Brian S Wherrett*	2363
D5	Optical information storage *John Marsh*	2389
D5.1	Optical data storage *Tom D Milster*	2391
D5.2	Lasers in printing *Atsushi Kawamura, Seizo Suzuki and Yoshinori Hayashi*	2421

D6	Spectroscopy *Colin Webb*	2463
D6.1	Laser cooling and trapping *C S Adams and I G Hughes*	2465
D6.2	Ion trapping and laser applications to length and time metrology *P Gill and G P Barwood*	2485
D6.3	Time-resolved spectroscopy *Gavin D Reid and Klaas Wynne*	2507
D7	Earth and environmental sciences *Lance Thomas*	2529
D7.1	Satellite laser ranging *Roger Wood and Graham Appleby*	2531
D7.2	Lidar for atmospheric ozone remote sensing *Gérard Ancellet*	2563
D8	Lasers in astronomy *R C Powell*	2579
D8.1	Lasers in astronomy *Renaud Foy and Jean-Paul Pique*	2581
D9	Holography: holographic optical elements and computer-generated holography *Mohammad R Taghizadeh*	2625
D9.1	Holography: holographic optical elements—computer-generated holography—diffractive optics *Hans Peter Herzig*	2627
D10	High-intensity lasers for plasma studies *Colin Webb*	2643
D10.1	High-power lasers for plasma physics *M H R Hutchinson*	2645
D10.2	High-power lasers and the extreme conditions that they can produce *S J Rose*	2657
	Index	2665

Editorial and Scientific Advisory Board

Editor-in-Chief

Colin Webb

Clarendon Laboratory,
University of Oxford,
Oxford,
UK

Executive Editor

Julian Jones

Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

Editorial Board

Clive Ireland

Advanced Optical Technology Ltd,
Basildon,
UK

John Marsh

Intense Photonics Ltd,
High Blantyre,
UK

Minoru Obara

Department of Electronics and Electrical Engineering,
Keio University,
Japan

Richard Powell

Optical Sciences Center,
University of Arizona,
Tucson, AZ,
USA

Richard Shoemaker

Optical Sciences Center,
University of Arizona,
Tucson, AZ,
USA

Ian White

Cambridge University,
Cambridge,
UK

Advisory Board

Walter Goethels

Diamond Tools Group BV,
The Netherlands

Mike Green

Association of Industrial Laser Users,
Abingdon,
UK

Denis Hall

Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

Terry King

Department of Physics and Astronomy,
University of Manchester,
Manchester,
UK

Mohammed Taghizadeh

Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

Lance Thomas

Department of Physics,
University of Wales,
Aberystwyth,
UK

Kunihiko Washio

Control Systems Operations Unit,
NEC Corporation,
Sagamihara,
Japan

Brian Wilson

Ontario Cancer Institute,
Toronto, Ontario,
Canada

List of contributors

C S Adams (D6.1)
Department of Physics,
University of Durham,
Durham,
UK

M-C Amann (B2.2)
Walter Schottky Insitut,
Technische Universität München,
Garching,
Germany

G Ancellet (D7.2)
Service d'Aeronomie du CNRS,
Université Paris 6,
Paris,
France

S Anders (B2.7)
Institut für Festkörperelektronik,
Technische Universität Wien,
Vienna,
Austria

A J Annala (D3.4)
Department of Surgery and Minimally Invasive Surgical Technologies Institute,
Cedars-Sinai Medical Center,
Los Angeles, CA,
USA

G Appleby (D7.1)
NERC Space Geodesy Facility,
Hailsham,
UK

B D Barmashenko (B3.4.1)
Department of Physics,
Ben-Gurion University of the Negev,
Beer-Sheva,
Israel

N P Barnes (B1.4)
NASA Langley Research Center,
Hampton, VA,
USA

G P Barwood (C3.3, D6.2)
National Physical Laboratory,
Teddington,
UK

T T Basiev (B1.7, B1.8)
Laser Materials and Technology Research Center,
General Physics Institute,
Moscow,
Russia

L H J F Beckmann (C1.1, C4.2)
Delft,
The Netherlands

D J Binks (C3.1)
Department of Physics and Astronomy,
University of Manchester,
Manchester,
UK

R Birngruber (D3.2, D3.2.1)
Medizinisches Laserzentrum Lübeck,
Lübeck,
Germany

D Blondel (D2.8)
Dantec Dynamics A/S,
Skovlunde,
Denmark

P Blood (B2.5.1)
Department of Physics and Astronomy,
Cardiff University,
Cardiff,
UK

G Boulon (B1.1)
Physico Chimie des Matériaux Luminescents,
Université Claude Bernard Lyon 1,
Villeurbanne,
France

S G Bown (D3.2.3, D3.2.4)
National Medical Laser Centre,
University College London,
London,
UK

I Boyd (D1.8)
Department of Electronic and Electrical Engineering,
University College London,
London,
UK

R W Boyd (A4)
Institute of Optics,
University of Rochester,
Rochester, NY,
USA

C Buckberry (D2.4)
Melles Griot Ltd,
Ely,
UK

G Buller (A7)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

J Byatt (A3)
Photonic Components Division,
Melles Griot,
Irvine, CA,
USA

G Cennamo (D3.2.2)
Department of Ophthalmology,
University of Naples Federico II,
Naples,
Italy

L M F Chirovsky (D4.4)
Bell Laboratories,
Holmdel, NJ,
USA

D B Chrisey (D1.8)
Naval Research Laboratory,
Washington DC,
USA

A Coello-Vera (D4.5)
Département Recherche Technologies Bord,
Alcatel Space,
Toulouse,
France

M Copeland (D3.2.8)
Department of Neurosurgery,
Vanderbilt University,
Nashville, TN,
USA

K Cordina (B4.5)
Nortel Networks plc,
Harlow,
UK

G A Cranch (D2.5)
Optical Techniques Branch,
Naval Research Laboratory,
Washington DC,
USA

N Damaschke (D2.8)
Fachgebiet Strömungslehre und Aerodynamik,
Technische Universität Darmstadt,
Darmstadt,
Germany

M J Damzen (C1.3)
The Blackett Laboratory,
Imperial College,
London,
UK

W M Davies (D3.5)
Medical Physics and Bioengineering Department,
Singleton Hospital,
Swansea,
UK

C C Davis (A1)
Department of Electrical and Computer Engineering,
University of Maryland,
College Park, MD,
USA

M P Y Desmulliez (D4.6)
School of Engineering and Physical Sciences,
Heriot-Watt University,
Edinburgh,
UK

M H Dunn (B3.5)
School of Physics and Astronomy,
University of St Andrews,
St Andrews,
UK

N K Dutta (B2.1)
Department of Physics,
University of Connecticut,
Storrs, CT,
USA

M Ebrahimzadeh (C3.2)
ICFO—Institut de Ciencies Fotoniques,
Barcelona,
Spain

R C Eckardt (A2.1)
Cleveland Crystals Inc.,
Highland Heights, OH,
USA

G S Edwards (D3.2.8)
Department of Physics and FEL Laboratory,
Duke University,
Durham, NC,
USA

D L Farkas (D3.4)
Department of Surgery and Minimally Invasive Surgical Technologies Institute,
Cedars-Sinai Medical Center,
Los Angeles, CA,
USA

J Flügge (D2.1)
Physikalisch-Technische Bundesanstalt,
Braunschweig,
Germany

R Forte (D3.2.2)
Department of Ophthalmology,
University of Naples Federico II,
Naples,
Italy

R Foy (D8.1)
Observatoire de Lyon/CRAL,
Lyon,
France

W Gabella (D3.2.8)
W M Keck FEL Center,
Vanderbilt University,
Nashville, TS,
USA

P Gill (C3.3, D6.2)
National Physical Laboratory,
Teddington,
UK

K W Goosen (D4.4)
Bell Laboratories,
Holmdel, NJ,
USA

E Gornik (B2.7)
Institut für Festkörperelektronik,
Technische Universitt Wien,
Vienna,
Austria

J Gowar (D4.1)
Department of Electrical and Electronic Engineering,
University of Bristol,
Bristol,
UK

M Gower (D1.6)
Exitech Ltd,
Oxford,
UK

K T V Grattan (B2.6)
School of Engineering and Mathematical Sciences,
City University,
London,
UK

J M Green (C6.1)
Pro Laser,
Abingdon,
UK

A Greenaway (C1.2)
School of Engineering and Physical Sciences,
Heriot-Watt University,
Edinburgh,
UK

M Greenwood (C1.5.3)
Spectron Lasers,
Rugby,
UK

T Gutierrez (B3.5)
Coherent Inc.,
Santa Clara, CA,
USA

D R Hall (B3.1)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

N A Halliwell (D2.3)
Wolfson School of Mechanical and Manufacturing Engineering,
Loughborough University,
Loughborough,
UK

D P Hand (C4.1, C4.3)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

D Hanna (B4.1)
Optoelectronics Research Centre,
University of Southampton,
Southampton,
UK

Y Hayashi (D5.2)
Imaging Technology Development Department,
RICOH Company Ltd,
Tokyo,
Japan

C Headley (B4.3)
Photonic Devices Research Department,
OFS Labs,
Somerset, NJ,
USA

H P Herzig (D9.1)
Institut de Microtechnique,
Université de Neuchâtel,
Neuchâtel,
Switzerland

R Hibst (D3.2.7)
Institut für Lasertechnologien in der Medizin und Messtechnik,
Ulm,
Germany

C Hill (A2.2)
QinetiQ,
Malvern,
UK

W S Hobson (D4.4)
Bell Laboratories,
Murray Hill, NJ,
USA

H Hügel (D1.1)
Institut für Strahlwerkzeuge (IFSW),
Universität Stuttgart,
Stuttgart,
Germany

I G Hughes (D6.1)
Department of Physics,
University of Durham,
Durham,
UK

M H R Hutchinson (D10.1)
Central Laser Facility,
Rutherford Appleton Laboratory,
Chilton,
UK

C Ireland (C2, D1)
Advanced Optical Technology Ltd,
Basildon,
UK

D A Jackson (D2.6)
School of Physical Sciences,
University of Kent,
Canterbury,
UK

S Jacques (D3.1)
Oregon Medical Laser Center,
Portland, OR,
USA

E D Jansen (D3.2.8)
Department of Biomedical Engineering,
Vanderbilt University,
Nashville, TN,
USA

K L Johnson (D4.3)
Honeywell VCSEL Products Division,
Plymouth, MN,
USA

J D C Jones (B3, C1, C4, C5, D2)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

K Joos (D3.2.8)
Department of Ophthalmology,
Vanderbilt University,
Nashville, TN,
USA

H Kawaguchi (B2.8)
Faculty of Engineering,
Yamagata University,
Yonezawa,
Japan

A Kawamura (D5.2)
Imaging Technology Development Department,
RICOH Company Ltd,
Tokyo,
Japan

T A King (C3, D3)
Department of Physics and Astronomy,
University of Manchester,
Manchester,
UK

J Koch (C4.4)
BIAS—Bremen Institute of Applied Beam Technology,
Bremen,
Germany

A V Krishnamoorthy (D4.4)
Bell Laboratories,
Holmdel, NJ,
USA

H Kunzmann (D2.1)
Physikalisch-Technische Bundesanstalt,
Braunschweig,
Germany

S Lanigan (D3.2.5)
Lasercare Clinics,
City Hospital NHS Trust,
Birmingham,
UK

G Lawrence (A8)
Applied Optics Research,
Woodland, WA,
USA

J Lehman (D4.3)
Teradyne Connection Systems,
Nashua, NH,
USA

G K Lewis (D1.7)
Los Alamos National Laboratory,
Los Alamos, NM,
USA

J Lopata (D4.4)
Bell Laboratories,
Murray Hill, NJ,
USA

F Luecke (C1.4)
FrankDesign, LLC,
Crestwood, KY,
USA

M A Mackanos (D3.2.8)
Department of Biomedical Engineering,
Vanderbilt University,
Nashville, TN,
USA

C Magnusson (D1.2)
Division of Materials Processing,
Lulea University of Technology,
Lulea,
Sweden

M Maignan (D4.5)
Département Recherche Technologies Bord,
Alcatel Space,
Toulouse,
France

J Marsh (D4, D5)
Intense Photonics Ltd,
High Blantyre,
UK

R Martin (B2.5.2)
Department of Physics and Applied Physics,
University of Strathclyde,
Glasgow,
UK

T J McKee (D1.3)
Nepean, Ontario,
Canada

V A Mikhailov (B1.3)
General Physics Institute,
Russian Academy of Sciences,
Moscow,
Russia

D W Miller (C1.5.3)
GSI Lumonics,
Rugby,
UK

T D Milster (D5.1)
Optical Sciences Center,
The University of Arizona,
Tucson, AZ,
USA

S B Mirov (B1.8)
University of Alabama at Birmingham,
Birmingham, AL,
USA

H Moseley (D3.5)
The Photobiology Unit,
Ninewells Hospital,
Dundee,
UK

J B Murphy (B5.1)
National Synchrotron Light Source,
Brookhaven National Laboratory,
Brookhaven, NY,
USA

P J Nash (D2.5)
QinetiQ,
Winfrith Technology Centre,
Dorchester,
UK

S Okazaki (D1.5)

EUV Process Technology Research Department,
Association of Super-Advanced Electronics Technologies,
Tokyo,
Japan

P G O'Shea (B5.1)

Insitute for Research in Electronics and Applied Physics,
University of Maryland,
College Park, MD,
USA

A Ostendorf (C2.2)

Laser Zentrum Hannover e.V.,
Hannover,
Germany

M Pai (D3.2.6)

Royal United Hospital,
Bath,
UK

C Paterson (C1.3)

The Blackett Laboratory,
Imperial College,
London,
UK

M Patterson (D3.1)

McMaster University,
Hamilton, Ontario,
Canada

S A Payne (B1.2)

Lawrence Livermore National Laboratory,
University of California,
Livermore, CA,
USA

F Pellé (B1.6)

Groupe d'Optique des Terres Rares,
Ecole Nationale Supérieure de Chimie de Paris,
Paris,
France

J-P Pique (D8.1)

Laboratoire de Spectrométrie Physique,
Grenoble,
France

J Powell (D1.2)

Laser Expertise Ltd,
Nottingham,
UK

R C Powell (B1, B1.7, B4, D8)

Optical Sciences Center,
University of Arizona,
Tucson, AZ,
USA

W Prettl (B3.9)

Institut für Experimentelle und Angewandte Physik,
Universität Regensburg,
Regensburg,
Germany

G J Quarles (B1.5)

VLOC Inc.,
New Port Richey, FL,
USA

B M A Rahman (B2.6)

School of Engineering and Mathematical Sciences,
City University,
London,
UK

D T Reid (C2.3)

School of Engineering and Physical Sciences,
Heriot-Watt University,
Edinburgh,
UK

G D Reid (D6.3)

Department of Chemistry,
University of Leeds,
Leeds,
UK

F Riehle (D2.1)

Physikalisch-Technische Bundesanstalt,
Braunschweig,
Germany

J J Rocca (B5.2)

Department of Electrical and Computer Engineering,
Colorado State University,
Fort Collins, CO,
USA

A Rogers (A5)

Department of Electronic Engineering,
University of Surrey,
Guildford,
UK

S J Rose (D10.2)

Clarendon Laboratory,
University of Oxford,
Oxford,
UK

S Rosenwaks (B3.4.1)
Department of Physics,
Ben-Gurion University of the Negev,
Beer-Sheva,
Israel

R Savioli (C1.5.1)
Thorlabs Inc.,
Newton, NJ,
USA

B Schäfer (C5.4)
Laser-Laboratorium Göttingen,
Göttingen,
Germany

C Schinzel (D1.1)
Institut für Strahlwerkzeuge (IFSW),
Universität Stuttgart,
Stuttgart,
Germany

K Schulmeister (C6.1)
LS Strahlenschutz,
Austrian Research Centre Siebersdorf,
Siebersdorf,
Austria

L H Sentman (B3.4.2)
Department of Aeronautical and Astronautical Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL,
USA

I A Shcherbakov (B1.3)
General Physics Insitute,
Russian Academy of Sciences,
Moscow,
Russia

J H Shen (D3.2.8)
Department of Ophthalmology,
Vanderbilt University,
Nashville, TN,
USA

D P Shepherd (B4.6)
Optoelectronics Research Centre,
University of Southampton,
Southampton,
UK

R Shoemaker (A)
Optical Sciences Center,
University of Arizona,
Tucson, AZ,
USA

W T Silfvast (B3.7)
School of Optics/CREOL,
University of Central Florida,
Orlando, FL,
USA

I Smilanski (C1.5.2)
Cambridge, MA,
USA

J Smith (A7)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

G Stewart (A6)
Department of Electronic and Electrical Engineering,
University of Strathclyde,
Glasgow,
UK

G Strasser (B2.7)
Institut für Festkörperelektronik,
Technische Universität Wien,
Vienna,
Austria

S Suzuki (D5.2)
Imaging Technology Development Department,
RICOH Company Ltd,
Tokyo,
Japan

M R Taghizadeh (D9)
Department of Physics,
Heriot-Watt University,
Edinburgh,
UK

J R Taylor (B4.4)
Physics Department,
Imperial College,
London,
UK

L Thomas (D7)
Department of Physics,
University of Wales,
Aberystwyth,
UK

D H Titterton (B5.3, B5.4)
Sensors Department,
DSTL,
Farnborough,
UK

M S Tobin (B3.8)
AMSRL-SE-EI,
Army Research Laboratory,
Adelphi, MD,
USA

D Towers (D2.4)
Mechanical and Chemical Engineering,
Heriot-Watt University,
Edinburgh,
UK

C Townes (Foreword)
Department of Physics,
University of California,
Berkeley, CA,
USA

C Tropea (D2.2)
Fachgebiet Strömungslehre und Aerodynamik,
Technische Universität Darmstadt,
Darmstadt,
Germany

L Tsufura (B3.6)
Melles Griot Inc.,
Carlsbad, CA,
USA

A Tünnermann (B4.2)
Institut für Angewandte Physik,
Friedrich-Schiller-Universität Jena,
Jena,
Germany

R K Tyson (C5.3)
Department of Physics,
University of North Carolina at Charlotte,
Charlotte,
USA

S R Uhlhorn (D3.2.8)
Department of Biomedical Engineering,
Vanderbilt University,
Nashville, TN,
USA

P Unger (B2.4)
Department of Optoelectronics,
University of Ulm,
Ulm,
Germany

P Urquhart (D4.2)
Département Optique,
ENST Bretagne,
Brest,
France

P P Vasil'ev (B2.3)
P N Lebedev Physical Insitute,
Russian Academy of Sciences,
Moscow,
Russia

S Wachsmann-Hogiu (D3.4)
Department of Surgery and Minimally Invasive Surgical Techniques Institute,
Cedars-Sinai Medical Center,
Los Angeles, CA,
USA

B A Ward (C5.1)
Europtics,
Reading,
UK

K Washio (C2.1)
Control Systems Operations Unit,
NEC Corporation,
Sagamihara,
Japan

C E Webb (Introduction, B3.3, B5, C6, D6, D10)
Clarendon Laboratory,
University of Oxford,
Oxford,
UK

D J Webb (D2.6)
School of Engineering and Applied Science,
Aston University,
Birmingham,
UK

M Wedd (D2.8)
Malvern Instruments Ltd,
Malvern,
UK

K Weir (5.2)
Department of Physics,
Imperial College,
London,
UK

B S Wherrett (D4.6)
School of Engineering and Physical Sciences,
Heriot-Watt University,
Edinburgh,
UK

A D White (B3.6)
Berkeley Heights, NJ,
USA
(Formerly of ATT Bell Labs,
Murray Hill, NJ,
USA)

I White (B2)

Cambridge University,
Cambridge,
UK

A Whybrew (D2.7, D2.8)

Oxford Lasers Pacific,
Melbourne,
Australia

S Williams (D1.4)

Sowerby Research Centre,
British Aerospace,
Bristol,
UK

B C Wilson (D3, D3.2.3, D3.3)

Ontario Cancer Institute,
Toronto, Ontario,
Canada

W J Witteman (B3.2)

University of Twente,
Enschede,
The Netherlands

R Wood (D7.1)

NERC Space Geodesy Facility,
Hailsham,
UK

K Wynne (D6.3)

Department of Physics,
Strathclyde University,
Glasgow,
UK

A I Zagumennyi (B1.3)

General Physics Institute,
Russian Academy of Sciences,
Moscow,
Russia

H Zellmer (B4.2)

Institut für Angewandte Physik,
Friedrich-Schiller-Universität Jena,
Jena,
Germany

P G Zverev (B1.8)

Laser Materials and Technology Research Center,
General Physics Institute,
Moscow,
Russia

Foreword

This remarkable Handbook covers the broad field of laser science and technology. It does so with clarity so that it can be useful to individuals with a wide variety of backgrounds, and with care and completeness so that those interested in specialized applications of lasers should also find it useful. It is a large handbook, more than 130 chapters and 170 expert authors—too large for a briefcase, but so complete and useful that it may be a necessity for any scientific or technical library.

When Arthur Schawlow and I first completed a description of how a laser might be made and properties of its radiation, many scientific friends teased me by questioning whether it would be of any use. It clearly married optics and electronics, both of which have wide applications, but as a new breakthrough most of its virtues and applications weren't easily visualized by the technical community. By now, these friends and others have recognized and developed a wealth of striking applications to both technology and science. For science, lasers have provided powerful and precise new scientific tools which have already been used in a number of Nobel Prize winning discoveries. In technology, the variety of applications and new possibilities opened up are now stunning. Laser applications certainly have further to go, but the community of scientists and engineers have developed the field rapidly, so that very likely a large fraction of the ultimate uses are at least visualized even if not yet fully developed and exploited. In my view, the field is in its adolescence, i.e. a stage where its potential and possibilities are evident, but there is still much to come. The present Handbook provides the information and background both for uses and for further development of the field.

Lasers give us exceptional precision in the measurement of distances and of time. They provide very sensitive tools for control, flexible operations, and new manufacturing methods. They also provide fantastic power and energy intensity essentially more than any other available source. They cover a wide range of wavelengths, from infrared to x-rays, representing a wavelength ratio of about one million. The bandwidth available for information transmission is in principle large enough to put all present communications on a single beam. Laser pulses can be so short that they have measured the vibrations of molecules and fast molecular reactions. Radiation intensity produced by lasers brings us easily into the field of nonlinear optics, where photons interact with each other and produce a wide range of new phenomena. Today's lasers also come in many varieties—from small dots to systems the size of a large building, with costs varying from 'peanuts' to billions of dollars.

The technical fields strongly impacted by lasers are wide-ranging, from medicine, to communications, manufacturing, (cutting or shaping the hardest materials or the most delicate), production of temperatures from the coolest ever achieved to the hottest, and with still more to come. Lasers provide delicate and gentle control of molecules or micro-organisms but also create some of our hottest and most violent media. They range from everyday use at the counter of grocery stores or in the classroom to some of the most esoteric and advanced science and technology. Laser technology is thus already a large and varied industry of importance both to scientific discovery and to our daily lives. It will continue to grow, and the present Handbook can play an important role in this growth.

This Handbook covers essentially all known areas of technical applications of lasers, using simple and understandable language, but also sophistication and understanding which its many specialists have been able to provide. It should be useful both to those who want to explore the field or some related idea for the first time, and those looking for advanced and rigorous discussion along with references to scientific and technical papers on aspects of special interest.

I am delighted to see this excellent and useful summary of laser technology. I expect it to be much welcomed by, and useful to, the technical community and a significant factor in further generation of ideas and growth of the field.

Charles Townes

Introduction

The invention of the laser must surely be regarded one of the towering achievements of the 20th century and yet, for the first 10 years, many called it 'a solution in search of a problem'. The sheer amount of knowledge encapsulated in the three volumes that make up this Handbook, describing the applications and technologies that lasers have enabled, shows just how mistaken that early judgement was.

Laser technology touches almost every aspect of daily life in the 21st Century. From telecommunications to data storage, and everything from supermarket bar codes to eye surgery, much of modern technology depends on the capabilities that lasers have made possible.

The very diversity and ubiquity of lasers and their applications, evident from the scope of these three volumes, frustrates any attempt at a comprehensive editorial introduction to the subject. Every member of the professional laser community will have their own perspective, and I hope that the reader will forgive my presentation of a personal view.

My own research career started shortly after Schawlow and Townes published their 1959 paper, with its prediction that MASER[1] action should be possible at optical frequencies. I was the first graduate student outside the USA to start on a doctorate in the field of Optical Masers (the word LASER had not yet been defined, even as an acronym). I remember being asked by other new graduate students at the Clarendon Lab in Oxford in 1960 'What use will this optical maser be—even if it works?' My response was to repeat the words of my supervisor John Sanders—something along the lines of 'It might be used for fundamental length standards, or perhaps for transmitting messages by sending a beam from an optical maser to a receiving station'. As it turned out both these predictions have come true and both form subjects of chapters in this Handbook (chapter D2.1 for fundamental metrology and section D4 for communications). However, I don't think anyone could have foreseen the myriad of laser types and applications that have sprung from the device which in 1960 was often described as a mere scientific curiosity. It has been my privilege to spend my entire research career as both spectator and participant in the unfolding of this story over the past forty three years.

In compiling this Handbook Julian Jones and I, together with our very distinguished international board of subject editors, have sought to provide a balanced coverage of the many topics which can be described as belonging to 'Laser Technology and Applications'. Our objective has been to provide a reference work which will have enduring and practical value for those who are newcomers to the subject as well as experts and practitioners in the field who wish to extend their knowledge.

Of course, no work such as this can ever claim to be complete or final; the subject is constantly growing and evolving. However, we have tried to pull together those aspects that seem to us likely to form the fundamental building blocks for new developments. We hope that the contents will be as relevant to future researchers as to engineers and managers of technological enterprises today.

For the new researcher entering the field, the book will be especially valuable because of the extensive references to current literature accompanying each of the articles. This will provide the key to tracing topics back to their origins, and finding out the background and status of recent developments. In one way then, today's novices are to be regarded as fortunate to have such a wealth of information at their disposal—although I cannot help feeling that in some ways my own situation was somewhat simpler when I started when all there was to read was Born's *Optik*, Mitchell and Zemanski's *Resonance Radiation and Excited Atoms* and of

[1] 'Microwave amplification by stimulated emission of radiation'.

course the Schawlow and Townes paper. However the field is now so huge and there is no lack of challenging problems for the enterprising young researcher to take on.

Bringing this book from concept to fruition has required a great deal of effort by a large number of people. It is a pleasure to acknowledge the immense amount of help and guidance given by the members of the Editorial Board: Clive Ireland, John Marsh, Minaru Obara, Richard Powell, Richard Shoemaker and Ian White—and the members of the Advisory Board: Walther Goethals, Mike Green, Mohammed Tagizadeh, Lance Thomas, Kunihiro Washio, Godfrey Beddard, Denis Hall, Terry King and Brian Wilson. Without the rich pool of wisdom and experience they provided the task would have been impossible. I am also personally indebted to Steve Bown, Joe Taylor, Henry Hutchinson, Steve Rose, Patrick Gill, and Giovanni Cennamo for providing valuable contributions at very short notice when gaps in coverage became apparent at late stages in the preparation of the text.

The Commissioning Editor at IOP Publishing who started off this project was Nikki Dennis, handing over to David Morris who carried the project through its most substantial stages; but it has been finally brought to completion by Karen Donnison and John Navas, and taken into production by Sarah Plenty. It is a pleasure to thank all the members of the IOPP staff who have been concerned with the project—they have had to deal with the input from 170 authors. However, the member of IOPP staff who bore the main brunt of ensuring that the project did not abort throughout its three years of gestation was David Morris to whose patience, persistence and unfailing politeness these volumes are a testimony.

Finally my thanks above all are due to Julian Jones who, despite his heavy responsibilities of Head of Department at Heriot-Watt University, has undertaken the job of Executive Editor from the very beginning of the project. His dedication, unflappable attitude and abiding good humour, as well as his wealth of experience, allowed us to deal with the several crises that arose without our falling out even once.

Colin Webb
Editor-in-Chief
The Clarendon Laboratory, Oxford, June 2003

PART A

PRINCIPLES

A
Principles

Richard Shoemaker

Since the operation of the first laser in 1960, literally hundreds of different laser varieties have been developed and the light that they produce is being used in thousands of applications ranging from precision measurement to materials processing to medicine. Underlying all this variety, however, is a small set of basic physical principles upon which laser operation, laser beam propagation and the interaction of laser beams with matter depend. The explanation of these principles is the subject of this section.

The first chapter begins by explaining the basic physics that allow one to construct optical amplifiers, including discussions of energy levels and level populations, stimulated and spontaneous emission, optical lineshapes and gain saturation. It then discusses the principles that allow an optical amplifier to be turned into a laser (i.e. an optical oscillator) by the addition of feedback in the form of an optical resonator. The article closes with a discussion of the physics that determine the linewidth, coherence properties and power of the laser output.

The frequency and spatial distribution of a beam produced by a laser are largely determined by the laser resonator and, as a result, an understanding of optical resonators and their modes is key to understanding the properties of laser beams. Chapter A2.1 presents the principles of Gaussian beams, stable resonators, stable resonator axial and transverse modes, beam quality, mode matching, plane parallel resonators, unstable resonators and frequency selection. Chapter A2.2 supplements this material by discussing the principles governing hollow waveguide optical resonators, widely used for carbon dioxide lasers.

The purpose for which most lasers are purchased or built is to make use of the laser beam that it produces. In many applications, making effective use of this beam requires that it be properly controlled in time (e.g. pulsed lasers), space (e.g. beam focusing), frequency or amplitude. Chapter A3 covers the principles used in laser beam control, including beam focusing with lenses, beam transmission through apertures, the M^2 value, transverse and axial mode control, frequency stabilization, frequency selection, astigmatic beam shaping, Q-switching, mode locking, cavity dumping and spatial filtering.

One of the key features that make lasers so useful is their ability to produce optical fields having very high intensity. When these fields interact with matter, a great variety of nonlinear optical effects can occur. Some of these, such as optical frequency doubling, can be very useful, while others, such as optical damage, cause problems that need to be controlled. Chapter A4 introduces the basic principles of nonlinear optics and the various mechanisms by which nonlinearities arise. It then goes on to discuss harmonic generation, optical parametric oscillators, phase conjugation, self-focusing, solitons, bistability, optical switching, stimulated light scattering, multi-photon absorption, optical damage, and strong-field effects such as high-order harmonic generation.

Many applications of lasers rely upon the fact that the light produced by most lasers is coherent and thus can exhibit strong interference effects. Usually, the light is also highly polarized, and this polarization can be utilized to good effect in many other applications. Chapter A5 covers the basic principles of coherent wave interference, Mach–Zehnder interferometers, Michelson interferometers, Fabry–Pérot interferometers and partial coherence. The discussion then moves to polarization concepts including the polarization ellipse,

crystal optics, retarding wave-plates, polarizing prisms, circular birefringence, polarization analysis, and applications of polarization optics, including electro-optic and magneto-optic effects.

Some of the most economically important applications of lasers rely upon our ability to confine laser beams within optical waveguides where they can be modulated, amplified, split, switched and recombined in ways similar to those used to manipulate currents in electronic circuits. These capabilities together with the ability to transmit the light over long distances through optical fibres with very low loss make optical communication systems possible. Chapter A6 covers the theory of optical waveguides and fibres. The chapter first introduces the primary types of waveguides and their fabrication, and then presents the basic theory of planar and 2D waveguides. The second part of the article turns to optical fibres, beginning with basic fibre propagation theory and then turning to a variety of important propagation effects including attenuation, dispersion, birefringence and polarization, nonlinear effects and mode coupling.

Many applications of lasers would be severely limited or impossible if we were unable to accurately and sensitively detect the energy or intensity of the beam with some type of optical detector. Chapter A7 presents basic descriptions and operating principles of photomultipliers, p–n phodiodes, Schottky and avalanche diode detectors, photoconductive detectors and thermal detectors, including bolometers and pyroelectric detectors. The final sections of the article discuss noise in photodetection, including detector figures of merit, noise sources, and methods of minimizing detector noise.

Chapter A8, the final article in this section, discusses the numerical modelling of laser beam propagation within and outside the laser resonator. These models are important tools used by optical engineers in designing laser systems and laser applications. The article begins by discussing the representation of optical beams for numerical work, followed by descriptions of specific methods for handling beam propagation: the split step method, finite difference propagation and angular spectrum propagation. Numerical calculations of propagation in homogeneous media including issues of sampling and propagation control are then presented, followed by an elementary discussion of propagation through gain and nonlinear media including the use of Beer's Law, rate equations, the Franz–Nodvik solution, refractive index effects and the inclusion of spontaneous emission. The last section of the article discusses the selection and validation of laser modelling software packages.

Excluded from the discussion of numerical modelling in chapter A8 is a treatment of numerical modelling for semiconductor lasers. Although obviously important, these lasers are by far the most difficult laser systems to model, and the development of software that can do such modelling is currently an active research topic at a number of universities and companies. The essential physics needed to model these lasers properly include the complex nonlinear interactions between the multi-component electron–hole plasma that produces the laser radiation, the intense laser radiation within the waveguide resonator, and the several layers of semiconductor materials that form the laser. As a consequence, the gain and refractive index cannot be represented in parametric form using the laser rate equations discussed in chapter A8. The gain peak and the gain lineshape both change on the fly with changes in internal carrier density and temperature, and electrical and heat transport from the external contacts into the active region of the p–i–n structure also critically influence the optical properties by modifying the optical gain and refractive index.

A1
Basic laser principles[1]

Christopher C Davis

A1.1 Introduction

A laser is an oscillator that operates at *optical* frequencies. These frequencies of operation lie within a spectral region that extends from the very far infrared to the *vacuum ultraviolet* (VUV) or soft x-ray region. At the lowest frequencies at which they operate, lasers overlap with the frequency coverage of masers, to which they are closely related, and millimetre wave sources using solid state or vacuum tube electronics, such as TRAPATT, IMPATT and Gunn diodes, klystrons, gyroklystrons and travelling wave tube oscillators, whose principles of operation are quite different [2]. In common with electronic circuit oscillators, a laser is constructed using an amplifier with an appropriate amount of positive feedback. The acronym LASER, which stands for *light amplification by stimulated emission of radiation* is in reality, therefore, a slight misnomer.

Of central importance are the fundamental processes that allow amplification at optical frequencies to be obtained. These processes use the energy that is involved when the discrete particles making up matter, specifically atoms, ions and molecules, move from one energy level to another. These particles have energies that can have only certain discrete values. This discreteness, or *quantization*, of energy is intimately connected with the duality that exists in nature. Light sometimes behaves as if it were a wave and in other circumstances it behaves as if it were composed of particles. These particles, called *photons*, carry the discrete packets of energy associated with the wave. For light of frequency ν, the energy of each photon is $h\nu$, where h is Planck's constant—6.6×10^{-34} J s. The energy $h\nu$ is the *quantum* of energy associated with the frequency ν. On an atomic scale, the amplification of light within a laser involves the emission of such quanta. Thus, the term *quantum electronics* is often used to describe the branch of science that has grown from the development of the maser in 1954 and the laser in 1960. The widespread use of lasers and other optical devices in practical applications such as communications, signal processing, imaging and data storage has also led to the use of the term *photonics*. Whereas electronics uses electrons in various devices to perform analogue and digital functions, photonics aims to replace the electrons with photons. Because photons have zero mass, do not interact with each other to any significant extent and travel at the speed of light, photonic devices promise small size and high speed.

A1.2 The amplifier–oscillator connection

In 'conventional' electronics, where by the word 'conventional' for the present purposes we mean frequencies where solid state devices such as transistors or diodes will operate, say below 10^{11} Hz, an oscillator is conveniently constructed by applying an appropriate amount of positive feedback to an amplifier. Such an arrangement is shown schematically in figure A1.1. The input and output voltages of the amplifier are V_i

[1] This chapter is based on a longer and more detailed exposition of these principles in [1].

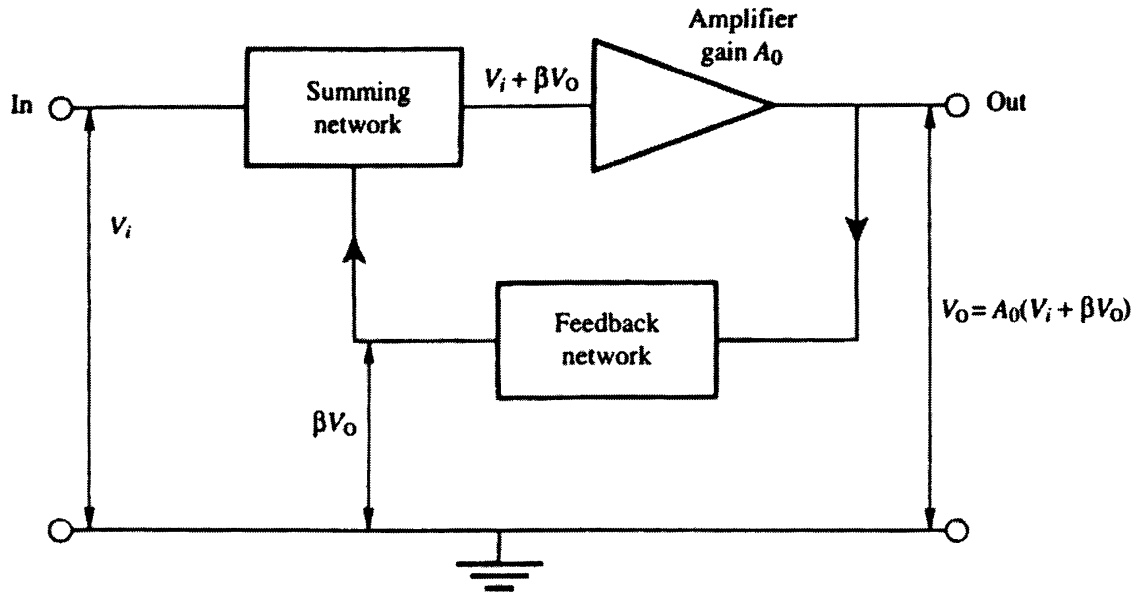


Figure A1.1. Circuit diagram of a simple amplifier with feedback.

and V_o respectively. The voltage gain of the amplifier is A_0 where, in the absence of feedback, $A_0 = V_o/V_i$. The feedback circuit returns part of the amplifier output to the input. The feedback factor $\beta = |\beta|e^{j\phi}$ is, in general, a complex number with amplitude $|\beta| \leq 1$ and phase ϕ.

$$V_o = A_0(V_i + \beta V_o) \tag{A1.1}$$

so

$$V_o = \frac{A_0 V_i}{1 - \beta A_0} \tag{A1.2}$$

and the overall voltage gain is

$$A = \frac{A_0}{1 - \beta A_0}. \tag{A1.3}$$

As βA_0 approaches $+1$, the overall gain of the circuit goes to infinity and the circuit would generate a finite output without any input. In practice, electrical 'noise', which is a random oscillatory voltage present to a greater or lesser extent in all electrical components in any amplifier system, provides a finite input. Because βA_0 is generally a function of frequency, the condition $\beta A_0 = +1$ is usually satisfied only at one frequency. The circuit oscillates at this frequency by amplifying the noise at this frequency present at its input. The output does not grow infinitely large, because as the signal grows, A_0 falls—this process is called saturation. This phenomenon is fundamental to all oscillator systems. A laser (or maser) is an optical (microwave) frequency oscillator constructed from an optical (microwave) frequency amplifier with positive feedback, shown schematically in figure A1.2. Light waves, which are amplified in passing through the amplifier, are returned through the amplifier by the reflectors and grow in intensity, but this intensity growth does not continue indefinitely because the amplifier saturates. The arrangement of mirrors (and sometimes other components) that provides the feedback is generally referred to as the laser cavity or resonator.

The characteristics of the device consisting of amplifying medium and resonator will be covered later, for the moment we must concern ourselves with the problem of how to construct an amplifier at optical frequencies, which range from 10^{11} Hz to beyond 10^{16} Hz. The operating frequencies of masers overlap this frequency range at the low-frequency end, the fundamental difference between the two devices is primarily one of scale. If the length of the resonant cavity which provides feedback is L, then for $L \gg \lambda$, where λ is

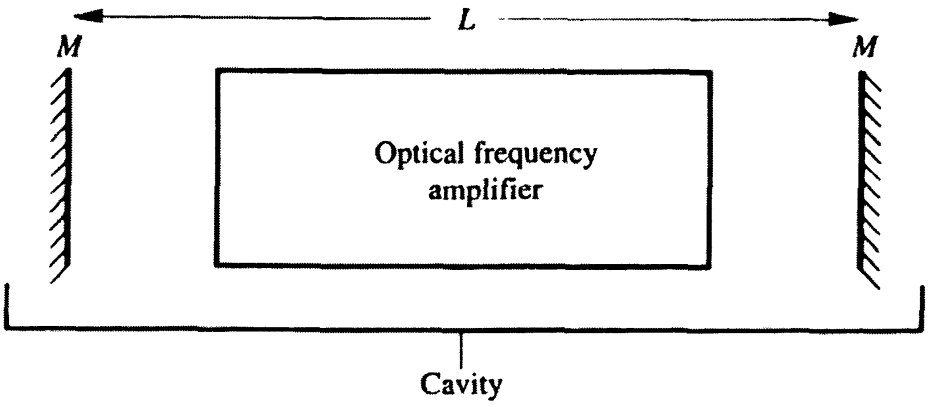


Figure A1.2. Schematic diagram of a basic laser structure with an amplifying medium in a resonant cavity formed by two feedback mirrors, M.

the wavelength at which oscillation occurs, we generally have a laser: for $L \simeq \lambda$ we usually have a maser, although the development of *microlasers*, which have small cavity lengths, has removed this easy way of distinguishing lasers from masers.

A1.3 The energy levels of atoms, molecules and condensed matter

All particles in nature have distinct states[2] that they can occupy. These states in general have different energies, although it is possible for particles in different states to have the same energy. The term 'energy level' is used to describe a particle with a specific, distinct energy, without implying any particular information about its (quantum) state. The lowest energy state in which a particle is stable is called the *ground state*. All higher energy states are called *excited* states. Excited states are intrinsically unstable and a particle occupying one will eventually lose energy and fall to lower energy states. When a particle falls from a higher energy state to a lower, energy is conserved. The energy ΔE lost by the particle can be emitted as a photon with energy $h\nu = \Delta E$: this is radiative energy loss. The particle can also lose energy *non-radiatively*, in which case the energy is dissipated into heating. Atomic systems have only electronic states, which in the simple Bohr model of the atom correspond to different configurations of electron orbits. The types of energy state that exist in a molecular system are more varied and include electronic, vibrational and rotational states.

In a molecule, changes in the internuclear separation of the constituent atoms give rise to *vibrational* energy states, which have quantized energies. The various characteristic vibrational motions of a molecule are called its normal modes, which for a molecule with N atoms number $3N$-6, unless the molecule is linear, in which case they number $3N$-5. The quantized energies of a normal mode can be written as [1]

$$E_{\text{vib}} \simeq (n + \tfrac{1}{2})h\nu_{\text{vib}} \tag{A1.4}$$

and form a ladder of (almost) equally spaced energy levels.

Molecules also have quantized rotational energy levels, whose energies can be written as

$$E_{\text{rot}} \simeq BJ(J + 1), \tag{A1.5}$$

where B is a rotational energy constant, and J is called the rotational quantum number. The overall energy state of a molecule thus has electronic, vibrational and rotational components. A molecule in a particular

[2] The term 'state' in quantum mechanics corresponds to a configuration with a particular 'state function', which often corresponds to a specific set of quantum numbers that identify the state.

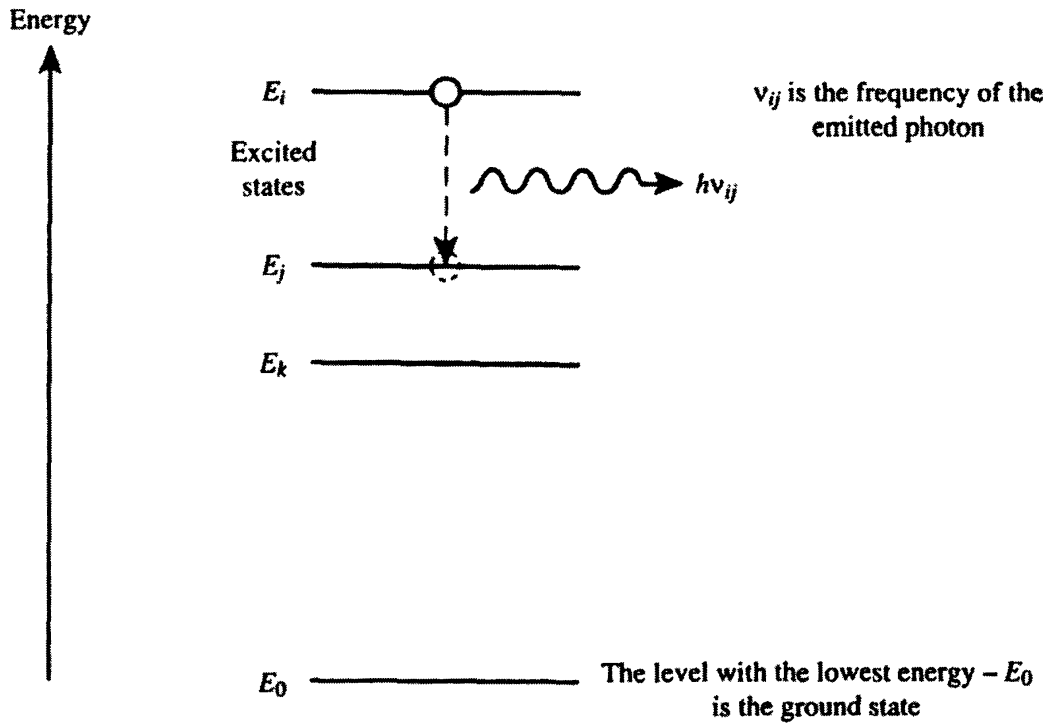


Figure A1.3. Simple energy level diagram for a particle.

combination of electronic and vibrational states is described as being in a *vibronic* state. A state with a specific combination of vibrational and rotational energies would be described as being in a *vibrot* state.

As a rough rule of thumb, transitions between different vibronic states where the electronic state changes lie in the visible spectrum with energy spacings[3] ~20 000 cm^{-1} and correspond to an energy spacing of 3×10^{10} h/λ J. Transitions between vibrot states where the electronic energy does not change but the vibrational state does are typically ~1000 cm^{-1}. Transitions between different rotational states where the electronic and vibrational states do not change are typically ~100 cm^{-1}. In practical terms, vibrational transitions are typically in the 3–20 μm range and rotational transitions are typically in the 50–1000 μm range.

In the gas phase, the energy levels of atoms or molecules are quite sharp and distinct, as shown schematically in figure A1.3, although we shall see later that even these precise energies are 'broadened'. This broadening occurs for several reasons but perhaps most importantly because of the interactions between neighbouring particles. In condensed matter, whether this be in the solid or liquid state, there are very many particles close to any individual particle of interest, and inter-particle interactions are strong. Consequently, the allowed energies of particles in the medium occupy broad, continuous ranges of energy called energy 'bands'. The lowest-lying energy band, which is analogous to the ground state of an isolated particle is called the *valence* band. The next highest band of allowed energies is called the *conduction* band. An energy band can be thought of as the result of very many sharp isolated energy states having their energies 'smeared' out so that they overlap. We will reserve further discussion of the energy bands in condensed matter until a little later and, for the moment, will consider the energy levels of particles as relatively sharp and not strongly influenced by inter-particle interactions.

[3] The cm^{-1} unit is often used to describe energy spacings. A transition at wavelength λ (cm) between two levels has an energy spacing characterized by $1/\lambda$ cm^{-1}.

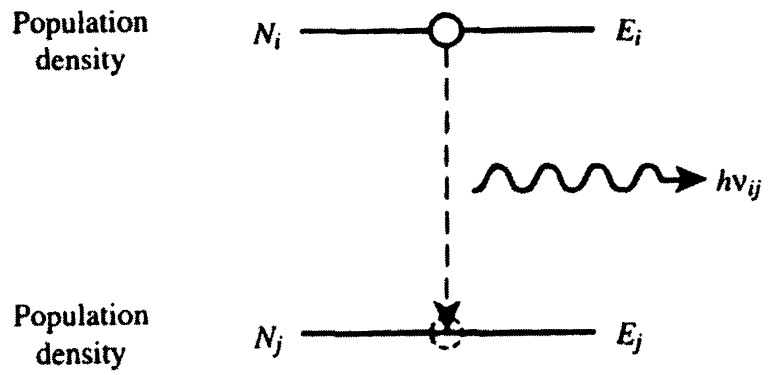


Figure A1.4. Representation of the spontaneous emission process for two levels of energy E_i and E_j.

A1.4 Spontaneous and stimulated transitions

To build an amplifier that operates at optical frequencies we use the energy delivered as the particles that constitute the amplifying medium make jumps between their different energy levels. The medium may be gaseous, liquid, a crystalline or glassy solid, an insulating material or a semiconductor. The particles of the amplifying medium, whether these are atoms, molecules or ions, can occupy only certain discrete energy levels. Consider such a system of energy levels, shown schematically in figure A1.3. Particles can make jumps between these levels in three ways. In the case of an atomic amplifier, these energy jumps involve electrons moving from one energy level to another.

A1.4.1 Spontaneous emission

When a particle spontaneously falls from a higher energy level to a lower one, as shown in figure A1.4, the emitted photon has frequency

$$\nu_{ij} = \frac{E_i - E_j}{h}. \tag{A1.6}$$

This photon is emitted in a random direction with arbitrary polarization (except in the presence of magnetic fields but this need not concern us here). The photon carries away momentum $h/\lambda = h\nu/c$ and the emitting particle (atom, molecule or ion) recoils in the opposite direction. The probability of a spontaneous jump within a small time interval is given quantitatively by the Einstein A coefficient defined by $A_{ij}\,\Delta t$ = 'probability' of a spontaneous jump from level i to level j during a short time interval Δt. A_{ij} has units of s^{-1}. To preserve the concept of $A_{ij}\,\Delta t$ as a true measure of the probability of a spontaneous emission, the time interval must be chosen so that $A_{ij}\,\Delta t \ll 1$.

For example, if there are N_i particles per unit volume in level i, then $N_i A_{ij}\,\Delta t$ make jumps to level j in a short time interval. The total rate at which jumps are made between the two levels is

$$\frac{\mathrm{d}N_{ij}}{\mathrm{d}t} = -N_i A_{ij}. \tag{A1.7}$$

There is a negative sign because the population of level i is decreasing.

Generally a particle can make jumps to more than one lower level, unless it is already in the first (lowest) excited level. The total probability that the particle will make a spontaneous jump to any lower level in a small time interval is $A_i\,\Delta t$ where

$$A_i = \sum_j A_{ij}. \tag{A1.8}$$

The summation runs over all levels j lower in energy than level i. The total rate at which the population of level i changes by spontaneous emission is

$$\frac{dN_i}{dt} = -N_i A_i \tag{A1.9}$$

which has the solution

$$N_i = N_i^0 e^{-A_i t} \tag{A1.10}$$

where N_i^0 is the population density of level i at time $t = 0$.

The population of level i falls exponentially with time as particles leave that level by spontaneous emission. The time in which the population falls to $1/e$ of its initial value is called the natural lifetime of level i, τ_i, where $\tau_i = 1/A_i$. The magnitude of this lifetime is determined by the actual probabilities of jumps from level i by spontaneous emission. Jumps which are likely to occur are called *allowed* transitions, those which are unlikely are said to be *forbidden. Allowed* transitions in the visible region typically have A_{ij} coefficients in the range 10^6–10^8 s^{-1}. *Forbidden* transitions in this region have A_{ij} coefficients below 10^4 s^{-1}. These probabilities decrease as the wavelength of the transition increases. Consequently, levels that can decay by allowed transitions in the visible have lifetimes generally shorter than 1 μs, similar forbidden transitions have lifetimes in excess of 10–100 μs. Although no jump turns out to be absolutely forbidden, some jumps are so unlikely that levels whose electrons can only fall to lower levels by such jumps are very long lived. Levels with lifetimes in excess of 1 h have been observed under laboratory conditions. Levels which can only decay slowly, and usually only by forbidden transitions, are said to be *metastable.*

A1.4.2 The lineshape function

When a particle loses energy spontaneously the emitted radiation is not, as might perhaps be expected, all at the same frequency. Real energy levels are not infinitely sharp, they are smeared out or *broadened.* A particle in a given energy level can actually have any energy within a finite range. The frequency spectrum of the spontaneously emitted radiation is described by the *lineshape function*, $g(\nu_0, \nu)$, where ν_0 is a reference frequency, usually the frequency where $g(\nu_0, \nu)$ has a maximum. The lineshape function is normalized so that

$$\int_{-\infty}^{\infty} g(\nu_0, \nu)\, d\nu = 1. \tag{A1.11}$$

$g(\nu_0, \nu)\, d\nu$ represents the probability that a photon will be emitted spontaneously in the frequency range $\nu + d\nu$. The lineshape function $g(\nu_0, \nu)$ is a true probability function for the spectrum of emitted radiation and is usually sharply peaked near the frequency ν_0, as shown in figure A1.5. Since negative frequencies do not exist in reality, the question might properly be asked: 'Why does the integral have a lower limit of minus infinity?' This is done because $g(\nu_0, \nu)$ can be viewed as the Fourier transform of a real function of time, so negative frequencies have to be permitted mathematically. In practice, $g(\nu_0, \nu_0)$ is only of significant value around a large value of ν_0 so

$$\int_{0}^{\infty} g(\nu_0, \nu)\, d\nu \simeq 1. \tag{A1.12}$$

The amount of radiation emitted spontaneously by a collection of particles can be described quantitatively by their *spectral radiant intensity* $I_e(\nu)$. The units of spectral radiant intensity are watts per hertz per steradian[4]. The total power (watts) emitted in a given frequency interval $d\nu$ is

$$W(\nu) = \int_S I_e(\nu)\, d\nu\, d\Omega, \tag{A1.13}$$

[4] The steradian is the unit of solid angle, Ω. The surface of a sphere encompasses a solid angle of 4π steradians.

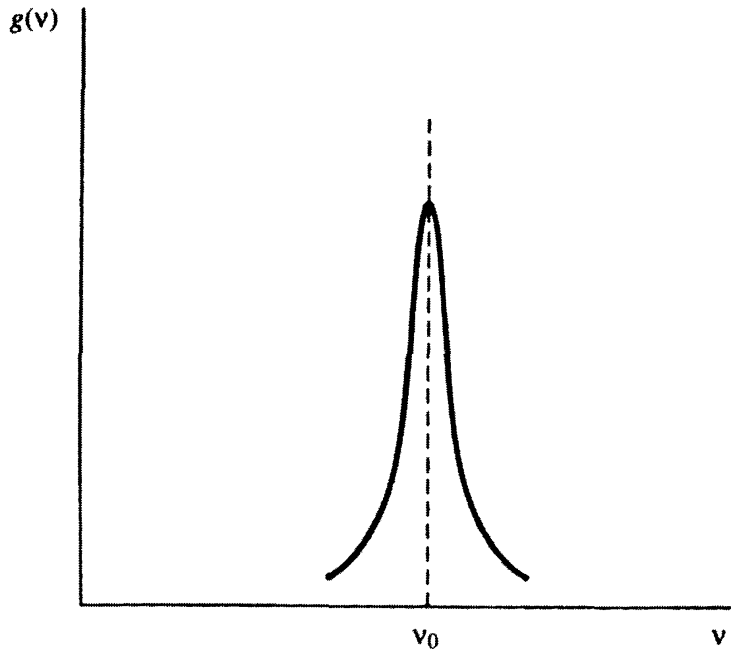


Figure A1.5. A lineshape function $g(\nu_0, \nu)$.

where the integral is taken over a closed surface S surrounding the emitting particles. The total power emitted is

$$W_0 = \int_{-\infty}^{\infty} W(\nu)\,\mathrm{d}\nu. \tag{A1.14}$$

$W(\nu)$ is closely related to the lineshape function

$$W(\nu) = W_0 g(\nu_0, \nu). \tag{A1.15}$$

For a collection of N_i identical particles, the total spontaneously emitted power per frequency interval (Hz) is

$$W(\nu) = N_i A_i h\nu g(\nu). \tag{A1.16}$$

Clearly this power decreases with time if the number of excited particles decreases.

For a plane electromagnetic wave we can introduce the concept of *intensity*, which has units of W m^{-2}. The intensity is the average amount of energy per second transported across a unit area in the direction of travel of the wave. The spectral distribution of intensity, $I(\nu)$, is related to the total intensity, I_0, by

$$I(\nu) = I_0 g(\nu_0, \nu). \tag{A1.17}$$

Although perfect plane waves do not exist, because such waves would have a unique propagation direction and infinite radiant intensity, they represent a useful, simple idealization. To a very good degree of approximation we can treat the light from a small source as a plane wave if we are far enough away from the source. The light coming from a star viewed outside the Earth's atmosphere is a good example.[5]

A1.4.3 Stimulated emission

As well as being able to make transitions from a higher level to a lower one by spontaneous emission, particles can also be stimulated to make these jumps by the action of an externally applied radiation field, as shown in figure A1.6.

[5] The plane waves coming from a star are distorted by the atmosphere because of density and refractive index fluctuations referred to as *atmospheric turbulence*.

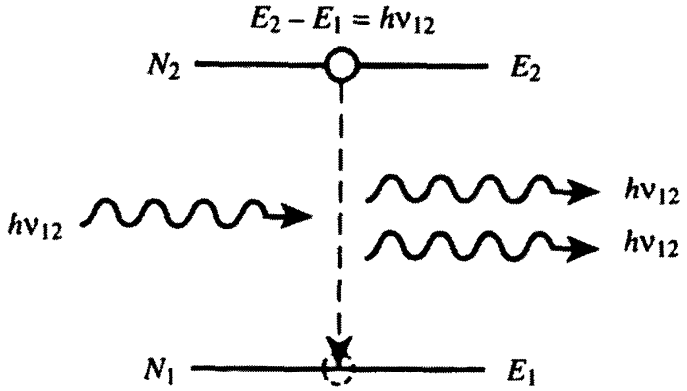


Figure A1.6. Representation of the stimulated emission process for two levels of energy E_2 and E_1.

The probability of the external radiation field causing stimulated emission depends on its energy density, which is written as $\rho(\nu)$ and is measured in J m^{-3} Hz^{-1}. The rate for stimulated emissions to occur within a small band of frequencies dν is

$$\frac{dN_2}{dt}(\nu)\,d\nu = N_2 B'_{21}(\nu)\rho(\nu)\,d\nu \text{ s}^{-1} \text{ m}^{-3}, \tag{A1.18}$$

where $B'_{21}(\nu)$ is a function specific to the transition between levels 2 and 1 and N_2 is the number of particles per unit volume in the upper level of the transition. Stimulated emission will occur if the external radiation field contains energy in a frequency range that overlaps the lineshape function. The frequency dependence of $B'_{21}(\nu)$ is the same as the lineshape function:

$$B'_{21}(\nu) = B_{21} g(\nu_0, \nu). \tag{A1.19}$$

B_{21} is called the Einstein B coefficient for stimulated emission. The total rate of change of particle concentration in level 2 by stimulated emission is

$$\begin{aligned}\frac{dN_2}{dt} &= -N_2 \int_{-\infty}^{\infty} B'_{21}(\nu)\rho(\nu)\,d\nu \\ &= -N_2 B_{21} \int_{-\infty}^{\infty} g(\nu_0, \nu)\rho(\nu)\,d\nu.\end{aligned} \tag{A1.20}$$

Note that, for the dimensions of both sides of equation (A1.20) to balance, B_{21} must have units m^3 J^{-1} s^{-2}. To evaluate the integral in equation (A1.20) we must consider how energy density is related to intensity and varies with frequency.

A1.4.4 The relation between energy density and intensity

The energy density of a radiation field $\rho(\nu)$ can be most easily related to its spectral intensity by examining the case of a plane electromagnetic wave. In figure A1.7 a plane wave propagating along carries energy across an area A oriented perpendicular to the direction of propagation. If the intensity of the wave is $I(\nu)$ W m^{-2} Hz^{-1}, then in one second the energy crossing A occupies a volume cA, where c is the velocity of light in the medium[6]. Clearly,

$$\rho(\nu) = \frac{I(\nu)}{c}. \tag{A1.21}$$

[6] $c = c_0/n$, where c_0 is the velocity of light in a vacuum and n is the *refractive index*.

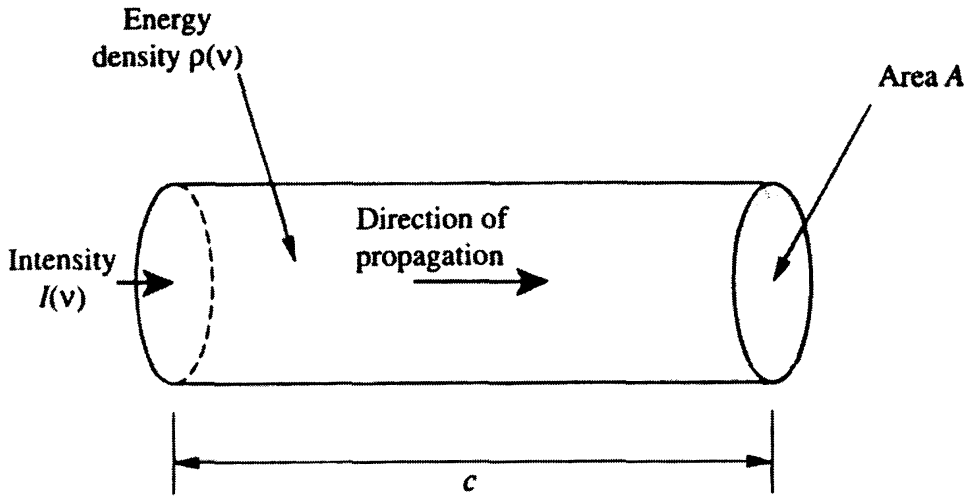


Figure A1.7. A volume of space swept through per second by part of a plane wave.

The energy density of a general radiation field $\rho(\nu)$ is a function of frequency ν. If $\rho(\nu)$ is independent of frequency, the radiation field is said to be *white*, as shown in figure A1.8. If the radiation field is *monochromatic* at frequency ν_{21}, its spectrum is as shown in figure A1.9. The ideal *monochromatic* radiation field has a δ-function energy density profile at frequency ν_{21}.

$$\rho(\nu) = \rho_{21}\delta(\nu - \nu_0). \tag{A1.22}$$

The δ-function has the property

$$\delta(\nu - \nu_{21}) = 0 \qquad \text{for } \nu \neq \nu_{21} \tag{A1.23}$$

and

$$\int_{-\infty}^{\infty} \delta(\nu - \nu_{21})\,\mathrm{d}\nu = 1. \tag{A1.24}$$

For a general radiation field the total energy stored per unit volume between frequencies ν_1 and ν_2 is $\int_{\nu_1}^{\nu} \rho(\nu_2)\,\mathrm{d}\nu$.

For a monochromatic radiation field, the total stored energy per unit volume is

$$\int_{-\infty}^{\infty} \rho(\nu)\,\mathrm{d}\nu = \int_{-\infty}^{\infty} \rho_{21}\delta(\nu - \nu_{21})\,\mathrm{d}\nu = \rho_{21}. \tag{A1.25}$$

The rate of stimulated emissions caused by a monchromatic radiation field can be calculated by using equation (A1.20) and is given by

$$\begin{aligned}\frac{\mathrm{d}N_2}{\mathrm{d}t} &= -N_2 B_{21} \int_{-\infty}^{\infty} g(\nu_0, \nu)\rho_{21}\delta(\nu - \nu_{21})\,\mathrm{d}\nu \\ &= -N_2 B_{21} g(\nu_0, \nu_{21})\rho_{21}.\end{aligned} \tag{A1.26}$$

It is very important to note that the rate of stimulated emissions produced by this input monochromatic radiation is directly proportional to the value of the lineshape function at the input frequency. The maximum rate of stimulated emission is produced, all else being equal, if the input radiation is at the line centre frequency ν_0.

If the stimulating radiation field has a spectrum that is broad, we can assume that the energy density $\rho(\nu)$ is constant over the narrow range of frequencies where $g(\nu_0, \nu)$ is significant. In this case equation (A1.20)

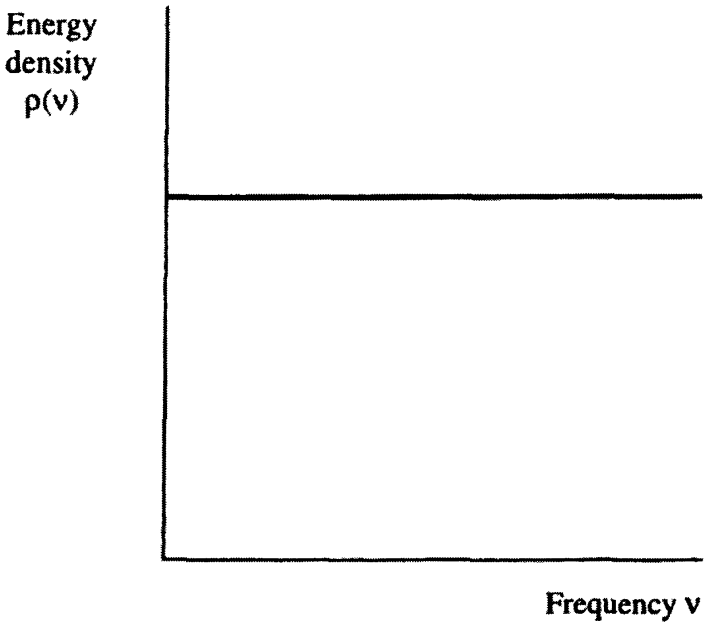


Figure A1.8. A 'white' energy density spectrum.

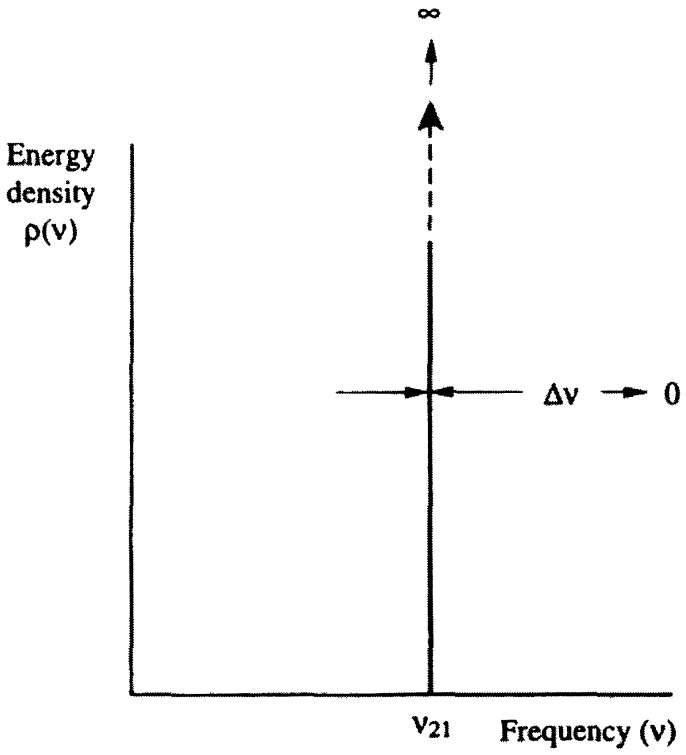


Figure A1.9. A monochromatic energy density spectrum.

gives

$$\frac{dN_2}{dt} = -N_2 B_{21} \rho(\nu) \tag{A1.27}$$

where $\rho(\nu) \simeq \rho(\nu_0)$ is the energy density in the frequency range where transitions take place.

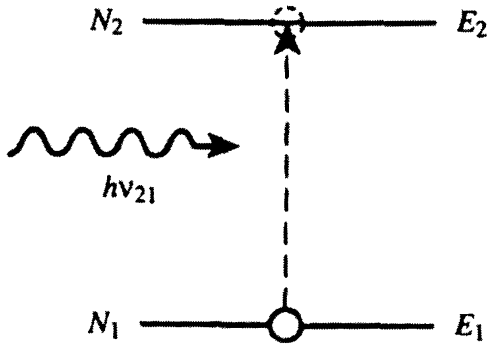


Figure A1.10. Representation of the stimulated absorption process for two levels of energy E_1 and E_2.

A1.4.5 Stimulated absorption

As well as making stimulated transitions in a downward direction, particles may make transitions in an upward direction between their energy levels by absorbing energy from an electromagnetic field, as shown in figure A1.10. The rate of such absorptions and the rate at which particles leave the lower level is $N_1\rho(\nu)B_{12}g(\nu_0, \nu)$ s^{-1} Hz^{-1} m^{-3}, which yields a result similar to equation (A1.20)

$$\frac{dN_1}{dt} = -N_1 B_{12} \int_{-\infty}^{\infty} g(\nu_0, \nu)\rho(\nu)\, d\nu. \tag{A1.28}$$

Once again B_{12} is a constant specific to the transition between levels 1 and 2 and is called the Einstein coefficient for stimulated absorption. Here again $\rho(\nu)$ is the energy density of the stimulating field. There is no analogue in the absorption process to spontaneous emission. A particle cannot spontaneously *gain* energy without an external energy supply. Thus, it is unnecessary for us to continue to describe the absorption process as stimulated absorption.

It is interesting to view both stimulated emission and absorption as photon–particle collision processes. In *stimulated* emission, the incident photon produces an identical photon by 'colliding' with the particle in an excited level, as shown in figure A1.11(*a*). After the stimulated emission process, both photons are travelling in the same direction and with the same polarization as the incident photon originally had. When light is described in particle terms, the polarization state describes the angular motion or spin of individual photons. Left- and right-hand circularly polarized light correspond in this particle picture to beams of photons that spin clockwise and counterclockwise, respectively, about their direction of propagation. Linearly polarized light corresponds to a beam of photons that has no net angular momentum about an axis parallel to their direction of propagation. In stimulated emission the stimulated photon has exactly the same frequency as the stimulating photon. In absorption the incident photon disappears, as shown in figure A1.11(*b*). In both stimulated emission and absorption, the particle recoils to conserve linear momentum.

A1.5 Transitions between energy levels for a collection of particles in thermal equilibrium

A collection of particles in thermal equilibrium is described by a common temperature T K. Although the collection of particles is described as being in 'equilibrium' this is a dynamic equilibrium. The processes of spontaneous emission, stimulated emission and absorption continuously occur because there is always a radiation field present. Even though no external radiation field is supplied, the thermal background radiation is always present. This radiation is called *black body* radiation and constantly interacts with each particle. Black-body radiation is so called because of the special characteristics of the radiation emitted and absorbed by a *black body*. Such a body absorbs with 100% efficiency all the radiation falling on it, irrespective of

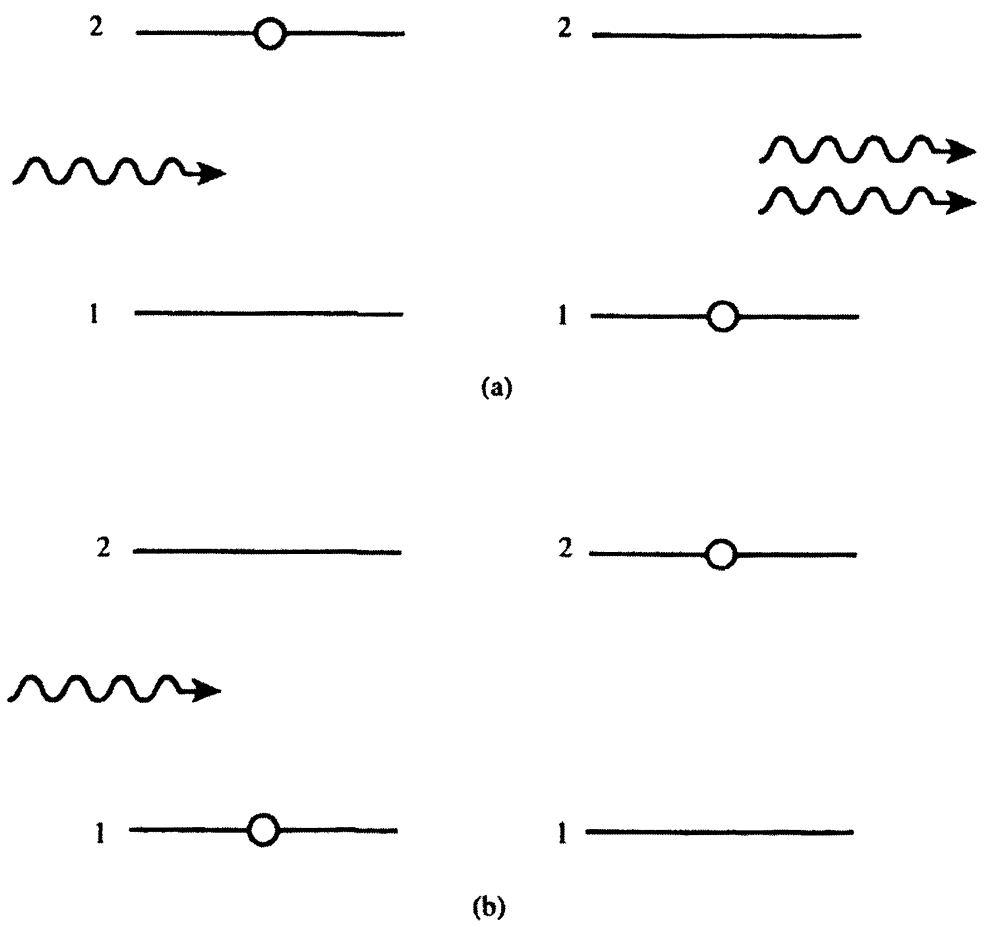


Figure A1.11. Photon–particle 'collision' pictures of the stimulated emission and absorption processes: (*a*) stimulated emission, (*b*) absorption.

the radiation frequency. A close approximation to a black body (absorber and emitter) is an enclosed cavity containing a small hole. Radiation that enters the hole has very little chance of escaping. If the inside of this cavity is in thermal equilibrium, it must lose as much energy as it absorbs and the emission from the hole is therefore characteristic of the equilibrium temperature T inside the cavity. Thus this type of radiation is often called 'thermal' or 'cavity' radiation. Black-body radiation has a spectral distribution as shown in figure A1.12.

Thermodynamically, the shape of the cavity should not influence the characteristics of the radiation, otherwise we could make a heat engine by connecting together cavities of different shapes. If, for example, two cavities of different shapes, but at the same temperature, were connected together with a reflective hollow pipe, we could imagine placing filters having different narrow frequency bandpass characteristics in the pipe. Unless the radiation emitted in each elemental frequency band from both cavities was identical, one cavity could be made to heat up and the other cool down, thereby violating the second law of thermodynamics.

In the latter part of the 19th century, experimental measurements of the spectral profile of black-body radiation had already been obtained and the data fitted to an empirical formula. Attempts had been made to explain the form of the data by treating the electromagnetic radiation as a collection of oscillators, each oscillator with its own characteristic frequency, however, these efforts had failed. It was a striking success of the new quantum theory and Planck's hypothesis that the radiation field was quantized, that led to a theoretical description of the energy density of black-body radiation. Planck's hypothesis was that an oscillator at frequency ν could only have energies

$$E_{n_\nu} = (n + \tfrac{1}{2})h\nu \qquad n = 0, 1, 2, 3 \ldots \tag{A1.29}$$

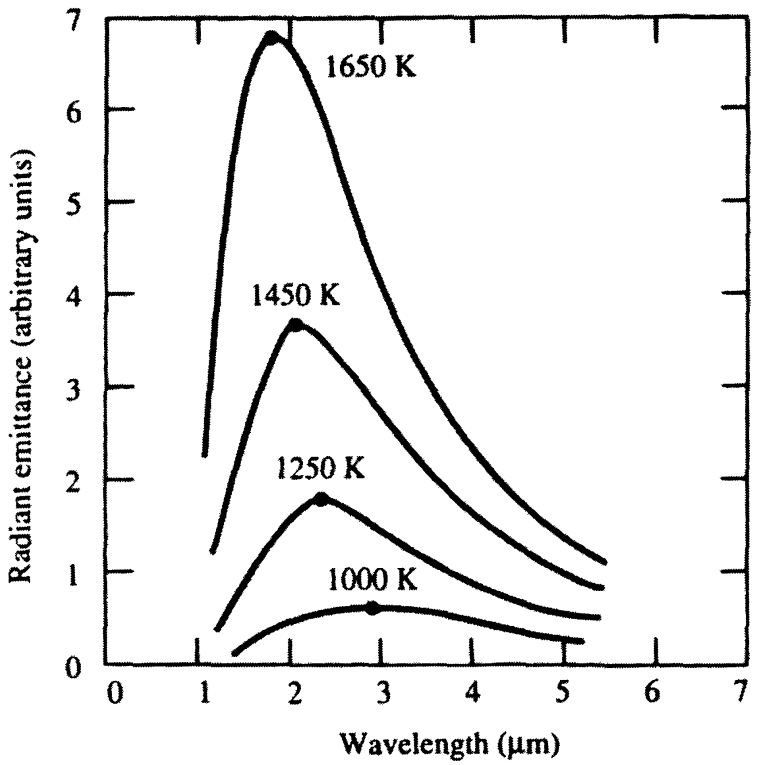


Figure A1.12. Spectral distribution of black-body radiation at different temperatures.

where the term $\frac{1}{2}h\nu$ is called the *zero-point energy*. Planck's hypothesis led to a theoretical prediction of the energy density of black-body radiation, which was

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3}\left(\frac{1}{2} + \frac{1}{e^{h\nu/kT} - 1}\right). \tag{A1.30}$$

The term arising from zero point energy corresponds to energy that cannot be released, so the available stored energy in the field is

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3}\left(\frac{1}{e^{h\nu/kT} - 1}\right). \tag{A1.31}$$

This formula predicts exactly the observed spectral character of black-body radiation. The term $8\pi h\nu^3/c^3$ is called the *density of states* or the number of modes of the radiation field per frequency interval. The term $\left(1/(e^{h\nu/kT} - 1)\right)$ is called the *occupation number*. It represents the average number of photons occupying a 'mode' of the radiation field at frequency ν.

A1.6 The relationship between the Einstein A and B coefficients

We can derive a useful relationship between Einstein's A and B coefficients by considering a collection of particles in thermal equilibrium inside a cavity at temperature T. The energy density of the radiation within the cavity is given by

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3}\left(\frac{1}{e^{h\nu/kT} - 1}\right) \tag{A1.32}$$

since in thermal equilibrium the radiation in the cavity will be black-body radiation. Although real particles in such a cavity possess many energy levels, we can restrict ourselves to considering the dynamic equilibrium

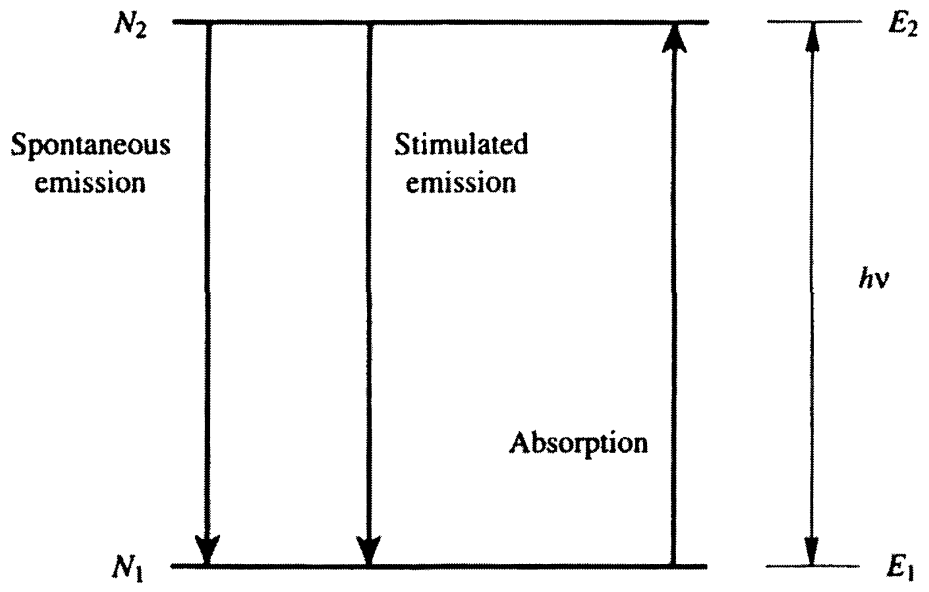


Figure A1.13. Radiative processes connecting two energy levels in thermal equilibrium at temperature T.

between any two of them, as shown in figure A1.13. The transitions that occur between two such levels as a result of interaction with radiation essentially occur independently of the energy levels of the system, which are not themselves involved in the transition.

In thermal equilibrium, the populations N_2 and N_1 of these two levels are constant, so

$$\frac{dN_2}{dt} = \frac{dN_1}{dt} = 0 \tag{A1.33}$$

and the rates of transfer between the levels are equal. Since the energy density of a black-body radiation field varies very little over the range of frequencies where transitions between levels 2 and 1 take place, we can use equations (A1.7) and (A1.27) and write

$$\frac{dN_2}{dt} = -N_2 B_{21}\rho(\nu) - A_{21}N_2 + N_1 B_{12}\rho(\nu). \tag{A1.34}$$

Therefore, substituting from equation (A1.32)

$$N_2\left[B_{21}\frac{8\pi h\nu^3}{c^3(e^{h\nu/kT}-1)} + A_{21}\right] = N_1\left[B_{12}\frac{8\pi h\nu^3}{c^3(e^{h\nu/kT}-1)}\right]. \tag{A1.35}$$

For a collection of particles that obeys Maxwell–Boltzmann statistics, in thermal equilibrium, energy levels of high energy are less likely to be occupied then levels of low energy. In exact terms the ratio of the population densities of two levels whose energy difference is $h\nu$ is

$$\frac{N_2}{N_1} = e^{-h\nu/kT}. \tag{A1.36}$$

So,

$$\frac{8\pi h\nu^3}{c^3(e^{h\nu/kT}-1)} = \frac{A_{21}}{B_{12}e^{h\nu/kT} - B_{21}}. \tag{A1.37}$$

This equality can only be satisfied if

$$B_{12} = B_{21} \tag{A1.38}$$

and

$$\frac{A_{21}}{B_{21}} = \frac{8\pi h\nu^3}{c^3} \tag{A1.39}$$

so a single coefficient A_{21} (say) will describe both stimulated emission and absorption. Equations (A1.38) and (A1.39) are called the *Einstein relations*. The stimulated emission rate is W_{21}, where

$$W_{21} = B_{21}\rho(\nu) = \frac{c^3 A_{21}}{8\pi h\nu^3}\rho(\nu) \tag{A1.40}$$

which is proportional to energy density. The spontaneous emission rate is A_{21}, which is independent of external radiation.

Although spontaneous emission would appear to be a different kind of radiative process from stimulated emission, in fact that is not really the case. Spontaneous emission can be shown to result from the zero-point energy of the radiation field, which was described in equation (A1.30).

A1.6.1 The effect of level degeneracy

In real systems containing atoms, molecules or ions, it frequently happens that different configurations of the system can have exactly the same energy. If a given energy level corresponds to a number of different arrangements specified by an integer g, we call g the *degeneracy* of the level. We call the separate states of the system with the same energy *sub-levels*. The levels 2 and 1 that we have been considering may consist of a number of degenerate sub-levels, where each sub-level has the same energy, as shown in figure A1.14, with g_2 sub-levels making up level 2 and g_1 sub-levels making up level 1. For each of the sub-levels of levels 1 and 2 with population n_1, n_2 respectively, the ratio of populations is

$$\frac{n_2}{n_1} = \mathrm{e}^{-h\nu/kT} \tag{A1.41}$$

and

$$N_1 = g_1 n_1 \qquad N_2 = g_2 n_2. \tag{A1.42}$$

Therefore,

$$\frac{n_2}{n_1} = \frac{g_1 N_2}{g_1 N_2} \tag{A1.43}$$

and

$$\frac{N_2}{N_1} = \frac{g_2}{g_1}\mathrm{e}^{-h\nu/kt}. \tag{A1.44}$$

From equations (A1.35) and (A1.44) it follows that in this case, where degenerate levels are involved, that the Einstein relations become

$$g_1 B_{12} = g_2 B_{21} \tag{A1.45}$$

and, as before,

$$\frac{A_{21}}{B_{21}} = \frac{8\pi h\nu^3}{c^3}. \tag{A1.46}$$

Note that

$$A_{21} = B_{21}\frac{8\pi h\nu^3}{c^3} = B_{21}\frac{8\pi\nu^2}{c^3}h\nu \tag{A1.47}$$

which can be described as B_{21} × no. of modes per unit volume per frequency interval × photon energy.

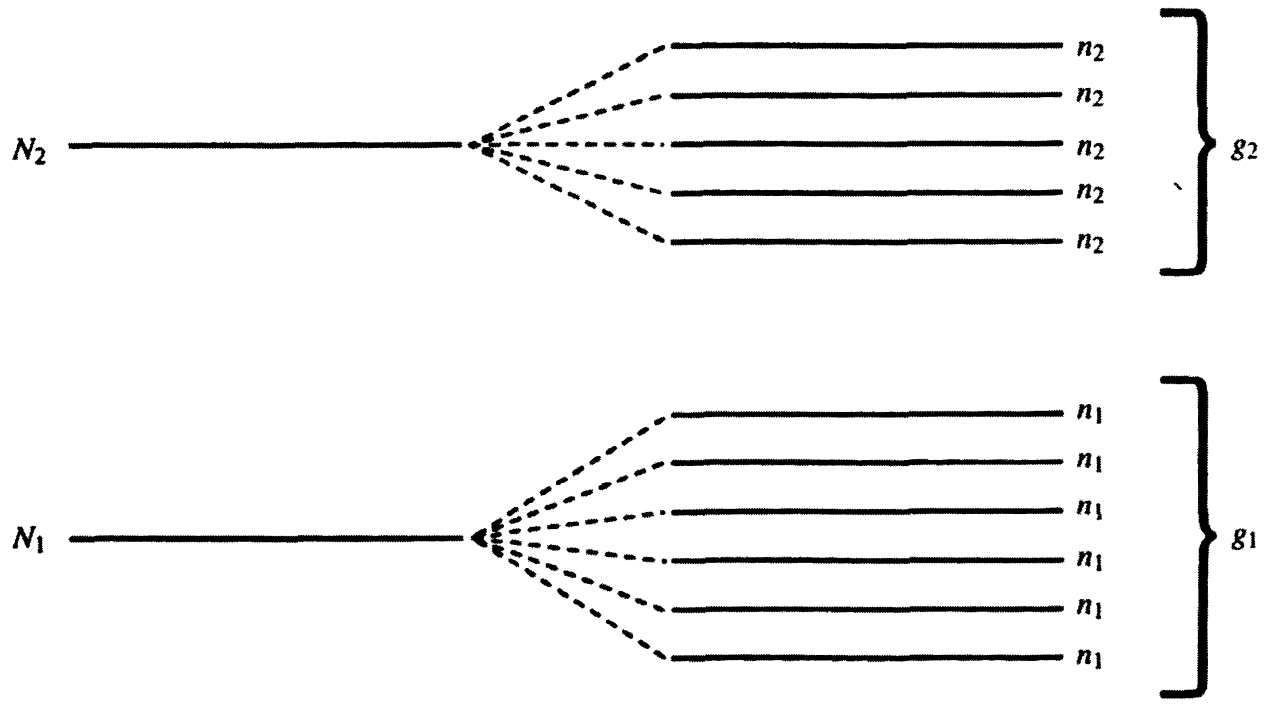


Figure A1.14. Two energy levels, each of which has a number of sub-levels of the same energy.

If there were only one photon in each mode of the radiation field, then the resulting energy density would be

$$\rho(\nu) = \frac{8\pi\nu^2}{c^3}h\nu. \tag{A1.48}$$

The resulting number of stimulated transitions would be

$$W_{21} = B_{21}\frac{8\pi\nu^2}{c^3}h\nu = A_{21} \tag{A1.49}$$

thus, the number of spontaneous transitions per second is equal to the number of stimulated transitions per second that would take place if there was just one photon excited in each mode.

A1.6.2 Ratio of spontaneous and stimulated transitions

It is instructive to examine the relative rates at which spontaneous and stimulated processes occur in a system in equilibrium at temperature T. This ratio is

$$R = \frac{A_{21}}{B_{21}\rho(\nu)}. \tag{A1.50}$$

We choose the $\rho(\nu)$ appropriate to a black-body radiation field, since such radiation is always present, to interact with an excited particle that is contained within an enclosure at temperature T.

$$R = \frac{A}{B\rho(\nu)} = (e^{h\nu/kT} - 1). \tag{A1.51}$$

If we use $T = 300$ K and examine the *microwave region*, $\nu = 10^{10}$ (say), then

$$\frac{h\nu}{kT} = \frac{6.626 \times 10^{-34} \times 10^{10}}{1.38 \times 10^{-23} \times 300} = 1.6 \times 10^{-3}$$

so

$$R = e^{0.0016} - 1 \approx 0.0016$$

and stimulated emission dominates over spontaneous. Particularly, in any microwave laboratory experiment

$$\rho(\nu)_{\text{laboratory created}} > \rho(\nu)_{\text{black-body}}$$

and spontaneous emission is negligible. However, spontaneous emission is still observable as a source of noise—the randomly varying component of the optical signal.

In the *visible region*,

$$\nu \approx 10^{15} \qquad \frac{h\nu}{kT} \approx 160 \qquad \text{and} \qquad A \gg B\rho(\nu)$$

so, in the visible and near-infrared region, spontaneous emission generally dominates unless we can arrange for there to be several photons in a mode. The average number of photons in a mode for black-body radiation is very small in the visible and infrared.

A1.7 Optical frequency amplifiers and line broadening

When an electromagnetic wave propagates through a medium, stimulated emissions increase the intensity of the wave, while absorptions diminish it. The overall intensity will increase if the number of stimulated emissions can be made larger than the number of absorptions. If we can create such a situation then we have built an amplifier that operates through the mechanism of stimulated emission. This *laser amplifier*, in common with electronic amplifiers, only has useful gain over a particular frequency bandwidth. Its operating frequency range will be determined by the lineshape of the transition and we expect the frequency width of its useful operating range to be of the same order as the width of the lineshape. It is very important to consider how this frequency width is related to the various mechanisms by which transitions between different energy states of a particle are smeared out over a range of frequencies. This *line broadening* affects in a fundamental way not only the frequency bandwidth of the amplifier but also its gain.

A laser amplifier can be turned into an oscillator by supplying an appropriate amount of positive feedback. The level of oscillation will stabilize because the amplifier saturates. Laser amplifiers fall into two categories, which saturate in different ways. The *homogeneously* broadened amplifier consists of a number of amplifying particles that are essentially equivalent while the *inhomogeneously* broadened amplifier contains amplifying particles with a distribution of amplification characteristics.

A1.7.1 Homogeneous line broadening

All energy states, except the lowest energy state of a particle (the ground state), cover a range of possible energies. This is reflected in the lineshape function, which shows the range of energies for a transition between one or two broadened energy states. At the fundamental level this smearing out of the energy is caused by the uncertainty involved in the energy measurement process. This gives rise to an intrinsic and unavoidable amount of line broadening called *natural broadening*.

A1.7.2 Natural broadening

This most fundamental source of line broadening arises, as just mentioned, because of uncertainty in the exact energy of the states involved. This uncertainty in measured energy, ΔE, arises from the time uncertainty, Δt, involved in making such a measurement. The product of these uncertainties is [3–5] $\Delta E \Delta t \sim \hbar$[7]. Because

[7] $\hbar = h/2\pi$.

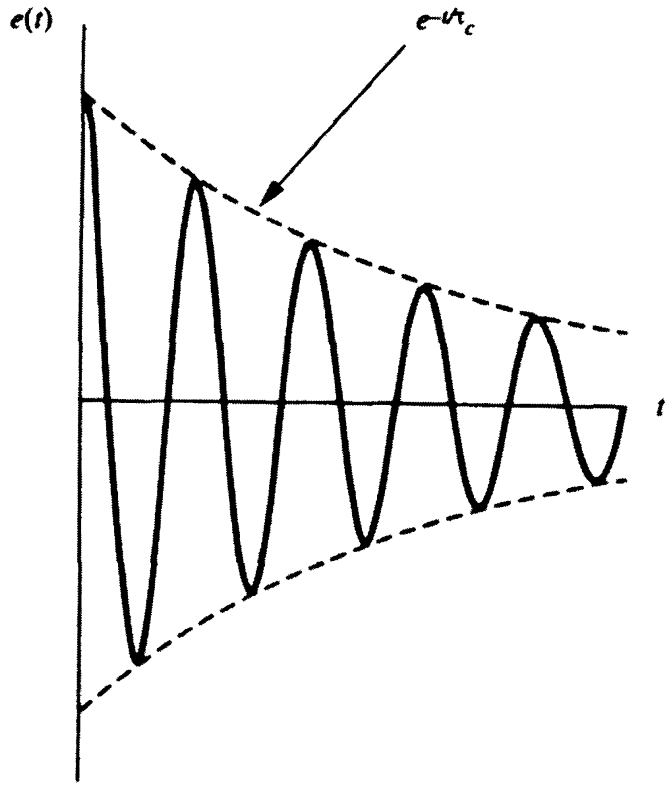


Figure A1.15. A damped oscillation used to represent the electric field produced by an excited particle as it decays.

an excited particle can only be observed for a time that is of the order of its lifetime, the measurement time uncertainty, Δt, is roughly the same as the lifetime, so

$$\Delta E \sim \hbar/\tau = A\hbar. \tag{A1.52}$$

The uncertainty in emitted frequency $\Delta\nu$ is $\Delta E/h$, so

$$\Delta\nu \sim A/2\pi. \tag{A1.53}$$

When the decay of an excited particle is viewed as a photon emission process, we can think of the particle, initially placed in the excited state at time $t' = 0$, emitting a photon at time t. The distribution of these times t among many such particles varies as $e^{-t/\tau}$. Our knowledge of when the photon is likely to be emitted with respect to the time origin restricts our ability to be sure of its frequency. For example, if a photon is observed at time t and is known to have come from a state with lifetime τ, we know that the *probable* time t' at which the atom became excited was $t - \tau < t' \leq t$. The longer the lifetime of the state is, the greater the uncertainty about when the particle acquired its original excitation becomes. In the limit as $\tau \to \infty$, our knowledge of the time of excitation becomes infinitely uncertain and we can ascribe a very well-defined frequency to the emitted photon, in this limit the electromagnetic waveform emitted by the atom approaches infinite length and is undamped.

We can put this approximate determination of $\Delta\nu$ on a more exact basis by considering the exponential intensity decay of a group of excited particles: The decay of each individual excited particle is modelled as an exponentially decaying (damped) cosinusoidal oscillation, as shown schematically in figure A1.15. It must be stressed that this is only a convenient way of *picturing* how an excited particle decays and emits electromagnetic radiation. It would not be possible in practice to observe such an electromagnetic field by watching a single excited particle decay. We can only observe a classical field by watching many excited

particles simultaneously. Within the framework of our model, we can represent the electric field of a decaying excited particle as

$$e(t) = E_0 e^{-t/\tau_c} \cos \omega_0 t. \tag{A1.54}$$

We need to determine the time constant τ_c that applies to this damped oscillation. The instantaneous intensity $i(t)$ emitted by an individual excited atom is

$$i(t) \propto |e(t)|^2 = E_0^2 e^{-2t/\tau_c} \cos^2 \omega_0 t. \tag{A1.55}$$

If we observe many such atoms the total observed intensity is

$$\begin{aligned} I(t) &= \sum_{\text{particles}} i(t) = \sum_i E_0^2 e^{-2t/\tau_c} \cos^2(\omega_0 t + \epsilon_i) \\ &= \sum_i \frac{E_0^2}{2} e^{-2t/\tau_c} [1 + \cos 2(\omega_0 t + \epsilon_i)] \end{aligned} \tag{A1.56}$$

where ϵ_i is the phase of the wave emitted by atom i. In the summation the cosine term gets smeared out because individual atoms are emitting with random phases. So, $I(t) \propto e^{-2t/\tau_c}$ However, we know that $I(t) \propto e^{-t/\tau}$, where τ is the lifetime of the emitting state, so the time constant τ_c is in fact $= 2\tau$. Thus:

$$e(t) = E_0 e^{-t/2\tau} \cos \omega_0 t. \tag{A1.57}$$

To find the frequency distribution of this signal we take its Fourier transform

$$E(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e(t) e^{-i\omega t} \, dt \tag{A1.58}$$

where

$$\begin{aligned} e(t) &= \tfrac{1}{2} E_0 (e^{i(\omega_0 + i/2\tau)t} + e^{-i(\omega_0 - i/2\tau)t}) \qquad \text{for } t > 0 \\ &= 0 \qquad \text{for} \quad t < 0. \end{aligned} \tag{A1.59}$$

The start of the period of observation at $t = 0$, taken for example at an instant when all the particles are pushed into the excited state, allows the lower limit of integration to be changed to 0, so

$$\begin{aligned} E(\omega) &= \frac{1}{2\pi} \int_0^{\infty} e(t) e^{-i\omega t} \, dt \\ &= \frac{E_0}{4\pi} \left[\frac{i}{(\omega_0 - \omega + i/2\tau)} - \frac{i}{(\omega_0 + \omega - i/2\tau)} \right]. \end{aligned} \tag{A1.60}$$

The intensity of emitted radiation is

$$I(\omega) \propto |E(\omega)|^2 = E(\omega) E^*(\omega) \propto \frac{1}{(\omega - \omega_0)^2 + (1/2\tau)^2}. \tag{A1.61}$$

Or, in terms of ordinary frequency

$$I(\nu) \propto \frac{1}{(\nu - \nu_0)^2 + (1/4\pi\tau)^2}. \tag{A1.62}$$

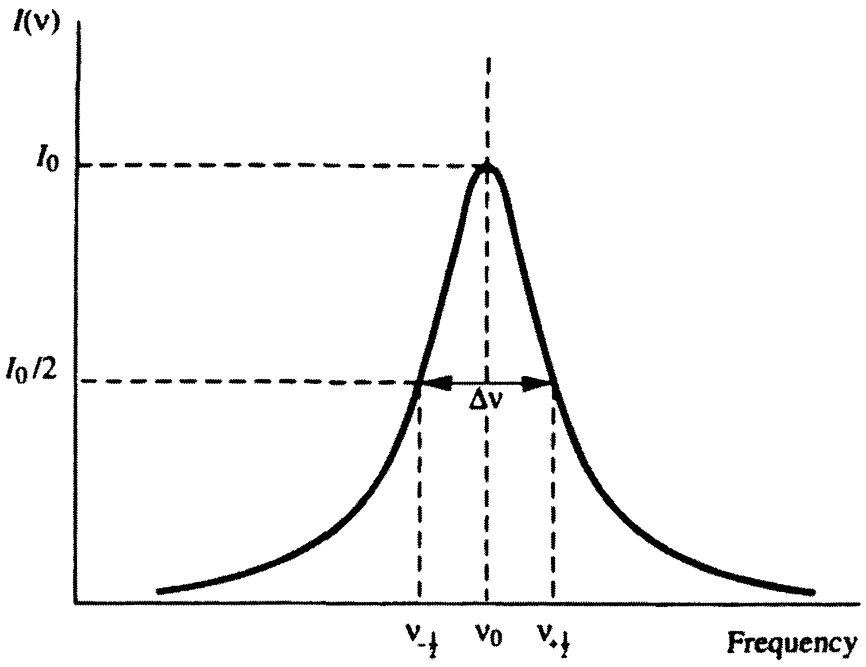


Figure A1.16. Lorentzian lineshape function for natural broadening.

The full width at half maximum height (FWHM) of this function is found from the half-intensity points of $I(\nu)$, which occur at frequencies $\nu_{\pm\frac{1}{2}}$ as shown in figure A1.16. This occurs where

$$\left(\frac{1}{4\pi\tau}\right)^2 = (\nu_{\pm\frac{1}{2}} - \nu_0)^2. \tag{A1.63}$$

The FWHM is $\Delta\nu = \nu_{+\frac{1}{2}} - \nu_{-\frac{1}{2}}$, which gives[8]

$$\Delta\nu = \frac{1}{2\pi\tau} = \frac{A}{2\pi} \tag{A1.64}$$

So, from equation (A1.62),

$$I(\nu) \propto \frac{1}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2}. \tag{A1.65}$$

The normalized form of this function is the lineshape function for natural broadening:

$$g(\nu_0, \nu)_N = \frac{(2/\pi\Delta\nu)}{1 + [2(\nu - \nu_0)/\Delta\nu]^2}. \tag{A1.66}$$

This type of function is called a Lorentzian. Since natural broadening is the same for each particle, it is said to be a *homogeneous* broadening mechanism.

A1.7.3 Other homogeneous broadening mechanisms

Besides natural broadening other mechanisms of homogeneous broadening exist, for example:

[8] A more exact treatment gives $\Delta\nu = (A_1 + A_2)/2\pi$ where A_2 and A_1 are the Einstein coefficients of the upper and lower levels of the transition.

(i) In a crystal the constituent particles of the lattice are in constant vibrational motion. This collective vibration can be treated as being equivalent to sound waves bouncing around inside the crystal. These sound waves, just like electromagnetic waves, can only carry energy in quantized amounts. These packets of acoustic energy are called *phonons*, and are analogous in many ways to photons. The principal differences between them are that phonons travel at the speed of sound and can only exist in a material medium. Collisions of phonons with the particles of the lattice perturb the phase of any excited, emitting particles present. This type of collision, which does not abruptly terminate the lifetime of the particle in its emitting state, is called a *soft* collision.

(ii) By *pressure* broadening, particularly in the gaseous and liquid phases interaction of an emitting particle with its neighbours causes perturbation of its emitting frequency and subsequent broadening of the transition. This interaction may arise in a number of ways:

  (a) Collisions with neutral particles, which may be *soft* or *hard*. A hard collision causes abrupt decay of the emitting species.

  (b) Collisions with charged particles. These collisions need not be very direct, but may involve a very small interaction that occurs when the charged particle passes relatively near, but perhaps as far as several tens of atomic diameters away from, the excited particle. In any case the relative motion of the charged and excited particles leads to a time-varying electric field at the excited particle that perturbs its energy states. This general effect in which an external electric field perturbs the energy levels of an atom (molecule or ion) is called the *Stark* effect; hence, line broadening caused by charged particles (ions or electrons) is called Stark broadening.

  (c) By van der Waals and resonance interactions (usually small effects). Resonance interactions occur when an excited particle can easily exchange energy with like neighbours, the effect is most important for transitions involving the ground state since, in this case, there are generally many particles near an excited particle for which the possibility of energy exchange exists. Broadening occurs because the *possibility* of energy exchange exists, not because an actual emission/reabsorption process occurs.

A1.8 Inhomogeneous broadening

When the environment or properties of particles in an emitting sample are non-identical, *inhomogeneous* broadening can occur. In this type of broadening, the shifts and perturbations of emission frequencies differ from particle to particle.

For example, in a real crystal the presence of imperfections and impurities in the crystal structure alters the physical environment of atoms from one lattice site to another. The random distribution of lattice point environments leads to a distribution of particles whose centre frequencies are shifted in a random way throughout the crystal.

A1.8.1 Doppler broadening

In a gas the random distribution of particle velocities leads to a distribution in the emission centre frequencies of different emitting particles seen by a stationary observer. For an atom whose component of velocity towards the observer is v_x, the observed frequency of the transition, whose stationary centre frequency is ν_0, is

$$\nu = \nu_0 + \frac{v_x}{c}\nu_0 \tag{A1.67}$$

where c is the velocity of light in the gas.

The Maxwell–Boltzman distribution of atomic velocities for particles of mass M at absolute temperature T is [6, 7]

$$f(v_x, v_y, v_z) = \left(\frac{M}{2\pi kT}\right)^{3/2} \exp\left[-\frac{M}{2kT}\left(v_x^2 + v_y^2 + v_z^2\right)\right]. \tag{A1.68}$$

The number of particles per unit volume that have velocities simultaneously in the range $v_x \rightarrow v_x + \mathrm{d}v_x$, $v_y \rightarrow v_y + \mathrm{d}v_y$, $v_z \rightarrow v_z + \mathrm{d}v_z$ is $Nf(v_z, v_y, v_z)\mathrm{d}v_z\mathrm{d}v_y\mathrm{d}v_z$ where N is the total number of particles per unit volume. The $(M/2\pi kT)^{3/2}$ factor is a normalization constant that ensures that the integral of $f(v_x, v_y, v_z)$ over all velocities is equal to unity, i.e.

$$\int_{-\infty}^{\infty}\int\int f(v_x, v_y, v_z)\,\mathrm{d}v_x\,\mathrm{d}v_d\,\mathrm{d}v_z = 1. \tag{A1.69}$$

The normalized one-dimensional velocity distribution is

$$f(v_x) = \sqrt{\frac{M}{2\pi kT}}\mathrm{e}^{-Mv_i^2/2kT}. \tag{A1.70}$$

This Gaussian-shaped function is shown in figure A1.17. It represents the probability that the velocity of a particle towards an observer is in the range $v_x \rightarrow v_x + \mathrm{d}v_x$. This is the same as the probability that the frequency be in the range

$$\nu_0 + \frac{v_x}{c}\nu_0 \rightarrow \nu_0 + \left(\frac{v_x + \mathrm{d}v_x}{c}\right) = \nu_0 + \frac{v_x}{c}\nu_0 + \frac{\mathrm{d}v_x}{c}\nu_0. \tag{A1.71}$$

The probability that the frequency lies in the range $\nu \rightarrow \nu + \mathrm{d}\nu$ is the same as the probability of finding the velocity in the range $(\nu - \nu_0)c/\nu_0 \rightarrow (\nu - \nu_0)c/\nu_0 + c\,\mathrm{d}\nu/\nu_0$, so the distribution of the emitted frequencies is

$$g(\nu) = \frac{c}{\nu_0}\sqrt{\frac{M}{2\pi kT}}\exp\left[\left(-\frac{M}{2kT}\right)\left(\frac{c^2}{\nu_0^2}\right)(\nu - \nu_0)^2\right]. \tag{A1.72}$$

This is already normalized (since $f(v_x)$ was normalized). It is called the normalized Doppler-broadened lineshape function. Its FWHM is

$$\Delta\nu_\mathrm{D} = 2\nu_0\sqrt{\frac{2kT\ln 2}{Mc^2}}. \tag{A1.73}$$

It can be seen that this increases with $\sqrt{T}$ and falls with atomic mass as $1/\sqrt{M}$. In terms of the Doppler-broadened linewidth $\Delta\nu_\mathrm{D}$, the normalized Doppler lineshape function is

$$g(\nu_0, \nu) = \frac{2}{\Delta\nu_\mathrm{D}}\sqrt{\frac{\ln 2}{\pi}}\,\mathrm{e}^{-[2(\nu-\nu_0)/\Delta\nu_\mathrm{D}]^2\ln 2}. \tag{A1.74}$$

This is a Gaussian function. It is shown in figure A1.18 compared with a Lorentzian function of the same FWHM. The Gaussian function is much more sharply peaked while the Lorentzian has considerable intensity far away from its centre frequency, in its *wings*.

Example. The 632.8 nm transition of neon is the most important transition to show laser oscillation in the HeNe laser (see chapter B3.6). The atomic mass of neon is 20. Therefore, using

$$M = 20 \times 1.67 \times 10^{-27}\ \mathrm{kg}$$
$$\nu_0 = 3 \times 10^8/632.8 \times 10^{-9}\ \mathrm{Hz}$$
$$T = 400\ \mathrm{K}$$

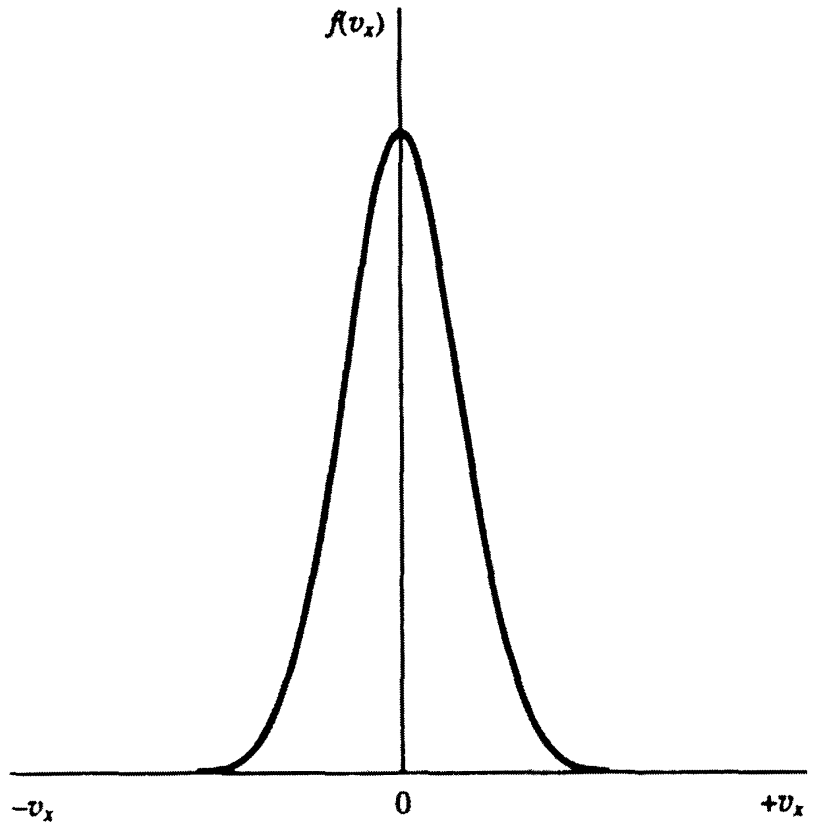


Figure A1.17. The Gaussian distribution of velocities in the x-direction for particles in a gas.

which is an appropriate temperature to use for the gas in a HeNe laser, the Doppler width is $\Delta\nu_D \sim 1.5$ GHz.[9] Doppler broadening usually dominates over all other sources of broadening in gaseous systems, except occasionally in very heavy gases at high pressures and in highly ionized plasmas of light gases; in the latter case Stark broadening frequently dominates.

Doppler broadening gives rise to a Gaussian lineshape and is a common source of inhomogeneous broadening. The inhomogeneous lineshape covers a range of frequencies because many *different* particles are being observed. The particles are *different* in the sense that they have different velocities and, consequently, different centre frequencies. Homogeneous broadening also always occurs at the same time as inhomogeneous broadening, to a greater or lesser degree. To illustrate this, imagine a hypothical experiment in which only those particles in a gas within a certain narrow velocity range are observed. The centre frequencies of these particles are confined to a narrow frequency band and, in this sense, there is no inhomogeneous broadening—all the *observed* particles are the same. However, a broadened lineshape would still be observed—the homogeneous lineshape resulting from natural and pressure broadening. This is illustrated in figure A1.19. When all particles are observed, irrespective of their velocity, an *overall* lineshape called a *Voigt* profile is observed. This overall lineshape results from the superposition of Lorentzian lineshapes spread across the Gaussian distribution of Doppler-shifted centre frequencies, as shown in figure A1.20. If the constituent Lorentzians have FWHM $\Delta\nu_L \ll \Delta\nu_D$, then the overall lineshape remains Gaussian and the system is properly said to be *inhomogeneously* broadened. However if all observed particles are identical, or almost identical, so that $\Delta\nu_D \ll \Delta\nu_L$, then the system as a whole is *homogeneously* broadened.

In solid materials, inhomogeneous broadening, when it is important, results from lattice imperfections and impurities that cause the local environment of individual excited particles to differ in a random way. We shall assume that the broadening that thereby results also gives rise to a Gaussian lineshape of appropriate

[9] The gigahertz (GHz) is a unit of frequency $\equiv 10^9$ Hz, another designation of high frequency is the terahertz (THz) $\equiv 10^{12}$ Hz.

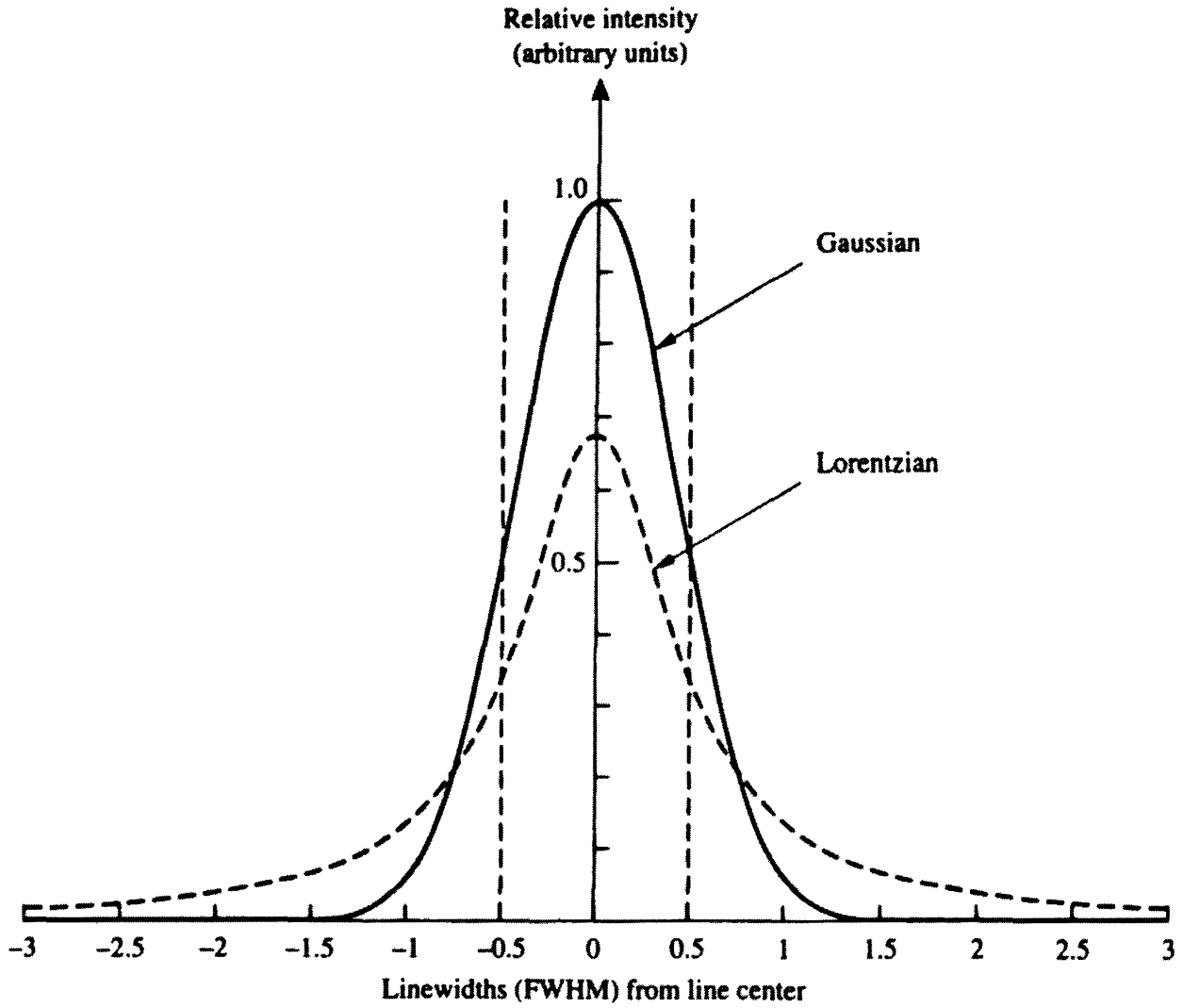


Figure A1.18. Comparison of normalized Gaussian and Lorentzian lineshapes.

FWHM, which we shall also designate as $\Delta\nu_D$.[10]

A1.8.2 Energy bands in condensed matter

Energy bands are the result of extensive line broadening in condensed matter and also occur because the energy states of electrons in such material reflect the de-localization of electrons throughout the material. The outermost electrons of atoms in condensed matter, the *valence* electrons, can no longer be thought of as being bound to a particular atom, rather they 'wander' throughout the material. Because of the Pauli exclusion principle, which states that no two electrons can oocupy the exact same energy state, even in the ground state these valence electrons separate themselves into very many, closely spaced energy states. These closely spaced states overlap and form the valence band. At absolute zero all energy states in this band are filled with electrons.

The next highest band of energies results from the electrons in the lowest excited states of the individual atoms becoming delocalized, interacting strongly and forming a band, called the conduction band. At absolute zero all the energy states in the conduction band are empty. At any temperature above absolute zero some electrons are thermally excited into the conduction band. It is the accessibility of the energy states in the

[10] The designation $\Delta\nu_D$ originates from Doppler broadening but is used generally to designate inhomogeneous linewidth.

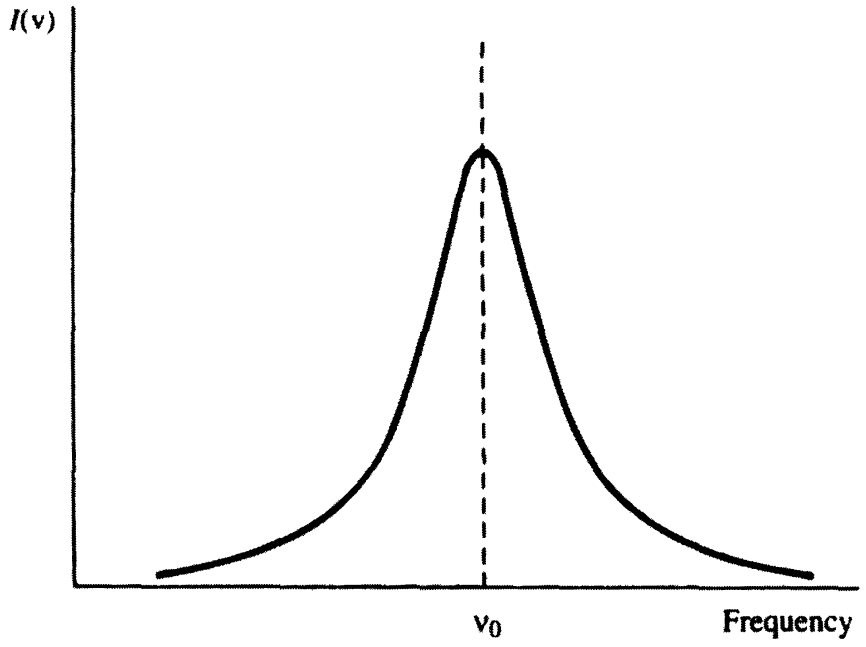


Figure A1.19. Homogeneous broadening of a group of particles in a gas that share the same velocity.

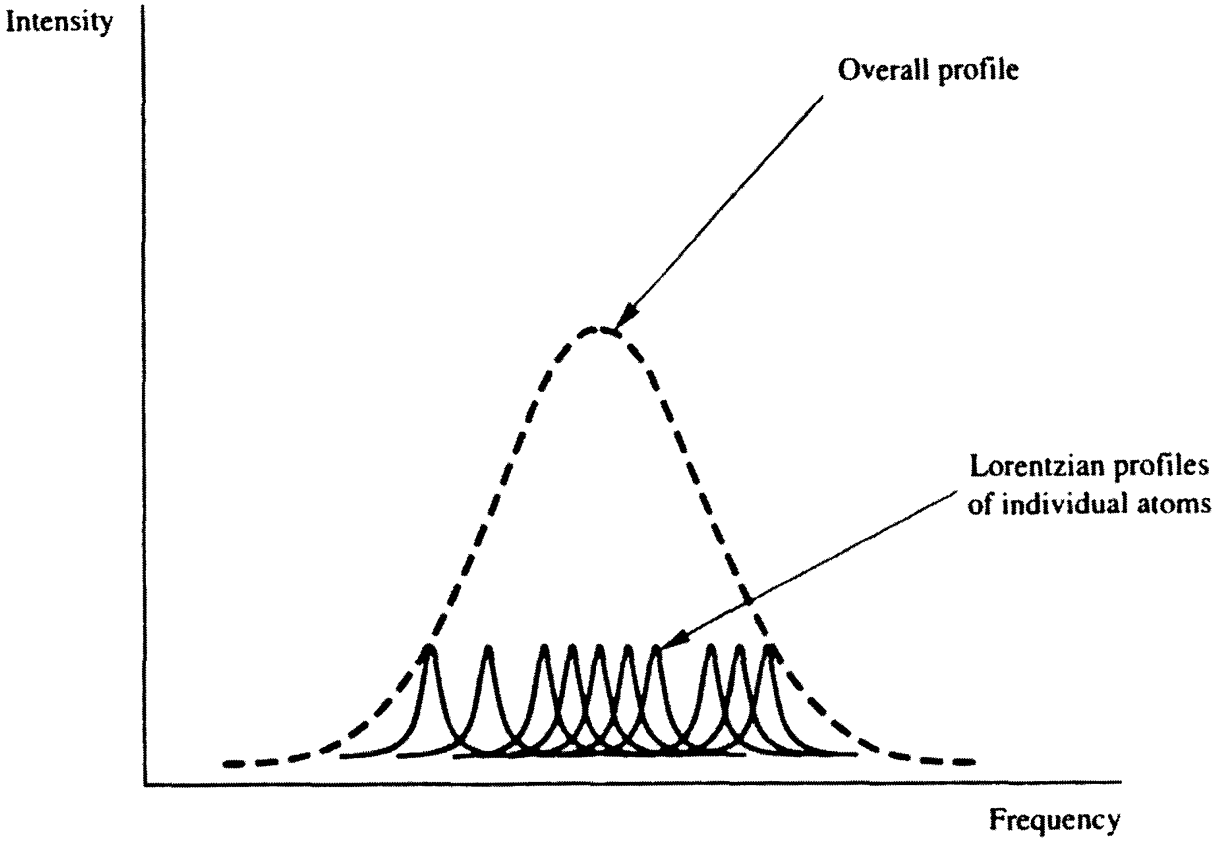


Figure A1.20. A Doppler-broadened distribution of Lorentzian lineshapes.

conduction band that characterizes the difference between *insulators*, *metals* and *semiconductors*. This is illustrated in figure A1.21. The gap in energy between the top of the valence band and the bottom of the conduction band is called the *energy gap*, E_g. If $E_g \gg kT$, where kT is the characteristic thermal energy

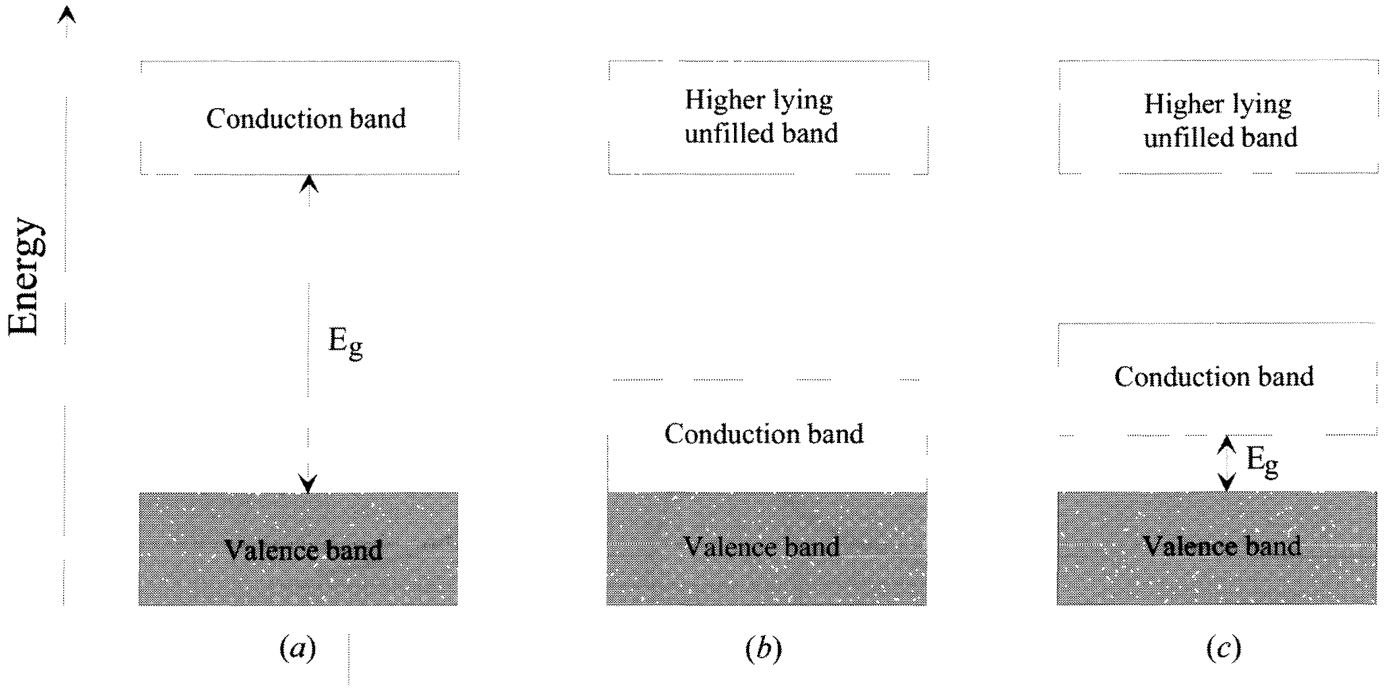


Figure A1.21. Schematic energy band diagram for (*a*) an insulator, (*b*) a metal, and (*c*) a semiconductor. E_g is the energy gap.

at temperature T, then electrons do not readily jump across the bandgap into the conduction band. The probability of such a jump depends on the Boltzmann factor $e^{-E_g/kt}$. In a metal, however, E_g is either a lot less than kT or zero. In a semiconductor, $E_g \simeq kT$. It is the presence of electrons in the conduction band, and also the absence of some electrons from the valence band (the 'holes'), which gives rise to electrical conductivity. In a simple sense, a material conducts electricity if there are available energy states for electrons to move into, since conduction corresponds to directed electron motion, which implies that an electron has moved into a different energy state. There are higher-lying energy bands above the conduction band, which are analogous to the higher-lying energy states of the isolated particles. They result from these states interacting, spreading over a range of energies and forming a band.

Some energy levels in condensed matter can remain relatively sharp if inter-particle interactions do not lead to extensive broadening. For example, for some levels of an impurity in a solid material, the spacing between adjacent impurity atoms can, on average, be as large as it might be in the gas phase, and broadening of the energy levels of these impurities may not lead to broad energy bands. In a laser using condensed matter, the laser transition involves a jump between energy bands or sometimes between relatively sharper levels that have not been extensively broadened. In a semiconductor laser, the laser transition generally involves electrons jumping across the energy gap.

A1.9 Optical frequency amplification with a homogeneously broadened transition

In an optical frequency amplifier, we are generally concerned with the interaction of a monochromatic radiation field with a transition between two energy states whose centre frequency is at, or near, the frequency of the monochromatic field. The magnitude of this interaction with each particle is controlled by the homogeneous lineshape function of the transition.

In the general case, the monochromatic radiation field and the centre frequency of the transition are not the same. This situation is shown schematically in figure A1.22. The stimulating radiation field is taken to be at frequency ν whilst the centre frequency of the transition is at ν'. The closer ν is to ν', the greater

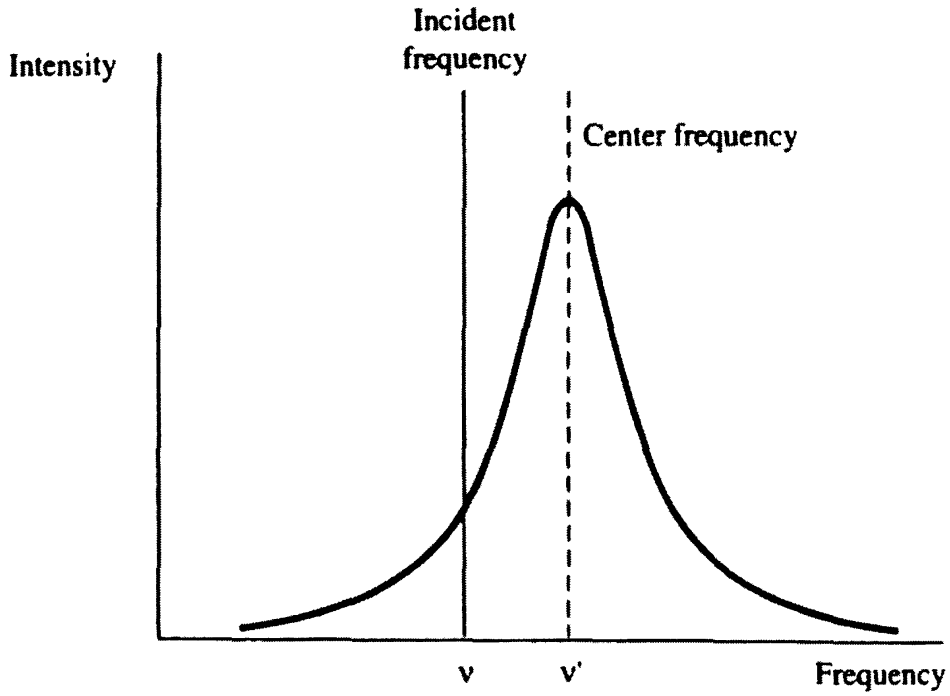


Figure A1.22. A monochromatic field interacting with a homogeneously broadened lineshape.

the number of transitions that can be stimulated. The stimulated transitions occur at frequency ν, since this is the frequency of the stimulating radiation. The number of stimulated transitions is proportional to the homogeneous lineshape function $g(\nu', \nu)$ and can be written as

$$N_S = N_2 B_{21} \rho(\nu) g(\nu', \nu). \tag{A1.75}$$

We have written $g(\nu', \nu)$ to indicate that this lineshape function has its centre frequency at ν' but is being evaluated at frequency ν. It is important to stress that the lineshape function $g(\nu', \nu)$ that is used here is the *homogeneous* lineshape function of the individual particles in the system, even though the contribution of homogeneous broadening to the overall broadening in the system may be small, for example when the overall broadening is predominantly inhomogeneous. The important point about the interaction of a particle with radiation is that an excited atom, molecule or ion can only interact with a monochromatic radiation field that overlaps its homogeneous (usually Lorentzian-shaped) lineshape profile. For example, consider the case of two excited atoms with different centre frequencies, these may be the different centre frequencies of atoms with different velocities relative to a fixed observer. The homogeneous lineshapes of these two atoms are shown in figure A1.23 together with a monochromatic radiation field at frequency ν. Particle A with centre frequency ν_A and *homogeneous width* $\Delta\nu$ can interact strongly with the field while the interaction of particle *B* is negligible.

We can analyse the interaction between a plane monochromatic wave and a collection of homogeneously broadened particles with reference to figure A1.24. As the wave passes through the medium it grows in intensity if the number of stimulated emissions exceeds the number of absorptions. The change in intensity of the wave in travelling a small distance dz through the medium is

$$\begin{aligned} \mathrm{d}I_\nu &= (\text{number of stimulated emissions} - \text{number of absorptions})/\text{vol} \\ &\quad \times h\nu \times \mathrm{d}z \\ &= \left(N_2 B_{21} g(\nu', \nu) \frac{I_\nu}{c} - N_1 B_{12} g(\nu', \nu) \frac{I_\nu}{c} \right) h\nu \, \mathrm{d}z. \end{aligned} \tag{A1.76}$$

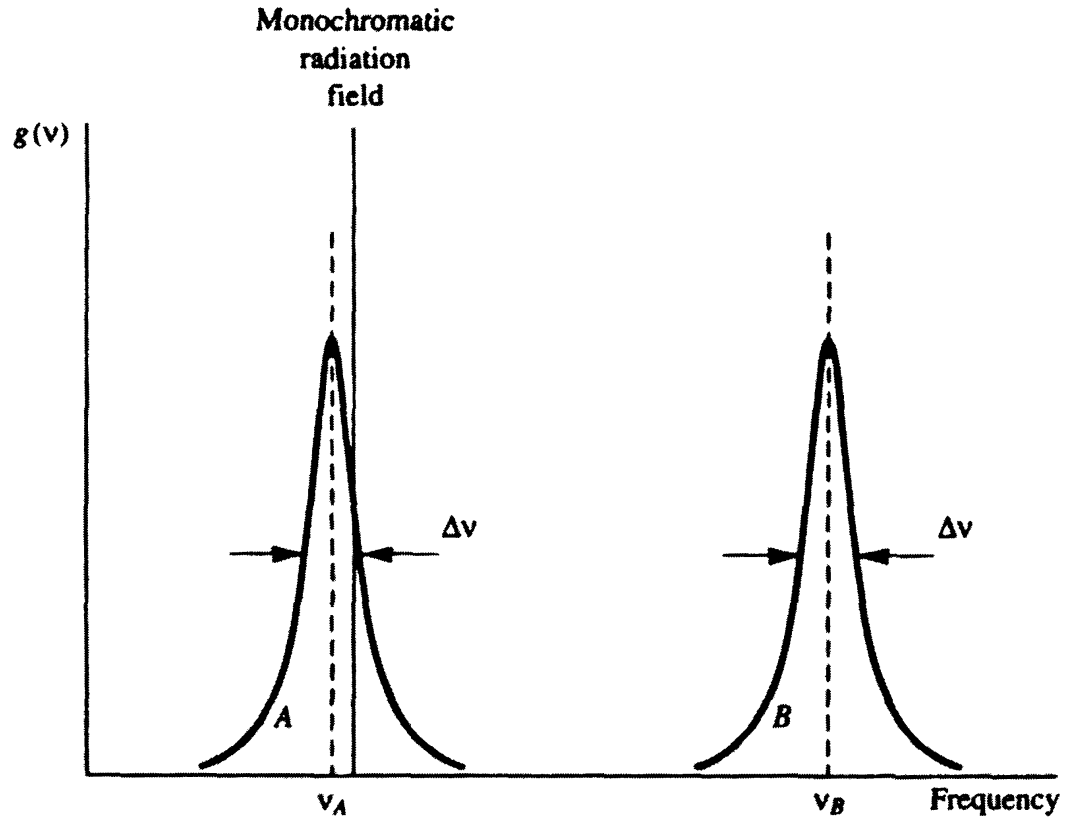


Figure A1.23. A monochromatic radiation field interacting with two homogeneously broadened lineshapes whose centre frequencies are different.

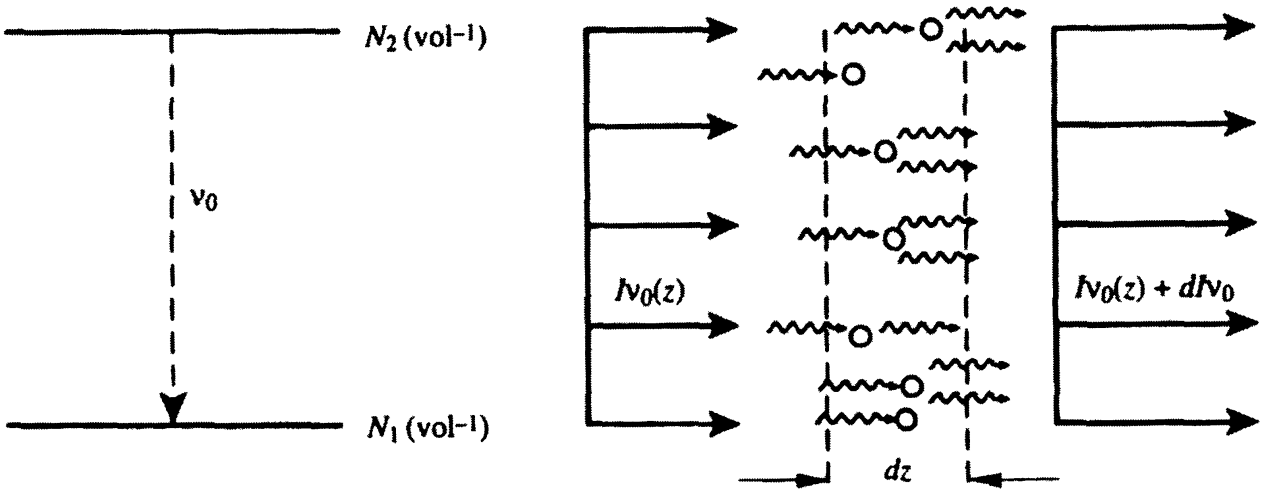


Figure A1.24. A plane wave travelling through and interacting with a collection of homogeneously broadened particles.

Use of the Einstein relations, equations (A1.38) and (A1.39), gives

$$\mathrm{d}I_\nu = \frac{I_\nu}{c}\left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{c^3 A_{21}}{8\pi h\nu^3} h\nu g(\nu', \nu)\,\mathrm{d}z. \tag{A1.77}$$

Therefore,

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}z} = \left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{c^2 A_{21}}{8\pi\nu^2} g(\nu', \nu) I_\nu \tag{A1.78}$$

which has the solution

$$I_\nu = I_\nu(0)\mathrm{e}^{\gamma(\nu)z} \tag{A1.79}$$

where $I_\nu(0)$ is the initial intensity at $z = 0$ and

$$\gamma(\nu) = \left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{c^2 A_{21}}{8\pi\nu^2}g(\nu', \nu). \tag{A1.80}$$

$\gamma(\nu)$ is called the gain coefficient of the medium and has the same frequency dependence as $g(\nu', \nu)$. If $N_2 > (g_2/g_1)N_1$, then $\gamma(\nu) > 0$ and we have an optical frequency amplifier. If $N_2 < (g_2/g_1)N_1$, then $\gamma(\nu) < 0$ and net absorption of the incident radiation occurs.

For a system in thermal equilibrium,

$$\frac{N_2}{N_1} = \frac{g_2}{g_1}\mathrm{e}^{-h\nu/kT} \tag{A1.81}$$

and for $T > 0$, $\mathrm{e}^{-h\nu/kT} < 1$, which implies that, in thermal equilibrium at positive temperatures, we cannot have positive gain. If we allow the formal existence of a negative temperature, at least for our system of two levels, then we can have $N_2 > (g_2/g_1)N_1$. Such a situation, which is essential for the construction of an optical frequency amplifier, is called a state of *population inversion* or *negative temperature*. This is not a true state of thermal equilibrium and can only be maintained by feeding energy into the system.

In the previous discussion, we have neglected the occurrence of spontaneous emission: this is reasonable for a truly plane wave, as the total number of spontaneous emissions into the zero solid angle subtended by the wave is zero. If the wave being amplified were diverging into a small solid angle $\delta\omega$, then $N_2 A_{21}\delta\omega/4\pi$ spontaneous emissions per second per unit volume would contribute to the increase in intensity of the wave. However, such emissions, being independent of the incident wave, are not in a constant phase relationship with this wave as are the stimulated emissions. These spontaneous emissions constitute a kind of 'noise' superimposed on the beam of identical photons created by stimulated emission.

A1.9.1 The stimulated emission rate in a homogeneously broadened system

The stimulated emission rate $W_{21}(\nu)$ is the number of stimulated emissions per particle per second caused by a monochromatic input wave at frequency ν:

$$W_{21}(\nu) = B_{21}g(\nu', \nu)\rho_\nu. \tag{A1.82}$$

$\rho_\nu(\mathrm{J\ m^{-3}})$ is the energy density of the stimulating radiation. Equation (A1.82) can be rewritten in terms of more practical parameters as

$$W_{21}(\nu) = \frac{A_{21}c^2 I_\nu}{8\pi h\nu^3}g(\nu', \nu). \tag{A1.83}$$

$W_{21}(\nu)$ has units $\mathrm{s^{-1}}$ per particle. Note that the frequency variation of $W_{21}(\nu)$ follows the lineshape function. The total number of stimulated emissions is

$$N_\mathrm{s} = N_2 W_{21}(\nu). \tag{A1.84}$$

A1.9.2 Optical frequency amplification with inhomogeneous broadening included

Although in our discussion so far we have been restricting our attention to homogeneous systems, we can show that equations (A1.76)–(A1.80) hold generally, even in a system with inhomogeneous broadening, if we take $g(\nu', \nu)$ as the *total* lineshape function.

In an inhomogeneously broadened system, we can divide the atoms up into classes, each class consisting of particles with a certain range of centre emission frequencies and the same homogeneous lineshape. For example, the class with centre frequency ν'' in the frequency range $d\nu''$ has $Ng_D(\nu', \nu'')\,d\nu''$ particles in it, where $g_D(\nu', \nu'')$ is the normalized inhomogeneous distribution of centre frequencies—the inhomogeneous lineshape function centred at ν'. This class of particles contributes to the change in intensity of a monochromatic wave at frequency ν as

$$\Delta(dI_\nu)(\text{from the group of particles in the band } d\nu'') = \left(N_2 B_{21} g_D(\nu', \nu'')\,d\nu'' g_L(\nu'', \nu)\frac{I_\nu}{c} - N_1 B_{12} g_D(\nu', \nu'')\,d\nu'' g_L(\nu'', \nu)\frac{I_\nu}{c}\right) h\nu\,dz \quad \text{(A1.85)}$$

where $g_L(\nu'', \nu)$ is the homogeneous lineshape function of a particle at centre frequency ν''. Equation (A1.85) is equivalent to equation (A1.76). The increase in intensity from all the classes of particles is found by integrating over these classes, that is over the range of centre frequencies ν'', so equation (A1.85) becomes

$$dI_\nu = \frac{I_\nu}{c}(N_2 B_{21} - N_1 B_{12})\left[\int_{-\infty}^{\infty} g_D(\nu', \nu'') g_L(\nu'', \nu)\,d\nu''\right]\nu\,dz \quad \text{(A1.86)}$$

which, in a similar fashion to equations (A1.76)–(A1.80), gives

$$\gamma(\nu) = \left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{c^2 A_{21}}{8\pi\nu^2} g(\nu', \nu) \quad \text{(A1.87)}$$

where $g(\nu', \nu)$ is now the overall lineshape function defined by the equation

$$g(\nu', \nu) = \int_{-\infty}^{\infty} g_D(\nu', \nu'') g_L(\nu'', \nu)\,d\nu''. \quad \text{(A1.88)}$$

In other words, the overall lineshape function is the convolution [8] of the homogeneous and inhomogeneous lineshape functions. The convolution integral in equation (A1.88) can be put in more familiar form if we measure frequency relative to the centre frequency of the overall lineshape, that is we put $\nu' = 0$, and equation (1.9.14) becomes

$$\begin{aligned} g(0, \nu) &= \int_{-\infty}^{\infty} g_D(0, \nu'') g_L(\nu'', \nu)\,d\nu'' \\ &= \int_{-\infty}^{\infty} g_D(0, \nu'') g_L(0, \nu - \nu'')\,d\nu'' \end{aligned} \quad \text{(A1.89)}$$

which can be written in simple form as

$$g(\nu) = \int_{-\infty}^{\infty} g_D(\nu'') g_L(\nu - \nu'')\,d\nu''. \quad \text{(A1.90)}$$

This is recognizable as the standard convolution integral of two functions $g_D(\nu)$ and $g_L(\nu)$.

If $g_D(\nu', \nu'')$ is indeed a normalized Gaussian lineshape as in equation (A1.74) and $g_L(\nu'', \nu)$ is a Lorentzian, then equation (A1.88) can be written in the form

$$g(\nu', \nu) = \frac{2}{\Delta\nu_D}\sqrt{\frac{\ln 2}{\pi}}\,\frac{y}{\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{y^2 + (x - t)^2}\,dt, \quad \text{(A1.91)}$$

where $y = \Delta\nu_L\sqrt{\ln 2}/\Delta\nu_D$ and $x = 2(\nu - \nu')\sqrt{\ln 2}/\Delta\nu_D$.

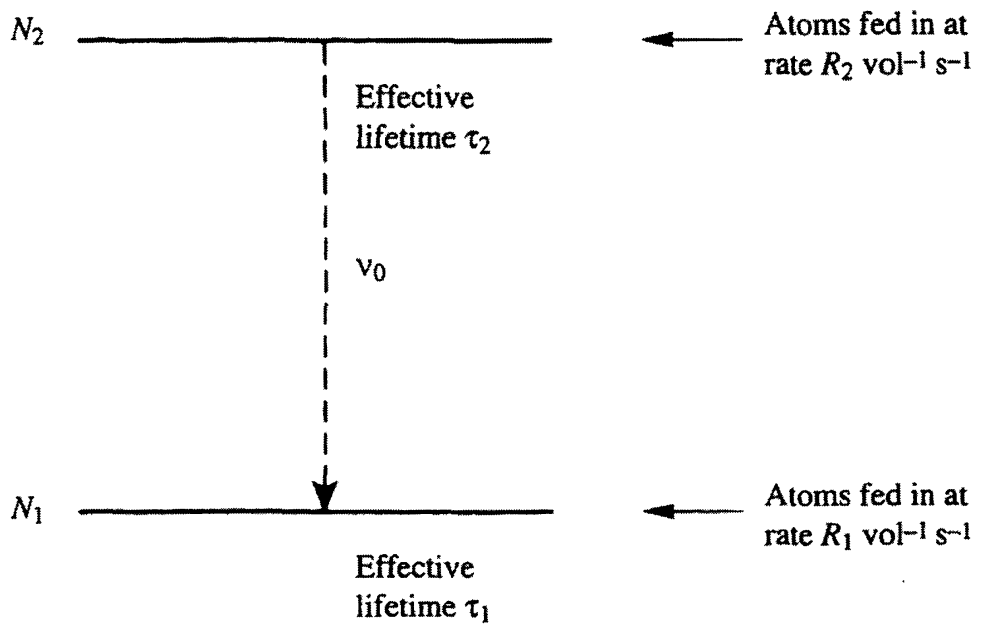


Figure A1.25. Pair of energy levels used in discussion of amplifier saturation.

This is one way of writing a normalized Voigt profile [9, 10]. The integral in equation (A1.91) cannot be evaluated analytically but must be evaluated numerically. For this purpose the Voigt profile is often written in terms of the error function for complex argument $W(z)$, which is available in tabulated form [11]

$$g(\nu', \nu) = \frac{2}{\Delta\nu_D}\sqrt{\frac{\ln 2}{\pi}}\mathcal{R}[W(z)] \tag{A1.92}$$

where $z = x + iy$ and $\mathcal{R}[W(z)]$ denotes the real part of the function.

A1.10 Optical frequency oscillation—saturation

If we can prepare a medium in a state of population inversion for a pair of its energy levels, then the transition between these levels can be used to make an optical frequency amplifier. To produce an oscillator, we need to apply appropriate feedback by inserting the amplifying medium between a pair of suitable mirrors. If the overall gain of the medium exceeds the losses of the mirror cavity and ancillary optics then oscillation will result. The level at which this oscillation stabilizes is set by the way in which the amplifier saturates.

A1.10.1 Homogeneous systems

Consider an amplifying transition at centre frequency ν_0 between two energy levels of a particle. We maintain this pair of levels in population inversion by feeding in energy. In equilibrium in the absence of an external radiation field, the rates R_2 and R_1 at which particles are fed into these levels must be balanced by spontaneous emission and non-radiative loss processes (such as collisions). The population densities and effective lifetimes of the two levels are N_2, τ_2 and N_1, τ_1 respectively as shown in figure A1.25. These effective lifetimes include the effect of non-radiative deactivation. If X_{2j} is the rate per particle per unit volume by which collisions depopulate level 2 and cause the particle to end up in a lower state j we can write

$$\frac{1}{\tau_2} = \sum_j (A_{2j} + X_{2j}). \tag{A1.93}$$

In equilibrium,

$$\frac{dN_2^o}{dt} = R_2 - \frac{N_2^o}{\tau_2} = 0 \tag{A1.94}$$

where the term N_2^o/τ_2 is the total loss rate per unit volume from spontaneous emission and other deactivation processes, so

$$N_2^o = R_2\tau_2 \tag{A1.95}$$

where the superscript o indicates that the population is being calculated in thc absence of a radiation field. Similarly, for the lower level of the transition, in equilibrium

$$\frac{dN_1^o}{dt} = R_1 + N_2^o A_{21} - \frac{N_1^o}{\tau_1} = 0. \tag{A1.96}$$

The term N_1^o/τ_1 is the total loss rate per unit volume from the level by spontaneous emission and other deactivation processes, while the term $N_2^o A_{21}$ is the rate at which atoms are feeding into level 1 by spontaneous emission from level 2. So from equations (A1.95) and (A1.96)

$$N_1^o = (R_1 + N_2^o A_{21})\tau_1 = (R_1 + R_2\tau_2 A_{21})\tau_1. \tag{A1.97}$$

The population inversion is

$$\left(N_2^o - \frac{g_2}{g_1}N_1^o\right) = \Delta N^o = R_2\tau_2 - \frac{g_2}{g_1}\tau_1(R_1 + R_2\tau_2 A_{21}) \tag{A1.98}$$

or, when the level degeneracies are equal,

$$\Delta N^o = R_2\tau_2 - \tau_1(R_1 + R_2\tau_2 A_{21}). \tag{A1.99}$$

If we now feed in a monochromatic (or other) signal, then stimulated emission and absorption processes will occur. We take the energy density of this signal at some point within the medium to be $\rho(\nu) = I(\nu)/c$. The rate at which this signal causes stimulated emissions is

$$W_{21}(\nu) = \int_{-\infty}^{\infty} B_{21} g(\nu_0, \nu)\rho(\nu)\,d\nu \qquad \text{per particle per second,} \tag{A1.100}$$

where $g(\nu_0, \nu)$ is the *homogeneous* lineshape function. If the input radiation was *white*, which in this context means that $\rho(\nu)$ is a constant over the range of frequencies spanned by the lineshape function, then

$$W_{21}(\nu) = B_{21}\rho(\nu)\int_{-\infty}^{\infty} g(\nu_0, \nu)\,d\nu = B_{21}\rho(\nu). \tag{A1.101}$$

The total rate at which a monochromatic plane wave causes stimulated transitions is

$$\begin{aligned} W_{21}(\nu) &= \int_{-\infty}^{\infty} B_{21} g(\nu_0, \nu)\rho(\nu)\,d\nu = \frac{I_\nu}{c} B_{21} g(\nu_0, \nu)\int_{-\infty}^{\infty} \delta(\nu - \nu'')\,d\nu'' \\ &= B_{21} g(\nu_0, \nu)\frac{I_\nu}{c}. \end{aligned} \tag{A1.102}$$

In the presence of a radiation field, in equilibrium the population densities of the pair of energy levels shown in figure A1.25 is

$$\frac{dN_2}{dt} = R_2 - \frac{N_2}{\tau_2} - N_2 B_{21} g(\nu_0, \nu)\rho(\nu) + N_1 B_{12} g(\nu_0, \nu)\rho(\nu) = 0 \tag{A1.103}$$

and

$$\frac{dN_1}{dt} = R_1 + N_2A_{21} - \frac{N_1}{\tau_1} + N_2B_{21}g(\nu_0, \nu)\rho(\nu) - N_1B_{12}g(\nu_0, \nu)\rho(\nu) = 0. \tag{A1.104}$$

If we write $B_{21}g(\nu_0, \nu)\rho(\nu) = W_{12}(\nu)$, the stimulated emission rate at frequency ν per particle, and neglect degeneracy factors so that we can assume that $W_{12}(\nu) = W_{21}(\nu) = W$ (equivalent to $B_{12} = B_{21}$), then we have

$$R_2 - \frac{N_2}{\tau_2} - N_2W + N_1W = 0 \tag{A1.105}$$

and

$$R_1 + N_2A_{21} - \frac{N_1}{\tau_1} + N_2W - N_1W = 0. \tag{A1.106}$$

From equation (A1.105),

$$N_2 = \frac{N_1W + R_2}{1/\tau_2 + W} \tag{A1.107}$$

and from equation (A1.106),

$$N_2 = \frac{N_1/\tau_1 + N_1W - R_1}{A_{21} + W} \tag{A1.108}$$

so

$$N_1 = \frac{R_1/\tau_2 + R_2A_{21} + W(R_1 + R_2)}{1/\tau_1\tau_2 + W(1/\tau_1 + 1/\tau_2 - A_{21})}. \tag{A1.109}$$

The population inversion in the system is now

$$N_2 - N_1 = \frac{N_1W + R_2}{1/\tau_2 + W} - N_1 = \frac{R_2 - N_1/\tau_2}{1/\tau_2 + W}. \tag{A1.110}$$

From equation (A1.109), this gives

$$\begin{aligned} N_2 - N_1 &= \frac{1/\tau_2(R_2/\tau_1 - R_2A_{21} - R_1/\tau_2) + W(R_2/\tau_1 - R_2A_{21} - R_1/\tau_2)}{(1/\tau_2 + W)(1/\tau_1\tau_2 + W/\tau_2 + W/\tau_1 - WA_{21})} \\ &= \frac{R_2/\tau_1 - R_1/\tau_2 - R_2A_{21}}{1/\tau_1\tau_2 + W(1/\tau_2 + 1/\tau_1 - A_{21})}. \end{aligned} \tag{A1.111}$$

Multiplying the numerator and denominator of equation (A1.111) by $\tau_1\tau_2$, we get

$$N_2 - N_1 = \frac{R_2\tau_2 - R_1\tau_1 - R_2\tau_1\tau_2A_{21}}{1 + W\tau_2(1 + \tau_1/\tau_2 - A_{21}\tau_1)}. \tag{A1.112}$$

The numerator of this expression is just the population inversion in the absence of any light signal, ΔN^o. Thus,

$$N_2 - N_1 = \frac{\Delta N^o}{1 + W\tau_2(1 + \tau_1/\tau_2 - A_{21}\tau_1)} \tag{A1.113}$$

or with the substitution

$$\phi = A_{21}\tau_2\left[1 + (1 - A_{21}\tau_2)\frac{\tau_1}{\tau_2}\right] \tag{A1.114}$$

$$N_2 - N_1 = \frac{\Delta N^o}{1 + \phi W/A_{21}}. \tag{A1.115}$$

Now from equation (A1.83),

$$W = \frac{c^2 A_{21}}{8\pi h \nu^3} I_\nu g(\nu_0, \nu). \tag{A1.116}$$

If we define

$$I_s(\nu) = \frac{8\pi h \nu^3}{c^2 \phi g(\nu_0, \nu)} \tag{A1.117}$$

then

$$N_2 - N_1 = \frac{\Delta N^o}{1 + I_\nu / I_s(\nu)} \tag{A1.118}$$

and $I_s(\nu)$, called the saturation intensity, is the intensity of an incident light signal (power area^{-1}) that reduces the population inversion to half its value when no signal is present. Note that the value of the saturation intensity depends on the frequency of the input signal relative to the line centre.

Returning to our expression for the gain constant of a laser amplifier,

$$\gamma(\nu) = (N_2 - N_1) \frac{c^2 A_{21}}{8\pi \nu^2} g(\nu_0, \nu); \tag{A1.119}$$

the gain as a function of intensity is, in a *homogeneously* broadened system,

$$\gamma(\nu) = \frac{\Delta N^o}{[1 + I_\nu / I_s(\nu)]} \frac{c^2 A_{21}}{8\pi \nu^2} g(\nu_0, \nu) \tag{A1.120}$$

which is reduced, that is *saturates* as the strength of the amplified signal increases. A good optical amplifier should have a large value of saturation intensity, from equation (A1.117) this implies that ϕ should be a minimum. In such systems often $A_{21} \simeq 1/\tau_2$ so $\phi \simeq 1$.

A1.10.2 Inhomogeneous systems

The problem of gain saturation in inhomogeneous media is more complex. For example, in a gas, a plane monochromatic wave at frequency ν interacts with a medium whose individual particles have Lorentzian homogeneous lineshapes with FWHM $\Delta\nu_N$, but whose centre frequencies are distributed over an inhomogeneous (Doppler) broadened profile of width (FWHM) $\Delta\nu_D$. The Lorentzian contribution to the overall lineshape is

$$g_L(\nu', \nu) = \frac{(2/\pi \Delta\nu_N)}{1 + [2(\nu - \nu')/\Delta\nu_N]^2} \tag{A1.121}$$

where ν' is the centre frequency of a particle set by its velocity relative to the observer. The Doppler-broadened profile of all the particles is

$$g_D(\nu_0, \nu') = \frac{2}{\Delta\nu_D} \sqrt{\frac{\ln 2}{\pi}} e^{-[(2(\nu' - \nu_0)/\Delta\nu_D)^2 \ln 2]} \tag{A1.122}$$

where ν_0 is the centre frequency of a particle at rest. The overall lineshape (from all the particles) is a sum of Lorentzian profiles spread across the particle velocity distribution, as shown in figure A1.20.

If $\Delta\nu_N \gg \Delta\nu_D$, the overall profile remains approximately Lorentzian and the observed behaviour of the system will correspond to *homogeneous* broadening. Such a situation is likely to arise for long wavelength transitions in a gas, particularly if this has a high atomic or molecular weight, at pressures where pressure broadening (which is a homogeneous process) is important, and frequently in solid materials. If $\Delta\nu_D \gg \Delta\nu_N$ (as is often the case in gases) the overall lineshape remains Gaussian and the system is *inhomogeneously* broadened.

Once again we reduce the problem to a consideration of the interaction of a plane electromagnetic wave with a two-level system as shown in figure A1.25. As the wave passes through the system, its intensity changes according to whether the medium is amplifying or absorbing. We take the intensity of the monochromatic wave to be $I(\nu, z)$ at plane z within the medium. The individual particles of the medium have a distribution of emission centre frequencies (or absorption centre frequencies) because of their random velocities (or, for example, their different crystal environments). We take the population density functions (particles vol^{-1} Hz^{-1}) in the upper and lower levels whose centre frequency is at ν' to be $N_2(\nu', z)$ and $N_1(\nu', z)$, respectively, at plane z. Particles are fed into levels 2 and 1 at rates $R_2(\nu')$ and $R_1(\nu')$. These rates are assumed to be uniform throughout the medium. $N_2(\nu', z)$, $N_1(\nu', z)$, $R_2(\nu')$ and $R_1(\nu')$ are assumed to follow the Gaussian frequency dependence set by the particle velocity distribution, and are normalized so that, for example, the total pumping rate of level 2 is

$$R_2 = \int_{-\infty}^{\infty} R_2(\nu')\,\mathrm{d}\nu' = R_{20}\int_{-\infty}^{\infty} \mathrm{e}^{-[2(\nu'-\nu_0)/\Delta\nu_D]^2 \ln 2}\,\mathrm{d}\nu'. \tag{A1.123}$$

In practice, the primary pumping process may not have this Gaussian dependence but, even when this is the case, the effect of collisions among particles that have been excited will be to *smear out* any non-Gaussian pumping process into a near-Gaussian form. This conclusion is justified by observations of Doppler-broadened lines under various excitation conditions where deviations from a true Gaussian lineshape are found to be minimal. From equation (A1.123)

$$R_2 = \sqrt{\frac{\pi}{\ln 2}}\frac{\Delta\nu_D}{2}R_{20} \tag{A1.124}$$

where R_{20} is a pumping rate constant and the total population density of level at plane z is

$$N_2 = \int_{-\infty}^{\infty} N_2(\nu', z)\,\mathrm{d}\nu'. \tag{A1.125}$$

The rate equations for the atoms whose centre frequencies are at ν' are

$$\frac{\mathrm{d}N_2}{\mathrm{d}t}(\nu', z) = R_2(\nu') - N_2(\nu', z)\left[\frac{1}{\tau_2} + B'_{21}(\nu', \nu)\frac{I(\nu, z)}{c}\right] + N_1(\nu', z)B'_{12}(\nu', \nu)\frac{I(\nu, z)}{c} \tag{A1.126}$$

and

$$\frac{\mathrm{d}N_1}{\mathrm{d}t}(\nu', z) = R_1(\nu') + N_2(\nu', z)\left[A_{21} + B'_{21}(\nu', \nu)\frac{I(\nu, z)}{c}\right] - N_1(\nu', z)\left[\frac{1}{\tau_1} + B'_{12}(\nu', \nu)\frac{I(\nu, z)}{c}\right]. \tag{A1.127}$$

Here we have used the modified Einstein coefficients $B'_{21}(\nu', \nu)$, $B'_{12}(\nu', \nu)$ that describe stimulated emission processes when the stimulating radiation is at frequency ν and the particle's centre emission frequency is at ν'. Written out in full,

$$B'_{21}(\nu', \nu) = B_{21}g(\nu', \nu) \tag{A1.128}$$

where $g(\nu', \nu)$ is the *homogeneous* lineshape function. The rate of change of intensity of the incident wave due to atoms with centre frequencies in a small range $\mathrm{d}\nu'$ at ν' is

$$\left[\frac{\mathrm{d}I}{\mathrm{d}z}(\nu, z)\right]_{\mathrm{d}\nu'} = h\nu\frac{I(\nu, z)}{c}[B'_{21}(\nu', \nu)N_2(\nu', z) - B'_{12}(\nu', \nu)N_1(\nu', z)]\,\mathrm{d}\nu'. \tag{A1.129}$$

The total rate of change of intensity due to all the particles, that is from all possible centre frequencies ν', is

$$\frac{\mathrm{d}I(\nu,z)}{\mathrm{d}z} = \frac{h\nu I(\nu,z)}{c}\int_{-\infty}^{\infty}[B'_{21}(\nu',\nu)N_2(\nu',z) - B'_{12}(\nu',\nu)N_1(\nu',z)]\,\mathrm{d}\nu'. \tag{A1.130}$$

In the steady state,

$$\frac{\mathrm{d}N_2}{\mathrm{d}t}(\nu',z) = \frac{\mathrm{d}N_1}{\mathrm{d}t}(\nu',z) = 0 \tag{A1.131}$$

so from equations (A1.126) and (A1.127)

$$B'_{21}(\nu',\nu)N_2(\nu') - B'_{12}(\nu',\nu)N_1(\nu') = \frac{B'_{21}(\nu',\nu)\left[\frac{R_2(\nu')}{1/\tau_2} - \left(\frac{g_2}{g_1}\right)\frac{(R_2(\nu')A_{21}+R_1(\nu')/\tau_2)}{1/\tau_1\tau_2}\right]}{1+\left[\frac{(g_1/g_2)(1/\tau_2-A_{21})}{1/\tau_1\tau_2}+\tau_2\right]B'_{21}(\nu',\nu)\frac{I(\nu,z)}{c}}. \tag{A1.132}$$

We note that

$$R_2(\nu') = R_{20}\mathrm{e}^{-[2(\nu'-\nu_0)/\Delta\nu_D]^2\ln 2}. \tag{A1.133}$$

Substituting in equation (A1.130) from (A1.132) and (A1.133) and bearing in mind that $B'_{21}(\nu',\nu)$ has a Lorentzian form,

$$B'_{21}(\nu',\nu) = \frac{B_{21}\frac{2}{\pi\Delta\nu_N}}{1+\left[\frac{2(\nu-\nu')}{\Delta\nu_N}\right]^2} \tag{A1.134}$$

where $\Delta\nu_N$ is the *homogeneous* FWHM of the transition, gives

$$\frac{1}{I(\nu,z)}\frac{\mathrm{d}I(\nu,z)}{\mathrm{d}z} = \gamma(\nu) = \frac{\gamma_0\int_{-\infty}^{\infty}\mathrm{d}\nu'\left(\frac{2}{\pi\Delta\nu_N}\right)\{1+[2(\nu-\nu')/\Delta\nu_N]^2\}^{-1}\exp\left(-\left[\frac{2(\nu'-\nu)}{\Delta\nu_D}\right]^2\ln 2\right)}{1+\eta I(\nu,z)\left(\frac{2}{\pi\Delta\nu_N}\right)[1+[2(\nu-\nu')/\Delta\nu_N]^2]^{-1}}. \tag{A1.135}$$

We have made the substitutions

$$\gamma_0 = \frac{h\nu}{c}B_{21}\left[R_{20}\tau_2 - \frac{g_2}{g_1}\left(\frac{R_{20}A_{21}+R_{10}A_2}{1/\tau_1\tau_2}\right)\right] \tag{A1.136}$$

$$\eta = \left[\frac{g_2}{g_1}\frac{(1/\tau_2 - A_{21})}{1/\tau_1\tau_2} + \tau_2\right]\frac{B_{21}}{c}. \tag{A1.137}$$

Equation (A1.135) can be written

$$\gamma(\nu) = \frac{2\gamma_0\Delta\nu_N}{\pi}\int_{-\infty}^{\infty}\frac{\exp\left(-[2(\nu'-\nu_0)/\Delta\nu_D]^2\ln 2\right)\mathrm{d}\nu'}{4(\nu-\nu')^2+\Delta\nu_N^2[1+2\eta I(\nu,z)/\pi\Delta\nu_N]}. \tag{A1.138}$$

Although it is fairly clear from equation (A1.138) that the gain of the amplifier falls as $I(\nu,z)$ increases, it is not easy to see from the integral exactly how this occurs. If the intensity is small the gain approaches its *unsaturated* value

$$\gamma_0(\nu) = \frac{2\gamma_0\Delta\nu_N}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-[2(\nu'-\nu_0)\Delta\nu_D]^2\ln 2}\,\mathrm{d}\nu'}{4(\nu-\nu')^2+\Delta\nu_N^2}. \tag{A1.139}$$

If equation (A1.138) is examined closely, for frequencies ν' close to the input frequency ν, the integrand can be written approximately as

$$\frac{\mathrm{e}^{-[2(\nu'-\nu_0)/\Delta\nu_D]^2\ln 2}}{\Delta\nu_N^2[1+2\eta I(\nu,z)/\pi\Delta\nu_N]}.$$

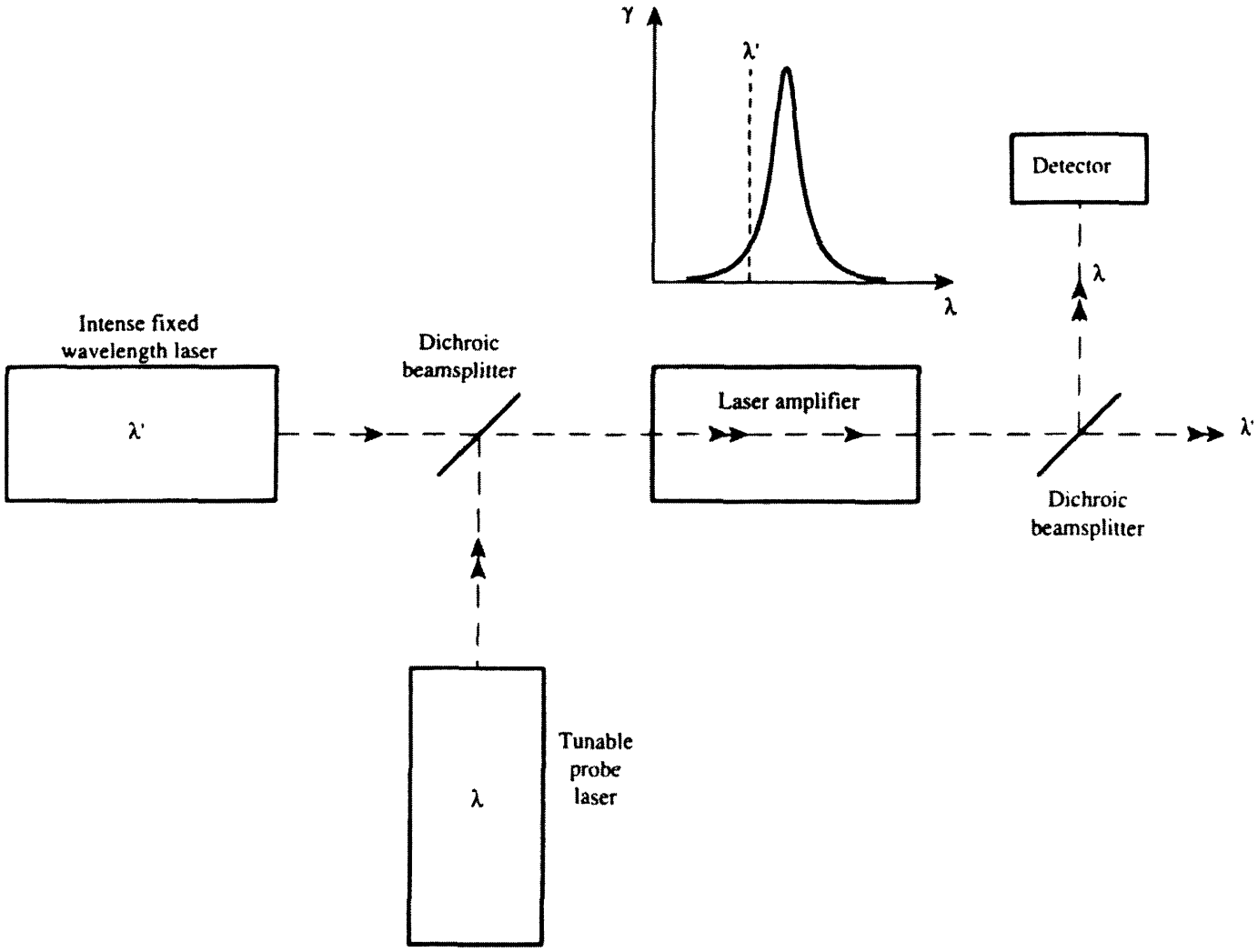


Figure A1.26. Schematic experimental arrangement for measuring the frequency dependence of the gain of an amplifier that is experiencing saturation from a strong fixed monochromatic signal.

However, for frequencies ν' far from the input frequency, the integrand can be written approximately as

$$\frac{e^{-[2(\nu'-\nu_0)/\Delta\nu_D]^2 \ln 2}}{4(\nu-\nu')^2+\Delta\nu_N^2}$$

which can be seen to be identical to the integrand in the unsaturated gain expression, equation (A1.139). Thus, we conclude that, in making their contribution to the overall gain, particles whose frequencies are far from the input frequency are relatively unaffected by the input radiation, whereas particles whose frequencies are close to that of the input show strong saturation effects. The gain in the system comes largely from those particles whose frequencies are within (roughly) a homogeneous linewidth of the input radiation frequency. The consequences of this are best illustrated by considering a hypothetical experiment, shown schematically in figure A1.26, in which the small-signal gain of a predominantly inhomogeneously broadened amplifier is measured with and without a strong saturating signal simultaneously present. Without the presence of a strong signal at a fixed frequency ν_S, the observed (small-signal) gain follows the Gaussian curve of the overall line profile of the amplifier, as shown in figure A1.27(*a*). However, if we perform this experiment again when a strong fixed frequency field is also present, which causes saturation of the gain, we find the gain is reduced locally by the saturating effect of the strong field as shown in figure A1.27(*b*). This phenomenon

is called *hole burning* [12]. The width of the *hole* that is thus produced is determined by the quantity

$$\Delta\nu_N^2\left[1+\frac{2\eta I(\nu,z)}{\pi\Delta\nu_N}\right].$$

If

$$\Delta\nu_N\sqrt{1+\frac{2\eta I(\nu,z)}{\pi\Delta\nu_N}} \ll \Delta\nu_D$$

for example, in a gaseous system where Doppler broadening is the largest contribution to the total observed line broadening, equation (A1.138) can be integrated by bringing the much less sharply peaked exponential factor outside the integral. In this case, the sharply peaked Lorentzian lineshape makes the integrand largest for frequencies ν' near to ν; over a small range of frequencies ν' near ν the exponential factor remains approximately constant, so equation (A1.138) can be written

$$\gamma(\nu)=\frac{2\gamma_0\Delta\nu_N}{\pi}e^{-[2(\nu-\nu_0)\Delta\nu_D]^2\ln 2} \times\int_{-\infty}^{\infty}\frac{d\nu'}{4(\nu-\nu')^2+\Delta\nu_N^2[1+2\eta I(\nu,z)/\pi\Delta\nu_N]}. \tag{A1.140}$$

Now the integral can be evaluated to give

$$\gamma(\nu)=\gamma_0\left[1+\frac{2\eta I(\nu,z)}{\pi\Delta\nu_N}\right]^{-\frac{1}{2}}e^{-[2(\nu-\nu_0)/\Delta\nu_D]^2\ln 2} \tag{A1.141}$$

which gives

$$\gamma(\nu)=\gamma_0\left[1+\frac{I(\nu,z)}{I_s'(\nu)}\right]^{-\frac{1}{2}}e^{-[2(\nu-\nu_0)/\Delta\nu_D]^2\ln 2} \tag{A1.142}$$

where $I_s'(\nu)=\pi\Delta\nu_N/2\eta$ is called the saturation intensity for inhomogeneous broadening. Note that γ_0 is the small-signal gain at line centre of the inhomogeneously broadened line, which can be written in the form

$$\gamma_0=\frac{1}{4\pi}\sqrt{\frac{\ln 2}{\pi}}\frac{\lambda^2A_{21}}{\Delta\nu_D}\left(N_2-\frac{g_2}{g_1}N_1\right). \tag{A1.143}$$

When Doppler broadening dominates in a system, incident radiation at frequency ν cannot interact with those particles whose Doppler-shifted frequency is different from ν by much more than $\Delta\nu_N$.

If the amplifier is homogeneously broadened, that is if

$$\Delta\nu_N\sqrt{1+\frac{\eta I(\nu,z)}{\pi\Delta\nu_N}} \gg \Delta\nu_D,$$

equation (A1.138) can be integrated by bringing the less-sharply peaked Lorentzian factor outside the integral to give

$$\gamma(\nu)=\frac{\gamma_0\Delta\nu_D}{\Delta\nu_N\sqrt{\pi\ln 2}}\left\{\left[\frac{2(\nu-\nu_0)}{\Delta\nu_N}\right]^2+1+\frac{I(\nu,z)}{I_s'(\nu)}\right\}^{-1} =\frac{\Delta\nu_D}{2}\sqrt{\pi/\ln 2}\frac{\gamma_0 g(\nu_0,\nu)}{[1+I(\nu,z)/I_s(\nu)]} \tag{A1.144}$$

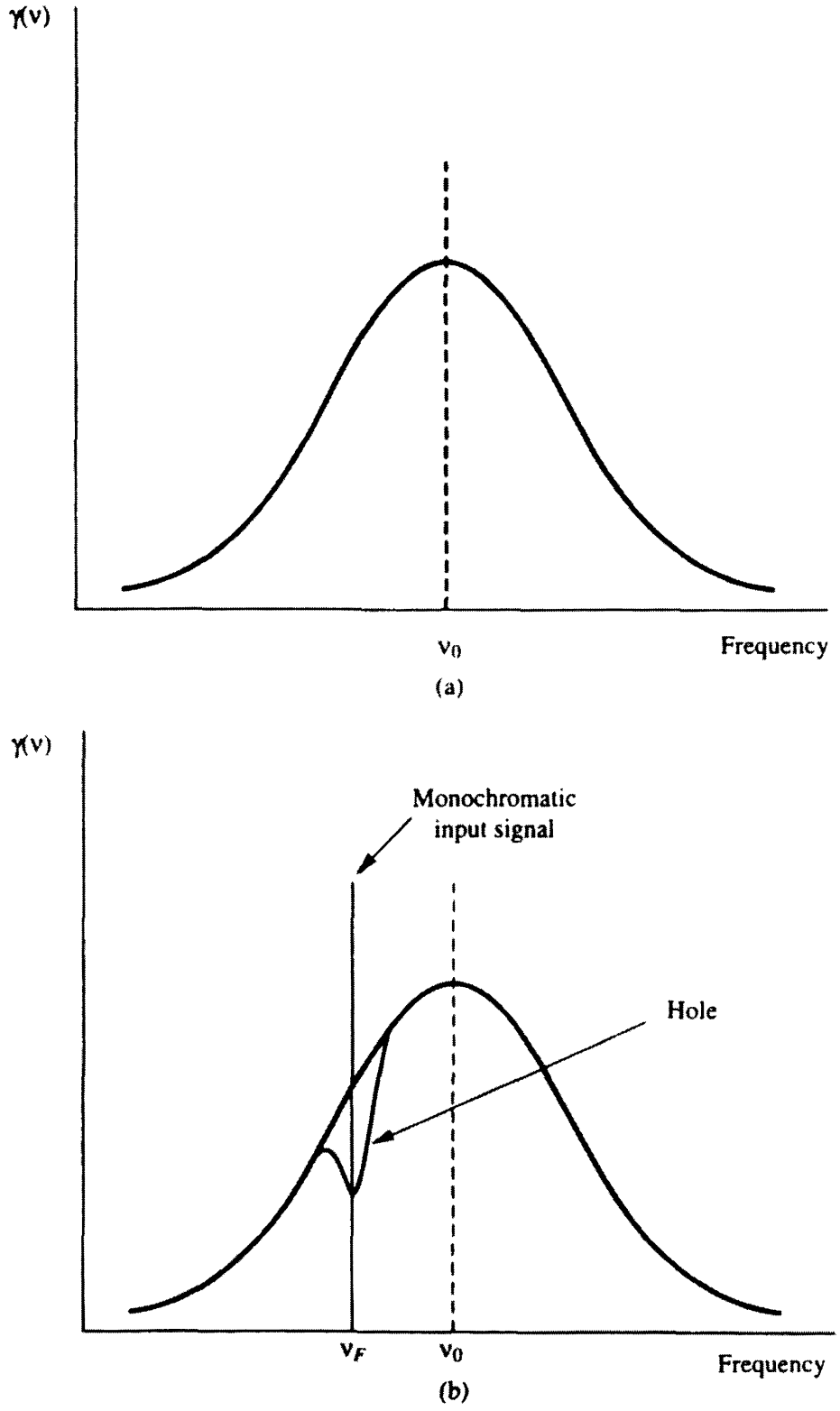


Figure A1.27. Gain as a function of frequency in an inhomogeneously broadened amplifier. (*a*) Small-signal situation when no saturation has occurred. (*b*) Showing the production of a 'hole' in the gain curve by a strong monochromatic input at frequency ν_F.

where $g(\nu_0, \nu)$ is the homogeneous lineshape function

$$g(\nu_0, \nu) = \frac{(2/\pi\,\Delta\nu_N)}{1 + [2(\nu - \nu_0)\Delta\nu_N]^2} \tag{A1.145}$$

and $I_s(\nu)$ is the saturation intensity for homogeneous broadening given by

$$I_s(\nu) = \frac{2I_s'(\nu)}{\pi\,\Delta\nu_N g(\nu_0, \nu)} = \frac{1}{\eta g(\nu_0, \nu)}. \tag{A1.146}$$

It can be seen from equation (A1.137) that for $g_2 = g_1$, η reduces to the expression

$$\eta = \left[\left(\frac{1/\tau_2 - A_{21}}{1/\tau_1\tau_2}\right) + \tau_2\right]\frac{B_{21}}{c} \tag{A1.147}$$

and $I_s(\nu)$ reduces to the expression obtained previously as the saturation intensity for homogeneous broadening. Namely,

$$I_s(\nu) = \frac{8\pi h\nu^3}{c^2\phi g(\nu_0, \nu)} \tag{A1.148}$$

where

$$\phi = A_{21}\tau_2\left[1 + (1 - A_{21}\tau_2)\frac{\tau_1}{\tau_2}\right]. \tag{A1.149}$$

A1.11 Power output from a laser amplifier

For a laser amplifier of length ℓ and gain coefficient $\gamma(\nu)$ the output intensity for a monochromatic input intensity of I_0 (W m^{-2}) at frequency ν is

$$I = I_0 e^{\gamma(\nu)\ell} \tag{A1.150}$$

if saturation effects are neglected. If saturation effects cannot be neglected then the differential equation that describes how intensity increases must be re-examined. This is

$$\gamma(\nu) = \frac{1}{I}\frac{dI}{dz}. \tag{A1.151}$$

For a homogeneously broadened amplifier with saturation an explicit solution to this equation can be found. In this case, if $\gamma_0(\nu)$ is the small-signal gain,

$$\gamma(\nu) = \frac{\gamma_0(\nu)}{1 + I/I_s(\nu)} = \frac{1}{I}\frac{dI}{dz} \tag{A1.152}$$

which can be rewritten in the form

$$\frac{dI}{I} + \frac{dI}{I_s(\nu)} = \gamma_0(\nu)\,dz. \tag{A1.153}$$

The solution to this equation is

$$I = I_0 e^{\gamma_0(\nu)\ell - (I - I_0)/I_s(\nu)}. \tag{A1.154}$$

equation (A1.154) must be solved iteratively. The solution will be somewhere between the input intensity and the output intensity when saturation is neglected.

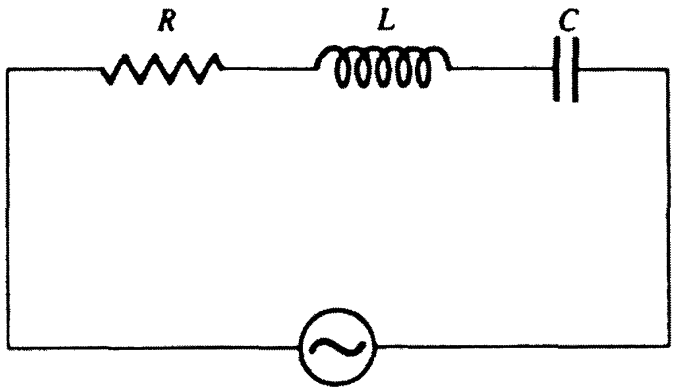


Figure A1.28. RLC circuit.

A1.12 The electron oscillator model of a radiative transition

When a particle decays from a higher energy level to a lower, we can model the resultant electric field as a damped oscillation. There is an analogy between this decay of an excited particle and the damped oscillation of an electric circuit. For example, for the RLC circuit shown in figure A1.28, the resonant frequency is $\nu_0 = 1/2\pi\sqrt{LC}$. If the sinusoidal driving voltage is disconnected from the circuit, then the oscillation of the circuit decays exponentially—provided the circuit is underdamped. The power spectrum of the decaying electric current is Lorentzian, just as it is for a spontaneous transition. The FWHM of the circuit resonance, is ν_0/Q, where Q is called the *quality factor* of the circuit, analogous to the homogeneously broadened linewidth $\Delta\nu_N$. A transition between levels almost always has $\Delta\nu_N \ll \nu_0$ so clearly has a very high Q.

In the classical theory of how a particle responds to electromagnetic radiation, each of the n electrons attached to the particle is treated as a damped harmonic oscillator. For example, when an electric field acts on an atom, the nucleus, which is positively charged, moves in the direction of the field while the electron cloud, which is negatively charged, moves in the opposite direction to the field. The resultant separation of the centres of positive and negative charge causes the atom to become an elemental dipole. If the separation of the nucleus and electron cloud is d, then the resultant dipole has magnitude ed and points from the negative towards the positive charge.[11] As the frequency of the electric field that acts on the atom increases, the amount of nuclear motion decreases much more rapidly than that of the electrons. At optical frequencies we generally neglect the motion of the nucleus, its great inertia compared to the electron cloud prevents it following the rapidly oscillating applied electric field. If the vector displacement of the ith electron on the atom from its equilibrium position is $\boldsymbol{x}_i$ then at any instant the atom has acquired a dipole moment

$$\mu = -\sum_{i=1}^{n} e\boldsymbol{x}_i \tag{A1.155}$$

where the summation runs over all the n electrons on the atom. The magnitude of the displacement of each electron depends on the value of the electric field $\boldsymbol{E}_i$ at the electron

$$k_i\boldsymbol{x}_i = -e\boldsymbol{E}_i \tag{A1.156}$$

where k_i is a force constant. A time-varying field $\boldsymbol{E}$ leads to a time-varying dipole moment. This dipole moment can become large if there is a resonance between the applied field and a particular electron on the atom. This happens if the frequency of the field is near the natural oscillation frequency of a particular electron. Classically, the resonance frequency of electron i is, by analogy with a mass attached to a spring, $\omega_i = \sqrt{k_i/m}$.

[11] The magnitude of the electronic charge is $e \simeq 1.6 \times 10^{-19}$ C, the charge on an electron is $-e$.

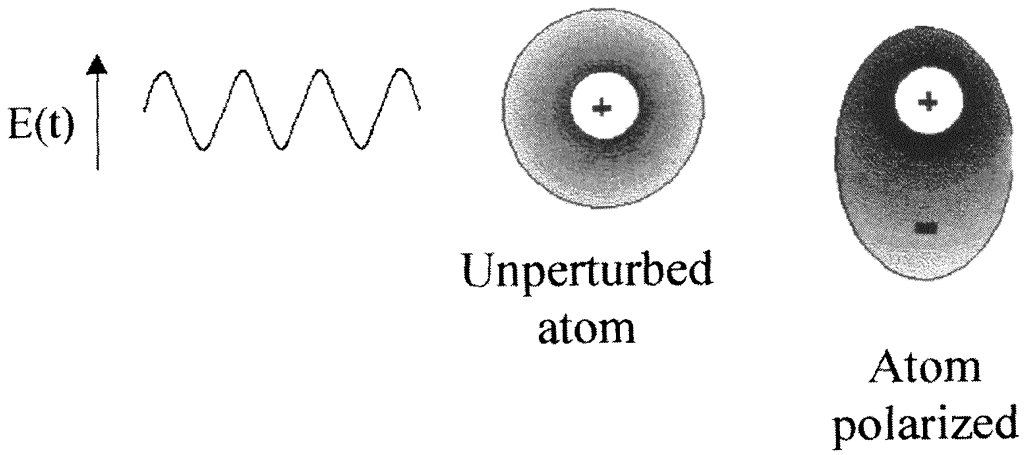


Figure A1.29. An atom exposed to an external electric field. The electron cloud and nucleus are displaced in opposite directions.

If the applied electric field is near this frequency, one electron in the summation in equation (A1.155) makes a dominant contribution to the dipole moment and we can treat the atom as a single electron oscillator. The physical significance of the resonant frequencies of the electrons is that they correspond to the frequencies of transitions that the electrons of the atom can make from one energy level to another. If we confine our attention to one of these resonances then we can treat an n-electron atom as a one-electron classical oscillator. A time-varying electric field $E(t)$, which we assume *a priori* to be at a frequency near to an atomic resonance, perturbs only a single electron to a significant degree, thereby inducing a dipole moment which varies at the same frequency as the applied field. The electron and nucleus are perturbed in opposite directions by the field as shown in figure A1.29.

The motion of the electrons on each of the particles in the medium, which for simplicity can be assumed to be identical, obeys the differential equation

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\Gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{k}{m}x = -\frac{e}{m}E(t). \tag{A1.157}$$

The terms on the left-hand side of the equation represent, reading from left to right: the acceleration of the electron, a damping term proportional to the electron velocity and a restoring force. These terms are balanced by the effect of the applied electric field $E(t)$. The restoring force is analogous to the restoring force acting on a mass suspended from a spring and given a small displacement from its equilibrium position. The damping can be regarded as a viscous drag that the moving electron experiences because of its interaction with the other electrons on the particle.

We take $E(t) = \mathcal{R}(E\mathrm{e}^{\mathrm{i}\omega t})$ and $x(t) = \mathcal{R}[X(\omega)\mathrm{e}^{\mathrm{i}\omega t}]$. The possibility that $E(t)$ and $x(t)$ are not in phase is taken into account by allowing the function $X(\omega)$ to include a phase factor. If we define a resonant frequency by $\omega_0 = \sqrt{k/m}$, then the differential equation (A1.157) becomes

$$(\omega_0^2 - \omega^2)X + 2\mathrm{i}\omega\Gamma X = -\frac{e}{m}E \tag{A1.158}$$

giving

$$X(\omega) = \frac{-(e/m)E}{\omega_0^2 - \omega^2 + 2\mathrm{i}\omega\Gamma}. \tag{A1.159}$$

This is the amplitude of the displacement of the electron from its equilibrium position as a function of the frequency of the applied field.

Near resonance $\omega \simeq \omega_0$, so

$$X(\omega \simeq \omega_0) = \frac{-(e/m)E}{2\omega_0(\omega_0 - \omega) + 2i\omega_0\Gamma}. \tag{A1.160}$$

The dipole moment of a single particle is $\mu(t) = -e[x(t)]$, which arises from the separation of the electron charge cloud and the nucleus.

If there are N electron oscillators per unit volume, there results a net polarization (dipole moment per unit volume) of:

$$P(t) = -Nex(t) = \mathcal{R}[P(\omega)e^{i\omega t}] \tag{A1.161}$$

where $P(\omega)$ is the complex amplitude of the polarization

$$\begin{aligned} P(\omega) = -NeX(\omega) &= \frac{(Ne^2/m)E_0}{2\omega_0(\omega_0 - \omega) + 2i\omega_0\Gamma} \\ &= \frac{-i(Ne^2/(2m\omega_0\Gamma))}{1 + i(\omega - \omega_0)/\Gamma}E_0 \end{aligned} \tag{A1.162}$$

The electronic susceptibility $\chi(\omega)$ is defined by the equation

$$P(\omega) = \epsilon_0\chi(\omega)E_0 \tag{A1.163}$$

where ϵ_0 is the permittivity of free space [12]. The susceptibility $\chi(\omega)$ is complex and can be written in terms of its real and imaginary parts as

$$\chi(\omega) = \chi'(\omega) - i\chi''(\omega) \tag{A1.164}$$

where the use of the negative sign is a common convention.

The polarization is

$$P(t) = \mathcal{R}[\epsilon_0\chi(\omega)E_0e^{i\omega t}] = \epsilon_0E_0\chi'(\omega)\cos\omega t + \epsilon_0E_0\chi''\sin\omega t. \tag{A1.165}$$

Therefore, $\chi'(\omega)$, the real part of the susceptibility, is related to the in-phase polarization, while $\chi''(\omega)$, the complex part, is related to the out of phase (quadrature) component. From equations (A1.162) and (A1.163),

$$\chi(\omega) = -i\left(\frac{Ne^2}{2m\omega_0\Gamma\epsilon_0}\right)\frac{1}{1 + i(\omega - \omega_0)/\Gamma} \tag{A1.166}$$

so

$$\chi'(\omega) = \left(\frac{Ne^2}{2m\omega_0\Gamma\epsilon_0}\right)\frac{(\omega_0 - \omega)/\Gamma}{1 + (\omega - \omega_0)^2/\Gamma^2} \tag{A1.167}$$

$$\chi''(\omega) = \left(\frac{Ne^2}{2m\omega_0\Gamma\epsilon_0}\right)\frac{1}{1 + (\omega - \omega_0)^2/\Gamma^2}. \tag{A1.168}$$

Changing to conventional frequency, $\nu = \omega/2\pi$, and putting $\Delta\nu = \Gamma/\pi$, which is the FWHM of the Lorentzian shape that describes $\chi''(\omega)$, we obtain

$$\chi''(\nu) = \left(\frac{Ne^2}{16\pi^2m\nu_0\epsilon_0}\right)\frac{\Delta\nu}{(\Delta\nu/2)^2 + (\nu - \nu_0)^2} \tag{A1.169}$$

$$\chi'(\nu) = \frac{2(\nu_0 - \nu)}{\Delta\nu}\chi''(\nu) = \left(\frac{Ne^2}{8\pi^2m\nu_0\epsilon_0}\right)\frac{\nu_0 - \nu}{(\Delta\nu/2)^2 + (\nu - \nu_0)^2} \tag{A1.170}$$

[12] $\epsilon_0 = 8.854 \times 10^{-12}$ Fm^{-1}

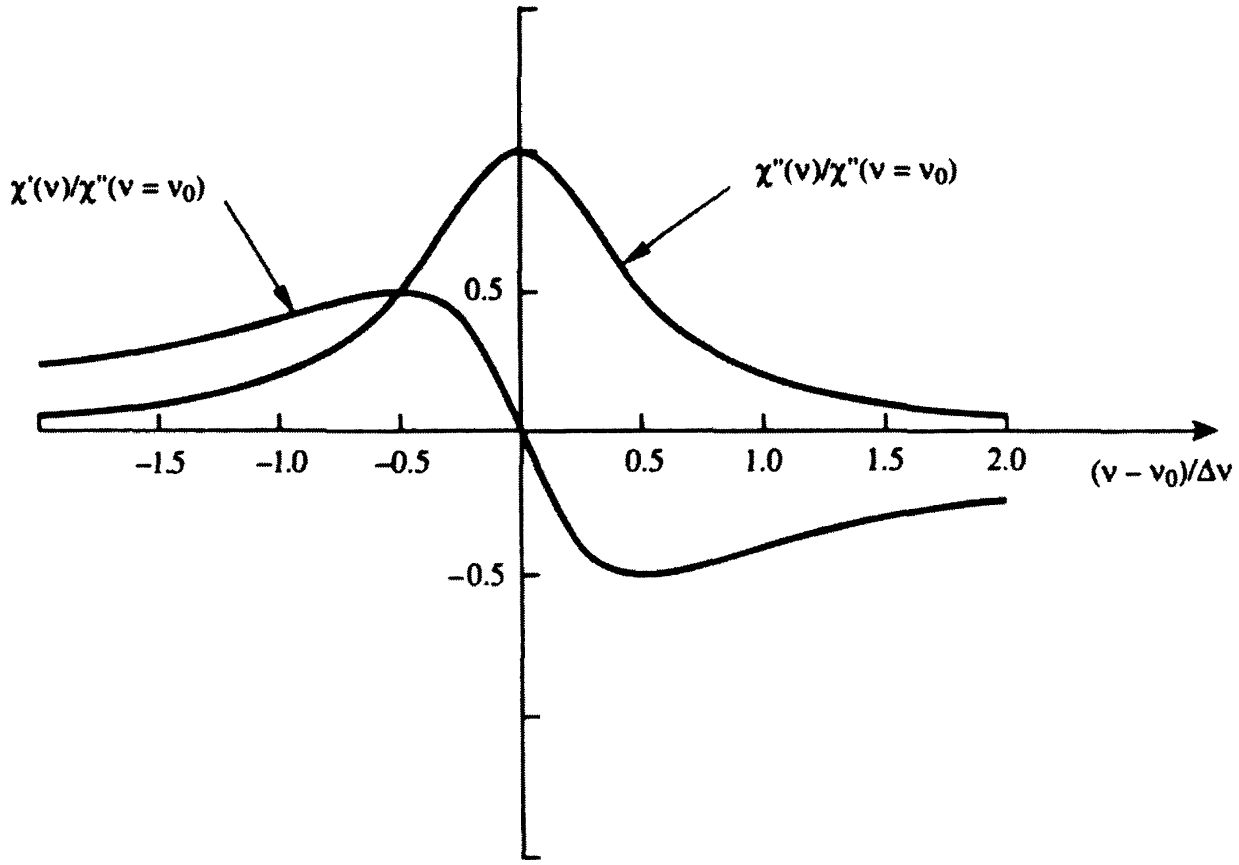


Figure A1.30. Frequency variation of the real, $\chi'(\nu)$, and imaginary, $\chi''(\nu)$, parts of the susceptibility calculated using the electron oscillator model.

Figure A1.30 is a plot of χ'' and χ' normalized to the peak value χ_0'' of χ''. Note that χ'' has the characteristic Lorentzian shape common to the frequency response of RLC circuits and homogeneously broadened spectral lines.[13]

A1.12.1 The connection between the complex susceptibility, gain and absorption

The complex susceptibility of a particle, or a medium containing a collection of particles, is closely related to the change in amplitude of an electromagetic wave passing through such a medium. We can make this connection by considering some fundamental concepts in classical electromagnetic theory.

The relationship between the applied electric field $\boldsymbol{E}$ and the electron displacement vector $\boldsymbol{D}$ is

$$\begin{aligned} \boldsymbol{D} &= \epsilon_0 \boldsymbol{E} + \boldsymbol{P} \\ &= \epsilon_0 (1 + \chi) \boldsymbol{E} \end{aligned} \tag{A1.171}$$

which, by introducing the dielectric constant $\epsilon_r = 1 + \chi$, can be written as

$$\boldsymbol{D} = \epsilon_0 \epsilon_r \boldsymbol{E}. \tag{A1.172}$$

The refractive index n of the medium is related to ϵ_r by $n = \sqrt{\epsilon_r}$.[14]

[13] It can be shown that although equations (A1.169) and (A1.170) have the correct frequency dependence for the susceptibility of a real particle, they have incorrect magnitudes. This arises because of inadequacies in the classical electron oscillator model and these inadequacies should be borne in mind before the classical model is used to make detailed predictions about the behaviour of a particle.

[14] In a magnetic material with relative magnetic permeability $\mu_r \neq 1$, $n = \sqrt{\mu_r \epsilon_r}$.

When an external electric field interacts with a group of particles, there are two contributions to the induced polarization, a macroscopic contribution $\boldsymbol{P}_{\mathrm{m}}$ from the collective properties of the particles, for example their arrangement in a crystal lattice, and a contribution from the polarization $\boldsymbol{P}_{\mathrm{t}}$ associated with transitions in the medium, so, in general

$$\boldsymbol{P} = \boldsymbol{P}_{\mathrm{m}} + \boldsymbol{P}_{\mathrm{t}}. \tag{A1.173}$$

Usually there are many transitions possible for the particles of the medium but only one will be in near resonance with the frequency of an applied field. $\boldsymbol{P}_{\mathrm{t}}$ is dominated by the contribution of this single transition near resonance. Far from any such resonance $\boldsymbol{P}_{\mathrm{t}}$ is negligible and $\boldsymbol{P} = \boldsymbol{P}_{\mathrm{m}}$, which allows us to define the macroscopic dielectric constant ϵ_{r} (far from resonance) from

$$\boldsymbol{P} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}_{\mathrm{m}} = \epsilon_0 \boldsymbol{E} + \chi_{\mathrm{m}} \epsilon_0 \boldsymbol{E} = \epsilon_{\mathrm{r}} \epsilon_0 \boldsymbol{E}. \tag{A1.174}$$

where χ_{m} is the macroscopic, non-resonant susceptibility. However, if the frequency of the electric field is near the frequency of a possible transition within the medium then there is a significant contribution to $\boldsymbol{P}$ from this transition. Other possible transitions far from resonance do not contribute and we can write

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}_{\mathrm{m}} + \boldsymbol{P}_{\mathrm{t}} = \epsilon_{\mathrm{r}} \epsilon_0 \boldsymbol{E} + \boldsymbol{P}_{\mathrm{t}}. \tag{A1.175}$$

$\boldsymbol{P}_{\mathrm{t}}$ is related to the complex susceptibility that results from the transition according to $\boldsymbol{P}_{\mathrm{t}} = \epsilon_0 \chi(\omega) \boldsymbol{E}$. Therefore, we can rewrite equation (A1.171) as

$$\boldsymbol{D} = \epsilon_0 \left[\epsilon_{\mathrm{r}} + \chi(\omega)\right] \boldsymbol{E} = \epsilon_0 \epsilon_{\mathrm{r}}^* \boldsymbol{E} \tag{A1.176}$$

so the complex susceptibility modifies the effective dielectric constant from ϵ_{r} to ϵ_{r}^*.

When an electromagnetic wave propagates through a medium with a complex susceptibility, both the amplitude and phase velocity of the wave are affected. This can be illustrated easily for a plane wave propagating in the z direction with a field variation $\sim \mathrm{e}^{i(\omega t - kz)}$, where the propagation constant, k is given by the expression

$$k = \omega \sqrt{\mu \epsilon} = \omega \sqrt{\mu_{\mathrm{r}} \mu_0 \epsilon_{\mathrm{r}} \epsilon_0}. \tag{A1.177}$$

The relative permeability μ_{r} is generally unity for optical materials. For a complex dielectric constant, equation (A1.177) can be rewritten as

$$k' = \omega \sqrt{\mu_0 \epsilon_{\mathrm{r}} \epsilon_0} \sqrt{\frac{1 + \chi(\omega)}{\epsilon_{\mathrm{r}}}} = k \sqrt{1 + \frac{\chi(\omega)}{\epsilon_{\mathrm{r}}}} \tag{A1.178}$$

where k' is now the new propagation constant, which differs from the non-resonant propagation constant k because of the complex susceptibility resulting from a transition. If $|\chi(\omega)| \ll \epsilon_{\mathrm{r}}$ equation (A1.178) can be simplified by the use of the binomial theorem to give

$$k' = k\left[1 + \frac{\chi(\omega)}{2\epsilon_{\mathrm{r}}}\right] = k\left[1 + \frac{\chi'(\omega)}{2\epsilon_{\mathrm{r}}} - \frac{\mathrm{i}\chi''(\omega)}{2\epsilon_{\mathrm{r}}}\right]. \tag{A1.179}$$

The wave now propagates through the medium as $\mathrm{e}^{-\mathrm{i}k'z}$. Written out in full, the electric field varies as

$$\begin{aligned} E &= E_0 \exp\left(\mathrm{i}\left\{\omega t - k\left[1 + \frac{\chi'(\omega)}{2\epsilon_{\mathrm{r}}} - \frac{\mathrm{i}\chi''(\omega)}{2\epsilon_{\mathrm{r}}}\right] z\right\}\right) \\ &= E_0 \exp\left(\mathrm{i}\left\{\omega t - k\left[1 + \frac{\chi'(\omega)}{2\epsilon_{\mathrm{r}}}\right]\right\}\right) \exp\left[-\frac{k\chi''(\omega)}{2\epsilon_{\mathrm{r}}} z\right]. \end{aligned} \tag{A1.180}$$

Clearly, this is a wave whose phase velocity is

$$c' = \frac{\omega}{k[1 + \chi'(\omega)/2\epsilon_r]} = \frac{\omega}{k + \Delta k} \tag{A1.181}$$

and whose field amplitude changes exponentially with distance. If we write

$$\gamma = -\frac{k\chi''(\omega)}{\epsilon_r} \tag{A1.182}$$

then the wave changes its electric field amplitude with distance according to $e^{(\gamma/2)z}$.

The intensity of the wave is $I \propto E(z, t)E^*(z, t)$, which changes as the wave passes through the medium as $I \propto e^{\gamma z}$. We can identify γ as the familiar gain coefficient of the medium, which was calculated previously by considering the spontaneous and stimulated radiative jumps between two energy levels.

Now from equation (A1.182)

$$\gamma(\nu) = -\frac{k\chi''(\nu)}{n^2} = -\left(\frac{2\pi\nu_0 n}{c_0}\right)\frac{\chi''(\nu)}{n^2} \tag{A1.183}$$

which, from equation (A1.184), obtained by consideration of the system as a collection of classical oscillators, is

$$\gamma(\nu) = -\left(\frac{2\pi\nu}{nc_0}\right)\left(\frac{Ne^2}{16\pi^2 m\nu_0\epsilon_0}\right)\frac{\Delta\nu}{(\Delta\nu/2)^2 + (\nu - \nu_0)^2} \tag{A1.184}$$

which is *always negative*. Clearly, this is an incorrect result since we have previously shown that

$$\gamma(\nu) = \left(N_2 - \frac{g_2}{g_1}N_1\right)\frac{c^2 A_{21}}{8\pi\nu^2}g(\nu_0, \nu) \tag{A1.185}$$

which can be positive or negative depending on the sign of $N_2 - (g_2/g_1)N_1$. Thus, the classical electron oscillator model appears to predict only absorption of incident radiation. It is possible, however, within the framework of the classical electron oscillator model to show that, in certain conditions, stimulated emission can occur. Although the classical electron oscillator model is instructive, it is not entirely adequate in describing the interaction between particles and radiation. It is better to accept that $\gamma(\nu) = -k\chi''(\nu)/n^2$ and use equation (A1.185) as the expression for $\gamma(\nu)$. In this case we find that the imaginary part of the complex susceptibility of the medium is

$$\chi''(\nu) = -\frac{n^2\gamma(\nu)}{k} = -\left(\frac{n^2 c}{2\pi\nu}\right)\gamma(\nu) \tag{A1.186}$$

which, from equation (A1.183), gives

$$\chi''(\nu) = -\frac{[N_2 - (g_2/g_1)N_1]n^2 c^3 A_{21}}{8\pi^3\nu^3\Delta\nu}\frac{1}{1 + [2(\nu - \nu_0)/\Delta\nu]^2}. \tag{A1.187}$$

This quantum mechanical susceptibility is negative or positive depending on whether $N_2 - (g_2/g_1)N_1$ is positive or not. A *negative* value of $\chi''(\nu)$ corresponds to a system in population inversion.

A1.12.2 The classical oscillator explanation for stimulated emission

If there were no applied electric field acting on the electron, then, from equation (A1.157), the position of the electron would satisfy the equation

$$x(t) = x_0 e^{-\Gamma t}\cos(\omega_0 t + \phi_0) \tag{A1.188}$$

where ω_0 is the resonant frequency and ϕ_0 is a phase factor set by the initial conditions. If at time $t = 0$ the position and velocity of the electron are a_0, v_0, respectively, then

$$\begin{aligned} a_0 &= x_0 \cos\phi_0 \\ v_0 &= -\Gamma x_0 \cos\phi_0 - \omega_0 x_0 \sin\phi_0. \end{aligned} \tag{A1.189}$$

When the electric field is applied, we have already seen that, in the steady state, energy is apparently only absorbed. However, this impression is erroneous. It neglects that, in reality, no electromagnetic field interacts indefinitely with an electron. Therefore, we must consider what happens when an electron, which can already be regarded as oscillating if it is in an excited state, is suddenly subjected to the additional perturbation of an applied field. We are going to be interested in the behaviour of the electron over the first few cycles of the applied field so we neglect damping and write

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \omega_0^2 = -\frac{eE_0}{2m}(\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}) \tag{A1.190}$$

where the applied field is $E_0 \cos\omega t$ and has been written in its complex exponential form. By introducing a new variable $z = \dot{x} + \mathrm{i}\omega_0 x$, where the dot indicates differentiation with respect to time, equation (A1.190) can be rewritten in the form

$$\frac{\mathrm{d}z}{\mathrm{d}t} - \mathrm{i}\omega_0 z = -\frac{eE_0}{2m}(\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}). \tag{A1.191}$$

This equation can be solved by multiplying each term by $\mathrm{e}^{-\mathrm{i}\omega_0 t}$ and then integrating to give

$$z\mathrm{e}^{-\mathrm{i}\omega_0 t} = -\frac{eE_0}{2m}\int (\mathrm{e}^{\mathrm{i}(\omega-\omega_0)t} + \mathrm{e}^{-\mathrm{i}(\omega+\omega_0)t})\,\mathrm{d}t. \tag{A1.192}$$

This gives

$$\frac{\mathrm{d}x}{\mathrm{d}t} + \mathrm{i}\omega_0 x = -\frac{eE_0}{2m}\left[-\frac{\mathrm{i}\mathrm{e}^{\mathrm{i}\omega t}}{(\omega-\omega_0)} + \frac{\mathrm{i}\mathrm{e}^{-\mathrm{i}\omega t}}{(\omega+\omega_0)} + A\mathrm{e}_0^{\mathrm{i}\omega t}\right] \tag{A1.193}$$

where A is a constant of integration. By integrating a second time in a similar manner and introducing the initial values of position and velocity, it is straightforward to show that the final solution is

$$x(t) = -\frac{eE_0}{m}\left[\cos\omega_0 t - \frac{\cos\omega t}{(\omega^2-\omega_0^2)}\right] + \sqrt{\left(\frac{v_0}{\omega_0}\right)^2 + x_0^2}\cos(\omega_0 t + \phi) \tag{A1.194}$$

where $\tan\phi = -v_0/\omega_0 x_0$.

By the use of the trigonometrical identity,

$$\cos X - \cos Y = -2\sin\left(\frac{X+Y}{2}\right)\sin\left(\frac{X-Y}{2}\right) \tag{A1.195}$$

and assuming that the applied frequency is close to resonance, equation (A1.194) can be written

$$x(t) = -\frac{eE_0}{2m\omega_0} t\sin\omega_0 t + \sqrt{\left(\frac{v_0}{\omega_0}\right)^2 + x_0^2}\cos(\omega_0 t + \phi). \tag{A1.196}$$

Thus, near resonance the amplitude of oscillation will increase linearly with time, which is, of course, a consequence of our neglect of damping. It is more interesting, however, to use equation (A1.196) to calculate the work done during the first n cycles of the applied field. This work is calculated as the work done by the

electric field in polarizing the medium: the polarization **P** is proportional to electron displacement. The work done is $\boldsymbol{E} \cdot \partial \boldsymbol{P}/\partial t$ [1]. During the first n cycles, the total work done by the field is

$$W = \int_0^{2n\pi/\omega_0} \boldsymbol{E} \cdot \frac{\partial \boldsymbol{P}}{\partial t}\,\mathrm{d}t = -NeE_0 \int_0^{2n\pi/\omega_0} (\cos\omega_0 t)\dot{x}(t)\,\mathrm{d}t \tag{A1.197}$$

where N is the total number of particles per unit volume. Writing $(v_0/\omega_0)^2 + x_0^2 = a^2$ and substituting from equation (A1.196) gives

$$\begin{aligned} W = & -NeE_0 \int_0^{2n\pi/\omega_0} \left[-\frac{eE_0}{m\omega_0}\sin 2\omega_0 t - \frac{eE_0 t}{4m}(1+\cos 2\omega_0 t) \right. \\ & \left. -\frac{a\omega_0}{2}\sin\phi + \frac{a\omega_0}{2}\sin(2\omega_0 t + \phi) \right] \mathrm{d}t \end{aligned} \tag{A1.198}$$

Clearly, the first and last terms of the integrand average to zero over a whole number of cycles. The remaining terms can be integrated to give

$$W = \frac{Ne^2E_0^2}{m}\left(\frac{n^2\pi^2}{2\omega_0^2} + \frac{n\pi ma}{eE_0}\sin\phi\right). \tag{A1.199}$$

This work done by the applied field is negative, implying that the oscillating electrons supply energy to the field if $\sin\phi < 0$ and $|\sin\phi| > eE_0 n\pi/2ma\omega_0^2$. This is the condition set by classical theory for stimulated emission to occur. Because the charge on the electron is negative, stimulated emission can only result if the electron velocity when the applied field is turned on is in the direction of the field. If the electron velocity is in the same direction as the field, the electron is decelerated by the field and, consequently, radiates energy. If the electron were accelerated by the field then absorption of energy from the field would occur.

There is a maximum number of cycles of the applied field after which the oscillating electrons start, and continue indefinitely, to absorb energy. This is set by the condition

$$n < 2ma\omega_0^2/eE_0\pi. \tag{A1.200}$$

After a long enough time, the motion of the electron is dominated by the first term in equation (A1.196) and can be written

$$x(t) = -\frac{eE_0}{2m\omega_0} t\sin\omega_0 t \tag{A1.201}$$

and the electron velocity is, for large enough t,

$$\dot{x}(t) \simeq -\frac{eE_0}{2m} t\cos\omega_0 t. \tag{A1.202}$$

The electron now has a velocity that is oppositely directed from the applied field, the electron is being accelerated and absorbs energy from the field.

We can conclude by saying that when an electric field near resonance is applied to an already oscillating electron, stimulated emission can occur at early times provided the initial velocity of the electron is in the same direction as the field.

A1.13 From amplifier to oscillator—the feedback structure

A feedback structure is used to channel output emissions from a laser amplifier back through the amplifier. The simplest example consists of a pair of plane mirrors placed parallel to each other but at opposite ends

of the ampifying medium. The structure is called the laser resonator or optical cavity (see section A2.1). It is often referred to as a Fabry–Pérot resonator[15]. The start of laser oscillation begins with spontaneous emission that occurs close to the axis of the resonator system: the direction perpendicular to both mirrors. This spontaneous emission is redirected by the mirrors back through the amplifying medium and its intensity grows. The fact that only photons that are travelling close to the axis direction can make many passes through the amplifying medium, especially if this is of small width, explains the directional character of laser beams.

The frequency spectrum of laser radiation is controlled by the interaction between the lineshape function of the amplifying transition and the properties of the optical resonator. A plane wave bouncing back and forth between the two flat mirrors of an empty cavity will 'resonate' if the wavelength of the radiation satifies the simple condition, with m integer,

$$\frac{m\lambda}{2} = \ell \tag{A1.203}$$

where ℓ is the axial spacing between the two parallel mirrors. This is the simple condition for the bouncing waves to be in phase and develop a maximum amplitude standing wave between the two mirrors, much like the acoustic resonances of an open or closed pipe. When an amplifying medium is present this resonance condition must still be satisfied and determines the precise frequency (or frequencies) at which laser oscillation can occur. In this case the effective 'length' ℓ of the cavity is modified by the presence of the amplifier.

Before exploring this further, let us consider at what frequency a laser would oscillate if the resonator did not interact with the gain profile of the amplifying medium in any way. Suppose the amplifying medium has a gain profile (gain/frequency response) of a Gaussian form, as shown in figure A1.31. Such a gain profile occurs in a gaseous amplifying medium where the individual homogeneous lineshapes of the particles are significantly narrower than the overall Doppler width of the spontaneous transition.

The maximum gain of the medium is at frequency ν_0, the line centre, so it is perhaps logical to expect that oscillation will build up at this frequency rather than at any other. If we view the buildup of oscillation as a process triggered by spontaneous emission we can see why this is so. A photon travelling in a direction that keeps it bouncing back and forth within the resonator is more likely to be emitted in a narrow band of frequencies $\Delta\nu_0$ near ν_0 than in some other band $\Delta\nu_1$ at frequency ν_1. As oscillation builds up, one can imagine photons spontaneously emitted at all points of the lineshape being amplified to some extent but oscillation at ν_0 builds up fastest. As its intensity grows it depletes the atomic population by causing sufficient stimulated emission that the medium ceases to be amplifying at frequencies near ν_0 (within a few homogeneous widths, say). If the medium has a homogeneous (Lorentzian) gain profile, then since photons oscillating at ν_0 can stimulate emission from all the atoms in the medium, it is easy to see that oscillation at frequency ν_0 can suppress oscillation at any other frequency under the gain profile. The possibility of additional oscillation at frequencies far away from ν_0 in an inhomogeneously broadened gaseous amplifier is not precluded by this discussion: this, in fact, oftens happens, as we shall see later.

The monochromatic character of the oscillation can be predicted by a simple consideration of the shape of the gain profile of the amplifier. In the early stages of oscillation, photons with a frequency distribution $g(\nu_0, \nu)$ (the total lineshape) are being amplified in a material whose gain/frequency response is $\gamma(\nu)$ (proportional to $g(\nu_0, \nu)$). The amplification process changes the lineshape of the emitted photons circulating in the cavity by a process that is dependent on the product of $g(\nu_0, \nu)$ and $\gamma(\nu)$, that is on $[g(\nu_0, \nu)]^2$. The resulting profile of the laser radiation is dependent on higher powers of $[g(\nu_0, \nu)]^2$ as the oscillation is dependent on many passes of photons back and forth through the amplifying medium. For Gaussian lineshapes like $e^{-[2(\nu-\nu_0)/\Delta\nu_D]^2 \ln 2}$, which is like e^{-x^2/σ^2}, the product of two lineshapes produces a narrower profile, for example

$$[e^{-x^2/\sigma^2}]^2 = e^{-2x^2/\sigma^2} = e^{-x^2/(\sigma/\sqrt{2})^2}, \tag{A1.204}$$

[15] Named for Marie P A C Fabry and Jean B G G A Pérot, the inventors of the plane-parallel mirror structure developed originally as an optical instrument for interferometry and first described in 1899.

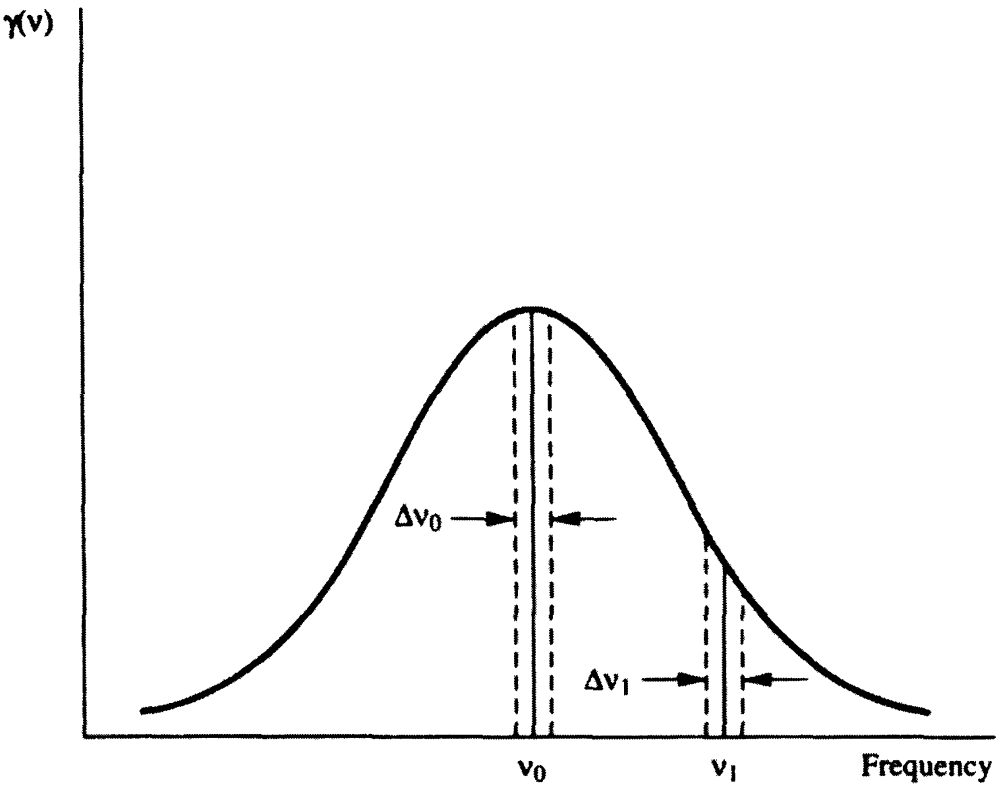


Figure A1.31. To illustrate how laser oscillation might build up at two different frequencies in a noninteractive laser cavity.

a function that has a width $1/\sqrt{2}$ of than the original. The same can also be shown to be true for Lorentzian profiles. In both cases, the gain of the medium causes a narrowing of the original, spontaneously emitted lineshape. Thus, we can see that in a non-interactive laser cavity, the laser oscillation will be highly monochromatic and at the line centre.

A1.14 Optical resonators containing an amplifying media

When an optical resonator is filled with an amplifying medium, laser oscillation will occur at specific frequencies if the gain of the medium is large enough to overcome the loss of energy through the mirrors and by other loss mechanisms within the laser medium. The onset of laser oscillation and the frequency, or frequencies, at which it occurs is governed by threshold amplitude and phase conditions. Once laser oscillation is established, it stabilizes at a level that depends on the saturation intensity of the amplifying medium and the reflectance of the laser mirrors.

Figure A1.32 represents an optical (Fabry–Pérot) resonator, whose interior is filled with an amplifying medium and which has plane mirrors. We consider the complex amplitudes of the waves bouncing backwards and forwards normally between the resonator mirrors. These waves result from an incident beam with electric vector E_0 at the first mirror, as shown in figure A1.32; where E_0 is the complex amplitude at some reference point. The reflection and transmission coefficients in the various directions at the mirrors are as shown in the figure. For example, t is the transmission coefficient for electromagnetic fields passing through the left-hand reflector and r_1 is the reflection coefficient for fields striking the left-hand reflector from inside the resonator. A, A_1 and A_2 are absorption coefficients, which lead to energy dissipation in the resonator reflectors. These coefficients are not used explicitly in the analysis that follows but they modify the values of the reflection and transmission coefficients. If the mirrors are lossless, then their reflectances, R_1, R_2 and transmittances,

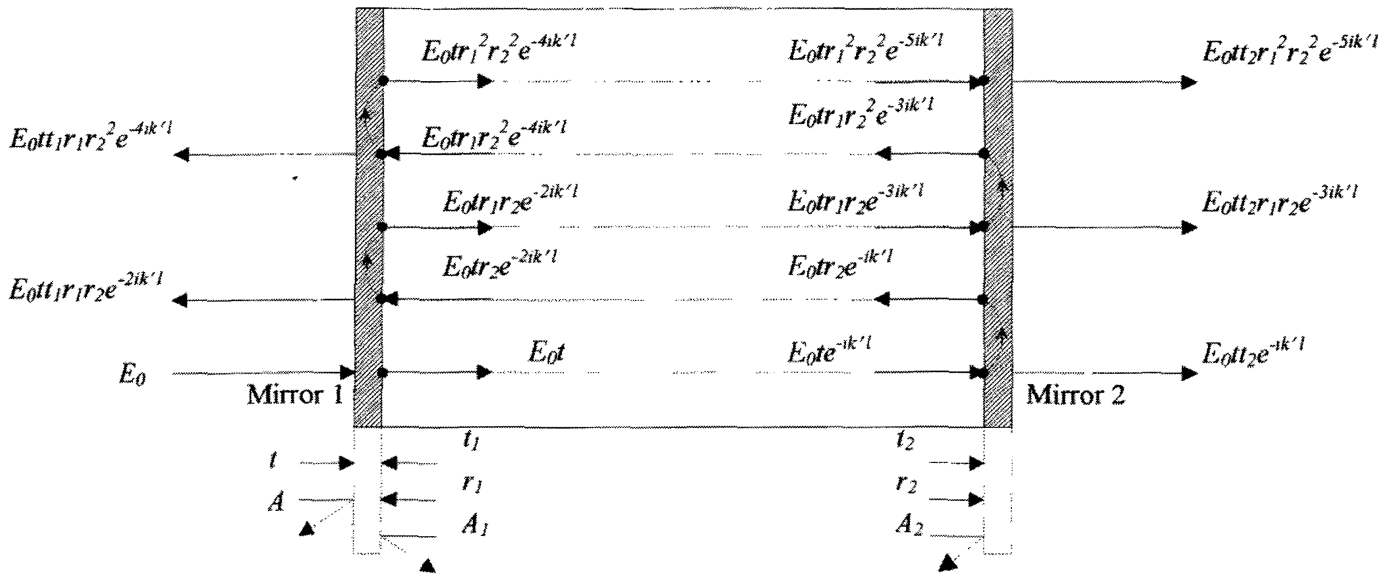


Figure A1.32. The amplitude of the electric field vectors of the successively transmitted, amplified and reflected waves in a Fabry–Pérot resonator system containing an amplifying (or absorbing) medium. The absorption factors A, A_1 and A_2 are not used explicitly in the analysis given in the text, but they modify the values of t, t_1 and, t_2.

T_1, T_2, obey $R_1 + T = 1$, etc and $|r_1|^2 = R_1$, $|r_2|^2 = R_2$, $|t|^2 = |t_1|^2 = T_1$, $|t_2|^2 = T_2$. The absorption coefficients, A, A_1, A_2, at the two mirrors include any reflection losses or scattering that send energy out of the resonator: we do not at this stage include diffraction losses, which result from the finite lateral dimensions of the mirrors or medium.

If there were nothing inside the resonator then a wave propagating between the mirrors would propagate as $E_0 e^{i(\omega t - kz)}$ to the right and $E_0 e^{i(\omega t + kz)}$ to the left. The presence of a gain medium changes the otherwise passive propagation factor k to

$$k'(\omega) = k + \Delta k = k + \frac{k\chi'(\omega)}{2n^2} \tag{A1.205}$$

and the gain coefficient $\gamma(\omega) = -k\chi''(\omega)/n^2$ causes the amplitude of the wave to change with distance as $e^{(\gamma/2)z}$. We allow for the possible existence of a distributed loss per pass given by an absorption coefficient α. Such absorption causes a fractional change in intensity for a single pass through the medium of $e^{-\alpha\ell}$. Such a distributed loss could, for example, arise from scattering by imperfections in a solid laser medium. This distributed loss modifies the complex amplitude by a factor $e^{-i\alpha\ell/2}$ per pass. Therefore, the full propagation constant of the wave in the presence of both gain and loss is

$$k'(\omega) = k + k\frac{\chi'(\omega)}{2n^2} - \frac{ik\chi''(\omega)}{2n^2} - \frac{i\alpha}{2} \tag{A1.206}$$

and the wave propagates as $e^{i(\omega t \pm k'z)}$.

A wave travelling to the right with complex amplitude E_0 at plane $z = 0$ in the resonator, the left-hand mirror, has at plane ℓ, the right-hand mirror, become

$$E = E_0 e^{i(\omega t - k\ell)} = E_0 e^{-ik\ell} e^{i\omega t} = E_0' e^{i\omega t}.$$

This wave then begins to propagate to the left as

$$E = E_0' e^{i(\omega t + kz)}.$$

At plane $-\ell$, the left-hand mirror, with the right-hand mirror now taken as the origin, it has become once more a wave travelling to the right

$$E = E_0 e^{ik\ell} e^{i(\omega t - k\ell)} = E_0 e^{-2ik\ell} e^{i\omega t}.$$

In this way we can write down the complex amplitudes of successive rays travelling at normal incidence between the two reflectors, as shown in figure A1.32.

The output beam through the right-hand mirror arises from the transmission of waves travelling to the right: its total electric field amplitude is

$$\begin{aligned} E_t &= E_0 t t_2 e^{-ik'\ell} + E_0 t t_2 r_1 r_2 e^{-3ik'\ell} + \cdots \\ &= E_0 t t_2 e^{-ik'\ell}(1 + r_1 r_2 e^{-2ik'\ell} + r_1^2 r_2^2 e^{-4ik'\ell} + \cdots) \\ &= \frac{E_0 t t_2 e^{-ik'\ell}}{1 - r_1 r_2 e^{-2ik'\ell}} \\ &= \frac{E_0 t t_2 e^{-i(k+\Delta k)\ell} e^{(\gamma-\alpha)\ell/2}}{1 - r_1 r_2 e^{-2i(k+\Delta k)\ell} e^{(\gamma-\alpha)\ell}} \end{aligned} \tag{A1.207}$$

where

$$\gamma(\nu) = \left[N_2 - \left(\frac{g_2}{g_1}\right) N_1\right] \left(\frac{c^2 A_{21}}{8\pi\nu^2}\right) g(\nu_0, \nu). \tag{A1.208}$$

The ratio of input to output intensities is

$$\left(\frac{E_t}{E_0}\right) = \frac{I_t}{I_0} = \frac{t^2 t_2^2 e^{(\gamma-\alpha)\ell}}{(1 - r_1 r_2 e^{-2i(k+\Delta k)\ell} e^{(\gamma-\alpha)\ell})(1 - r_1 r_2 e^{2i(k+\Delta k)\ell} e^{(\gamma-\alpha)\ell})} \tag{A1.209}$$

which becomes

$$\frac{I_t}{I_0} = \frac{t^2 t_2^2 e^{(\gamma-\alpha)\ell}}{1 + r_1^2 r_2^2 e^{2(\gamma-\alpha)\ell} - 2 r_1 r_2 e^{(\gamma-\alpha)\ell}[\cos 2(k + \Delta k)\ell]}. \tag{A1.210}$$

In a passive resonator, which has no gain γ or loss α, $\Delta k = 0$, and if $|r_1|^2 = |r_2|^2 = R$

$$\frac{I_t}{I_0} = \frac{T^2}{1 + R^2 - 2R\cos 2k\ell}. \tag{A1.211}$$

This function is shown in figure A1.33 for a few values of the variable $2k\ell = \delta$ and for three different values of the mirror reflectance R. Resonances in the transmittance (I_t/I_0) of this structure satisfy $2k\ell = m\pi$, and correspond to wavelengths that satisfy the condition $m\lambda/2 = \ell$. These wavelengths also correspond to the largest standing wave fields between the two mirrors. If R is close to 1, then the resonances in figure A1.33 are very sharp. It is easy to show [1] that they approximate a Lorentzian shape with a FWHM defined as

$$\Delta\nu_{\frac{1}{2}} = \Delta\nu / F \tag{A1.212}$$

where $\Delta\nu = c/2\ell$, and $F = \pi\sqrt{R}/(1 - R)$. $\Delta\nu$ is called the *free spectral range* of the resonator, and F is called its *finesse*.

In a resonator containing an active medium, as $\gamma - \alpha$ increases from zero, the denominator of equation (A1.207) approaches zero and the whole expression blows up when

$$r_1 r_2 e^{-2i(k+\Delta k)\ell} e^{(\gamma-\alpha)\ell} = 1. \tag{A1.213}$$

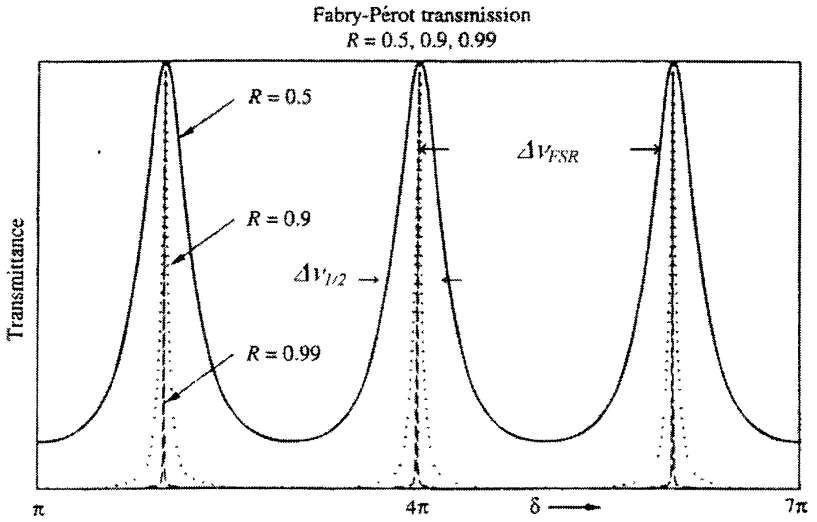


Figure A1.33. Theoretical transmittance (I_t/I_0) characteristic of a laser resonator calculated from equation (A1.211) for different values of (equal) mirror reflectance R. The theoretical reflection characteristic that corresponds to these curves can be viewed by turning the picture upside down.

When this happens we have an infinite amplitude transmitted wave for a finite amplitude incident wave. In other words, a finite amplitude transmitted wave for zero incident wave—*oscillation*. Physically, equation (A1.213) is the condition that must be satisfied for a wave to make a complete round trip inside the resonator and return to its starting point with the same amplitude and, apart from a multiple of 2π, the same phase.

Equation (A1.213) provides an amplitude condition for oscillation that gives an expression for the threshold gain constant, $\gamma_t(\nu)$,

$$r_1 r_2 e^{[\gamma_t(\omega)-\alpha]\ell} = 1. \tag{A1.214}$$

To satisfy equation (A1.213), $e^{-2i(k+\Delta k)\ell}$ must be real, which provides us with the phase condition

$$2[k + \Delta k(\nu)]\ell = 2\pi m \qquad m = 1, 2, 3 \ldots \tag{A1.215}$$

The threshold gain coefficient can be written

$$\gamma_t(\nu) = \alpha - \frac{1}{\ell} \ln r_1 r_2 \tag{A1.216}$$

which from the gain equation (A1.185) gives the population inversion needed for oscillation

$$\left(N_2 - \frac{g_2}{g_1} N_1\right)_t = \frac{8\pi}{g(\nu_0, \nu) A_{21} \lambda^2} \left(\alpha - \frac{1}{\ell} \ln r_1 r_2\right). \tag{A1.217}$$

For a homogeneously broadened transition, the parametric variation of equation (A1.217) that depends on the gain medium can be written as

$$\left(N_2 - \frac{g_2}{g_1} N_1\right)_t \propto \frac{\Delta\nu}{A_{21}\lambda^2}. \tag{A1.218}$$

Whereas, for an inhomogeneously broadened transition, since $\Delta\nu_D \propto 1/\lambda$,

$$\left(N_2 - \frac{g_2}{g_1} N_1\right)_t \propto \frac{1}{A_{21}\lambda^3}. \tag{A1.219}$$

Clearly, lower inversions are needed to achieve laser oscillation at longer wavelengths. It is much easier to build lasers that oscillate in the infrared than at visible, ultraviolet or x-ray wavelengths. For example, in an inhomogeneously broadened laser, a population inversion 10^6 times greater would be required for oscillation at 200 nm than at 20 μm (all other factors such as A_{21} being equal). In practice, since A_{21} factors generally increase at shorter wavelengths, the difference in population inversion may not need to be as great as this.

In a resonator such as is shown in figure A1.32, if $R_1 = r_1^2 \simeq 1$, $R_2 = r_2^2 \simeq 1$ and distributed losses are small, a wave starting with intensity I inside the resonator will, after one complete round trip, have intensity $IR_1R_2e^{-2\alpha\ell}$, the change in intracavity intensity after one round trip is

$$dI = (R_1R_2e^{-2\alpha\ell} - 1)I. \tag{A1.220}$$

This loss occurs in a time $dt = 2\ell/c$. So,

$$\frac{dI}{dt} = cI[R_1R_2e^{-2\alpha\ell} - 1]/2\ell. \tag{A1.221}$$

This equation has the solution

$$I = I_0 \exp\{-[1 - R_1R_2e^{-\alpha\ell}]ct/2\ell\} \tag{A1.222}$$

where I_0 is the intensity at time $t = 0$. The time constant for intensity (energy) loss is

$$\tau_0 = \frac{2\ell}{c(1 - R_1R_2e^{-2\alpha\ell})}. \tag{A1.223}$$

Now if $R_1R_2e^{-2\alpha\ell} \simeq 1$, with α small as we have assumed here, then

$$(1 - R_1R_2e^{-2\alpha\ell}) \simeq -\ln(R_1R_2e^{-2\alpha\ell}) = -\ln(R_1R_2) + 2\alpha\ell \tag{A1.224}$$

and we get

$$\tau_0 = \frac{2\ell}{c(2\alpha\ell - \ln R_1R_2)} = \frac{1}{c[\alpha - (1/\ell)\ln r_1r_2]}. \tag{A1.225}$$

Thus, the threshold population inversion can be written

$$N_t = \frac{8\pi}{A_{21}\lambda^2 g(\nu)c\tau_0}. \tag{A1.226}$$

Threshold population inversion—numerical example. For the 488 nm transition in the argon ion laser (see chapter B3.5)

$$\lambda = 488 \text{ nm} \qquad c = 3 \times 10^8 \text{ ms}^{-1} \qquad A_{21} \simeq 10^9 \text{ s}^{-1} \qquad \Delta\nu_D \sim 3 \text{ GHz}.$$

Take $\ell = 1$ m, $R_1 = 100\%$, $R_2 = 90\%$ (typical values for a practical device). Since this is a gas laser internal losses are easily kept small so $\alpha \simeq 0$. In this case

$$\begin{aligned}\tau_0 &= 2\ell/c(1 \quad R_1R_2) \\ &= 66.67 \text{ ns}.\end{aligned}$$

For oscillation at or near line centre,

$$g(\nu_0, \nu_0) = \frac{2}{\Delta\nu_D}\sqrt{\frac{\ln 2}{\pi}} = \frac{0.94}{\Delta\nu_D} \sim \frac{1}{\Delta\nu_D}.$$

The threshold inversion is, from equation (A1.226),

$$N_t = \frac{8\pi \times 3 \times 10^9}{10^9 \times (488 \times 10^{-9})^2 \times 3 \times 10^8 \times 66.67 \times 10^{-9}} = 1.58 \times 10^{13} \text{ m}^{-3}.$$

A1.15 The oscillation frequency

To determine the frequency at which laser oscillation can occur we return to the phase condition, equation (A1.215). This phase condition was

$$(k + \Delta k)\ell = m\pi \tag{A1.227}$$

which, from equation (A1.205), gives

$$k\ell\left[1 + \frac{\chi'(\nu)}{2n^2}\right] = m\pi. \tag{A1.228}$$

Now, from equation (A1.170),

$$\chi'(\nu) = \frac{2(\nu_0 - \nu)}{\Delta\nu}\chi''(\nu) \tag{A1.229}$$

where ν_0 is the line-centre frequency and $\Delta\nu$ is its *homogeneous* FWHM and

$$\gamma(\nu) = -\frac{k\chi''(\nu)}{n^2}. \tag{A1.230}$$

So we must have

$$\frac{2\pi\nu\ell}{c}\left[1 - \frac{(\nu_0 - \nu)}{\Delta\nu}\frac{\gamma(\nu)}{k}\right] = m\pi \tag{A1.231}$$

and on rearranging,

$$\nu\left[1 - \frac{(\nu_0 - \nu)}{\Delta\nu}\frac{\gamma(\nu)}{k}\right] = \frac{mc}{2\ell} = \nu_m \tag{A1.232}$$

where ν_m is the mth resonance of the passive laser resonator in normal incidence as calculated previously. equation (A1.232) can be rewritten as

$$\nu = \nu_m - (\nu - \nu_0)\frac{\gamma(\nu)c}{2\pi\Delta\nu}. \tag{A1.233}$$

We expect the actual oscillation frequency ν to be close to ν_m so we can write $(\nu - \nu_0) \simeq (\nu_m - \nu_0)$ and $\gamma(\nu) \simeq \gamma(\nu_m)$, to give

$$\nu = \nu_m - (\nu_m - \nu_0)\frac{\gamma(\nu_m)c}{2\pi\Delta\nu}. \tag{A1.234}$$

At threshold

$$\gamma_t(\nu_m) = \alpha - \frac{1}{\ell}\ln r_1 r_2 \tag{A1.235}$$

and if $\alpha \simeq 0$, $r_1 = r_2 = \sqrt{R}$

$$\gamma_t(\nu_m) = \frac{1 - R}{\ell}. \tag{A1.236}$$

Now the FWHM of the passive resonances (the transmission intensity maxima of the Fabry–Pérot resonator) is

$$\Delta\nu_{1/2} = \frac{\Delta\nu_{FSR}}{F} = \frac{c(1 - R)}{2\pi\ell\sqrt{R}} \tag{A1.237}$$

which, with $R \simeq 1$, gives

$$\Delta\nu_{1/2} = \frac{c(1 - R)}{2\pi\ell} \tag{A1.238}$$

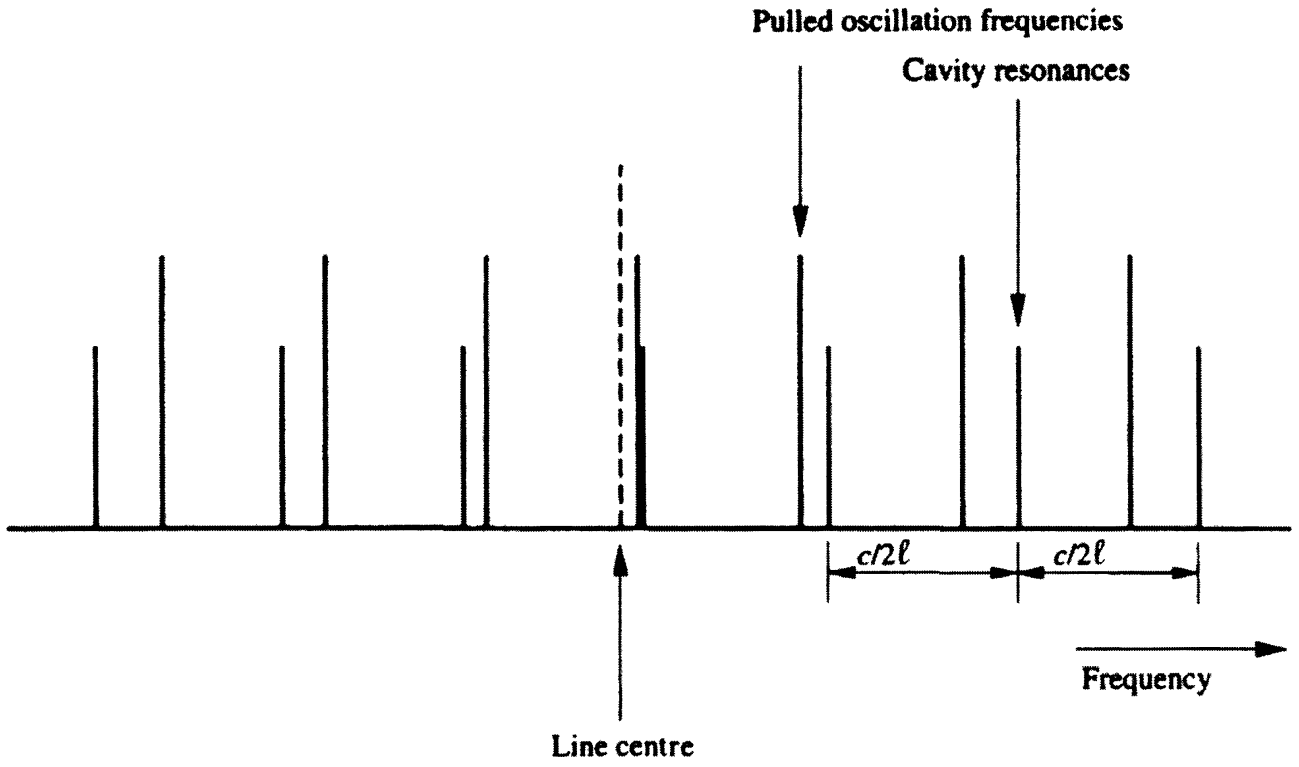


Figure A1.34. Relative position of line centre, cavity resonances and pulled oscillation frequencies that satisfy the phase condition (A1.227).

and, finally,

$$\nu = \nu_m - (\nu_m - \nu_0)\frac{\Delta\nu_{1/2}}{\Delta\nu}. \tag{A1.239}$$

Thus, if ν_m coincides with the line centre, oscillation occurs at the line centre. If $\nu_m \neq \nu_0$, oscillation takes place near ν_m but is shifted slightly towards ν_0. This phenomenon is called 'mode-pulling' and is illustrated in figure A1.34.

A1.15.1 Multi-mode laser oscillation

We have seen that for oscillation to occur in a laser system the gain must reach a threshold value $\gamma_t(\nu) = \alpha - (1/\ell)\ln r_1 r_2$. For gain coefficients greater than this oscillation can occur at, or near (because of mode-pulling effects), one or more of the passive resonance frequencies of the Fabry–Pérot laser cavity. The resulting oscillations of the system are called longitudinal modes. As oscillation at a particular one of these mode frequencies builds up, the growing intracavity energy density depletes the inverted population and gain saturation sets in. The reduction in gain continues until

$$\gamma(\nu) = \gamma_t(\nu) = \alpha - \frac{1}{\ell}\ln r_1 r_2. \tag{A1.240}$$

Further reduction of $\gamma(\nu)$ below $\gamma_t(\nu)$ does not occur, otherwise the oscillation would cease. Therefore, the gain is stabilized at the loss

$$\alpha - \frac{1}{\ell}\ln r_1 r_2.$$

Usually α, r_1 and r_2 are nearly constant over the frequency range covered by typical amplifying transitions, so, over such moderate frequency ranges, 10^{11} Hz say, $\alpha - (1/\ell)\ln r_1 r_2$ as a function of frequency is a straight

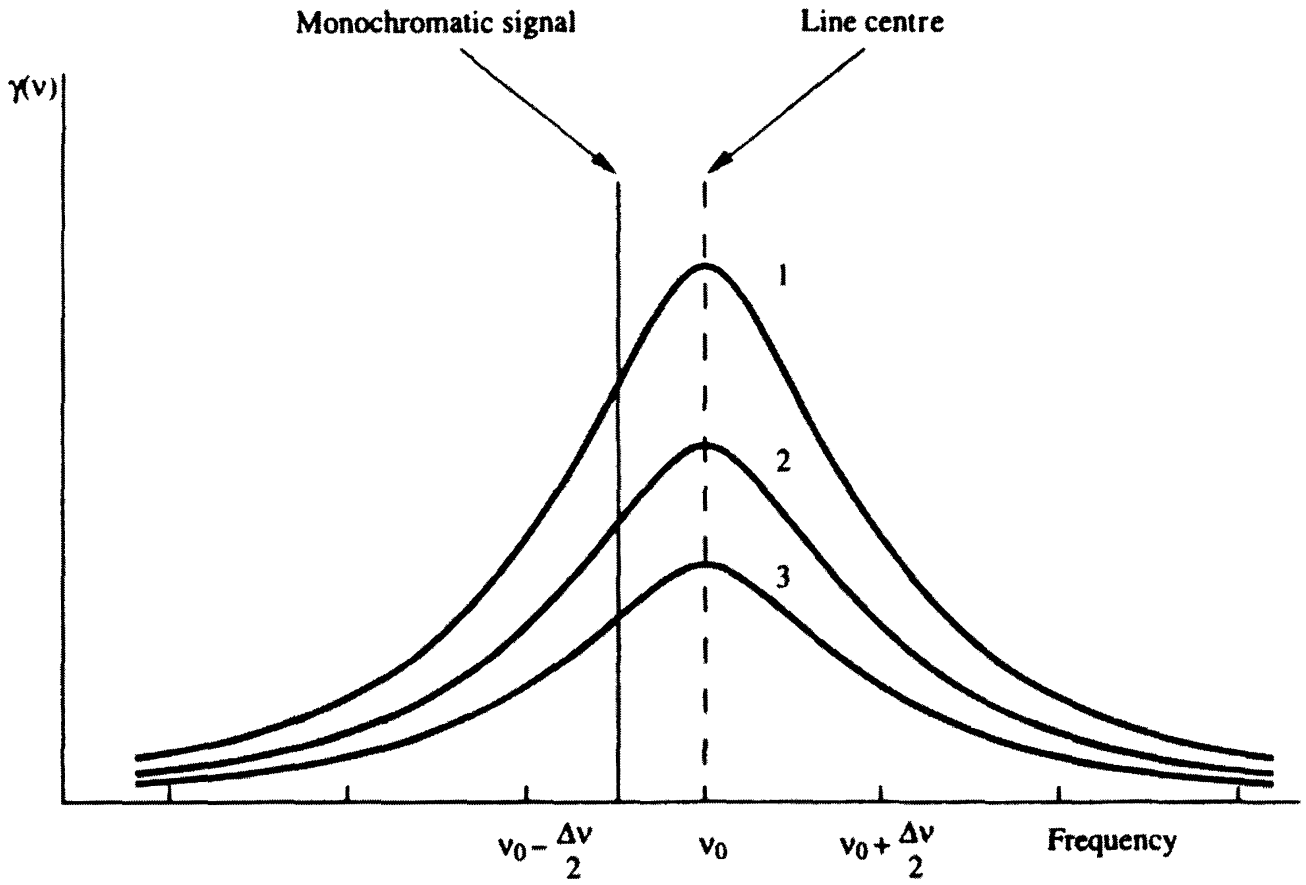


Figure A1.35. Saturation of gain of a homogeneously broadened transition produced by a monochromatic signal whose intensity increases from 1→2→3.

line parallel to the frequency axis. This line is called the *loss line*. At markedly different frequencies α, r_1 and r_2 can be expected to change: for example, a laser mirror with high reflectivity in the red region of the spectrum could have quite low reflectivity in the blue.

In a homogeneously broadened laser, because the reduction in gain caused by a monochromatic field is uniform across the whole gain profile, the clamping of the gain at $\gamma_t(\nu)$ leads to final oscillation at only one of the cavity resonance frequencies, the one where the original unsaturated gain was highest. We can show this schematically by plotting $\gamma(\nu)$ at various stages as oscillation builds up. Remember first the effect on $\gamma(\nu)$ produced by a monochromatic light signal of increasing intensity as shown in figure A1.35. Note that the gain profile is depressed uniformly even though the saturating signal is not at the line centre, as predicted by equation (A1.120).

In a laser, as oscillation begins, several such monochromatic fields start to build up at those cavity resonances where gain exceeds loss, as shown in figure A1.36. The oscillation stabilizes when the highest (small-signal) gain has been reduced to the loss line by saturation as shown in figures A1.37 and A1.38. Thus, in a *homogeneously* broadened laser, oscillation only occurs at *one* longitudinal mode frequency.

In an inhomogeneously broadened laser, the onset of gain saturation due to a monochromatic signal only reduces the gain locally over a region which is of the order of a homogeneous width. Only particles whose velocities (or environments in a crystal) make their centre emission frequencies lie within a homogeneous width of the monochromatic field can interact strongly with it. Schematically, the effect of an increasing intensity monochromatic field on the gain profile is as shown in figure A1.39. A localized dip, or *hole*, in the gain profile occurs. If only one cavity resonance has a small-signal gain above the loss line then only this longitudinal mode oscillates. The stabilization of the oscillation might be expected to occur schematically

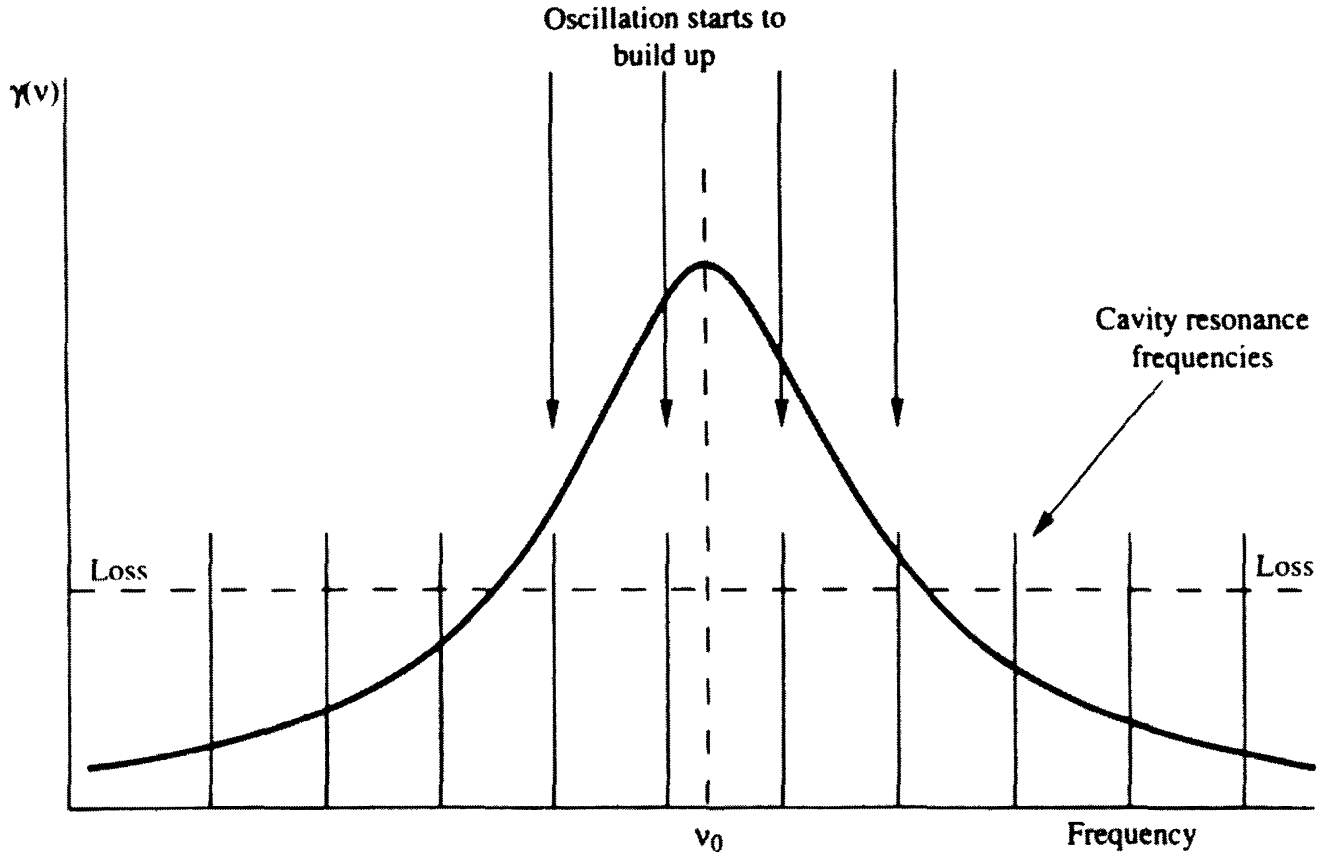


Figure A1.36. Schematic illustration of the onset of oscillation at cavity resonances that lie above the loss line in a homogeneously broadened laser.

as shown in figure A1.40. However, the situation is not quite as simple as this. Oscillation at this single longitudinal mode frequency implies waves travelling in both directions inside the laser cavity. These waves can be represented by

(a) the wave travelling to the right $\sim E_0 e^{i(\omega t - kz)}$ and

(b) the wave travelling to the left $\sim E_0 e^{i(\omega t + kz)}$

where we choose for convenience that $\omega = 2\pi\nu < 2\pi\nu_0$. Wave (a) can interact with particles whose centre frequency is near ν. These particles are, as far as their Doppler shifts are concerned, moving away from an observer looking into the laser from right to left. Their centre frequencies satisfy $\nu = \nu_0 - |v| \nu_0/c$, where positive atom velocities correspond to particles moving from left to right. Wave (b) which is travelling in the opposite direction (to the left) and is monitored, still at frequency $\nu(< \nu_0)$, by a second observer looking into the laser from left to right cannot interact with the same velocity group of particles as wave (a). The particles which interacted with wave (a) were moving away from the first observer and were Doppler shifted to lower frequencies so as to satisfy

$$\nu = \nu_0 - \frac{|v|}{c}\nu_0. \tag{A1.241}$$

The second observer sees these particles approaching and their centre frequency as

$$\nu = \nu_0 + \frac{|v|}{c}\nu_0 \tag{A1.242}$$

so they cannot interact with wave (b). Wave (b) interacts with particles moving away from the second observer

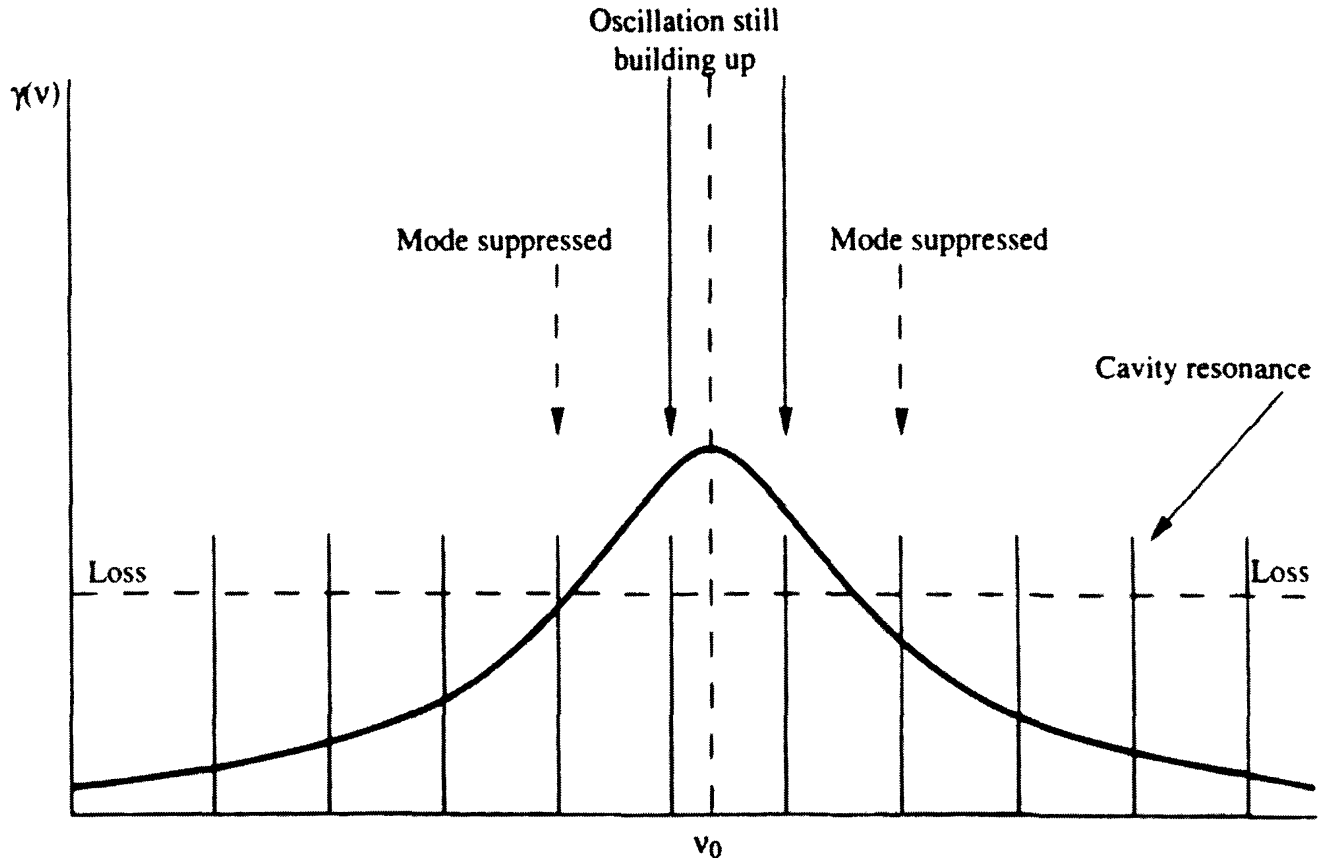


Figure A1.37. Oscillation building up in a homogeneously broadened laser. Gain saturation has already suppressed oscillation at two of the cavity modes that were above the loss line in figure A1.36.

so that their velocity would be the solution of

$$\nu = \nu_0 - \frac{|v|}{c}\nu_0. \tag{A1.243}$$

These particles would be monitored by the first observer at centre frequency

$$\nu = \nu_0 + \frac{|v|}{c}\nu_0. \tag{A1.244}$$

So the oscillating waves interact with two velocity groups of particles as shown in figure A1.41. This leads to saturation of the gain by a single laser mode in an inhomogeneously broadened laser both at the frequency of the mode ν and at a frequency $\nu_0 + (\nu_0 - \nu)$, which is equally spaced on the opposite side of the line centre, as shown in figure A1.42). The power output of the laser (strictly the intracavity power) comes from those groups of particles that have gone into stimulated emission and left the two holes. The combined area of these two holes gives a measure of the laser power.

If the frequency of the oscillating mode is moved in towards the line centre, the main hole and image hole begin to overlap. This corresponds physically to the left and right travelling waves within the laser cavity beginning to interact with the same velocity group of particles. As the oscillating mode moves in towards the line centre, the holes overlap further, the combined area decreases and the laser output power falls, reaching a minimum at the line centre. This phenomenon is called the Lamb dip, named after Willis E Lamb, Jr, who first predicted the effect [13], and is illustrated in figure A1.43. When the cavity resonance is at the line centre frequency ν_0, both travelling waves are interacting with the same group of atoms—those with near-zero directed velocity along the laser resonator axis.

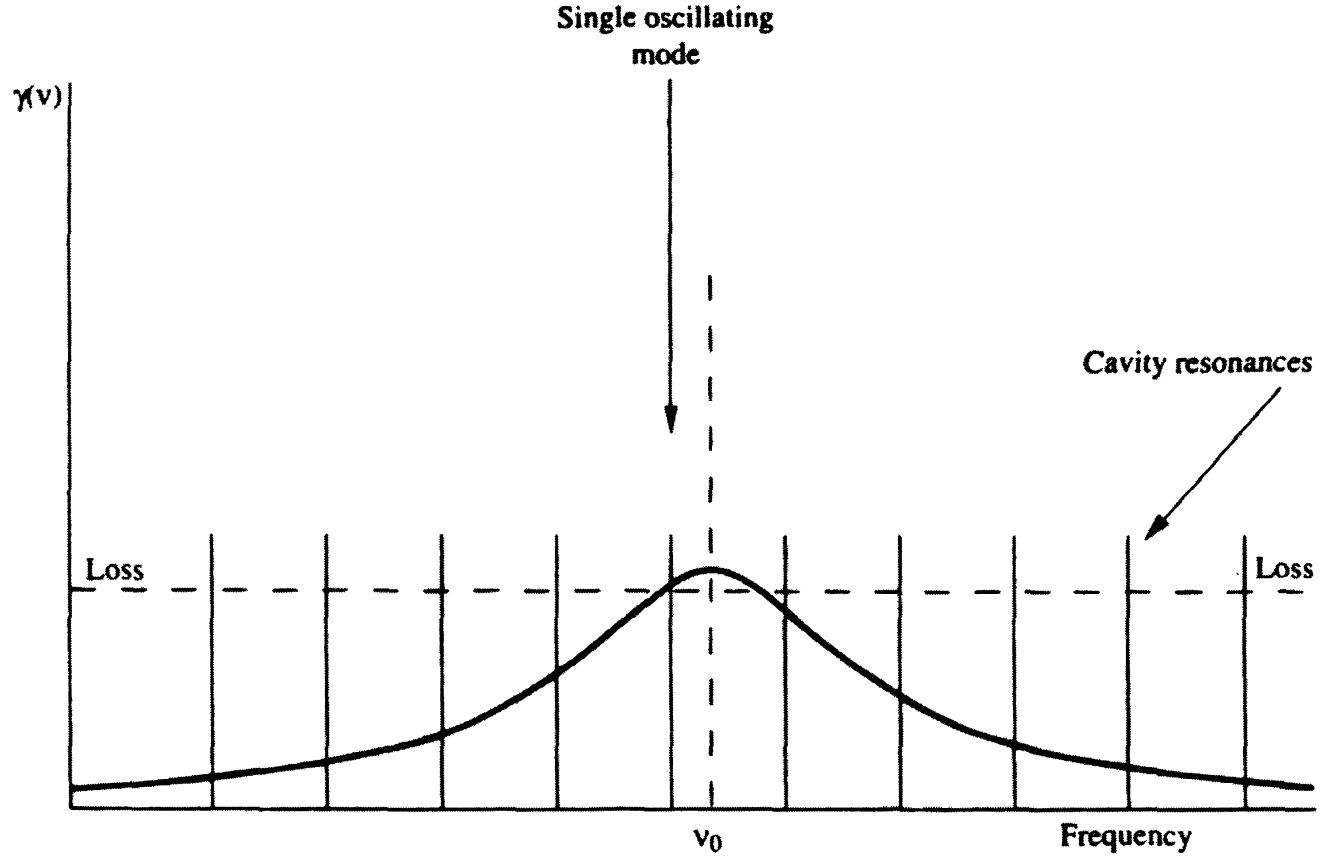


Figure A1.38. Oscillation stabilized in a homogeneously broadened laser. The gain has been uniformly reduced by saturation until only one mode remains at the loss line.

Because hole burning in gain saturation in inhomogeneously broadened lasers is localized near the frequency of a cavity mode, one oscillating mode does not reduce the gain at other cavity modes, so simultaneous oscillation at several longitudinal modes is possible. If several such modes have small-signal gains above the loss line the oscillation stabilizes in the manner shown in figure A1.44(*a*). The output frequency spectrum from the laser would appear as is shown in figure A1.44(*b*). This simultaneous oscillation at several closely spaced frequencies ($c/2\ell$ apart) can be observed with a high resolution spectrometer—for example a scanning Fabry–Pérot interferometer. The multiple modes are almost exactly $c/2\ell$ in frequency apart but are not exactly equally spaced because of mode pulling. This effect can be observed in the beat spectrum observed with a square-law optical detector (which means most optical detectors). Such a detector responds to the intensity, not the electric field of an incident light signal.

A1.15.2 Mode beating

Suppose we shine the light from a two-mode laser on a square-law detector. The incident electric field is

$$E_i = \mathcal{R}(E_1 e^{i\omega t} + E_2 e^{i(\omega+\Delta\omega)t}) \tag{A1.245}$$

where E_1 and E_2 are the complex amplitudes of the two modes and $\Delta\omega$ is the frequency spacing between them. Using real notation for these fields, the output current i from the detector is

$$\begin{aligned} i &\propto \{|E_1|\cos(\omega t + \phi_1) + |E_2|\cos[(\omega + \Delta\omega)t + \phi_2]\}^2 \\ &\propto |E_1|^2\cos^2(\omega t + \phi_1) + |E_2|^2\cos^2[(\omega + \Delta\omega)t + \phi_2] \end{aligned}$$

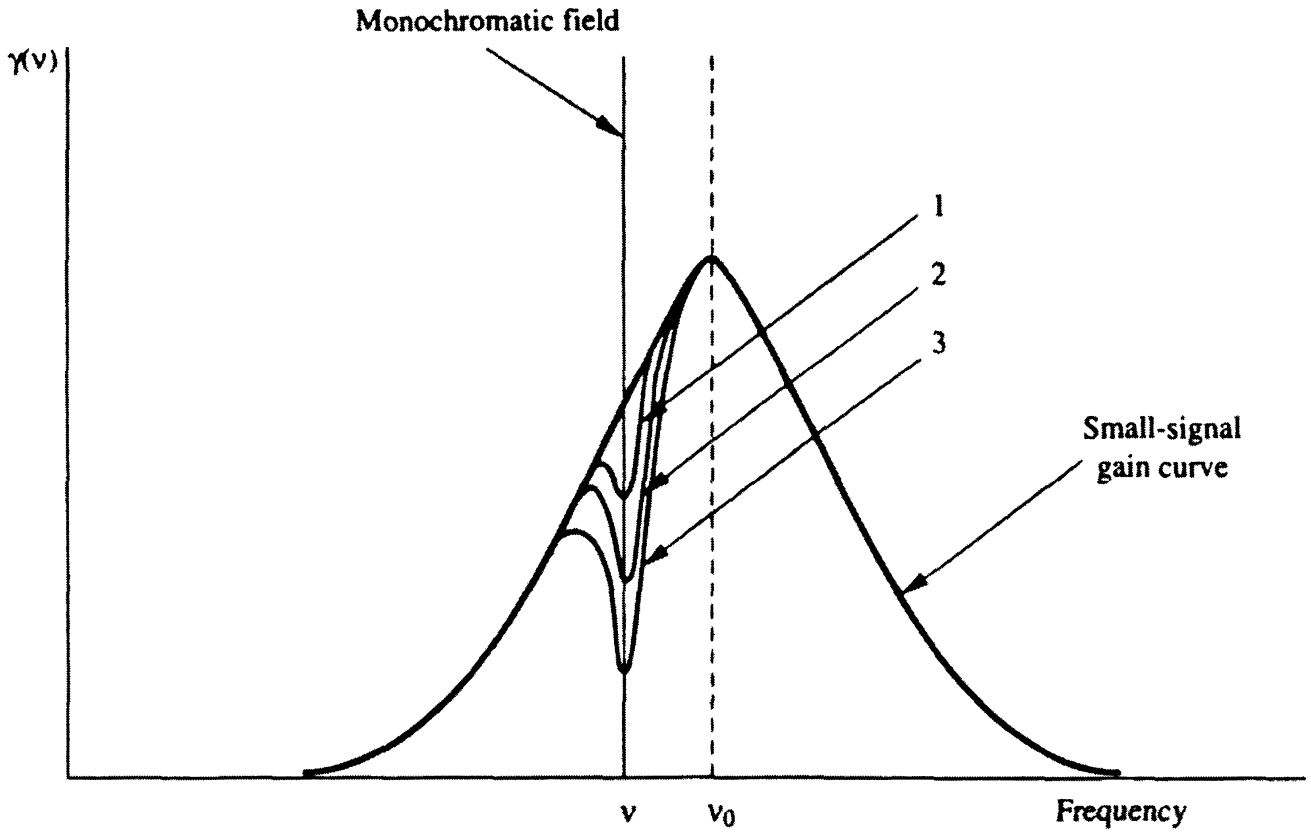


Figure A1.39. Localized gain saturation in an inhomogeneously broadened amplifier produced by a monochromatic signal whose intensity increases from 1→2→3.

$$\begin{aligned} &+ 2|E_1||E_2|\cos(\omega t + \phi_1)\cos[\omega + \Delta\omega t + \phi_2] \\ &\propto |E_1|^2\cos^2(\omega t + \phi_1) + |E_2|^2\cos^2[(\omega + \Delta\omega)t + \phi_2] \\ &+ |E_1||E_2|\cos[(2\omega + \Delta\omega)t + \phi_1 + \phi_2] \\ &+ |E_1||E_2|\cos(\Delta\omega t + \phi_2 - \phi_1). \end{aligned} \tag{A1.246}$$

And since, for example,

$$|E_1|^2\cos^2(\omega t + \phi_1) = \tfrac{1}{2}|E_1|^2[1 + \cos 2(\omega t + \phi_1)] \tag{A1.247}$$

the output frequency spectrum of the detector appears to contain the frequencies 2ω, $2(\omega + \Delta\omega)$, $2\omega + \Delta\omega$ and $\Delta\omega$. However, the first three of these frequencies are very high, particularly for light in the visible and infrared regions of the spectrum, and do not appear in the output of the detector. It is as if the high frequency terms are averaged to zero by the detector time response to give

$$i \propto \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} + |E_1||E_2|\cos(\Delta\omega t + \phi_2 - \phi_1) \tag{A1.248}$$

so only the difference frequency beat $\Delta\omega$ is observed.

If the output from the square-law detector is analysed with a radio frequency spectrum analyser (because it is in this frequency range where the difference frequencies between longitudinal laser modes are usually observed), different displays are obtained according to how many longitudinal modes of a multi-mode laser are simultaneously oscillating. Figure A1.45 gives some examples. Because equation (A1.239) is not quite exact, the beat frequencies can split as shown because of nonlinear mode-pulling. This splitting will only

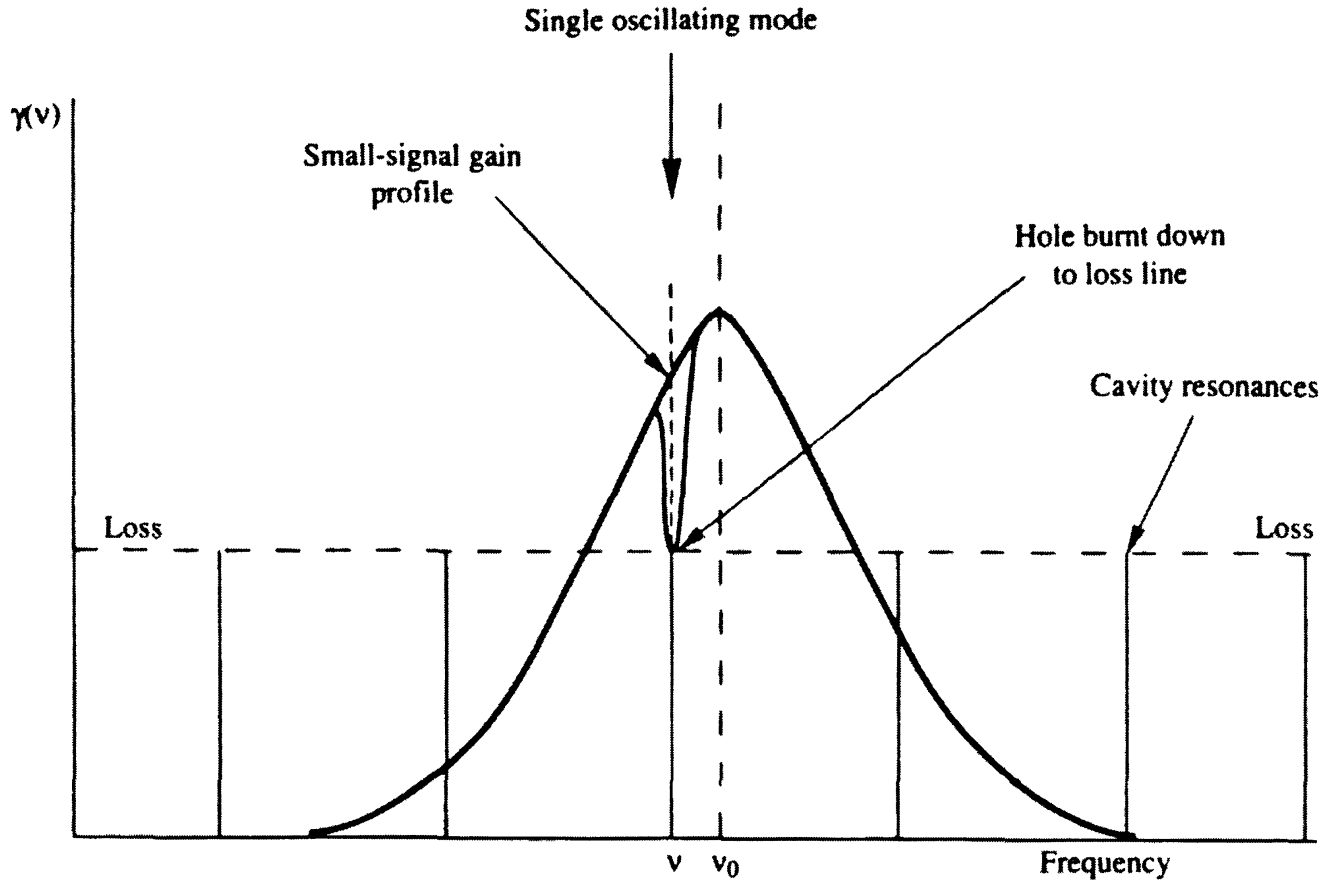


Figure A1.40. Simplified illustration of how saturation stabilizes oscillation at a single longitudinal mode in an inhomogeneously broadened laser.

be observed if nonlinear mode-pulling is large and the spectrum analyser that analyses the output of the photo-detector has high resolution.

If a predominantly inhomogeneously broadened laser also has a significant amount of homogeneous broadening, the holes burnt in the gain curve can start to overlap, for example, when $\Delta\nu \gtrsim c/2\ell$. If $\Delta\nu$ is large enough this causes neighbouring oscillating modes to compete, and may lead to oscillation on a strong mode suppressing its weaker neighbours, as shown in figure A1.46. This effect has been observed in several laser systems, for example in the argon ion laser, where an increase in the strength of the oscillation can lead to the successive disappearance, first of every other mode, then two modes out of every three, and so on.

A1.16 The characteristics of laser radiation

Laser radiation has special properties that distinguish it from ordinary light. We have already seen that laser radiation should be very directional[16]. Laser radiation is, in general, very monochromatic: the spectrum of a laser longitudinal mode is confined to a narrow spectral range. Laser radiation is generally very *coherent* compared to the light from traditional light sources. This coherence is a measure of the temporal and spatial phase relationships that exist for the fields associated with laser radiation (see section A5.2.2).

The special nature of laser radiation is graphically illustrated by the ease with which the important optical phenomena of interference and diffraction are demonstrated using it. Interference effects demonstrate the

[16] Diffraction effects cause the laser beams from lasers with small lateral dimensions, such as semiconductor lasers, to diverge substantially after they leave the laser resonator.

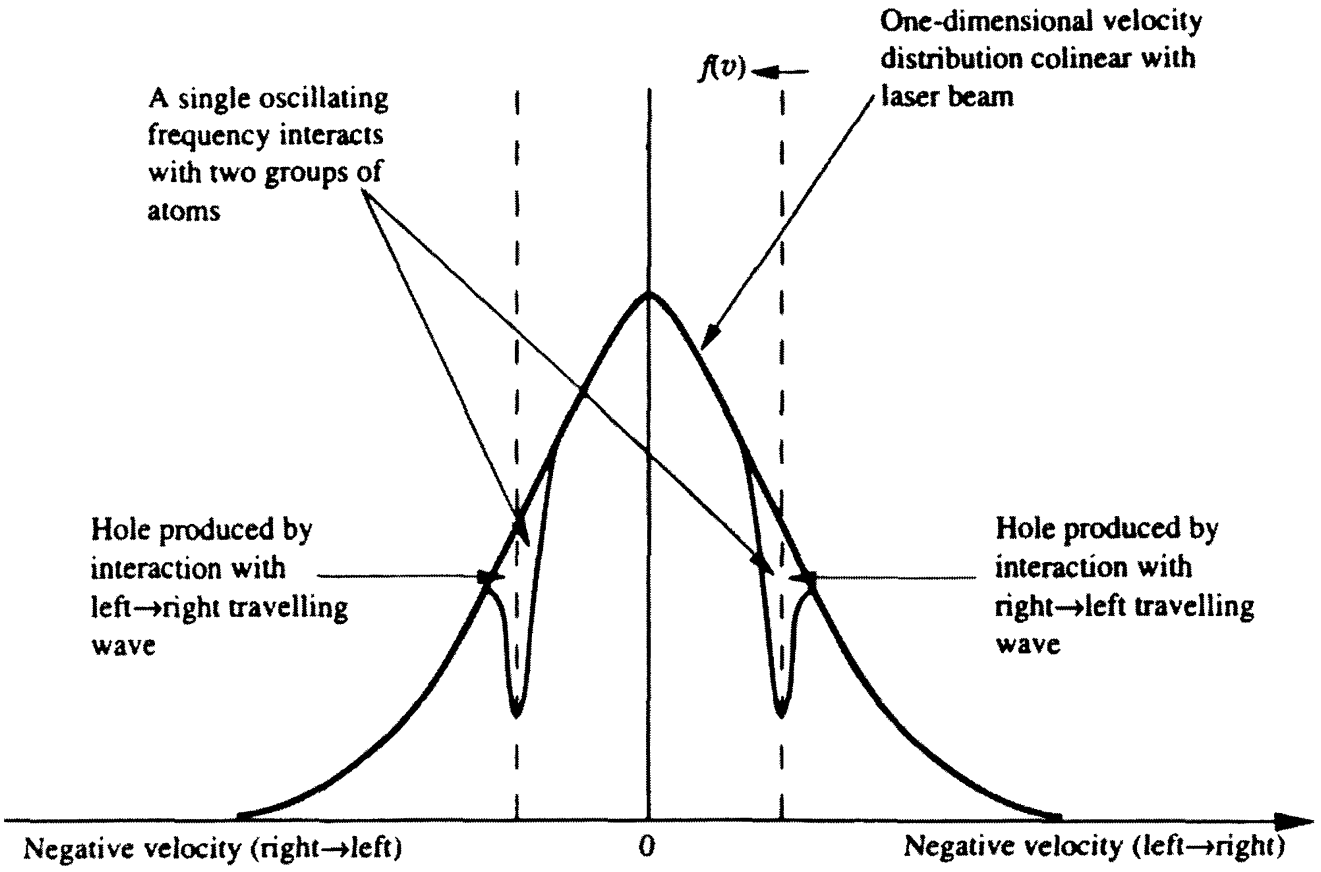


Figure A1.41. Production of two holes in the velocity distribution of a collection of amplifying particles by a single cavity mode.

coherence properties of laser radiation, while diffraction effects are intimately connected with the beam-like properties that make this radiation special.

A1.16.1 Laser modes

When a laser oscillates, it emits radiation at one or more frequencies that lie close to passive resonant frequencies of the cavity. These frequencies are called *longitudinal* modes (see section A2.1.9). In our initial discussion of these modes we treated them as plane waves reflecting back and forth between two plane laser mirrors. In practice, laser mirrors are not always plane. Usually at least one of the laser mirrors will have concave spherical curvature. The use of spherical mirrors relaxes the alignment tolerance that must be maintained for adequate feedback to be achieved. Even if the laser mirrors are plane, the waves reflecting between them cannot be plane, as true plane waves can only exist if there is no lateral restriction of the wave fronts. Practical laser mirrors are of finite size so any wave reflecting from them will spread out because of diffraction [14]. Diffraction results whenever a wave is restricted laterally, for example by passing it through an aperture. Reflection from a finite-size mirror produces equivalent effects. We can explain this phenomenon qualitatively by introducing the concept of Huygens secondary wavelets [17]. If a plane mirror is illuminated by a plane wave then each point on the mirror can be treated as a source of a spherical wave called a secondary wavelet. The overall reflected wave is the envelope of the sum total of secondary wavelets originating from every point on the mirror surface, as shown in figure A1.47. This construction shows that the reflected wave from a finite-size mirror is not a plane wave.

[17] Christian Huygens (1629–1695) was a Dutch astronomer who first suggested the concept of secondary wavelets.

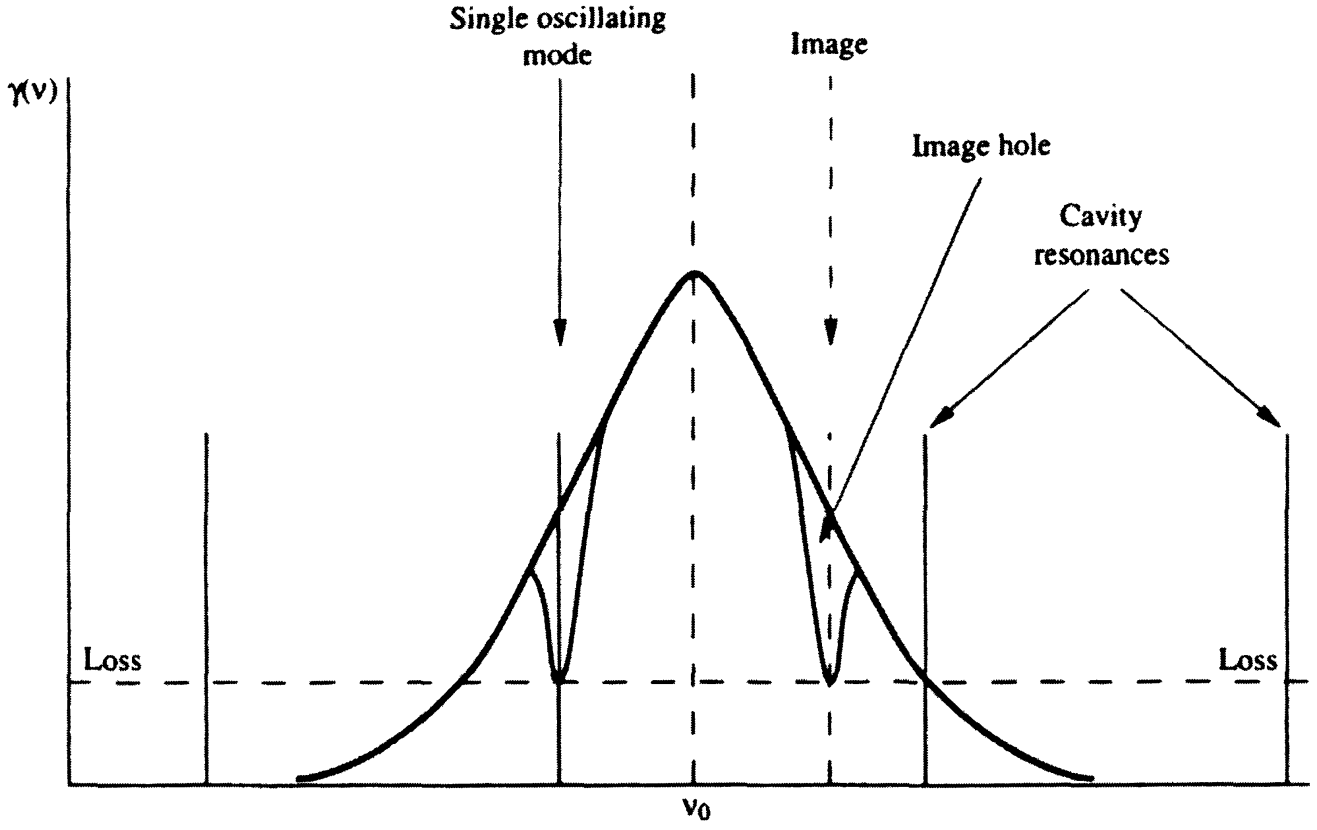


Figure A1.42. Stable saturation of a single longitudinal mode in an inhomogeneously broadened laser.

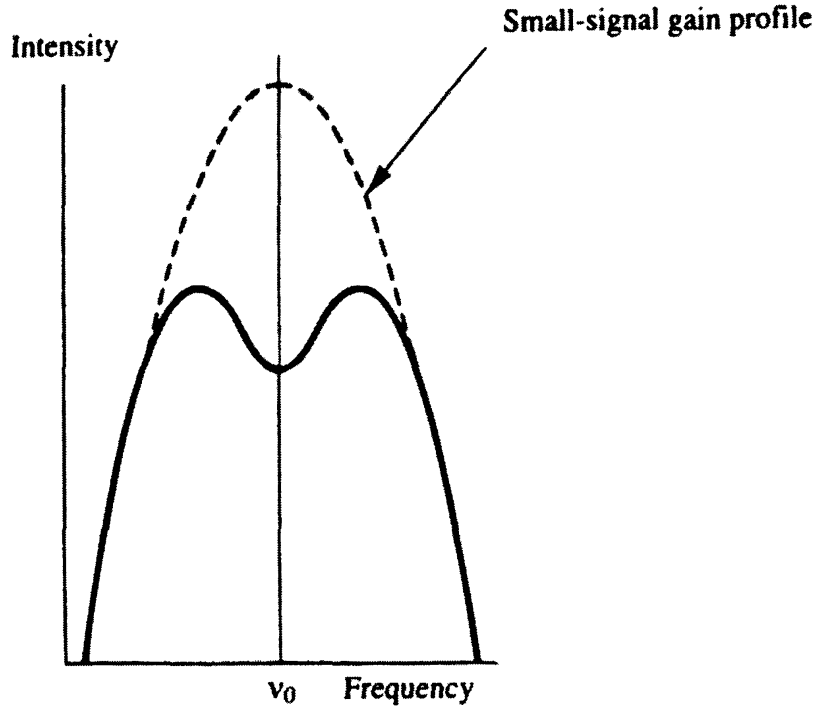


Figure A1.43. The Lamb dip—a reduction in the intensity of a single oscillating longitudinal mode in an inhomogeneously broadened laser as its frequency is scanned through line centre.

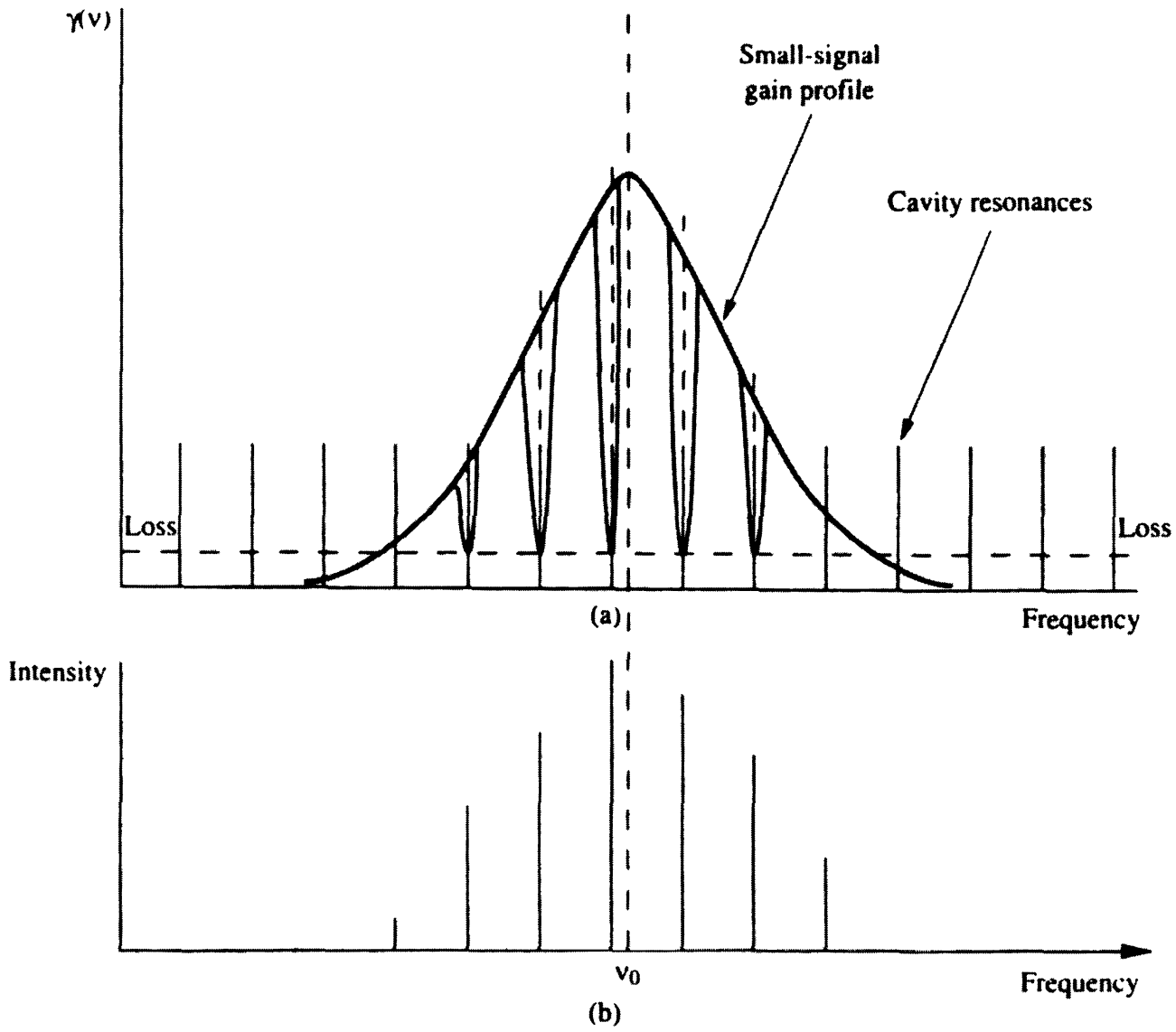


Figure A1.44. Multi-longitudinal-mode oscillation in an inhomogeneously broadened laser. (*a*) Only the primary holes are shown burnt down to the loss line. The image holes are not shown. (*b*) Schematic laser output spectrum.

The existence of diffraction in a laser resonator places restrictions on the minimum size of mirrors that can be used at a given wavelength λ and spacing ℓ. If the diameters of the resonator mirrors are d_1, d_2, respectively and the resonator has length ℓ, then the resonator will have high loss unless the *Fresnel condition* is satisfied [1]: namely

$$\frac{d_2 d_2}{\lambda \ell} \geq 1. \tag{A1.249}$$

The actual waves that reflect back and forth between the mirrors of a laser resonator are not plane waves. They have characteristic spatial patterns of electric (and magnetic) field amplitude and are called *transverse modes* (see section A2.1.4). To be amplified effectively such modes must correspond to rays which make substantial numbers of specular reflections before being lost from the cavity. A transverse mode is a field configuration on the surface of one reflector that propagates to the other reflector and back, returning in the same pattern, apart from a complex amplitude factor that gives the total phase shift and loss of the round trip. To each of these transverse modes there corresponds a set of longitudinal modes spaced by approximately $c/2\ell$. A more detailed treatment of these transverse modes is given in chapter A2.

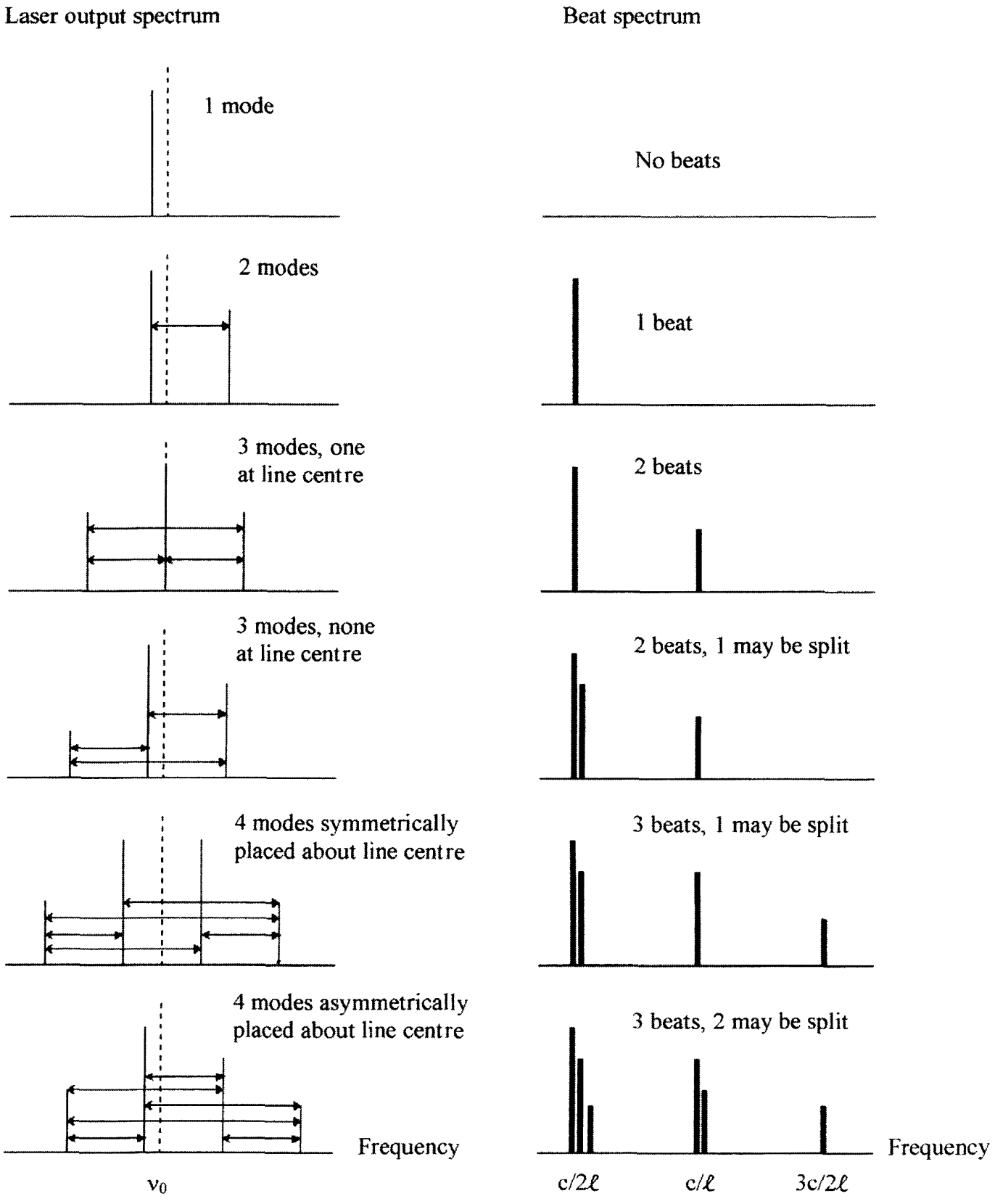


Figure A1.45. Schematic mode-beating spectra observed with a square-law detector and a multi-mode laser.

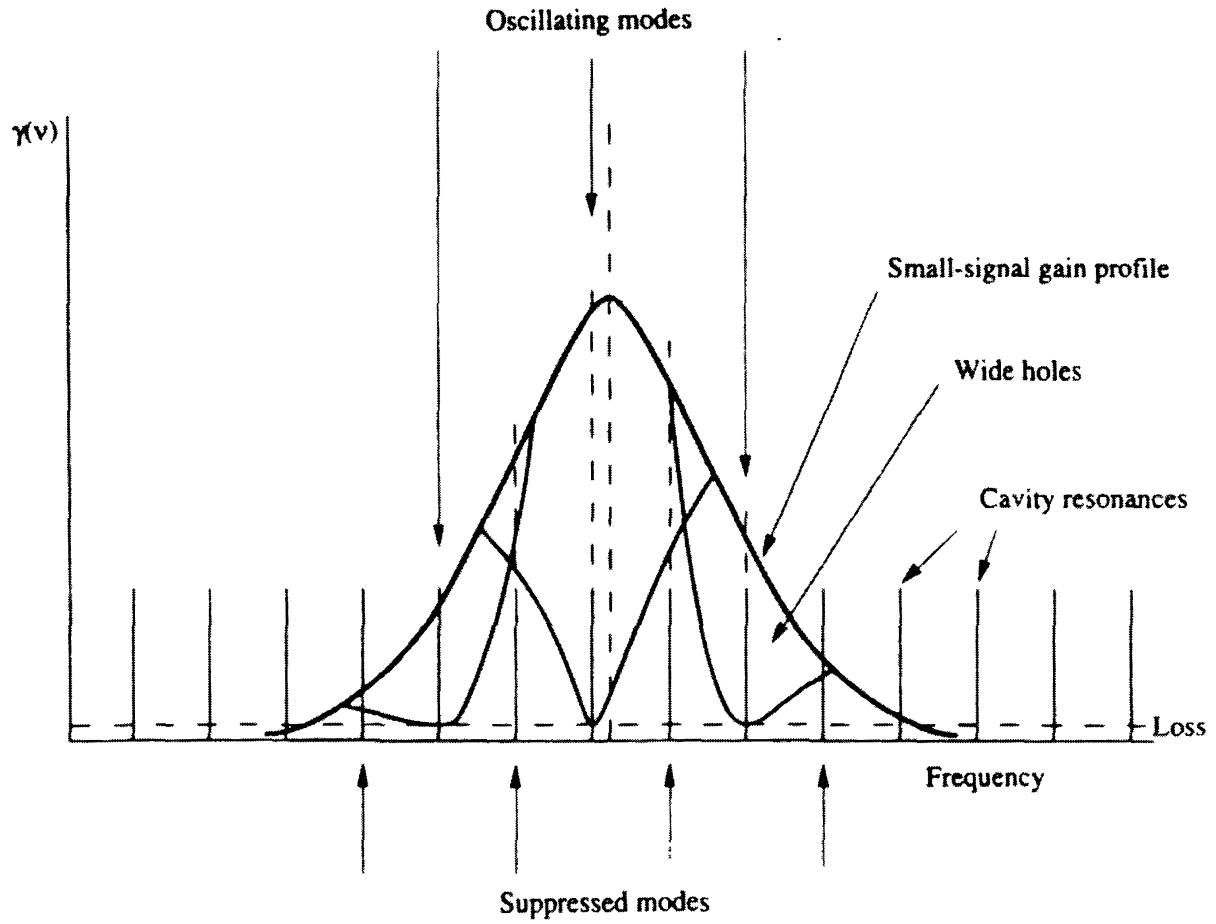


Figure A1.46. Schematic illustration of mode competition in an inhomogeneously broadened laser in which there is significant homogeneous broadening.

A1.16.2 Beam divergence

Since the oscillating field distributions inside a laser are not plane waves, when they propagate through the mirrors as output beams they spread by diffraction. The semi-vertical angle of the cone into which the output beam diverges is[18]

$$\theta_{\text{beam}} = \tan^{-1}\left(\frac{\lambda}{\pi w_0}\right) \approx \frac{\lambda}{\pi w_0} \tag{A1.250}$$

where λ is the wavelength of the output beam and w_0 is a parameter called the 'minimum spot size' that characterizes the transverse mode.

Let us take the specific example of a 530 nm laser beam with $w_0 = 1$ mm for which $\theta_{\text{beam}} = 169\ \mu\text{rad} \simeq 1$ millidegree. This is a highly directional beam but the beam does become wider the further it goes away from the laser. Such a beam is, however, highly useful in providing the perfect straight line reference. Over a 100 m distance, the laser beam just described would have expanded to a diameter of 34 mm. After travelling the distance to the moon ($\sim$ 390 000 km) the beam would be $\simeq$ 132 km in diameter.

A1.16.3 Linewidth of laser radiation

A single longitudinal mode of a laser is an oscillation resulting from the interaction of a broadened gain curve with a passive resonance of the Fabry–Pérot laser cavity. The frequency width of the gain curve is $\Delta\nu$,

[18] For a derivation of this result see section A2.1.2.

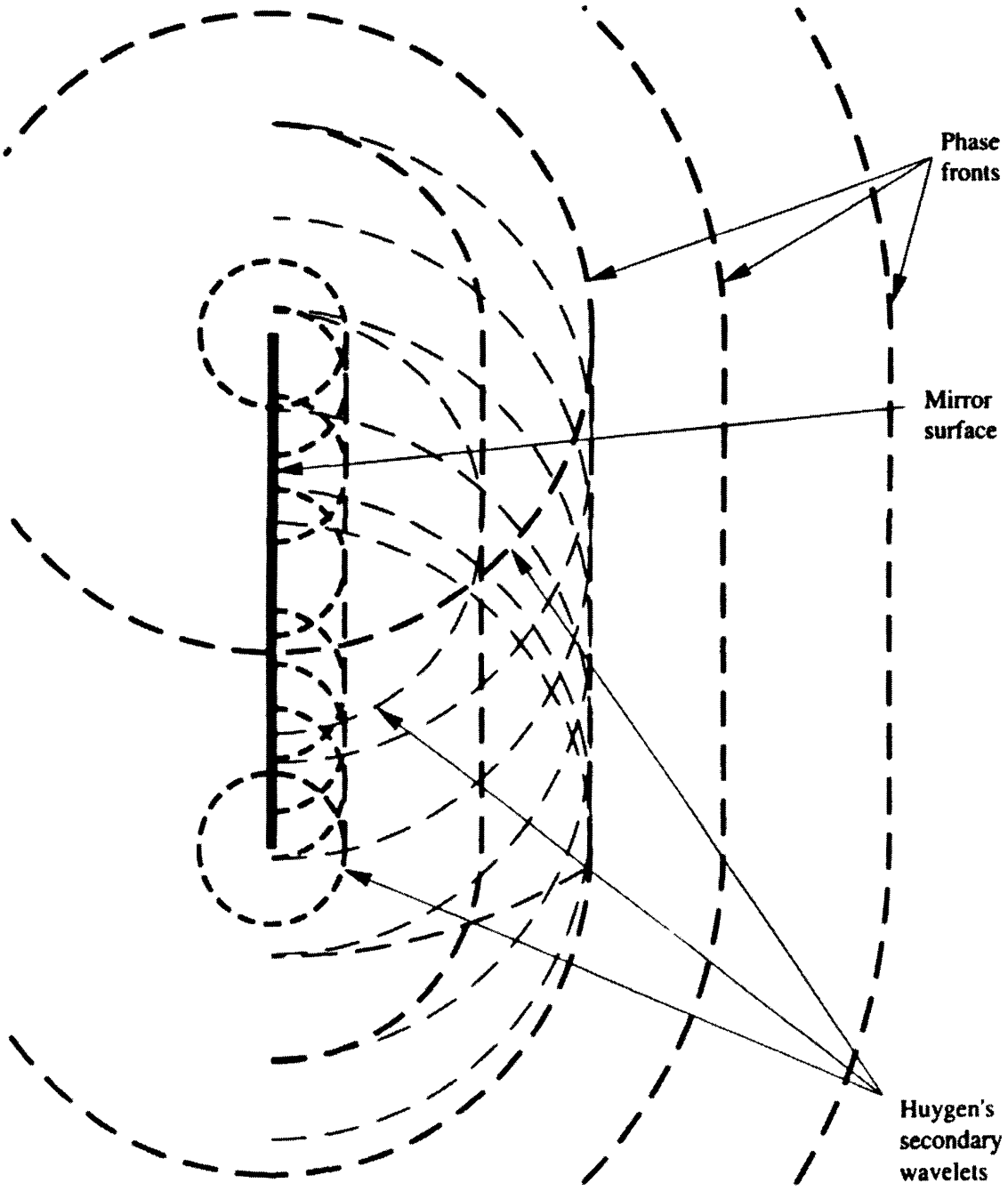


Figure A1.47. Secondary wavelets originating from a finite plane mirror.

the frequency width of the passive cavity resonance is $\Delta\nu_{1/2} = \Delta\nu_{\mathrm{FSR}}/F$. We expect the linewidth of the resulting oscillation to be narrower than either of these widths, as shown schematically in figure A1.48. It can be shown that the frequency width of the laser oscillation itself is [10, 15–17]

$$\Delta\nu_{\mathrm{laser}} = \frac{\pi h\nu_0(\Delta\nu_{1/2})^2}{P}\frac{N_2}{[N_2 - (g_2/g_1)N_1]_{\mathrm{threshold}}} \tag{A1.251}$$

here P is the output power.

Equation (A1.252) predicts very low linewidths for many lasers. For a typical He–Ne laser with 99%

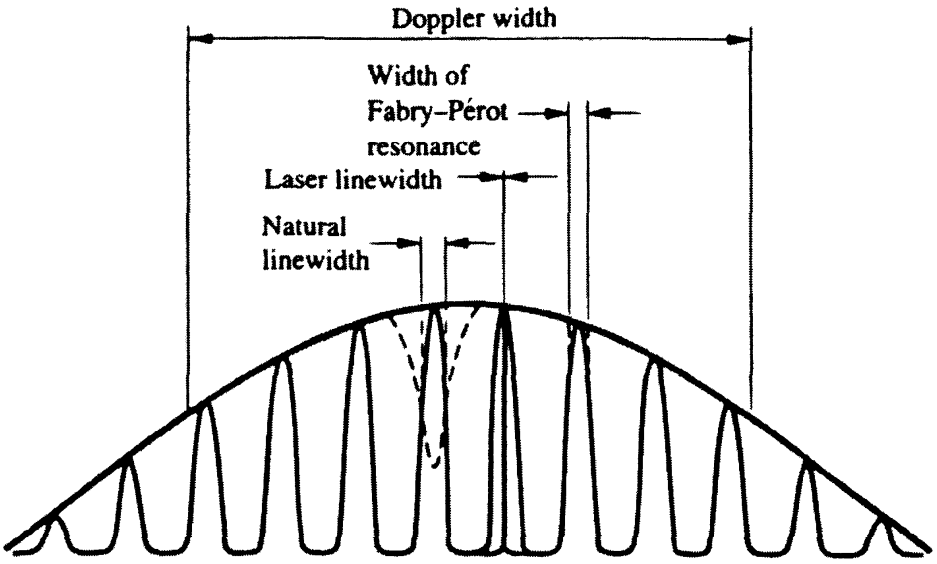


Figure A1.48. Linewidth factors in a laser.

reflectance mirrors and a cavity 30 cm long

$$\Delta\nu_{1/2} = \frac{c(1-R)}{2\pi\ell} = 1.59\ \text{MHz}.$$

The factor $N_2/[N_2 - (g_2/g_1)N_1]_{\text{threshold}}$ is close to unity for a typical low power, say 1 mW, laser. Consequently,

$$\Delta\nu_{\text{laser}} = \frac{\pi \times 6.626 \times 10^{-34} \times 3 \times 10^8 \times 1.59^2 \times 10^{12}}{10^{-3} \times 632.8 \times 10^{-9}} = 2.5 \times 10^{-3}\ \text{Hz}.$$

Such a small linewidth is never observed in practice because thermal instabilities and acoustic vibrations lead to variations in resonator length that further broaden the output radiation lineshape. The best observed minimum linewidths for highly stabilized gas lasers operating in the visible region of the spectrum are around 10^3 Hz. Even if macroscopic thermal and acoustic vibrations could be eliminated from the system, a fundamental limit to the resonator length stability would be set by Brownian motion of the mirror assemblies. For example, consider two laser mirrors mounted on a rigid bar. The mean stored energy in the Brownian motion of the whole bar is $\overline{E} = kT$.

The frequency spread of the laser output that thereby results is

$$\Delta\nu_{\text{Brownian}} = \nu\sqrt{\frac{2kT}{YV}} \tag{A1.252}$$

where Y is Young's modulus of the bar material and V is the volume of the mounting bar. Typical values of $\Delta\nu_{\text{Brownian}}$ are ~ 2 Hz.

A1.17 Coherence properties

A1.17.1 Temporal coherence

Because of its extremely narrow output linewidth, the output beam from a laser exhibits considerable *temporal coherence* (longitudinal coherence). To illustrate this concept, consider two points A and B a distance L apart in the direction of propagation of a laser beam, as shown in figure A1.49. If a definite and fixed phase relationship exists between the wave amplitudes at A and B, then the wave shows temporal coherence for a

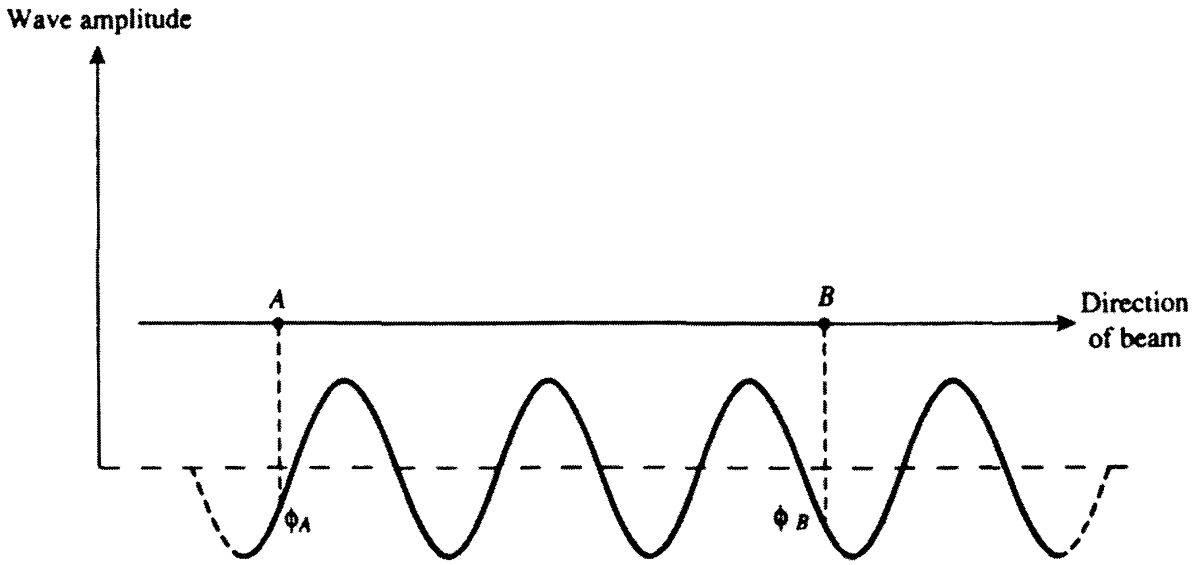


Figure A1.49. To illustrate the concept of temporal coherence. If an unbroken wave train connects the points A and B then the phase difference $(\phi_B - \phi_A)$ will have a constant value.

time c/L. The further apart A and B can be, while still maintaining a fixed phase relation with each other, the greater is the temporal coherence of the output beam. The maximum separation at which the fixed phase relationship is retained is called the *coherence length*, L_c, which is a measure of the length of the continuous uninterrupted wave trains emitted by the laser. The coherence length is related to the *coherence time* τ_c by $L_c = c\tau_c$. The coherence time itself is a direct measure of the monochromaticity of the laser, since by Fourier transformation, done in an analogous manner to the treatment of natural broadening,

$$\tau_c \simeq \frac{1}{\Delta\nu_L} \qquad \text{and} \qquad L_c \simeq \frac{c}{\Delta\nu_L}. \tag{A1.253}$$

The coherence length and time of a laser source are considerably better than a conventional monochromatic source (a spontaneously emitted line source). The greatly increased coherence can be demonstrated in a Michelson interferometer experiment, which allows interference between waves at longitudinally different positions in a wavefront to be studied [14, 18, 19].

A1.17.2 Laser speckle

Perhaps the most easily observable consequence of the high temporal coherence of most lasers is the occurence of *speckle* [20–22]. When a laser beam strikes an object, unless the object is *very* flat (to better than a faction of a wavelength over the illuminated area) light that is scattered from the object will be observed to have a 'grainy' texture. If the scattered light falls directly onto a screen, then an *objective* speckle pattern will be observed. This manifests itself as a pattern of bright and dark patches with a 'salt and pepper' appearance. If the speckle pattern is collected by a lens, or enters the eye, then a *subjective* speckle pattern is observed. Speckle results from the roughness of the illuminated object and, on a scale of visible wavelengths, most objects are intrinisically very rough. A piece of paper provides a good object to observe the production of a speckle pattern. When a rough surface is illuminated with temporally coherent light, each point on the surface serves as a scattering centre for the production of an outgoing spherical wave. At a point in space these many scattered waves arrive with different phases, because they have come from the 'hills and valleys' of the surface roughness. If these waves add substantially in phase, then bright illumination results at the point. However, if destructive interference occurs, a dark region will result.

The spatial structure of the speckle pattern results largely from the size of the illuminated region on the object. If this region is large, an objective speckle pattern with fine spatial scale results. If the region illuminated is small, then a speckle pattern with a coarser spatial scale is observed. This is related to the diffraction angle associated with the diameter D of the illuminated region. At a dark spot in the speckle pattern, the waves that have come from the two halves of the illuminated region can be thought of as being π radians out of phase. At an adjacent bright spot, the waves can be thought of as being in phase. It is easy to see that the angular separation of adjacent bright and dark spot is $\sim \lambda/D$.

For a fixed object the objective speckle pattern is stationary in space. If this pattern of light and dark enters the eye, then because of the natural and continuous small motions of the eye, an apparent twinkling effect will be perceived as the eye moves through the regions of bright and dark illumination of the speckle pattern.

A1.17.3 Spatial coherence

A laser also possesses *spatial* (lateral) coherence, which implies a definite fixed-phase relationship between points separated by a distance L transverse to the direction of beam propagation. The transverse coherence length, which has similar physical meaning to the longitudinal coherence length, is

$$L_{tc} \sim \frac{\lambda}{\theta_{beam}} \sim \pi\omega_0 \tag{A1.254}$$

for a laser source.

The existence of spatial coherence in a wavefront and the limit of its extent can be demonstrated in a classic Young's slits interference experiment [14, 18, 19]. A pair of thin, parallel slits or a pair of pinholes spaced a distance d apart is illuminated normally with a spatially coherent monochromatic plane wave. If an interference pattern of bright and dark bands is observed on the other side of the slits, then the illumination laser beam is spatially coherent over at least the distance d, if the slit variation separation were increased to beyond the lateral coherence length, so that $d > L_{tc}$, then the fringe pattern would disappear.

A1.18 The power output of a laser

When a laser oscillates, the intracavity field grows in amplitude until saturation reduces the gain to the loss line for each oscillating mode. What this means in practice can be best illustrated with reference to figure A1.50.

For an asymmetrical resonator, whose mirror reflectances are not equal, the distribution of standing-wave energy within the resonator is not symmetrical. For example, in figure A1.50, if $R_2 > R_1$ the distribution of intracavity travelling wave intensity will be schematically as shown, and

$$\frac{I_3}{I_2} = R_2 \qquad \frac{I_1}{I_4} = R_1. \tag{A1.255}$$

The left travelling wave, of intensity I_-, grows in intensity from I_3 to I_4 on a single pass. The right travelling wave, of intensity I_+, grows in intensity from I_1 to I_2 on a single pass. The total output intensity is

$$I_{out} = T_2 I_2 + T_1 I_4. \tag{A1.256}$$

However, calculation of I_2 and I_4 is not straightforward in the general case. I_2 grows from I_1 through a gain process that depends in a complex way on $I_+ + I_-$ as does the growth of I_3 to I_4. We can identify at least three scenarios in which the calculation proceeds differently:

(a) a homogeneously broadened amplifier and single-mode operation,

(b) an inhomogeneously broadened amplifier and single-mode operation and

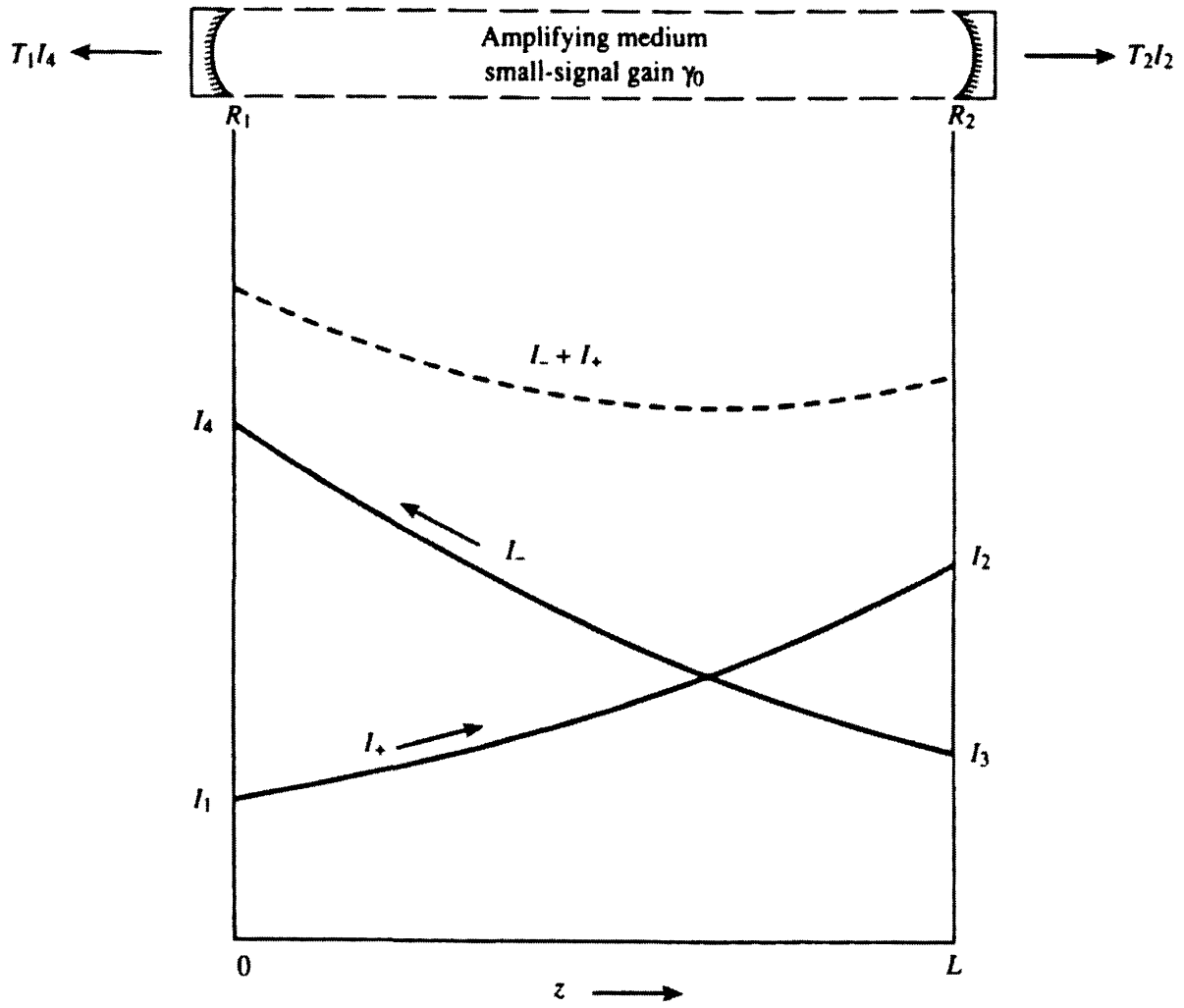


Figure A1.50. Distribution of bi-directional propagating wave intensities in an asymmetric laser cavity.

(c) an inhomogeneously broadened amplifier and multi-mode operation.

In both cases (b) and (c), the calculation of the output power becomes more complicated as the homogeneous contribution to the broadening grows more significant compared to $\Delta\nu_D$. This additional complexity arises because each oscillating mode burns both a primary and an image hole in the gain curve. The resultant distribution of overlapping holes makes the gain for each mode dependent not only on its own intensity but also on the intensity of the other simultaneously oscillating modes. The presence of distributed intracavity loss presents additional complications. We shall not attempt to deal with these complex situations here but will follow Rigrod [23] in dealing with a homogeneously broadened amplifier in which the primary intensity loss occurs at the mirrors. Inhomogeneously broadened systems and multi-mode operation have been discussed elsewhere by Smith [24].

In a purely homogeneously broadened system, the saturated gain in figure A1.50 is

$$\gamma(z) = \frac{\gamma_0}{1 + (I_+ + I_-)/I_s}. \tag{A1.257}$$

Both I_- and I_+ grow according to $\gamma(z)$

$$\frac{1}{I_+}\frac{dI_+}{dz} = -\frac{1}{I_-}\frac{dI_-}{dz} = \gamma(z). \tag{A1.258}$$

Consequently,

$$I_+ I_- = \text{constant} = C. \tag{A1.259}$$

From equation (A1.260)

$$I_4 I_1 = I_2 I_3 = C \tag{A1.260}$$

and, therefore, from equation (A1.255)

$$I_2/I_4 = \sqrt{R_1/R_2}. \tag{A1.261}$$

For the right travelling wave, using equations (A1.258) and (A1.259) gives

$$\frac{1}{I_+}\frac{\mathrm{d}I_+}{\mathrm{d}z} = \frac{\gamma_0}{1 + (I_+ + C/I_+)/I_s} \tag{A1.262}$$

which can be integrated to give

$$\gamma_0 L = \ln\left(\frac{I_2}{I_1}\right) + \frac{(I_2 - I_1)}{I_s} - \frac{C}{I_s}\left(\frac{1}{I_2} - \frac{1}{I_1}\right). \tag{A1.263}$$

In a similar way, for the left travelling wave

$$\gamma_0 L = \ln\left(\frac{I_4}{I_3}\right) + \frac{(I_4 - I_3)}{I_s} - \frac{C}{I_s}\left(\frac{1}{I_4} - \frac{1}{I_3}\right). \tag{A1.264}$$

Adding equations (A1.263) and (A1.264) and using equations (A1.255), (A1.260) , and (A1.261) gives

$$I_2 = \frac{I_s\sqrt{R_1}(\gamma_0 L + \ln\sqrt{R_1 R_2})}{(\sqrt{R_1} + \sqrt{R_2})(1 - \sqrt{R_1 R_2})}. \tag{A1.265}$$

From equation (A1.261)

$$I_4 = I_2\sqrt{\frac{R_2}{R_1}}. \tag{A1.266}$$

Now

$$T_1 = 1 - R_1 - A_1 \tag{A1.267}$$

$$T_2 = 1 - R_2 - A_2 \tag{A1.268}$$

so, from equations (A1.256) and (A1.265), if $A_1 = A_2 = A$

$$I_{\text{out}} = I_s\frac{(1 - A - \sqrt{R_1 R_2})}{1 - \sqrt{R_1 R_2}}(\gamma_0 L + \ln\sqrt{R_1 R_2}). \tag{A1.269}$$

If one mirror is made perfectly reflecting, say $T_1 = 0$, $R_1 = 1$, then

$$I_{\text{out}} = T_2 I_2 = \frac{T_2 I_s[\gamma_0 L + \frac{1}{2}\ln(1 - A_2 - T_2)]}{(A_2 + T_2)}. \tag{A1.270}$$

For a symmetrical resonator, defined by $R_1 R_2 = R^2$,

$$R = 1 - A - T \tag{A1.271}$$

the output intensity at each mirror is

$$\frac{I_{\text{out}}}{2} = \frac{I_s}{2}\frac{(1 - A - R)}{1 - R}(\gamma_0 L + \ln R). \tag{A1.272}$$

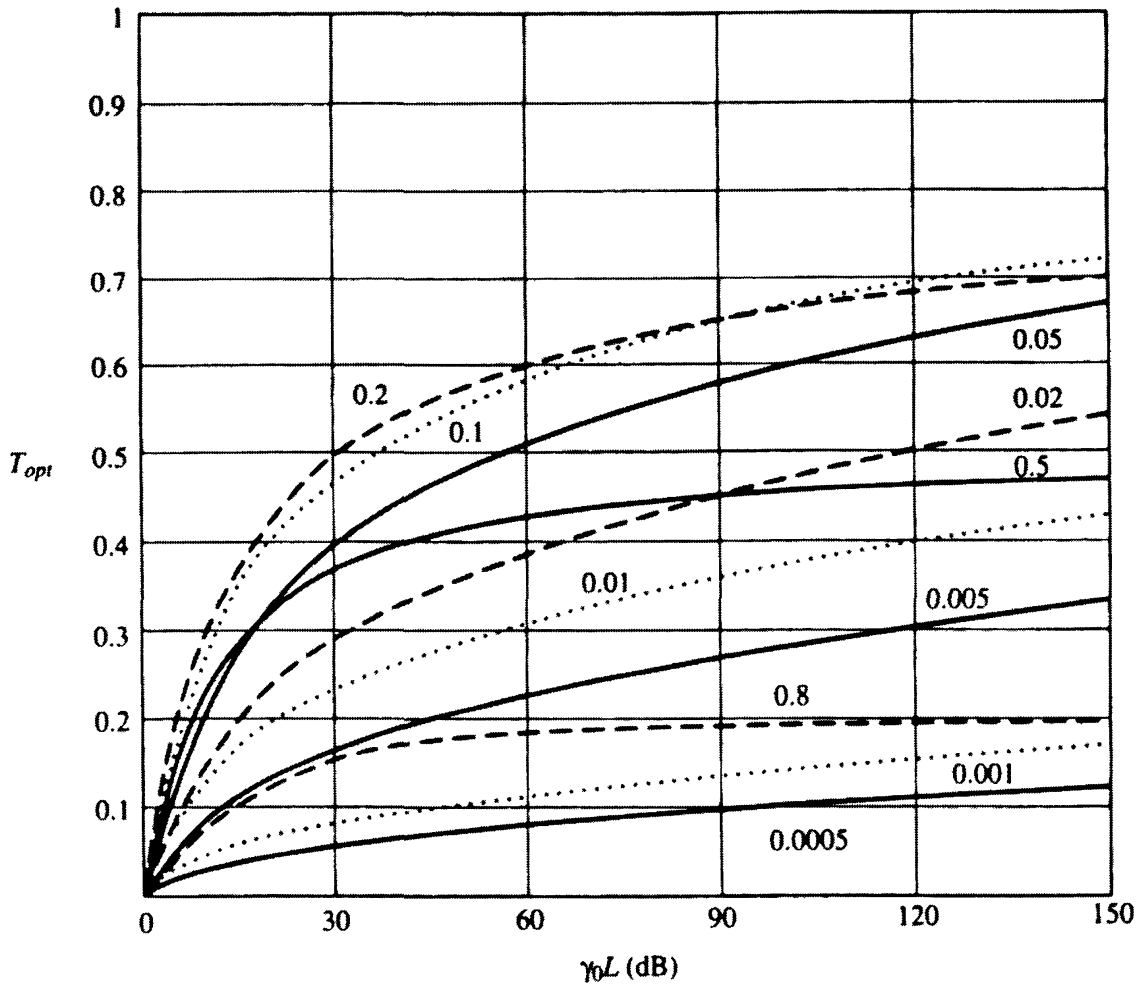


Figure A1.51. Calculated optimum coupling for a symmetrical resonator for various values of the loss paramater A and the unsaturated gain.

A1.18.1 Optimum coupling

To maximize the output intensity from the symmetrical resonator, we must find the value of R such that $\partial I_{\text{out}}/\partial R = 0$ which gives

$$\frac{T_{\text{opt}}}{A} = \left(\frac{1 - A - T_{\text{opt}}}{A + T_{\text{opt}}}\right)[\gamma_0 L + \ln(1 - A - T_{\text{opt}})]. \tag{A1.273}$$

For small losses, such that $A + T_{\text{opt}} \ll 1$ equation (A1.273) gives

$$\frac{T_{\text{opt}}}{A} = \sqrt{\frac{\gamma_0 L}{A}} - 1. \tag{A1.274}$$

Figure A1.51 shows the calculated optimum coupling for various values of the loss parameter A and the unsaturated gain in dB ($4.343\ \gamma_0 L$).

In practice, it should be pointed out, the optimum mirror transmittance in a laser system is generally determined empirically. For example, for the cw CO_2 laser, whose unsaturated gain varies roughly inversely with the tube diameter d, the optimum mirror transmittance has been determined to be $T \simeq L/500d$.

Acknowledgment

All figures except A1.21, 29, 32, 33 and 45 are reprinted by permission of Cambridge University Press.

References

[1] Davis C C 1996 *Lasers and Electro-Optics* (Cambridge: Cambridge University Press)
[2] Liao S Y 1985 *Microwave Solid-State Devices* (Englewood Cliffs, NJ: Prentice-Hall)
[3] Dicke R H and Wittke J P 1960 *Introduction to Quantum Mechanics* (Reading, MA: Addison-Wesley)
[4] Liboff L 2003 *Introductory Quantum Mechanics* 4th edn (San Francisco, CA: Benjamin-Cumings)
[5] White R L 1966 *Basic Quantum Mechanics* (New York: McGraw-Hill)
[6] Jeans Sir J H 1940 *An Introduction to the Kinetic Theory of Gases* (Cambridge: Cambridge University Press)
[7] Present R D 1958 *Kinetic Theory of Gases* (New York: McGraw-Hill)
[8] Champeney D C 1973 *Fourier Transforms and Their Physical Applications* (London: Academic)
[9] Mitchell A G C and Zemansky M W 1971 *Resonance Radiation and Excited Atoms* (Cambridge: Cambridge University Press)
[10] Milonni P W and Eberly J H 1988 *Lasers* (New York: Wiley)
[11] Abramowitz M and Stegun I A 1968 *Handbook of Mathematical Functions* (New York: Dover)
[12] Bennett W R Jr 1962 Hole-burning effects in a He–Ne optical maser *Phys. Rev.* **126** 580–93
[13] Lamb W E 1964 Theory of an optical maser *Phys. Rev.* A **134** 1429–50
[14] Born M and Wolf E 1999 *Principles of Optics Electromagnetic Theory of Propagation, Interference and Diffraction of Light* 7th edn (Cambridge: Cambridge University Press)
[15] Siegman A E 1986 *Lasers* (Mill Valley, CA: University Science Books)
[16] Yariv A 1991 *Introduction to Optical Electronics* 4th edn (New York: Holt, Rinehart and Winston)
[17] Yariv A 1989 *Quantum Electronics* 3rd edn (New York: Wiley)
[18] Ditchburn R W 1976 *Light* 3rd edn (New York: Academic)
[19] Hecht E 1998 *Optics* 3rd edn (Reading, MA: Addison-Wesley)
[20] Dainty J C (ed) 1975 *Laser Speckle and Related Phenomena (Topics in Applied Physics 9)* (Berlin: Springer)
[21] Francon M 1979 *Laser Speckle and Applications in Optics* (New York: Academic)
[22] Jones R and Wykes C 1989 *Holographic and Speckle Interferometry* 2nd edn (Cambridge: Cambridge University Press)
[23] Rigrod W W 1965 Saturation effects in high-gain lasers *J. Appl. Phys.* **36** 2487–90
See also Rigrod W W 1963 Gain saturation and output power of optical masers *J. Appl. Phys.* **34** 2602–9
Rigrod W W 1978 Homogeneously broadened CW lasers with uniform distributed loss *IEEE J. Quantum Electron.* **QE-14** 377–81
[24] Smith P W 1966 The output power of a 6328 Å He–Ne gas laser *IEEE J. Quantum Electron.* **QE-2** 62–8

A2.1
Free-space laser resonators

Robert C Eckardt

A2.1.1 Introduction

Resonators provide the optical structure in which laser oscillations are established. Passive optical resonators can also be used to increase locally the power of coherent optical radiation or to filter optical radiation. An understanding of laser resonators is necessary in analysing the spatial beam characteristics and temporal coherence properties of the light output of laser systems. Optimizing the design of a laser system requires resonator analysis. In addition to the operation and design of lasers, an understanding of optical resonators is important in coupling the laser output to the application in which it is to be used. This chapter reviews the fundamental aspects of laser resonators and discusses some of the techniques of laser operation.

Optical resonators have resonant modes and these modes are important to the analysis of laser operation. An optical resonator mode is a field distribution that is resonant with the structure and that is reproduced in phase and in relative intensity after a round-trip transit of the resonator. The intensity of a mode may decrease due to resonator losses or it may be increased by amplification in an active laser material or by light introduced from outside the resonator. Mode competition and selection usually occur when an oscillation builds up in a resonator. Typically, laser oscillations start from random quantum fluctuations and build to high intensity in a single mode or many simultaneous modes. After many resonator transits the mode or modes with minimum loss tend to dominate and to be reproduced after each cavity transit. An iterative numerical analysis that simulated this process was used in the seminal paper by Fox and Li [1] to analyse the field distributions of laser resonators. Numerical methods of resonator analysis augmented by fast Fourier transform techniques and computers of increasing capability are currently in wide use. The analysis of resonators that have large dynamic changes, nonlinearities or significant diffraction typically requires numerical methods rather than analytical techniques. However, certain analytical methods for resonator analysis also have wide application and are extensively used.

Many of the analytical techniques for describing stable open resonators originated in the work of Boyd and Gordon [2]. The Gaussian beams of stable open resonators can be characterized with analytical expressions and provide a widely used approach to the discussion of laser resonators. A Gaussian beam is the fundamental mode of a set of Hermite–Gaussian modes in rectangular coordinates, and the same Gaussian mode is also the fundamental mode of a set of Laguerre–Gaussian modes in cylindrical coordinates. Either of these sets of modes forms complete sets of orthogonal functions, which can be used in a series expansion to describe a transverse field distribution to an arbitrary degree of accuracy. These modes will usually be a very good approximation, although still an approximation, to the modes of any real laser resonator. Real resonators will always have some limiting aperture, which introduces a small loss, and possibly other perturbations that will slightly distort the modes and make them non-orthogonal.

A resonator is called 'stable' if a ray of a geometrical optical analysis always remains within the resonator rather than wandering off an increasing distance from the resonator centre with an increasing number of cavity

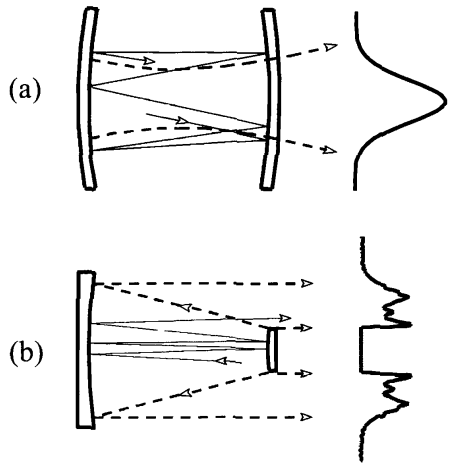


Figure A2.1.1. Schematic representation of a stable resonator (*a*) and an unstable resonator (*b*). A ray path (full line) is confined to remain on the mirrors in the stable resonator, but the geometrical ray walks off the mirrors in the unstable resonator. The curves on the right-hand side and the broken lines represent the intensity distributions of the fundamental resonator modes.

transits. Unstable resonators, which do not satisfy this condition, nonetheless have modes that are reproduced through diffraction from one cavity transit to the next. The term 'unstable resonator' is somewhat misleading because these resonators can be quite stable in operation and they are useful for efficient energy extraction in high-gain systems that can support the higher diffraction losses typical of unstable resonators. Siegman [3] prefers the descriptive name 'geometrically unstable' as being more accurate for these resonators. Stable resonators typically have small diffraction losses and support Gaussian transverse fundamental modes. They may also support many higher-order transverse modes giving a great deal of structure to the beam. Unstable resonators usually offer large loss discrimination between the lowest-loss transverse mode and other transverse modes. For this reason they often provide good spatial beam quality in large-diameter laser output beams. Stable and unstable resonators are illustrated schematically in figure A2.1.1.

Resonators simultaneously have axial modes in addition to their transverse modes. The condition that the phase distribution be reproduced after a round-trip cavity transit allows axial modes that cycle through a phase change of integer multiples of 2π on a round-trip transit. Higher-order transverse modes have more complex spatial amplitude distributions, which are also reproduced from one cavity transit to the next. The higher-order transverse modes have additional phase shifts that place them, in frequency, between the fundamental transverse modes. The axial or longitudinal resonator modes with a fundamental transverse distribution will be nearly equally spaced in frequency or wavenumber. The modes will only be precisely spaced if care is taken to compensate for dispersion and frequency-dependent phase shifts of the optical elements of the resonator. Such compensation becomes important for mode locking, in which case many axial modes are locked in phase to synthesize a single short pulse that propagates back and forth in the resonator. Other cavity control techniques include frequency selection or narrowing to restrict axial modes and spatial filtering to restrict transverse modes. Laser resonator losses can be controlled to hold off oscillation and then suddenly reduced in a technique called Q-switching, to produce an energetic pulse of only a few resonator transits duration.

The objective of this chapter is to present a basic description of laser resonators at a level that will allow the reader to apply the material to practical systems. Derivations are not given in this review but are covered in detail in the references. There have been many excellent papers and reviews on this subject. The books by Siegman [3], by Hall and Jackson [4] and by Hodgson and Weber [5] cover the topic thoroughly. Other more general books dealing with lasers also have good discussions of the topic [6–8]. Two review

papers by Siegman [9, 10] detail the development of the analysis of laser resonators and laser beams and include extensive bibliographies. Many of the results, terminology and notation developed in the earlier papers have become standard and are followed in this chapter. The basic types of resonators that will be discussed include geometrically stable two-mirror cavities, simple unstable resonators and resonators with plane-parallel mirrors.

The next section of this chapter reviews the optics of Gaussian light beams. These beams are solutions of the wave equation for propagation in a homogeneous, isotropic medium and are the fundamental modes of geometrically stable open resonators. The diffraction of Gaussian beams can be treated directly with analytical methods, simplifying calculations of laser-beam propagation and mode matching between resonators. The minimum diffraction of a Gaussian beam is used to define the diffraction-limited beam quality factor M^2 of one. The fundamental mode of a stable two-mirror cavity can be determined by matching the Gaussian-beam wavefront radii to the mirror radii. More complicated cavities are typically analysed with ABCD or ray-transfer-matrix techniques. Such analysis and application of the ray transfer matrices to the propagation of Gaussian beams are discussed here. A description of higher-order transverse modes of stable resonators follows and is used for a further description of beam quality.

Unstable resonators are discussed here both from the geometrical optics and diffraction points of view. An illustrative example of numerical resonator analysis is used to discuss transverse-mode selection in several types of unstable resonators. The calculation is seeded with a transverse distribution obtained by randomizing the phases of the initial Fourier-spatial-frequency components to simulate development from quantum fluctuations. The comparison includes plane-parallel-mirror resonators and collimated-output unstable resonators with hard-edged and soft mirrors. The use of variable-reflectivity laser mirrors is important to mitigate diffraction effects in unstable resonators. Gain-guided laser resonators, common in laser-pumped lasers, offer another technique for controlling the diffraction effects in lasers with plane-parallel resonators and in unstable resonators.

The discussion of axial modes starts with the frequency spacing of fundamental and higher-order transverse modes. The topic of cavity finesse deals with the frequency width of individual resonator modes. The bandwidth of the active gain medium of the resonator limits the number of axial modes or the frequency bandwidth of the oscillation. Components such as prisms, gratings and etalons are added to resonators to further narrow the bandwidth. Frequency stability is achieved first through mechanical and thermal stabilization of resonators. Higher levels of frequency stabilization require active stabilization to a frequency standard and careful control of the pumping stability. Modulation within the resonator is used to achieve mode locking and Q-switching. Techniques for injection seeding and injection locking can be used to transfer the properties of highly coherent low-power resonators to high-power systems. It is necessary to match the properties of a beam to the cavity into which it is injected. Examples of mode matching are presented with one lens and two lenses used to relay the laser beam. The optics of Gaussian beams is fundamental to all of these topics and that is where we begin.

A2.1.2 Gaussian beams

A2.1.2.1 Conventions and notation

The discussion here will be restricted to paraxial optics. In this limit the sine of an angle is approximated by that angle expressed in radians. Spherical surfaces can be represented by a parabolic approximation: $z = (x^2 + y^2)/2R$ is used to describe a spherical surface of radius of curvature R. A positive wavefront radius of curvature indicates a beam that is convex in the direction of propagation or diverging. A negative wavefront radius of curvature indicates a beam that is concave in the direction of propagation or converging (figure A2.1.2). Similarly, the radius of curvature of a concave spherical mirror surface is positive and that of a convex surface is negative. Coordinate rotations or reflections are assumed for refraction or reflection at

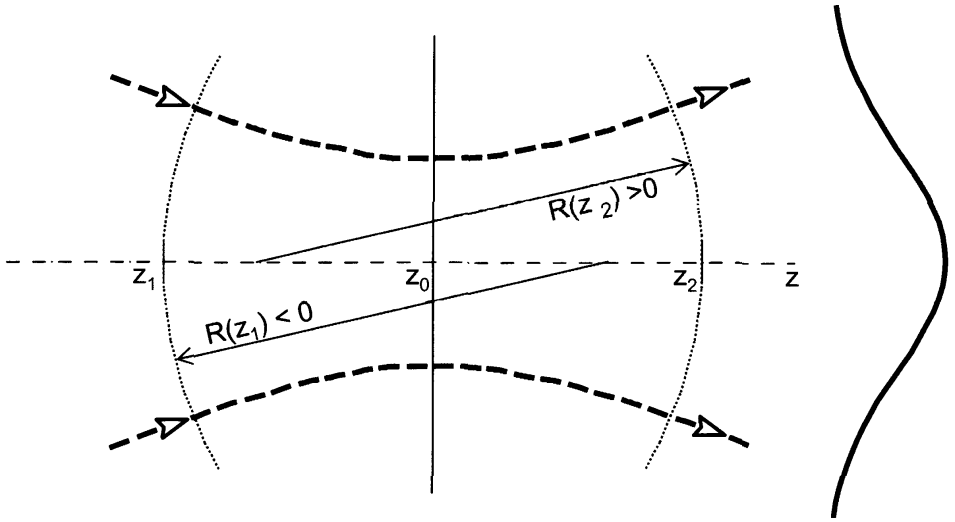


Figure A2.1.2. The sign convention for wavefront radii of curvature is positive when the beam is divergent and negative when the beam is convergent.

arbitrary angles such that the beam remains centered on the z-axis and propagating in the $+z$ direction. There is no coordinate translation or scaling in the direction of propagation and z remains the cumulative physical distance of propagation. A plane wave travelling in the $+z$ direction is expressed as $\exp\{-\mathrm{i}(kz - \omega t)\}$. If the complex conjugate of this expression is used to describe this plane wave, appropriate changes of sign may be required in some expressions.

A2.1.2.2 Description of Gaussian beams

The derivation of the properties of a Gaussian beam is discussed in detail in texts such as those by Siegman [3] and Hodgson and Weber [5]. The early review article by Kogelnik and Li [11] standardized much of the notation for the treatment of Gaussian beams and the article remains relevant. The results presented in these sources are reproduced here. Gaussian beams are solutions of the paraxial wave equation

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} - 2ik\frac{\partial E}{\partial z} = 0 \tag{A2.1.1}$$

and, equivalently, Fresnel's approximation to Huygen's integral

$$E(x, y, z) = \frac{i\mathrm{e}^{-\mathrm{i}k(z-z_0)}}{(z - z_0)} \iint E(x_0, y_0, z_0) \exp\left[-ik\frac{(x - x_0)^2 + (y - y_0)^2}{2(z - z_0)}\right] \mathrm{d}x_0\, \mathrm{d}y_0. \tag{A2.1.2}$$

Here E is the electric field, which is specified at coordinates (x_0, y_0, z_0) and calculated at (x, y, z), $\mathrm{i} = \sqrt{-1}$ and $k = 2\pi/\lambda$ where λ is the wavelength. For these equations to be valid it is necessary that there is no sharp discontinuity, such as an aperture edge or a phase step, near the plane at which the field is evaluated.

A Gaussian beam has transverse amplitude distributions proportional to $\exp\{-(r/w)^2\}$ and the transverse intensity distribution is proportional to $\exp\{-2(r/w)^2\}$. The radial coordinate is the distance from the centre of the beam: $r = \sqrt{x^2 + y^2}$. The parameter w is called the spot size and changes with distance as the beam propagates. Circularly symmetric beams are initially discussed here but later the discussion is expanded to include astigmatic beams with amplitude distributions expressed as $\exp\{-(x^2/w_x^2) - (y^2/w_y^2)\}$. Wavefronts of constant phase of circularly symmetric Gaussian beams form spherical surfaces and the wavefronts of the astigmatic beams are circular in cross section along the principal axes of the astigmatic components, which are assumed here to be aligned for all components in the beam.

The mathematical description of the electric field of a Gaussian light beam or TEM_{00} mode is a real function expressed as the sum of a complex expression and its complex conjugate:

$$E(x, y, z, t) = \frac{1}{2} E_0 \frac{w_0}{w(z)} \exp\left\{-\frac{r^2}{w^2(z)} - \mathrm{i}\left(\frac{r^2 k}{2R(z)} - \Phi(z)\right)\right\} \exp\{-\mathrm{i}(kz - \omega t)\} + \text{c.c.} \qquad \text{(A2.1.3)}$$

This is a solution to the wave equation representing a nearly collimated beam propagating in the $+z$ direction. The complex conjugate is not carried in the discussion and it is sufficient to treat the field as a complex parameter. Only in a few special cases, not encountered here, is it necessary to retain the real number representation of the electric field. It is assumed that the transverse dimension of the beam is much larger than the wavelength, that is $w_0 \gg \lambda$, a condition which is typical of most laser beams and laser resonators. The factor $\exp\{-\mathrm{i}(kz - \omega t)\}$, with $k = 2\pi n/\lambda_0$, n the index of refraction of the medium in which the beam is propagating, λ_0 the free-space wavelength, ω the angular frequency and t the time, represents the plane wave component of the distribution. The remaining portion of the expression describes differences in field distribution and phase from that of the plane wave. The parameter $w(z)$ is the spot size and $R(z)$ is the wavefront radius of curvature. The change due to propagation with diffraction for these two parameters is

$$w^2(z) = w_0^2 \left[1 + \left\{\frac{\lambda(z - z_0)}{\pi w_0^2}\right\}^2\right] = w_0^2[1 + \{(z - z_0)/z_R\}^2] \qquad \text{(A2.1.4)}$$

and

$$R(z) = (z - z_0)\left[1 + \left\{\frac{\pi w_0^2}{\lambda(z - z_0)}\right\}^2\right] = (z - z_0)[1 + \{z_R/(z - z_0)\}^2]. \qquad \text{(A2.1.5)}$$

The beam waist is located at z_0 where $w(z_0) = w_0$ has a minimum value. The parameter z_R is the Rayleigh range or Rayleigh length:

$$z_R = \pi w_0^2/\lambda. \qquad \text{(A2.1.6)}$$

The Rayleigh length is the distance necessary to travel from the beam waist for the spot size to increase by a factor of $\sqrt{2}$. The confocal parameter b of a Gaussian beam, a commonly used parameter, is twice the Rayleigh range:

$$b = kw_0^2 = 2\pi n w_0^2/\lambda_0. \qquad \text{(A2.1.7)}$$

The parameter $\Phi(z)$ is the Gouy phase shift and gives the departure of the on-axis phase from that of a plane wave:

$$\Phi(z) = \arctan\{\lambda(z - z_0)/(\pi w_0^2)\}. \qquad \text{(A2.1.8)}$$

When propagated into the far field, a Gaussian beam will expand with a divergence half-angle of $\theta = \lambda/(\pi w_0)$ (figure A2.1.3). The constant E_0 in equation (A2.1.3) is the peak electric field at the beam waist and is expressed in SI units in terms of the total power in the beam P or peak intensity I_0 as

$$E_0 = \sqrt{2I_0/(n\varepsilon_0 c)} = \sqrt{4P/(\pi w_0^2 n\varepsilon_0 c)} \qquad \text{(A2.1.9)}$$

where E_0 is in units of V m^{-1}, P is in W, I_0 is in W m^{-2}, $\varepsilon_0 \simeq 8.854 \times 10^{-12}$ CN m^{-2} is the permittivity of free space and $c \simeq 2.997 \times 10^8$ m s^{-1} is the speed of light.

Expressing a Gaussian beam in terms of w_0, the beam waist spot size, and z_0, the position of the beam waist, offers the advantage of explicitly stating the spot size and wavefront radius of curvature as a function of position. An alternate description is given in terms of $q(z)$ the 'complex beam parameter' or 'complex radius of curvature',

$$1/q(z) = 1/R(z) - \mathrm{i}\lambda/\{\pi w^2(z)\}. \qquad \text{(A2.1.10)}$$

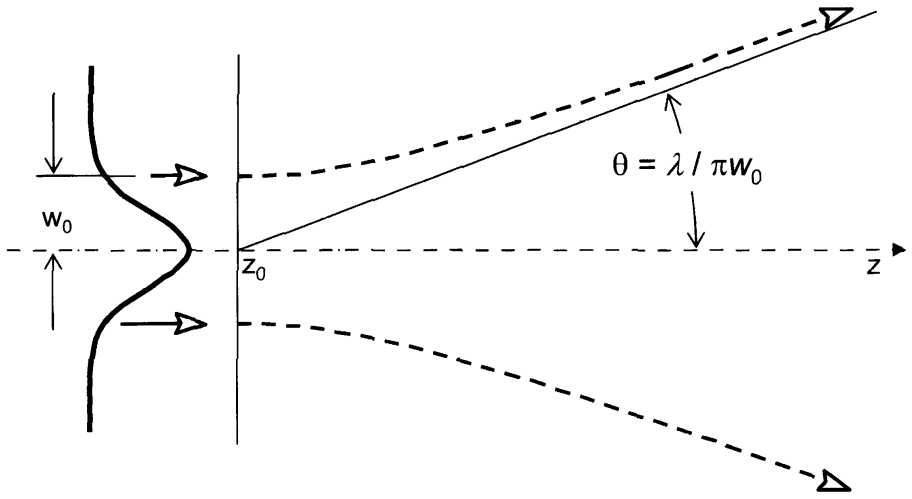


Figure A2.1.3. A Gaussian beam diffracts into a half-angle of $\theta = \lambda/\pi w_0$ as it propagates into the far field.

Using $q(z)$ the expression (A2.1.3) for the Gaussian beam becomes

$$E(x, y, z, t) = \frac{1}{2} E_0 \frac{q_0}{q(z)} \exp\left\{-\mathrm{i}\frac{kr^2}{2q(z)}\right\} \exp\{-\mathrm{i}(kz - \omega t)\} + \text{c.c.} \tag{A2.1.11}$$

where $q_0 = -\mathrm{i}\lambda/(\pi w_0^2)$. It follows that $q_0/q(z) = \exp(\mathrm{i}\Phi(z))w_0/w(z)$, and the two forms equation (A2.1.3) and equation (A2.1.11) are equivalent. Using the complex beam parameter offers advantages in terms of mathematical simplicity. For example, the change of q with propagation is simply

$$q(z) = q_0 + z - z_0. \tag{A2.1.12}$$

The use of the complex beam parameter and the 2 × 2 ABCD or ray transfer matrices greatly simplifies the analysis of complex resonators. Before discussing the ray transfer matrices, some additional relationships involving spot sizes and wavefront radii are given.

The parameters w_0 and z_0, along with the direction of propagation, wavelength, index of refraction and beam power specify a Gaussian beam. Usually the characterization of a beam depends on the determination of w_0 and z_0 with the other parameters known. Table A2.1.1 lists relationships that can be used to characterize a beam. Equations (A2.1.14) can be used to determine the fundamental mode of a two-mirror laser cavity by matching the wavefront radius of curvature to the radii of curvature of the cavity mirrors (figure A2.1.4). The stability condition for a two-mirror oscillator is implicit in equation (A2.1.14) but difficult to extract in a simple form. Use of the complex beam parameter equation (A2.1.10) and the ray transfer matrices provides this simplicity.

A2.1.2.3 Ray transfer matrices

Ray transfer matrices or ABCD matrices provide concise and useful representations of both the geometrical propagation of paraxial rays and the propagation with diffraction of Gaussian beams. The propagation of rays through simple optical components is described by

$$\begin{bmatrix} x_2 \\ x_2' \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_1 \\ x_1' \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_2 \\ y_2' \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ y_1' \end{bmatrix}. \tag{A2.1.17}$$

The ray is assumed to be propagating nearly parallel to the z-axis. The distance of the ray from the z-axis is given by (x, y), and the projections of the slope of the ray are $x' = \mathrm{d}x/\mathrm{d}z$ and $y' = \mathrm{d}y/\mathrm{d}z$ (figure A2.1.5).

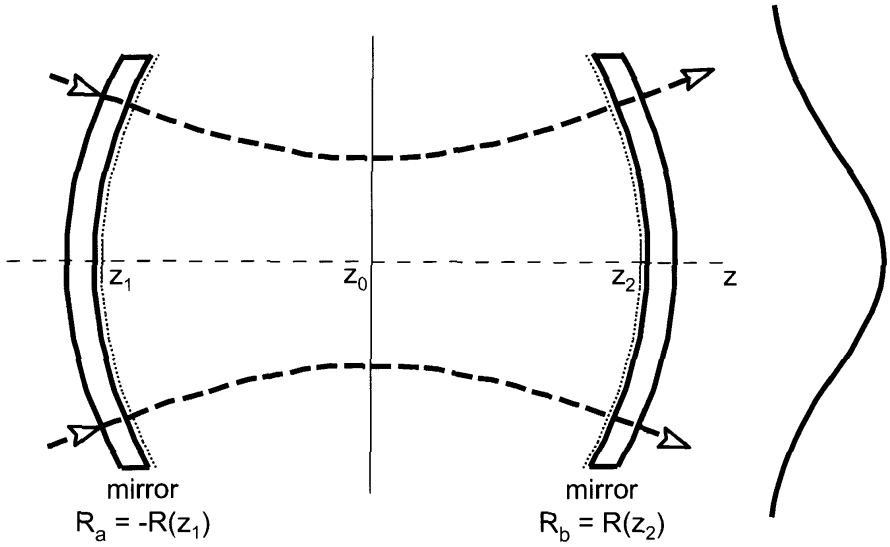


Figure A2.1.4. The modes of a two-mirror, stable resonator can be determined by matching wavefront radii of curvature to the mirror surfaces.

The ray position and slope before the optical element are identified by the subscript 1 and after the element by the subscript 2. Ray transfer matrices for six simple elements are given in table A2.1.2: propagation over a distance d; a thin lens of focal length f; propagation through a spherical interface from a medium of index n_1 to one of n_2; an element with flat parallel surfaces and an index distribution of $n(z) = n_0 - n_2(x^2 + y^2)/2$ inside and $n = 1$ outside; a spherical mirror; and a flat interface from index n_1 to index n_2. Ray transfer matrices can be combined by matrix multiplication to describe multiple-component systems:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} \begin{bmatrix} A_{n-1} & B_{n-1} \\ C_{n-1} & D_{n-1} \end{bmatrix} \cdots \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \tag{A2.1.18}$$

The multiplication must be performed as shown in equation (A2.1.18) with subscript 1 representing the first element encountered and n the last. When the index of refraction before the first element and after the last element of a system of components are equal, the determinant of the resultant matrix is unity:

$$AB - CD = 1. \tag{A2.1.19}$$

There is a generalization of the ray matrix formulation, namely multiplying the slope by the local index of refraction, that results in the unity determinant even if the index changes [3]. For optical resonators and propagation of one round-trip of the resonator, the index of refraction at the start and finish must be the same. For an arbitrary number of cavity transits of the resonator represented by an ABCD matrix, the distance of the ray from the z-azis is

$$r_s = r_{\max} \sin(s\theta + \delta) \tag{A2.1.20}$$

where $\theta = \cos^{-1}\{(A + D)/2\}$, δ and $r_{\max}$ are determined by initial conditions and s is the number of cavity transits. If the quantity $(A + D)/2$ is in the range

$$-1 < (A + D)/2 < 1 \tag{A2.1.21}$$

the ray will be bound within the resonator never exceeding some maximum distance from the z-axis. If this condition is not met, the relationship describing the distance from the axis becomes a hyperbolic trigonometric function and the distance eventually expands without limit. Resonators that satisfy condition (A2.1.21) are geometrically stable.

Table A2.1.1. Equations for Gaussian beam characterization.

if $w(z)$ and $R(z)$ are known at some position z:
$w_0^2 = w^2(z)\{1 + \pi w^2(z)/(\lambda R(z))\}, \qquad z_0 = z - R(z)\{1 + \lambda R(z)/(\pi w^2(z))\}$ (A2.1.13)

if $R_1 = R(z_1)$ and $R_2 = R(z_2)$ are known, $d = z_2 - z_1$:
$w_0^4 = \frac{\lambda^2 d(R_2-d)(R_1+d)(R_1-R_2+d)}{\pi^2(R_1-R_2+2d)^2}, \qquad z_0 = \frac{z_2-d(R_1+d)}{R_1-R_2+2d}$ (A2.1.14)

if $R_2 = R(z_2)$ and $w_1 = w(z_1)$ are known, $d = z_2 - z_1$:
$z_0 = z_1 + \left(b \pm \sqrt{b^2 - 4ac}\right)/(2a), \; w_0^2 = \lambda\sqrt{(z_2 - z_0)(R_2 - z_2 + z_0)}/\pi$ (A2.1.15)
where $a = (\lambda(R_2 - 2d)/\pi)^2 + w_1^2$, $b = w_1^4(2d - R_2) + 2\lambda^2 d(R_2 - 2d)(R_2 - d)/\pi^2$, and $c = w_1^4 d(d - R_2) + (\lambda d(R_2 - d)/\pi)^2$

if $w_1 = w(z_1)$ and $w_2 = w(z_2)$ are known at positions z_1 and z_2, $d = z_2 - z_1$:
$w_0^2 = \left(-b \pm \sqrt{b^2 - 4ac}\right)/(2a), \; z_0 = z_1 - \{(w_1^2 - w_1^2)(\pi w_0/\lambda)^2 - d^2\}/(2d)$ (A2.1.16)

where $a = 1 + \{(w_2^2 - w_1^2)\pi/(2d\lambda)\}^2$, $b = -(w_2^2 + w_1^2)/2$, and $c = (\lambda d/(2\pi))^2$.

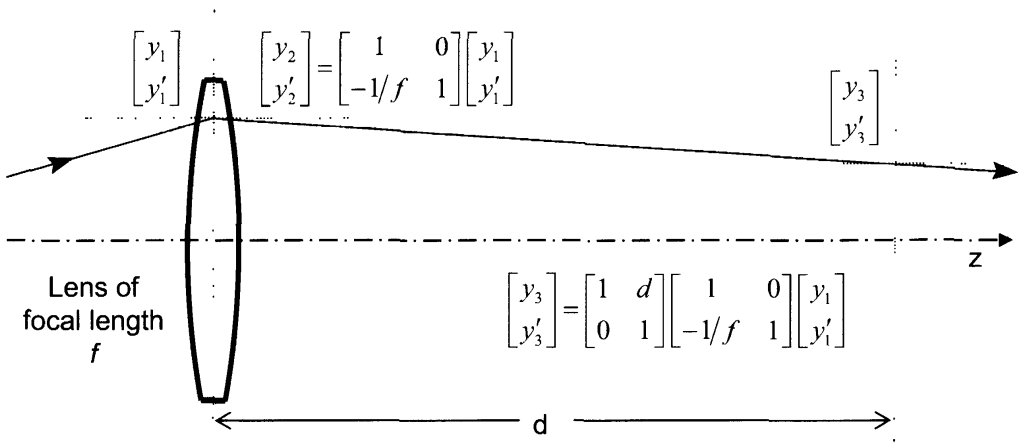


Figure A2.1.5. An application of ray transfer matrices. Subscript 1 refers to height and slope immediately before a thin lens, 2 immediately after and 3 after propagating a distance d beyond the lens.

The ray transfer matrix also describes the propagation of Gaussian beams by

$$q_2 = (Aq_1 + B)/(Cq_1 + D). \tag{A2.1.22}$$

Here A, B, C and D are the components of a ray transfer matrix for a single component or a combination of components. The complex beam parameter (A2.1.10) before the element or combination of elements is q_1 and after it is q_2. Equation A2.1.22 is called the ABCD law. It is easily confirmed using matrices from table A2.1.1 for propagation over a distance d or for a thin lens of focal length f; the respective results are $q_2 = q_1 + d$ and $1/R_2 = 1/R_1 - 1/f$.

A2.1.2.4 Gaussian resonant modes

Modes of stable two-mirror resonators can be found by fitting a Gaussian beam to the curvature and spacing of the resonator mirrors. More complex multiple-element resonators require the use of ABCD matrix techniques

Table A2.1.2. Ray transfer matrices for circularly symmetric components.

propagate distance d	thin lens of focal length f
$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}$
spherical interface	thick slab of graded index
of radius of curvature R	$n(z) = n_A - n_B(x^2 + y^2)/2$
going from index n_1 to n_2	with $n = 1$ outside
$\begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{n_2 R} & \frac{n_1}{n_2} \end{bmatrix}$	$\begin{bmatrix} \cos\left(d\sqrt{n_B/n_A}\right) & \frac{\sin\left(d\sqrt{n_B/n_A}\right)}{\sqrt{n_A n_B}} \\ -\sqrt{n_A n_B}\sin\left(d\sqrt{n_B/n_A}\right) & \cos\left(d\sqrt{n_B/n_A}\right) \end{bmatrix}$
spherical mirror with	flat interface
radius of curvature R_A	going from index n_1 to n_2
$\begin{bmatrix} 1 & 0 \\ \frac{-2}{R_A} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix}$

to determine the modes of a stable cavity. The ray transfer matrix for a resonator satisfies the condition that the initial index is the same as the final index, and the determinant is unity, that is, equation (A2.1.19) applies. If a resonator is to have a Gaussian beam as the fundamental mode, the complex beam parameter must be reproduced after each cavity round-trip transit. Starting at z_1 where $q_1 = q(z_1)$, and propagating through all n resonator components and returning to the same position requires

$$q_n = (Aq_1 + B)/(Cq_1 + D) = q_1. \tag{A2.1.23}$$

The ABCD matrix elements are obtained by ordered matrix multiplication as performed in equation (A2.1.18). The complex beam parameter at the position z_1 is obtained from equation (A2.1.22) using equation (A2.1.19):

$$\frac{1}{q_1} = \frac{D - A}{2B} - \frac{\mathrm{i}}{2|B|}\sqrt{4 - (A + D)^2} = \frac{1}{R(z_1)} - \frac{\mathrm{i}\lambda}{\pi w^2(z_1)}. \tag{A2.1.24}$$

The sign of the square root is chosen to obtain a positive value of w^2. The condition for w to be real is the same as equation (A2.1.21), i.e. that required for a geometrically stable resonator. It is also possible to write

$$q_1 = \frac{A - D}{2C} + \frac{\mathrm{i}\sqrt{(A + D)^2}}{2|C|} \tag{A2.1.25}$$

and obtain

$$z_0 = z_1 + (D - A)/(2C) \qquad \text{and} \qquad w_0^2 = \lambda\sqrt{4 - (A + D)^2}/(2\pi|C|). \tag{A2.1.26}$$

There are other cavity configurations that can support Gaussian modes, such as gain-guided resonators and those with apodized apertures but these are not treated with the simple ray transfer matrices used here. Siegman [3] describes higher-order matrices for generalized astigmatism and complex matrices to treat the apodized apertures. One generalization that is described later is an orthogonal astigmatic system in which the axes of the astigmatic components remain parallel and perpendicular. In this case the sagittal and tangential characteristics of the resonator modes are treated separately with appropriate ray transfer matrices. This is illustrated schematically for a mirror in figure A2.1.6. A tabulation of ray transfer matrices for astigmatic elements is given in table A2.1.3.

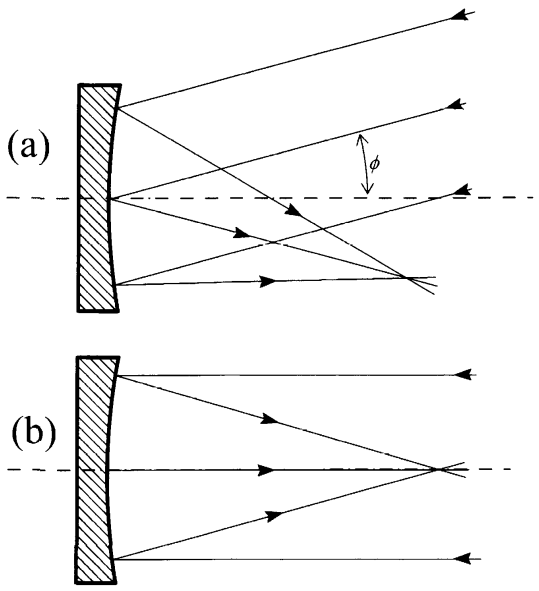


Figure A2.1.6. Top view showing the tangential focus (*a*) and side view showing the sagittal focus (*b*) of a spherical mirror. The axis of the mirror and incident beam are in a horizontal plane.

A2.1.3 Stable resonators

A2.1.3.1 Two mirror resonators

The previous section contains the analytic techniques necessary to characterize stable resonators. The simplest stable resonators have two mirrors. In a first approximation, the active laser material and other intracavity components can be accommodated by an effective cavity length change resulting from flat parallel plates of uniform refractive index. A graded index guide that approximates the effects of thermal loading in a laser rod could be used for a more accurate approximation. The resonator considered here simply has two spherical mirrors of radius of curvature R_A and R_B separated by distance d. The mirror radii are positive for concave mirrors and negative for convex mirrors. The first technique is to match the mirror curvatures to the wavefront curvatures by considering a Gaussian beam that passes through the two mirrors as shown in figure A2.1.4. Use equation (A2.1.14) with $R_1 = -R_A$ and $R_2 = R_B$ to obtain

$$w_0^4 = \frac{\lambda^2 d(R_B - d)(R_A - d)(d - R_A - R_B)}{\pi^2(2d - R_A - R_B)^2} \quad \text{and} \quad z_0 = \frac{z_2 - d(d - R_A)}{2d - R_A - R_B}. \tag{A2.1.27}$$

Here z_1 is the position of the first mirror, z_2 the position of the second mirror, $d = z_2 - z_1$ and z_0 is the location of the beam waist, which has spot size w_0. There will be a real value for the beam waist only when the right-hand side of the first equation of equation (A2.1.27) is positive. This is the condition for a stable cavity.

To perform a ray-transfer-matrix analysis of the same resonator pick a starting position, for example just before mirror R_B. Then the appropriate matrices from table A2.1.2 are used for reflection from the mirror R_B, propagation over the resonator length d, reflection from mirror R_A and propagation back to just before the right-hand mirror.

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2/R_A & \end{bmatrix}\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2/R_B & 1 \end{bmatrix} \\ &= \begin{bmatrix} \left(1 - \frac{2d}{R_A} - \frac{4d}{R_b} + \frac{4d^2}{R_A R_B}\right) & 2d\left(1 - \frac{d}{R_A}\right) \\ \left\{\frac{4d}{R_A R_B} - 2\left(\frac{1}{R_A} + \frac{1}{R_B}\right)\right\} & \left(1 - \frac{2d}{R_A}\right) \end{bmatrix}. \end{aligned} \tag{A2.1.28}$$

Table A2.1.3. Ray transfer matrices for astigmatic components at angle of incidence ϕ

Sagittal	Tangential
Thin lens of focal length f with index n_2 surrounded by medium of index n_1: $\begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{n_2^2-n_1^2\sin^2\phi}-n_1\cos\phi}{(n_2-n_1)f} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{n_2^2-n_1^2\sin^2\phi}-n_1\cos\phi}{(n_2-n_1)f\cos^2\phi} & 1 \end{bmatrix}$
Thin cylindrical lens of tangential focal length f and index n_2 in a medium of index n_1: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{n_2^2-n_1^2\sin^2\phi}-n_1\cos\phi}{(n_2-n_1)f\cos^2\phi} & 1 \end{bmatrix}$
Flat interface going from index n_1 to index n_2: $\begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix}$	$\begin{bmatrix} \frac{\sqrt{n_2^2-n_1^2\cos\phi}}{n_2\cos\phi} & 0 \\ 0 & \frac{n_1\cos\phi}{\sqrt{n_2^2-n_1^2\cos^2\phi}} \end{bmatrix}$
Concave spherical mirror of radius of curvature R: $\begin{bmatrix} 1 & 0 \\ -\frac{2\cos\phi}{R} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ -\frac{2}{R\cos\phi} & 1 \end{bmatrix}$
Spherical interface of radius of curvature R going from index n_1 to index n_2: $\begin{bmatrix} 1 & 0 \\ \frac{\sqrt{n_2^2-n_1^2\sin^2\phi}-n_1\cos\phi}{n_2R} & \frac{n_1}{n_2} \end{bmatrix}$	$\begin{bmatrix} \frac{\sqrt{n_2^2-n_1^2\cos\phi}}{n_2\cos\phi} & 0 \\ \frac{1}{R\cos\phi}-\frac{n_1/(n_2R)}{\sqrt{n_2^2-n_1^2\cos^2\phi}} & \frac{n_1\cos\phi}{\sqrt{n_2^2-n_1^2\cos^2\phi}} \end{bmatrix}$
Thick plate of thickness d and index n_2 with parallel surfaces and index n_1 outside: $\begin{bmatrix} 1 & \frac{n_1 d}{\sqrt{n_2^2-n_1^2\cos^2\phi}} \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \frac{dn_1n_2\cos\phi}{n_2^2-n_1^2\sin^2\phi} \\ 0 & 1 \end{bmatrix}$

The manipulation of the completed matrix multiplications in the second line of equation (A2.1.28) could be continued to obtain equation (A2.1.27). When the condition for a stable resonator (A2.1.21) is applied to equation (A2.1.28),

$$0 < (1 - d/R_A)(1 - d/R_B) < 1 \tag{A2.1.29}$$

is obtained. The quantities $(1-d/R_A)$ and $(1-d/R_B)$ are commonly referred to as the resonator g parameters g_1 and g_2 and the stability condition (A2.1.29) in terms of these quantities becomes

$$0 < g_1 g_2 < 1. \tag{A2.1.30}$$

The familiar stability diagram (shown in figure A2.1.7) first used by Boyd and Gardner [2] is based on equation (A2.1.29).

It is possible to include a large amount of information in the stability diagram. Resonators for which the g_1g_2 product is located in the unshaded area are stable and those in the shaded area are unstable. The straight

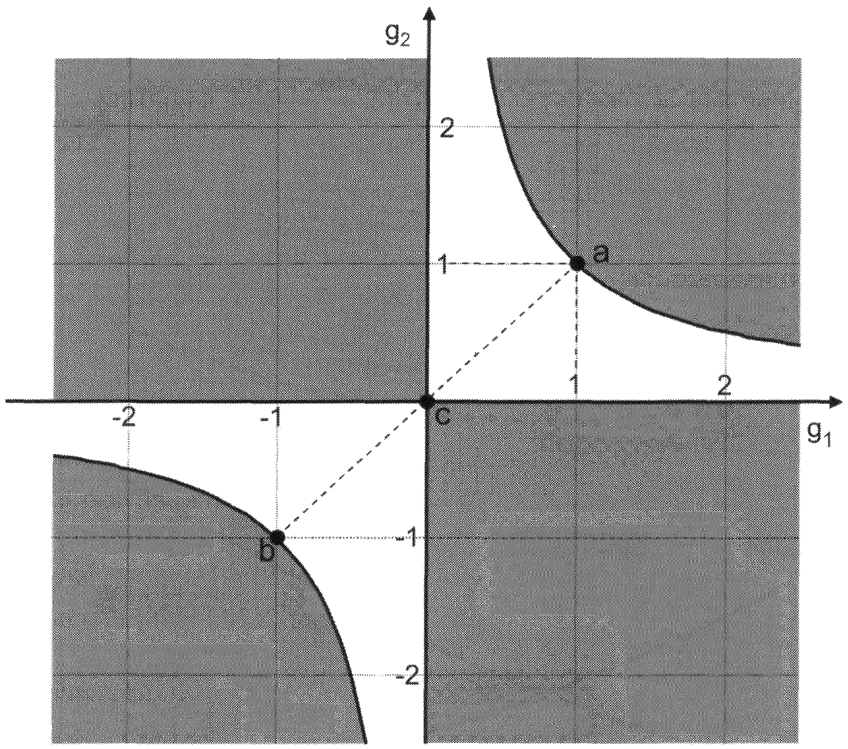


Figure A2.1.7. Resonator stability diagram. The two-mirror resonators for which $g_1 g_2$ falls in the unshaded area are stable. Point a represents the plane-parallel resonator, c a confocal resonator and b a concentric one.

line from point a ($g_1 = 1, g_2 = 1$) to point b ($g_1 = -1, g_2 = -1$) represents symmetric, stable cavities. The end points of this line, the plane-parallel cavity at point a and the concentric cavity at point b, are on the border of the stability region. It is not possible to implement these geometries as stable cavities because an infinite spot size is required at the cavity mirrors. The centre point c ($g_1 = 0, g_2 = 0$) representing the confocal cavity is also on the border of instability. A change from a symmetric geometry could make the confocal resonator unstable. The horizontal and vertical broken lines starting at point a ($g_1 = 1, g_2 = 1$) represent stable resonators with one flat mirror. If a resonator is chosen for stability, it is useful to have $g_1 g_2 \approx 1/2$. Other resonator choices could be based on other considerations such as the desired spot size at specific locations in the resonator. If a resonator is to accommodate a lensing that develops due to thermal loading of the active laser material, a good choice of resonator geometry might be to have the cold-cavity $g_1 g_2$ product near the appropriate end of the stable-symmetric-cavity line of figure A2.1.7, and have the thermal lensing bring the cavity toward the other end of the line. It is a geometrical property of stable resonators that the on-axis line from the surface of one mirror to its centre of curvature will partially overlap the corresponding line for the other mirror. If this condition is not met and there is no overlap or one line is completely contained within the other, the cavity is unstable.

A2.1.4 Higher-order modes of stable resonators

Stable lasers often run in multiple stable modes leading to mode beats. For laser operation on multiple longitudinal modes but restricted to a single transverse mode, the mode beating will result in temporal intensity and phase modulation that repeats with the cavity round-trip-transit frequency. If there are multiple transverse modes as well as multiple longitudinal modes, the mode beating becomes more complex. The transverse intensity and phase distribution will fluctuate as well. The average or time-integrated intensity distribution, however, may appear uniform, obscuring the instantaneous spatial structure.

Typically, special apertures or obstructions such as wires are required to force a laser resonator to operate in a single higher-order mode (see section A3.2.1). Modelling of higher-order modes is useful because actual laser output is often closely represented by a superposition of several resonator transverse modes and, therefore, it offers a technique for dealing with actual laser oscillations and beams. Higher-order modes for open stable resonators or free-space propagation usually are derived mathematically by substitution of a trial solution in the paraxial wave (equation A2.1.1) or Fresnel's paraxial approximation to Huygen's integral (equation A2.1.2). Depending on the form of trial solution, the result can be equations that describe the Hermite–Gaussian modes for rectangular symmetry or the Laguerre–Gaussian modes for cylindrical symmetry. This procedure determines the scaling factors and relationship between modes of different order. A fundamental-mode spot size $w(z)$ with propagation dependence given by equation (A2.1.4) and a wavefront radius of curvature $R(z)$ described by equation (A2.1.5) will carry over unchanged as parameters of both sets of modes. The higher-order modes retain their shape but expand in proportion to $w(z)$ with propagation. The Gouy phase shift changes with the order of the modes. The different phase shifts will cause the superposition of a series of modes to synthesize amplitude and phase distributions that change with propagation. Only individual transverse modes are guaranteed to retain their relative distribution with propagation. The more commonly used Hermite–Gaussian modes are described next, followed by the Laguerre–Gaussian modes.

A2.1.4.1 Cartesian coordinates

The Hermite–Gaussian modes, often called the transverse electromagnetic modes of order m and n or TEM_{mn}, have the form

$$E_{mn}(x, y, z, t) = \frac{1}{2} E_{mn0} \sqrt{\frac{w_{x0} w_{y0}}{w_x(z) w_y(z)}} \exp\{-\mathrm{i}(kz - \omega t)\} H_m\left(\frac{\sqrt{2}x}{w_x(z)}\right) H_n\left(\frac{\sqrt{2}y}{w_y(z)}\right)$$
$$\exp\left\{-x^2\left(\frac{1}{w_x^2(z)} + \frac{\mathrm{i}k}{2R_x(z)}\right) - y^2\left(\frac{1}{w_y^2(z)} + \frac{\mathrm{i}k}{2R_y(z)}\right) + i\Phi_{mn}(z)\right\} + \text{c.c.} \tag{A2.1.31}$$

The constant E_{mn0} in the in the above equation is given by

$$\sqrt{\frac{2P_{mn}}{n\varepsilon_0 c}} \sqrt{\frac{1}{w_{x0} w_{y0} \pi 2^{m+n-1} m! n!}} \tag{A2.1.32}$$

where P_{mn} is the power of the TEM_{mn} mode in watts. The Gouy phase shifts of the Hermite–Gaussian modes are given by

$$\Phi_{mn}(z) = (m + 1/2) \arctan\{\lambda(z - z_{x0})/(\pi w_{x0}^2)\} + (n + 1/2) \arctan\{\lambda(z - z_{y0})/(\pi w_{y0}^2)\}. \tag{A2.1.33}$$

Ellipticity and astigmatism are included in equations (A2.1.31)–(A2.1.33) by using separate spot sizes $w_x(z)$ and $w_y(z)$ and wavefront radii of curvature $R_x(z)$ and $R_y(z)$ for the orthogonal transverse directions. Implicit in the separate parameters for the two transverse directions is the possibility of different beam-waist positions z_{x0} and z_{y0}. The functions H_m and H_n are Hermite polynomials [12] and, for order up to four, are;

$$\begin{aligned} H_0 &= 1 \\ H_1(s) &= 2s \\ H_2(s) &= 4s^2 - 2 \\ H_3(s) &= 8s^3 - 12s \\ H_4(s) &= 16s^4 - 48s^2 + 12. \end{aligned} \tag{A2.1.34}$$

Higher-order Hermite polynomials can be obtained with the recurrence relation

$$H_{n+1}(s) = 2sH_n(s) - 2nH_{n-1}(s). \tag{A2.1.35}$$

The Hermite–Gaussian-polynomial orthogonality integrals are

$$\int_{-\infty}^{+\infty} \exp(-s^2)H_n(s)H_m(s)\,\mathrm{d}s = 0 \qquad (n \neq m) \tag{A2.1.36}$$

and

$$\int_{-\infty}^{+\infty} \exp(-s^2)H_n^2(s)\,\mathrm{d}s = \sqrt{\pi}2^n n!. \tag{A2.1.37}$$

The orthogonality integrals permit the evaluation of the constant E_{mn0} in equation (A2.1.31) describing the TEM_{mn} mode amplitude. Another integral that is required shortly is the second moment

$$\langle s^2 \rangle = \int_{-\infty}^{+\infty} s^2 \exp(-s^2)H_n^2(s)\mathrm{d}s \Big/ \int_{-\infty}^{+\infty} \exp(-s^2)H_n^2(s)\mathrm{d}s = (2n+1)/2. \tag{A2.1.38}$$

This is obtained using the recurrence relation and orthogonality integrals.

The Hermite–Gaussian modes form the familiar rectangular array of transverse distribution intensity peaks that are demonstrated in laser outputs with cavity perturbations that favour a single mode. Normalized mode amplitudes and relative intensities are plotted for one dimension in figure A2.1.8. The plots are for a beam waist or point where the beam is collimated to simplify the figure.

A2.1.4.2 Cylindrical coordinates

The modes of a resonator with strict circular symmetry can be described in cylindrical coordinates using generalized Laguerre polynomials. The fundamental mode in cylindrical coordinates is identical to the TEM_{00} Hermite–Gaussian mode with no astigmatism. Higher-order modes are given by

$$\begin{aligned} E_{pl}(r,\phi,z,t) = {} & \frac{1}{2}E_{pl0}\frac{w_0}{w(z)}\exp\{-\mathrm{i}(kz-\omega t)\}\left(\sqrt{2}\frac{r}{w(z)}\right)^{|l|} L_p^{|l|}(2r^2/w^2(z))\mathrm{e}^{-\mathrm{i}l\phi} \\ & \times \exp\left\{-r^2\left(\frac{1}{w^2(z)}+\frac{\mathrm{i}k}{2R(z)}\right)+\Phi_{pl}(z)\right\}+\mathrm{c.c.} \end{aligned} \tag{A2.1.39}$$

where (r,ϕ) are cylindrical coordinates and $L_p^{|l|}(\rho)$ is a generalized Laguerre polynomial with l an integer or zero and p zero or a positive integer. The parameters w_0, $w(z)$, k, ω and $R(z)$ are the same as those used with the Hermite–Gaussian modes. The first two generalized Laguerre polynomials are

$$L_0^{|l|}(\rho) = 1 \qquad L_1^{|l|}(\rho) = |l| + 1 - \rho. \tag{A2.1.40}$$

Higher-order generalized Laguerre polynomials can be obtained from the recurrence relation

$$(p+1)L_{p+1}^{|l|}(\rho) = (2p+|l|+1-\rho)L_p^{|l|}(\rho) - (p+|l|)L_{p-1}^{|l|}(\rho). \tag{A2.1.41}$$

The Gouy phase shifts for the cylindrical modes are given by

$$\Phi_{pl} = (2p+|l|+1)\tan^{-1}(\lambda z/\pi w_0^2). \tag{A2.1.42}$$

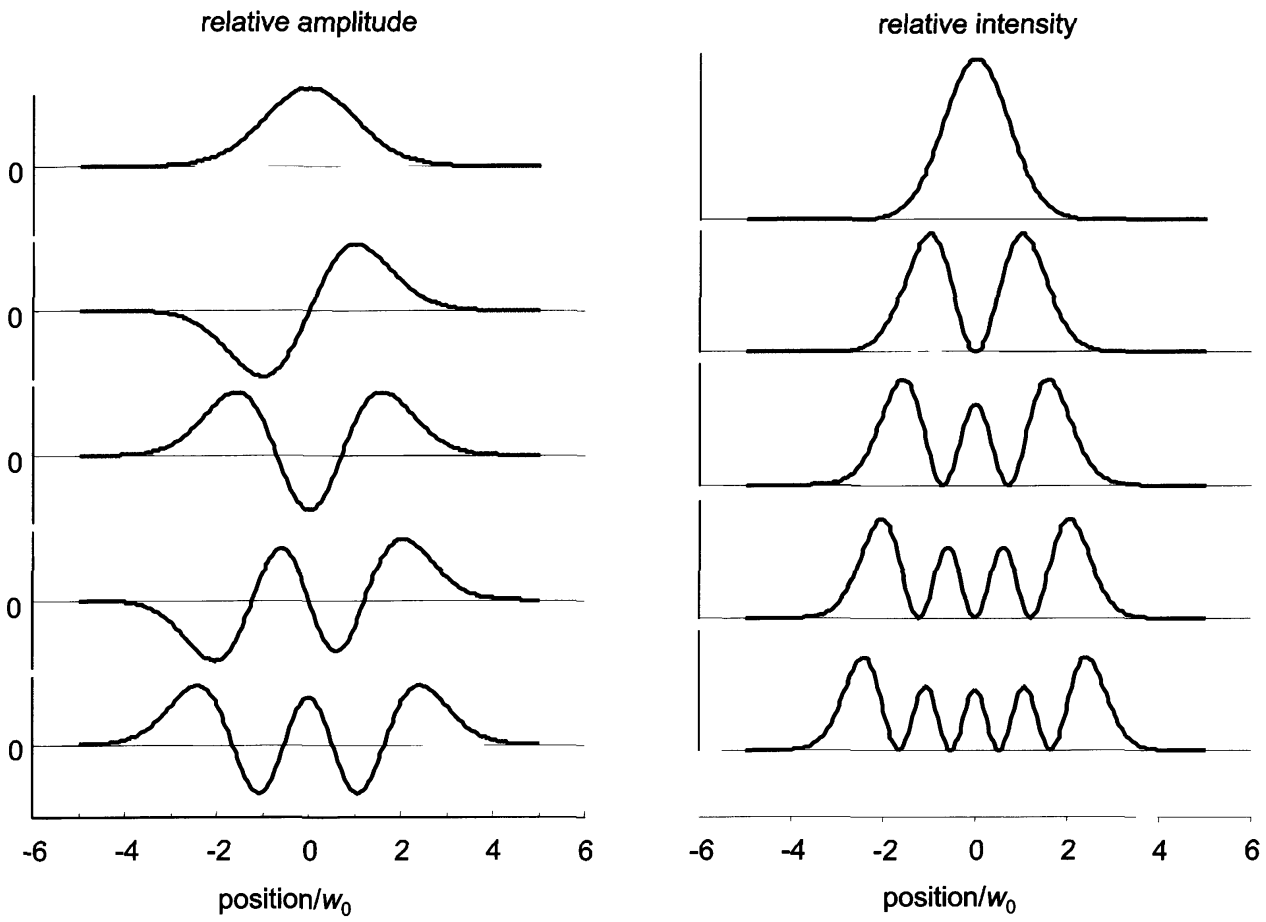


Figure A2.1.8. Relative amplitudes and intensities in one dimension of the five lowest-order Hermite–Gaussian modes.

The orthogonality integral for the generalized Laguerre polynomials is

$$\int_0^\infty \exp(-\rho)\rho^{\{(|l_1|+|l_2|)/2\}} L_{p_1}^{|l_2|}(\rho) L_{p_2}^{|l_2|}(\rho)\,\mathrm{d}\rho = \delta_{|l_1||l_2|}\delta_{p_1 p_2}(|l|+p)!/p! \tag{A2.1.43}$$

where the Kronecker delta $\delta_{\alpha\beta}$ is 1 for $\alpha = \beta$ and 0 if $\alpha \neq \beta$. The coefficient E_{pl0} in equation (A2.1.39) is given in terms of P_{pl}, the power of the p, l mode in watts, by

$$E_{pl0} = \sqrt{\frac{2P_{pl}}{n\varepsilon_0 c}\frac{2}{\pi w_0}\frac{p!}{(|l|+p)!}}. \tag{A2.1.44}$$

In practice, it is difficult to eliminate astigmatism completely and attain the degree of circular symmetry necessary to produce cylindrical modes.

A2.1.4.3 Beam quality

Actual resonators will have imperfections that distort a laser oscillation from the ideal Hermite–Gaussian modes, even though the distortion may be small. Laser oscillations may also consist of multiple transverse modes. Siegman's 'M squared' or M^2 parameter has been established as a measure of beam quality [13, 14]. M^2 is the ratio of the product of the square root of the second moment of the time-averaged transverse spatial distribution and the square root of the second moment of the angular distribution of a beam and the corresponding value for an ideal Gaussian beam. The transverse spatial distribution second moment or

variance changes with propagation, and the minimum values are used in the measurement of M^2. The M^2 value can be specified separately for the x and y coordinates as M_x^2 and M_y^2 or a combined value for the total beam can be specified. The variance for the x coordinate of the fundamental or TEM_{00} Gaussian mode is $\sigma_{x,m=0}^2(z) = w_x^2(z)/4$, which has a minimum value at the beam waist of $\sigma_{x,m=0}^2(z_0) = w_{x,0}^2/4$. The angular distribution can be obtained by Fourier transform or, equivalently, by propagating to the far field, where the angular halfwidth in the x direction at the $1/e$ maximum amplitude is $\theta_{x,m=0} = \lambda/(\pi w_{x,0})$. It follows that the angular variance of the fundamental Gaussian mode is $\sigma_{\theta x,m=0}^2 = \lambda^2/(4\pi^2 w_{x,0}^2)$ and the product that defines an M_x^2 of 1 is $\sigma_{x,m=0}(z_0)\sigma_{\theta x,m=0} = \lambda/(4\pi)$. For an arbitrary Hermite–Gaussian mode of order m, n,

$$\sigma_{x,m}(z_0)\sigma_{\theta x,m} = (2m+1)\lambda/(4\pi) \qquad \text{and} \qquad M_{x,m}^2 = 2m+1. \tag{A2.1.45}$$

Similar expressions apply for the y coordinate.

In numerical calculations, it is common to describe an arbitrary transverse distribution by the field amplitudes at an array of uniformly spaced sampling points. It is also possible to describe a transverse amplitude distribution as the sum of the orthogonal modes. The summation over a set of normalized Hermite–Gaussian modes can be expressed as

$$E(x, y, z, t) = \sum_m \sum_n c_{mn} E_{mn}(x, y, z, t) \tag{A2.1.46}$$

where the c_{mn} are the complex amplitudes of the normalized modes $E_{mn}(x, y, z, t)$. If the phases of the modes are random, a statistical average will yield

$$M_x^2 = \sum_m \sum_n (2m+1)|c_{mn}|^2 \Big/ \sum_m \sum_n |c_{mn}|^2 \tag{A2.1.47}$$

and

$$M_y^2 = \sum_m \sum_n (2n+1)|c_{mn}|^2 \Big/ \sum_m \sum_n |c_{mn}|^2. \tag{A2.1.48}$$

When Fourier transform techniques are used to describe beam propagation, the transverse amplitude distribution is typically described by its value at a rectangular array of $N_1 \times N_2$ discrete sampling points:

$$E_{n_1,n_2} = E(x_{n_1}, y_{n_2}, z, t) = E(n_1 \Delta x/N_1, n_2 \Delta y/N_2, z, t). \tag{A2.1.49}$$

The indices n_1 and n_2 are integers and range in value: $0 \le n_1 < N_1$ and $0 \le n_2 < N_2$. The array size is Δx in the x direction and Δy in the y direction. A discrete Fourier transform pair such as [15, 16]

$$\tilde{E}_{q_1,q_2} = \sum_{n_1}^{N_1-1} \sum_{n_2}^{N_2-1} \exp\{-i2\pi(q_1 n_1/N_1 + q_2 n_2/N_2)\} E_{n_1,n_2} \tag{A2.1.50}$$

$$E_{n_1,n_2} = (1/N_1 N_2) \sum_{q_1}^{N_1-1} \sum_{q_2}^{N_2-1} \exp\{i2\pi(q_1 n_1/N_1 + q_2 n_2/N_2)\} \tilde{E}_{q_1,q_2} \tag{A2.1.51}$$

is used to change from the spatial representation to the angular representation and back again. In the angular representation the individual components propagate at angles to the yz plane given by

$$\begin{aligned} &\theta_{x,q_1} = \lambda q_1/\Delta x \qquad \text{when } 0 \le q_1 < N_1/2, \\ \text{and} \quad &\theta_{x,q_1} = \lambda(q_1 - N_1)/\Delta x \qquad \text{when } N_1/2 \le q_1 < N. \end{aligned} \tag{A2.1.52}$$

A similar expression describes the angle with respect to the xz plane. The angular variance in the x direction is

$$\sigma_{\theta x}^2 = \sum_{q1}^{N_1-1} \sum_{q2}^{N_2-1} (\theta_{x,ql} - \bar{\theta}_x)^2 \tilde{E}_{q_1,q_2} \tilde{E}^*_{q_1,q_2} \Big/ \sum_{q1}^{N_1-1} \sum_{q2}^{N_2-1} \tilde{E}_{q_1,q_2} \tilde{E}^*_{q_1,q_2}. \tag{A2.1.53}$$

The x spatial variance is

$$\sigma_x^2 = \sum_{n1}^{N_1-1} \sum_{n2}^{N_2-1} \left(\frac{n_1 \Delta x}{N_1} - \bar{x} \right)^2 E_{n_1,n_2} E^*_{n_1,n_2} \Big/ \sum_{n1}^{N_1-1} \sum_{n2}^{N_2-1} E_{n_1,n_2} E^*_{n_1,n_2}. \tag{A2.1.54}$$

It is necessary to propagate to a z position where the spatial variance is minimum to obtain M^2 or, equivalently, to remove the spherical curvature [17] from the wavefront before calculation of the spatial and angular variances. When this is done, the M^2 values are given by

$$M_x^2 = 4\pi \sigma_{\theta x} \sigma_{x,\min}/\lambda \qquad \text{and} \qquad M_y^2 = 4\pi \sigma_{\theta y} \sigma_{y,\min}/\lambda. \tag{A2.1.55}$$

There are as many ways to obtain the M^2 of an actual beam experimentally as there are for numerically modelled beams. One experimental technique involves estimating a Gaussian spot size W_n of the beam at many positions z_n along the beam and fitting these to functions of the form:

$$W_x^2(z) = M_x^2 \{ w_{x,0}^2 + (z - z_{x,0})^2 (\lambda/\pi w_{x,0})^2 \} = W_{x,0}^2 \{ 1 + M_x^4 (z - z_{x,0})^2 (\lambda/\pi W_{x,0}^2)^2 \} \tag{A2.1.56}$$

to obtain the M^2 value, the imbedded Gaussian beam-waist spot size w_0 and the beam-waist position z_0. The capital letter W is used to indicate the spot of an actual beam, which is larger than the spot size of the imbedded Gaussian beam. In performing a least-squares fit to a set of experimental data, it is helpful to weight the individual measurements of W^2 by $1/W^2$ to place more significance on the measurements near the beam waist and reduce the possibility of a fit that predicts meaningless negative values of W^2. The spot sizes W_n may be estimated by taking the difference of knife-edge positions that transmit 16% and 84% of the total beam. A refinement could involve obtaining best-fit Gaussian distributions from several knife-edge positions at each of many propagation distances, z_n. Scanning pinholes, scanning slits and array detectors also can be used to measure beam distributions and second moments. It is necessary to consider the properties of the measurement and application. For example, a small amount of energy or error in measuring at a large distance from the central lobe of a beam will increase M^2. However, the small amount of energy at a large distance from the central lobe may or may not be significant in the application. The M^2 parameter provides both a measure of beam quality and a mechanism to use Gaussian beam propagation methods to deal with the propagation of actual laser beams.

A2.1.5 Mode matching

In many applications it is necessary to be precise in the spatial distribution of a laser beam delivered on a target. Knowledge of the beam's intensity distribution at the target is obtained from the beam power and the beam propagation parameters. The size and beam quality parameters are critically important when coupling laser beams into optical fibres. It is desirable to match the transverse distribution of a pump beam to the mode size of a resonator for laser-pumped lasers. In some applications, mode matching is extended to matching the confocal parameters of a pump beam and an external resonant cavity. With a single lens or spherical mirror it is possible to control either the size of the beam waist or its location. Two lenses adjustable in position are required to control both the beam-waist position and size simultaneously.

Configurations for single-lens mode matching with an ideal Gaussian beam are described next. The application of the beam quality parameter and the concept of an imbedded Gaussian beam allow these results to be extended to actual beams that are far from perfect. The results for single-lens mode matching can be cascaded to make predictions for two-lens mode matching.

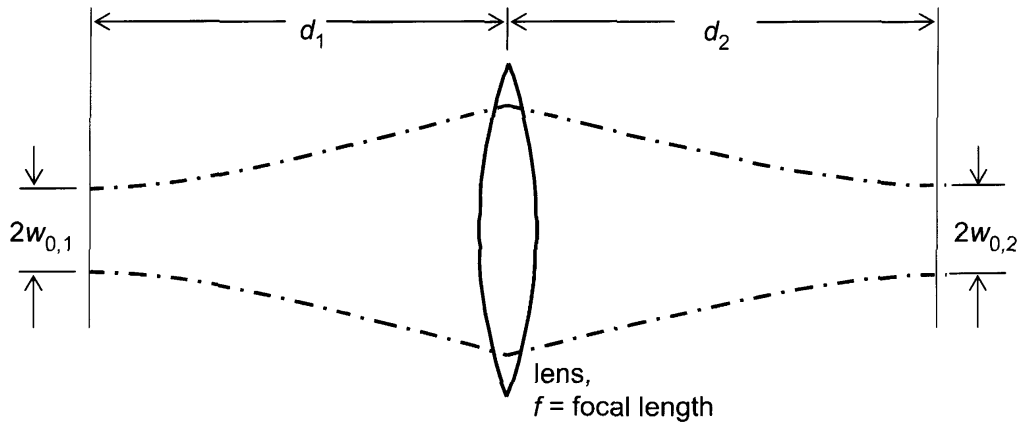


Figure A2.1.9. Parameters for mode matching with one lens.

A2.1.5.1 One lens approach

The standard approach is to consider the ABCD matrix that results when starting with a Gaussian beam waist of spot size $w_{0,1}$, propagating over a distance d_1 to a lens of focal length f, and finally propagating an additional distance d_2 to the new beam waist of spot size $w_{0,2}$ formed by the lens (figure A2.1.9). The ABCD law (A2.1.22) is applied to the complex beam parameter (A2.1.10). The real and imaginary parts of the resulting equation are separated to yield [11]

$$(d_1 - f)b_2 = (d_2 - f)b_1 \tag{A2.1.57}$$

and

$$(d_1 - f)(d_2 - f) = f^2 - b_1 b_2/4. \tag{A2.1.58}$$

Here $b_1 = 2\pi w_{0,1}^2/\lambda$ and $b_2 = 2\pi w_{0,2}^2/\lambda$ are the confocal parameters of the beam of wavelength λ before and after the lens. The quantity $b_1 b_2/4$ is sometimes labelled f_0^2. It is necessary that $f^2 \geq b_1 b_2/4 = f_0^2$ for there to exist distances d_1 and d_2 that will yield a confocal parameter b_2 from an initial beam with confocal parameter b_1; that is, the absolute value of the lens' focal length must be longer than a minimum value. In this case, when $w_{0,1}$, $w_{0,2}$ and f are specified, the distance from the first waist to the lens and the distance from the lens to the second waist are given by

$$d_1 = f \pm (w_{0,1}/w_{0,2})\sqrt{f^2 - b_1 b_2/4} \tag{A2.1.59}$$

and

$$d_2 = f \pm (w_{0,2}/w_{0,1})\sqrt{f^2 - b_1 b_2/4}. \tag{A2.1.60}$$

Here, either + or − signs should be used in both equations (A2.1.59) and (A2.1.60).

Another set of useful equations comes from combining the last two equations. What is the focal length of the required lens when the beam waists $w_{0,1}$ and $w_{0,2}$ and their separation $d = d_1 + d_2$ are specified? This will aid in the choice of a lens from available focal lengths to provide approximately the desired waist separation. The resulting equation is quadratic in f:

$$\{4 - (w_{0,1}^2 + w_{0,2}^2)^2/w_{0,1}^2 w_{0,2}^2\}f^2 - 4df + (w_{0,1}^2 + w_{0,2}^2)^2/4 + d^2 = 0 \tag{A2.1.61}$$

which has either two or no real solutions. Even with real solutions, it is necessary to check that the lens positions are physically realizable, e.g. not beyond the target. The position of a lens of focal length f determined in equation (A2.1.61) is

$$d_1 = (w_{0,1}^2(d - f) + w_{0,2}^2 f)/(w_{0,1}^2 + w_{0,2}^2). \tag{A2.1.62}$$

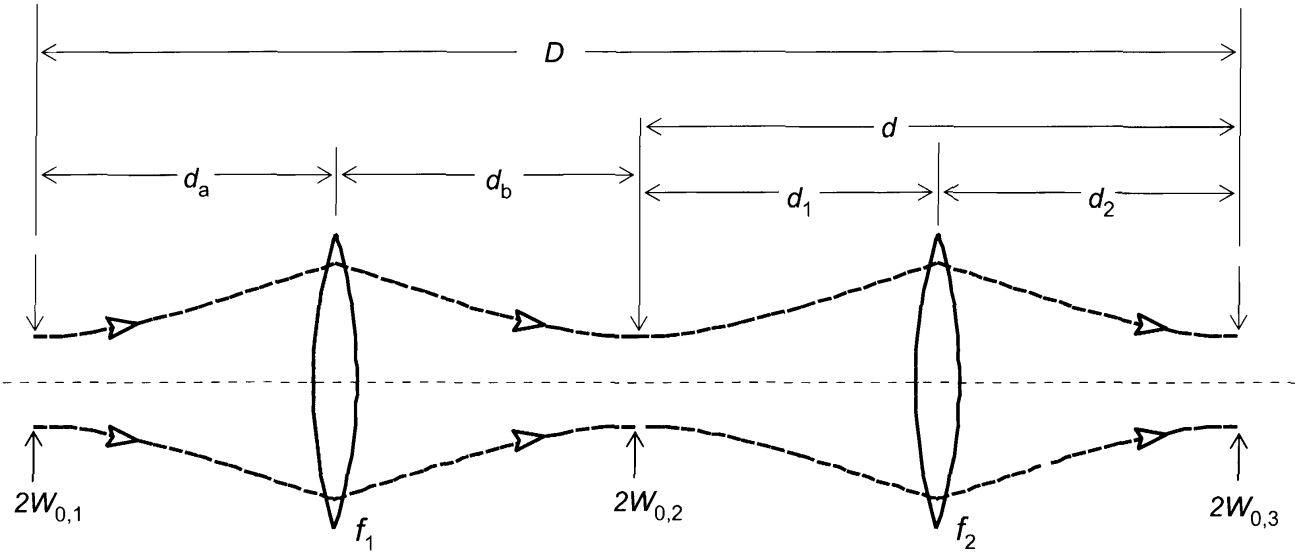


Figure A2.1.10. Mode matching with two lenses may be necessary when there is a fixed distance D between the initial and final beam waists.

A2.1.5.2 Two-lens mode matching

One way to approach calculations of two-lens mode matching is to step through a series of lens positions until the desired mode matching is found. Each iteration involves a different placement of the first lens, for which the position of the waist and the confocal parameter of the beam formed by the first lens of focal length f_1 are given by

$$d_b = f_1\{b_1^2 + 4(d_a - f_1)d_a\}/\{b_1^2 + 4(d_a - f_1)^2\} \tag{A2.1.63}$$

and

$$b_2 = b_1(d_b - f_1)/(d_a - f_1). \tag{A2.1.64}$$

This leaves a known distance d from the position of the waist formed by the first lens to the desired position of the final waist. The second lens is then numerically placed at a position that produces a third beam waist at the desired position. The numerical analysis is stepped through a range of positions for the first lens. The parameters are illustrated in figure A2.1.10. A cubic equation is obtained for the position of the second lens when the distance between beam waists $d(= d_1 + d_2)$, $w_{0,2}$, the size of the intermediate beam waist, and f_2, the focal length of the second lens, are specified:

$$d_2^3 - (f_2 + d)d_2^2 + (2f_2d + b_2^2/4)d_2 - (d - f_2)b_2^2/4 - f_2^2d = 0 \tag{A2.1.65}$$

There are either one or three real roots to equation (A2.1.65) and the confocal parameter of each of the resulting beams is given by

$$b_3 = b_2(f_2 - d_2)/(f_2 - d_1). \tag{A2.1.66}$$

Single-lens mode matching is simpler than two-lens mode matching. In cases where the distance between the initial and final beam waists can be adjusted, a single lens will work well. However, if that distance is fixed, two-lens mode matching may be required. It is possible to add two more adjustable parameters by tilting the lenses, but any astigmatism or ellipticity in the beam might best be removed before mode matching.

A2.1.6 Plane parallel resonators

The analysis of ideal stable resonators provides a useful background for understanding practical laser resonators including plane-parallel-mirror resonators and unstable resonators. We begin with plane-parallel

resonators in this section and continue with unstable resonators in the next section. Plane-parallel resonators are characterized by their Fresnel number:

$$N_F = a^2/(L\lambda) \tag{A2.1.67}$$

where a is the radius of the limiting aperture of the resonator, L is the propagation distance from one encounter of the limiting aperture to the next, and λ is the wavelength of the resonated light. The limiting aperture could be a laser rod, a cavity mirror or an actual aperture placed in the resonator. Resonators with the same Fresnel number will have equivalent diffraction properties. A circular aperture is used here and the resonator has circular symmetry. The circular symmetry is lost, however, when the Fourier components of the initial amplitude distribution are given a random phase.

The diffraction for various apertures and beams provides some insight into the significance of the Fresnel number. In many cases there is a diffraction spread of a collimated beam that is approximately λ/a rad. For example, the diffraction half-angle of a Gaussian beam is $\lambda/(\pi w_0)$. The centre to first minimum angle of the Airy diffraction pattern of a uniformly illuminated circular aperture is $1.22\lambda/(2a)$ and the full width at half maximum of the diffraction pattern from a slit of width $2a$ illuminated by a plane wave is $1.39\lambda/(\pi a)$. With a Fresnel number of $N_F = 1$, a nearly collimated beam will spread by diffraction in a resonator length L to slightly overfill the aperture for significant loss on the next encounter with the aperture. Numerical calculations show that a resonator with a Fresnel number of 1 will have about 18% loss for each cavity transit from aperture to aperture, whereas the loss will be approximately 0.88% for a Fresnel number of 10. The Fresnel number is a useful and general concept. For example, the N_F of a stable resonator is approximately the number of Hermite–Gaussian modes that the resonator will support.

Iterative computer techniques are commonly used to determine the cavity modes of plane-parallel resonators. In the original paper by Fox and Li [1], iterative solutions to Huygen's integral in cylindrical symmetry starting with a uniform plane wave were used. It is more common now to use Hankel transform techniques for cylindrical symmetry and Fourier transform techniques for rectangular symmetry [15, 16]. A discrete Fourier transform pair such as equations (A2.1.50) and (A2.1.51) is used. The spatial amplitude and phase distribution transmitted through the aperture is transformed into a sum of plane waves propagating at regularly spaced directions with respect to the central direction of propagation. The propagation of the plane waves to the next encounter with the aperture is straightforward, each having a relative phase shift dependent on the direction of propagation. The plane waves are then summed to synthesize the spatial distribution, which is modified by transmission through the aperture. Routines for calculating the Fourier transformations are available [18, 19].

Calculated diffraction losses as a function of Fresnel number for plane-parallel resonators of circular symmetry are shown in figure A2.1.11. Starting with a uniform distribution and propagating through 300 cavity transits yielded the plotted values. The power in the beam was re-normalized after each cavity transit. The plotted line shows transmission on the 300th transit. The process is more slowly converging for Fresnel numbers greater than ten, typically requiring more than 300 transits to converge. Virtual source techniques [20] are useful for the analysis of resonators with large Fresnel numbers.

A modification of the Fox and Li iterative cavity transit technique was used for the calculations illustrated in figures A2.1.12–A2.1.14. A two-dimensional FFT technique was used to calculate beam propagation. In this case, the initial spatial frequency or angular components all had equal power but the phases were random. A Fresnel number of 10 was used for the calculation. An initial intensity distribution transmitted through an aperture is shown in figure A2.1.12(a). This is intended to simulate a laser oscillation that is growing out of zero-point quantum fluctuations. The initial distribution is limited by the number of spatial sampling points used in the calculation and the distribution changes for a new calculation with different random phases. After each transit, the intensity was again re-normalized. After 10 cavity transits (figure A2.1.12(b)) the high spatial frequencies are greatly attenuated but the beam is still strongly structured with an M^2 of 9.5. Spatial filtering continues with fewer high frequency components and an M^2 of 5.3 after 20 transits (figure A2.1.12(c)).

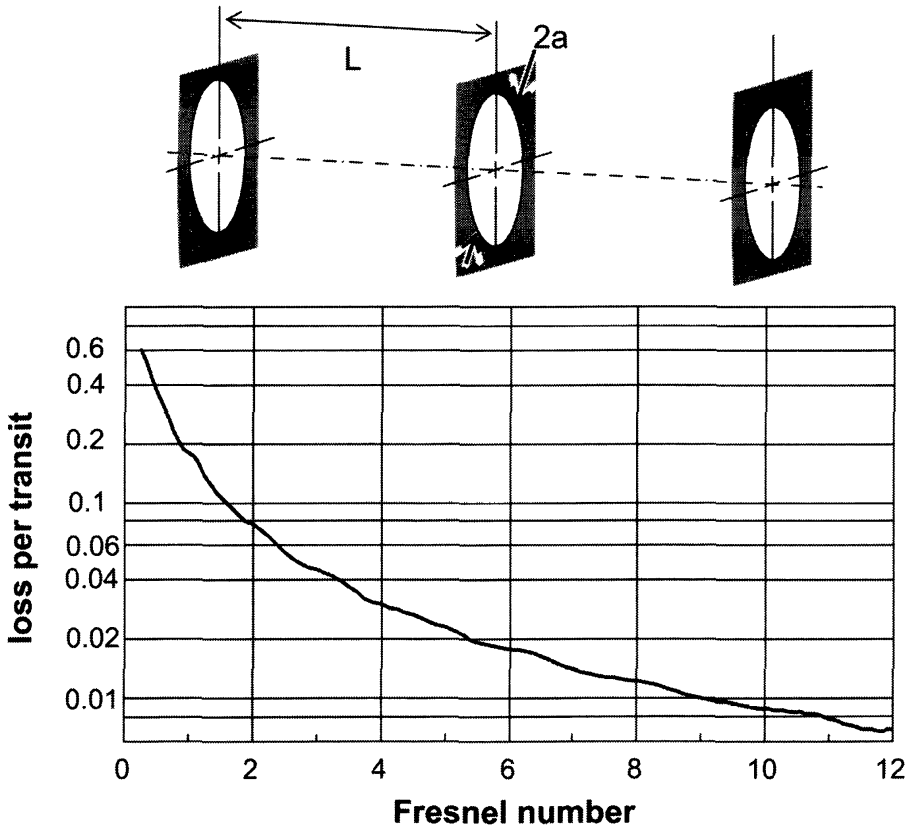


Figure A2.1.11. Calculated loss per transit in a plane-parallel resonator with circular mirrors as a function of Fresnel number $N_F = a^2/(L\lambda)$.

Many high-gain Q-switched lasers reach their peak output power in 10 or 20 cavity transits and this type of distribution could be present in an instantaneous sampling of the output. Averaging over the total Q-switched pulse could give the appearance of a uniform intensity beam with an M^2 between 5 and 10. The loss per transit behaviour and the evolution of M^2 during the iterative calculation is illustrated in figure A2.1.13.

After several hundred cavity transits, the intensity and phase distribution become constant, as shown in figure A2.1.14. At this point the beam has an M^2 of 1.5 and the intensity loss per transit is 0.88%. In practice, such a distribution would be difficult to obtain, even with hundreds of cavity transits, due to the sensitivity of the plane-parallel resonator to misalignment. Injection seeding of a field distribution close to that of the fundamental mode of the resonator would quickly produce a dominant oscillation of that mode in the resonator. The use of injection seeding, however, is more commonly used to achieve single-frequency oscillation in resonators with large gain. Selection of the fundamental transverse mode is easily achieved in unstable resonators.

A2.1.7 Unstable resonators

Unstable resonators offer advantages of good energy extraction efficiency from a large-volume active laser material and reasonably good spatial beam quality in the laser output. Typically, unstable resonators have large loss or output coupling that must be offset by high laser gain. The high-gain requirement usually restricts operation to the pulsed mode because of the difficulty in maintaining high cw gain. Frequency control and narrow spectral bandwidth operation are more difficult with high laser gain. Unstable resonators have transverse modes that reproduce from one cavity transit to the next. These modes, however, are not orthogonal or even nearly orthogonal as in the case of stable resonators. The non-orthogonality of the modes

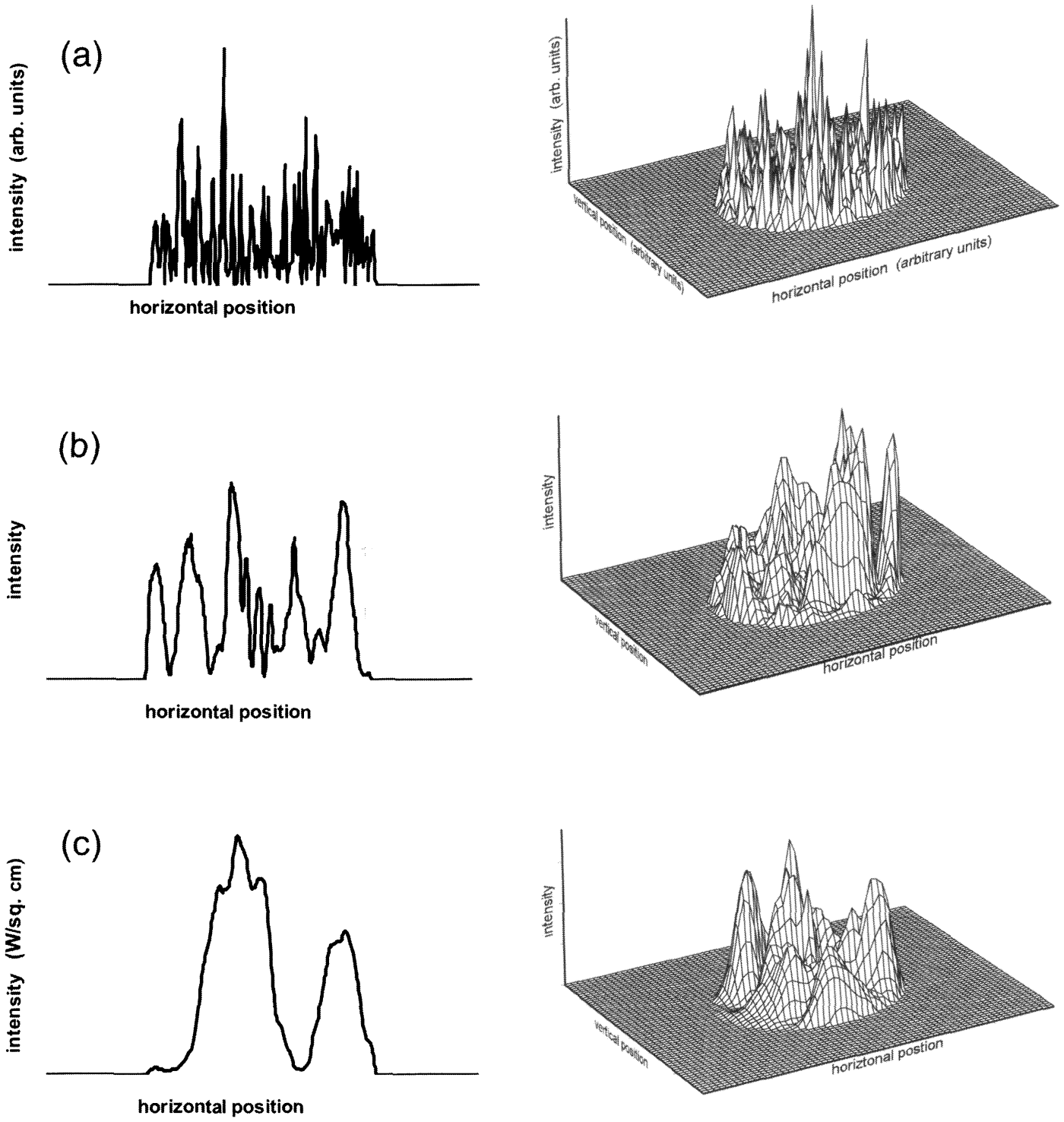


Figure A2.1.12. Intensity on a trace through the centre of the transverse distribution of a beam propagating in a plane-parallel circular-mirror resonator of Fresnel number 10. The calculation starts with randomly phased Fourier components (*a*) and is shown after 10 transits (*b*) and 20 transits (*c*).

leads to some perhaps surprising effects, such as a mode is most efficiently seeded with a conjugate beam; that is, the backward propagating beam, converging where the output beam was diverging, will most effectively seed the oscillation. Transverse mode discrimination is usually large and oscillations usually resolve to a single transverse mode in a small number of cavity transits.

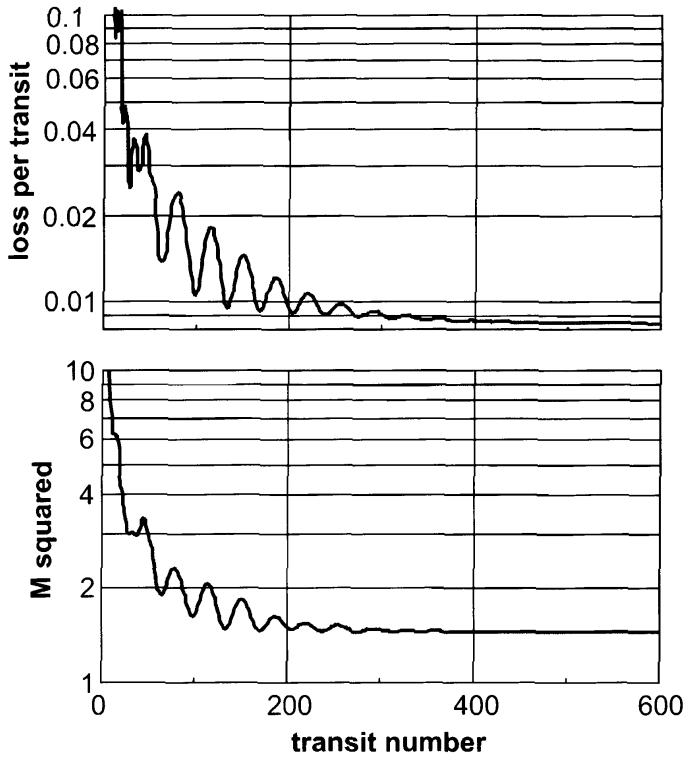


Figure A2.1.13. Both loss per transit and M^2 become smaller on approaching limiting values in the calculation of beam development. The results for a calculation starting with randomly phased Fourier components in a plane-parallel resonator of $N_F = 10$ are shown.

A2.1.7.1 Hard-edged apertures

Unstable resonators can be classified as either hard- or soft-edged. In a hard-edged resonator, output coupling is typically by expansion of the resonated beam beyond the edge of some limiting aperture. This could be a hole cut in the centre of a mirror used for output coupling, a high-reflectivity mirror smaller than the expanded beam or a high-reflectivity spot on a substrate used to reflect and output couple the resonated beam. Soft-edged resonators can be created with apodized apertures, variable reflectivity mirrors (VRMs) and gain guiding such as is common in laser-pumped lasers. Unstable resonators are also classified as either negative or positive branch according to the sign of the product $g_1 g_2 = (1 - d/R_A)(1 - d/R_B)$. Two-mirror negative-branch unstable resonators have a beam focus inside the resonator, whereas a two-mirror positive-branch resonator can have either none or two. The absence of an intracavity beam focus is an advantage for high-power laser oscillations. Confocal unstable resonators are useful because the resonated beam is collimated in the output part of the resonator round trip [21]. This can be useful in producing a collimated output or region of collimated propagation inside the resonator.

Unstable resonators are usually analyzed on two levels. Geometrical analysis of unstable resonators provides a fair first approximation. Some soft-edge unstable resonators can be analysed using Gaussian optics and an extension of the ABCD matrix techniques to include complex matrix elements. Numerical beam propagation calculations, however, are usually required for a detailed understanding of the diffraction effects. Fourier transform techniques are commonly used to perform these numerical calculations.

When equation (A2.1.24), $q = (Aq + B)/(Cq + D)$, is solved for $1/q$ in the case of an unstable resonator, the result is real:

$$\frac{1}{q} = \frac{D - A}{2B} \pm \frac{1}{B}\sqrt{\left(\frac{A + D}{2}\right)^2 - 1} = \frac{1}{R}. \tag{A2.1.68}$$

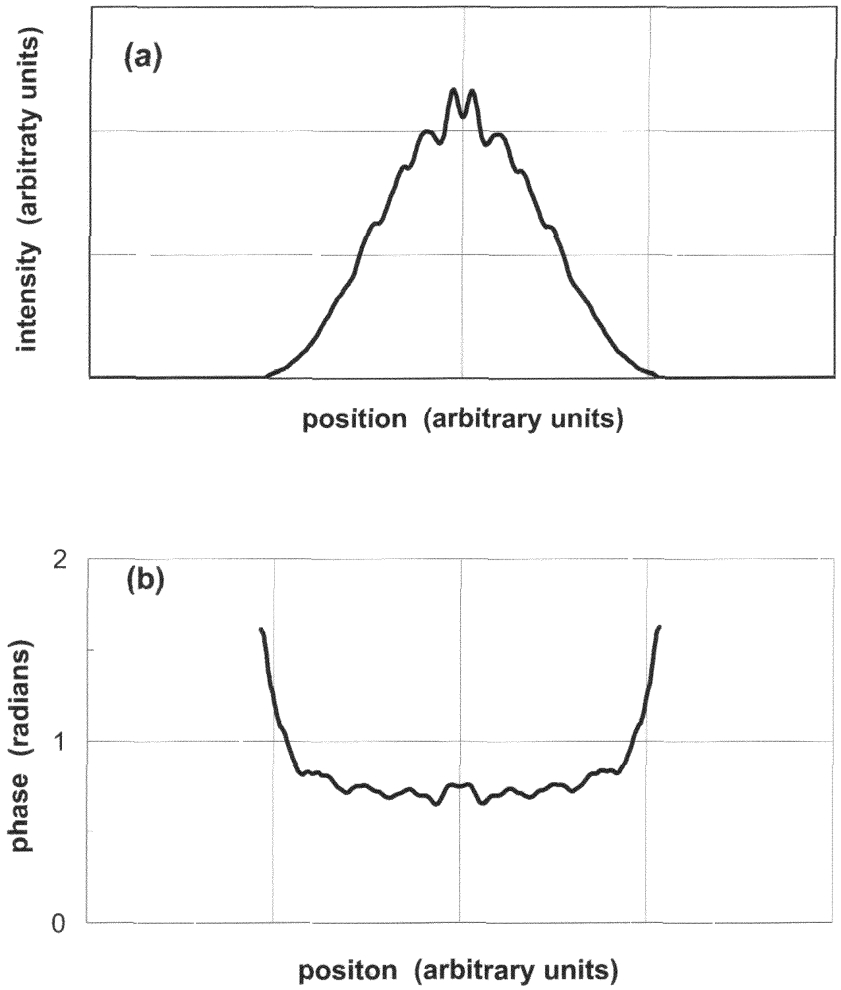


Figure A2.1.14. Intensity and phase on a trace through the centre of the transverse distribution. This is the calculated distribution of the same Fresnel number 10 resonator used in figures A2.1.12 and A2.1.13 after several hundred transits.

This result gives a non-physical interpretation of spherical waves since $1/w^2 = 0$. When the accurate solutions with limiting apertures and diffraction are considered, it is found that the two solutions for R represent a stable solution of a divergent beam and an unstable solution for a convergent beam. The convergent beam becomes smaller only for a few cavity transits until diffraction begins to dominate and it quickly changes to become a divergent beam.

Geometrical magnification of the unstable resonator is also obtained from the ABCD matrices. For positive-branch unstable resonators, $m = (A + B)/2 > 1$; and the values for magnification are:

$$M = m + \sqrt{m^2 - 1} \tag{A2.1.69a}$$

for the expanding beam and

$$1/M = m - \sqrt{m^2 - 1} \tag{A2.1.69b}$$

for the convergent beam. For negative-branch unstable resonators, $m = (A + B)/2 < -1$; and the values for magnification are:

$$M = m - \sqrt{m^2 - 1} < -1 \tag{A2.1.70a}$$

and

$$-1 < 1/M = m + \sqrt{m^2 - 1} < 0. \tag{A2.1.70b}$$

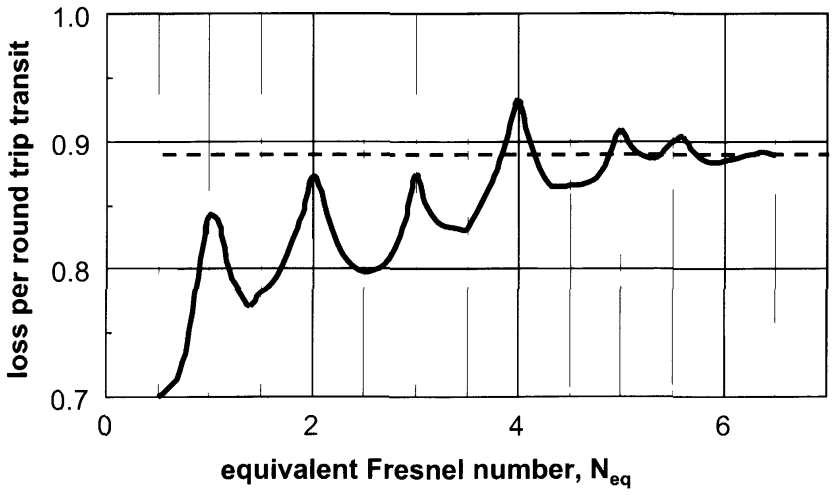


Figure A2.1.15. Loss as a function of equivalent Fresnel number for a positive-branch, confocal, unstable resonator with magnification $M = 3$. The broken line is the geometrical value and the full curve is obtained with a diffraction calculation.

A parameter important for characterization of hard-edged unstable resonators is the equivalent Fresnel number:

$$N_{\mathrm{eq}} = \left(\frac{M^2 - 1}{MB}\right)\frac{a^2}{2\lambda}. \tag{A2.1.71}$$

Here M is the magnification of the resonator as given earlier; B is the component of the ABCD matrix of the resonator; a is the radius of the limiting circular aperture; and λ is the wavelength of the light circulating in the resonator.

We now restrict the discussion to an illustrative example of a positive-branch, confocal, unstable resonator as shown in figure A2.1.1(b). This resonator is formed by a concave mirror of radius of curvature $R_{\mathrm{A}} > 0$ and a convex mirror of radius of curvature $R_{\mathrm{B}} < 0$ separated by distance d. The output mirror has a central highly reflecting spot of radius a and is transmitting for a radius greater than a. Such an output mirror could be a small suspended mirror, as shown, or a reflecting spot deposited on a meniscus substrate to preserve the collimation of the output beam. The confocal property of the resonator specifies that $R_{\mathrm{A}} = 2d - R_{\mathrm{B}}$. In the special case of a confocal unstable resonator, the magnification is given by $M = -R_{\mathrm{A}}/R_{\mathrm{B}}$ and the equivalent Fresnel number is $N_{\mathrm{eq}} = a^2/(\lambda|R_{\mathrm{B}}|)$. The magnification is a positive value because $R_{\mathrm{B}} < 0$.

A calculation of loss per round-trip transit as a function of N_{eq} for a positive-branch confocal resonator with magnification $M = 3$ is shown in figure A2.1.15. At larger values of N_{eq} the loss converges on the geometrical value of $(1 - 1/M^2) = 0.89$. The broken line in figure A2.1.15 represents this value. At smaller values, loss is minimum at half-integer values of N_{eq}, a feature common to hard-edged unstable resonators. This feature is due to the transverse-mode properties of the resonator. At half-integer values of N_{eq}, the loss difference between modes is large, and the oscillation of a single transverse mode dominates after a few cavity transits. At integer values of N_{eq}, there are two transverse modes of nearly equal loss and many cavity transits are required to obtain a reproducible loss and transverse field distribution.

The initial distribution and distribution after eight transits for a resonator with magnification $M = 3$ and equivalent Fresnel number $N_{\mathrm{eq}} = 5.5$ (figure A2.1.16) illustrates that the lowest loss mode can be resolved quickly both in calculations and in the actual build-up of oscillation in an unstable resonator laser. The calculation used to generate figure A2.1.16 used random phasing of the initial Fourier components all of equal power. The total power was normalized on each cavity transit, preserving the phase and relative intensity of the individual components in the method of the Fox and Li calculation. The lowest loss distribution of this resonator is numerically propagated outside the cavity in figure A2.1.17. In a distance equal to four

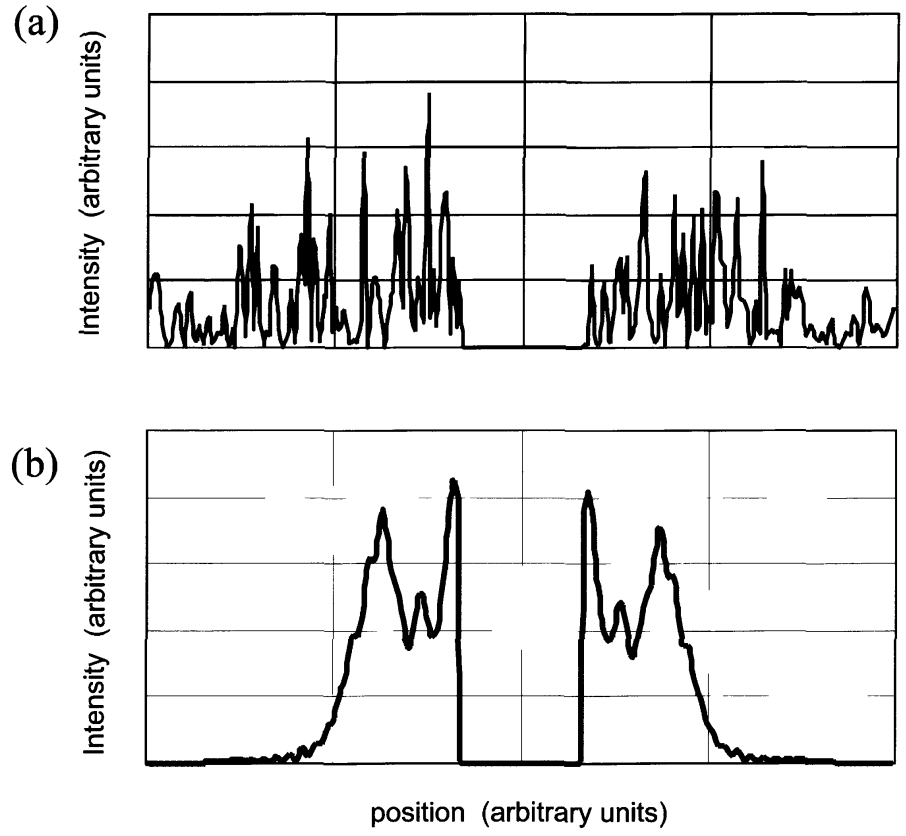


Figure A2.1.16. Calculated development of transverse intensity distribution in a positive-branch, confocal, unstable resonator with equivalent Fresnel number, $N_{eq} = 5.5$ and magnification $M = 3$: (*a*) initial distribution; (*b*) After eight cavity transits the distribution is close to the final form.

times the cavity length, a strong spot of Arago develops in the centre of the beam. The development of the spot of Arago, with a peak intensity many times that of the other portions of the beam, is a disadvantage of hard-edged unstable resonators.

A short description of the numerical techniques used in the calculations is appropriate (see section A8.1). A direct application of Fourier beam propagation methods to the unstable resonator would require prohibitively large arrays of sampling points to handle the diverging portions of beam propagation in the cavity. This problem is avoided with a simple transformation that reduces the problem of calculating the propagation of the collimated beam. The technique is illustrated with propagation in a Galilean telescope equivalent to reflections from the convex mirror followed by reflection from the concave mirror in our resonator. The beam, just before the convex mirror, is first magnified by setting

$$E_{mag}(x, y) = (1/M)E_{in}(x/M, y/M) \tag{A2.1.72}$$

where E_{in} is the incident electric field, E_{mag} is the magnified field, $M = -R_A/R_B$ is the magnification and the factor $1/M$ is needed for conservation of energy. Next, the field is propagated over an effective distance of the magnification times the mirror separation;

$$d_{eff} = M \times d. \tag{A2.1.73}$$

Finally, it is necessary to retrieve the spacing of the original array sampling points by some method such as interpolation using first, second and cross derivatives. Siegman [3] provides a justification of this transformation based on Fermat's principle. The technique is more general than that described here and can be applied to an open optical system described by an ABCD matrix.

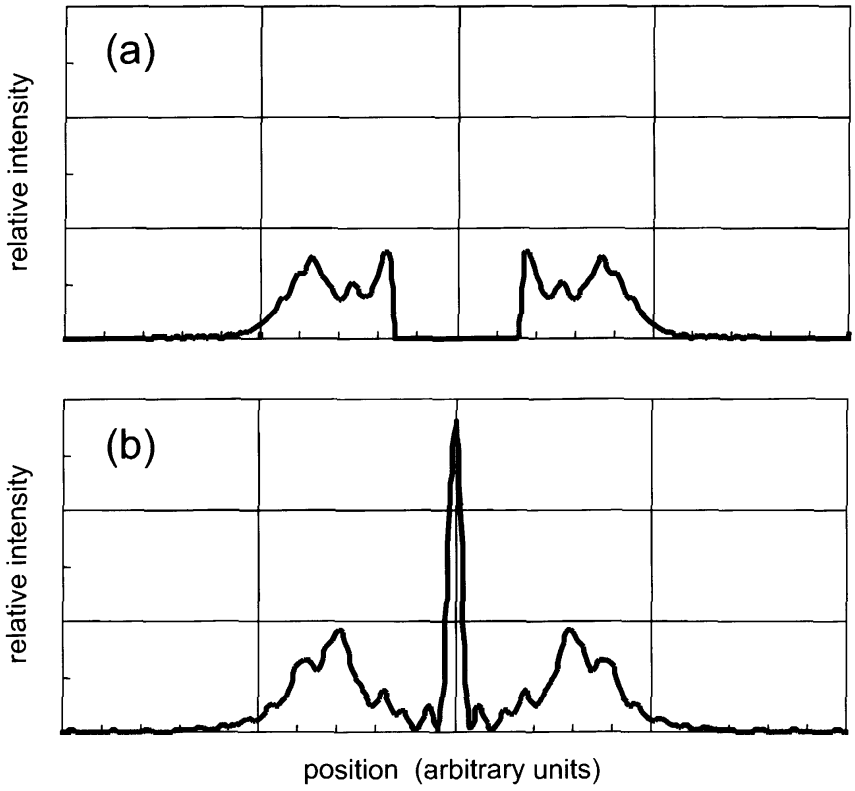


Figure A2.1.17. Development of the spot of Arago for the distribution described in figure A2.1.16: (*a*) at resonator output; (*b*) four cavity lengths from the output mirror.

The value of the calculated beam quality parameter M^2 is dependent on the sharpness of the edge of the aperture and on the number of sampling points used in the calculation. For example, the distributions shown in figures A2.1.16 and A2.1.17 were obtained using a rather sharp edge of intensity reflectivity on the output mirror given by $R_M = \exp\{-(r^2/a^2)^{64}\}$, and the value $M^2 = 7.2$ was obtained for the $N_{eq} = 5.5$ resonator. When the edge is softened to $R = \exp\{-(r^2/a^2)^{32}\}$, $M^2 = 5.6$ results. In the limit of an infinitely sharp edge the paraxial approximation will fail. The number of sampling points also limits the resolution of the sharpness of the edge. The improvement of beam quality with softening of the reflector or aperture edge leads to variable reflectivity mirrors.

A2.1.7.2 Soft-edged apertures

An example of a variable reflectivity mirror with an intensity reflectivity of $R = 0.34\exp(-r^2/a^2)$ is described here. The magnification of the resonator is reduced to 1.75 to yield a loss per transit of 0.89, the same as the resonator with the sharp-edged reflector discussed earlier. The reflectivity profiles of the hard-edged reflectors and the variable reflector used in these examples are shown in figure A2.1.18. A positive-branch confocal resonator with an equivalent Fresnel number of $N_{eq} = 2.3$ is used with the variable reflectivity mirror and a magnification of $M = 1.75$. There is nothing unique about these values. They were chosen for comparison because the cavity transit losses were the same as the hard-edged aperture example and the output distribution was reasonable. Again, the phases of the Fourier components are initially random and the initial amplitudes are equal; the total power is renormalized after each cavity transit. After only eight resonator transits, the calculated initensity distribution is near the final form with a beam quality of $M^2 = 1.4$ (figure A2.1.19). Good beam quality can develop in relatively few cavity transits in such a resonator.

These examples are for ideal conditions. An actual laser resonator would have a gain saturation that

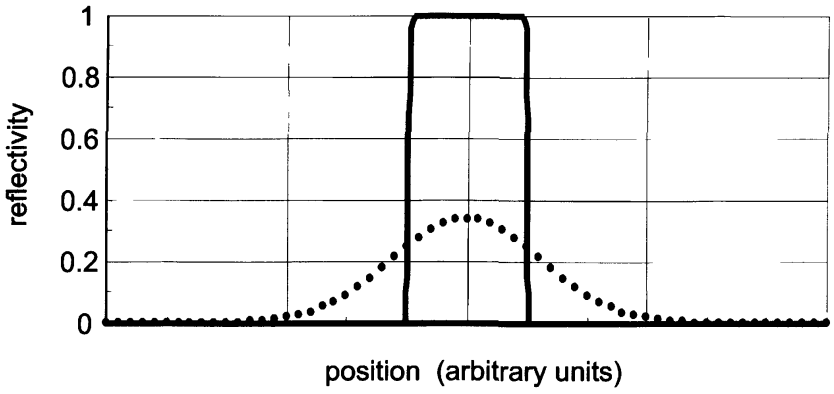


Figure A2.1.18. Reflectivity distributions used for hard-aperture (full line) and soft aperture (dotted curve) calculations.

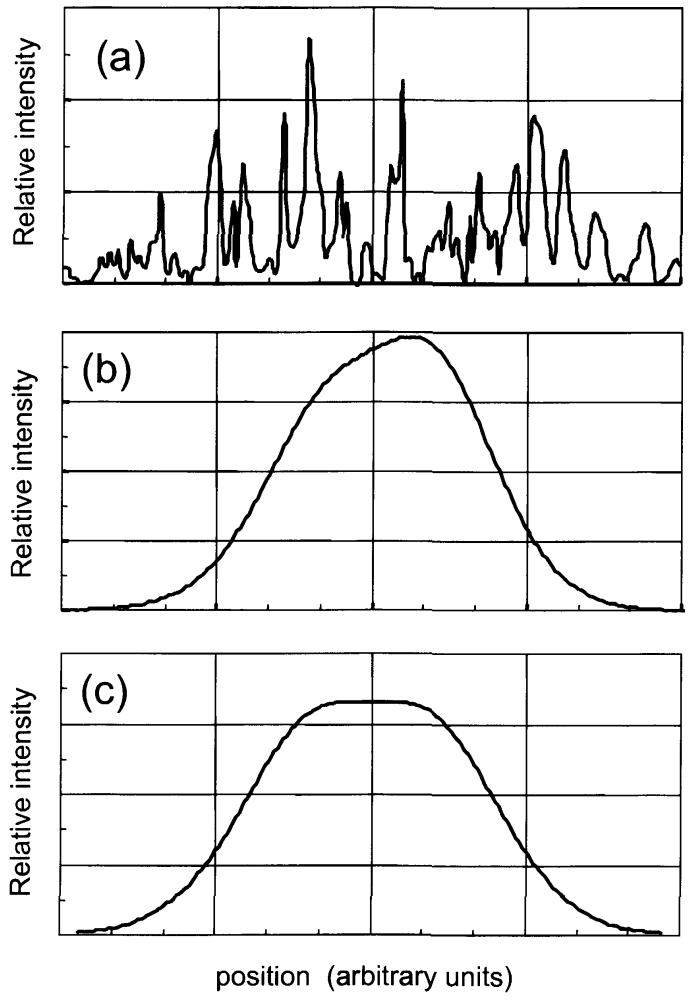


Figure A2.1.19. Transverse beam development in unstable resonator with variable reflectivity mirror shown in figure A2.1.18 (*a*) after first resonator transit; (*b*) after eight transits and (*c*) after 20 transits.

would make the distribution more 'top-hat' like and less like a Gaussian distribution. Laser-pumped laser resonators can also produce good beam quality. Some laser-pumped lasers have a transverse gain distribution that is nearly Gaussian.

A2.1.8 Distortion effects

There are a number of practical problems in actual lasers that lead to beam distortion. Heat deposition in the laser gain medium resulting from pumping can cause thermal lensing and thermally induced stress birefringence. Master oscillator power amplifier techniques may be useful. A high-quality beam is generated in a low-power oscillator where beam quality is more easily controlled. Power amplification is then obtained with a single or double pass through a high-power laser amplifier. This avoids multiple passes in a high-power resonator where greater beam distortion could accumulate. A technique that is finding wide use in these systems is phase-conjugate reflection (see section A3.2.3). After the first pass through a laser amplifier, the beam acquires sufficient intensity that relatively efficient phase-conjugate reflection in a Brillouin cell is possible. The reflection of the conjugate mirror traverses the path through the laser amplifier in reverse, cancelling the distortion acquired on the first transit. These topics are beyond the scope of this discussion and are discussed in later chapters. Next we turn to an overview of axial modes and temporal properties of laser resonators.

A2.1.9 Axial modes

A2.1.9.1 Stable-resonator axial-mode spectral separation

Up to this point we have only discussed the phase difference between transverse resonator modes and plane waves propagating in the same direction. For a discussion of axial modes, it is necessary to add the additional constraint that the resonator mode must reproduce itself both in amplitude and phase except for uniform amplification or attenuation. This means the wavelength of a resonator mode must satisfy the condition that the optical length of a round-trip cavity transit is an integer multiple of that wavelength with adjustment for Gouy and other possible phase shifts.

Often it is the case that the optical length of a resonator is not known precisely on the scale of a fraction of a wavelength of light. This is commonly the case for resonators used as interferometers to determine the relative spectral distribution of an optical beam. It may be sufficient to consider only the spectral spacing of modes. For example, the free spectral range of fundamental modes of an open, stable resonator is given to a high degree of accuracy by

$$\Delta\tilde{\nu}_{\mathrm{FSR}} = \Delta\lambda_{\mathrm{FSR}}/\lambda^2 = 1/\lambda_{n+1} - 1/\lambda_n \approx 1/(2d). \tag{A2.1.74}$$

Here λ_{n+1} and λ_n are the wavelengths of adjacent modes. The free spectral range is given as a wavenumber separation $\Delta\tilde{\nu}_{\mathrm{FSR}}$ and as a wavelength separation $\Delta\lambda_{\mathrm{FSR}}$ in equation (A2.1.74). The mirror separation of the resonator is d, and λ is the central or average wavelength. A symmetric confocal interferometer is commonly used in the spectral analysis of laser beams. In typical use there are four reflections before the beam path inside the resonator is closed (figure A2.1.20) and the free spectral range is $\Delta\tilde{\nu}_{\mathrm{FSR}} = 1/(4d)$. Usually the accuracy of these expressions is limited by the precision to which the mirror separation is known.

The optical length is the integral of the refractive index over the round-trip transit path followed by the centre of the transverse field distribution:

$$\text{optical length} = \oint_{\text{round trip}} n(\lambda, z)\,\mathrm{d}z. \tag{A2.1.75}$$

The condition that the phase of a resonator mode change by an integer multiple on a cavity round trip is

$$\frac{\text{optical length}}{\lambda} - \frac{\Phi}{2\pi} = \text{integer} \tag{A2.1.76}$$

where Φ represents the sum of the Gouy phase shifts on the complete cavity transit. The axial or longitudinal cavity modes associated with a transverse mode will be nearly equally spaced in wavenumber $\tilde{\nu} = 1/\lambda$

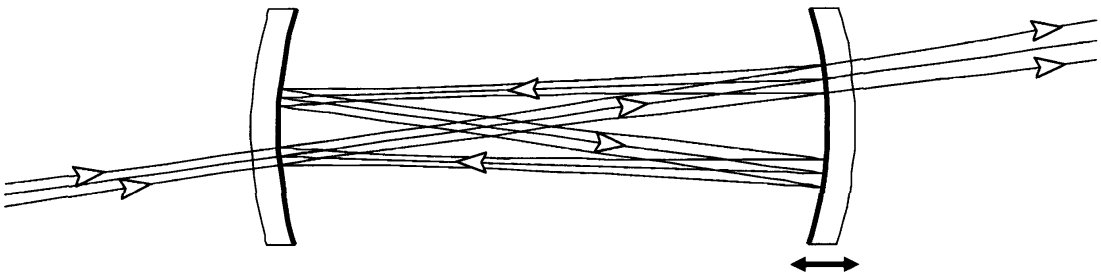


Figure A2.1.20. Schematic representation of the light path in a confocal interferometer.

but the spacing will not be exactly equal because of the dispersion of the refractive index. Transverse modes of different order will also usually be spectrally positioned between the fundamental transverse modes. Secondary changes such as additional wavelength-dependent phase shifts in optical components and wavelength-dependent path differences must be considered in critical applications such as the propagation of femtosecond-duration pulses.

The spectral width of a resonator mode depends on several factors. Losses typically determine the spectral width in a resonator used passively as a multipass interferometer. It is useful to consider an ideal case of a two-mirror stable resonator with identical mirror reflectivity R_m for lossless mirrors that have transmissions and reflections that sum to one: $T_m + R_m = 1$. The transmission of a monochromatic Gaussian beam mode matched to the resonator is

$$\frac{P_t}{P_i} = \frac{1}{1 + 4R_m \sin^2(\phi/2)/(1 - R_m)^2}. \tag{A2.1.77}$$

Here P_i is the incident power in the beam, P_t the transmitted power and ϕ the phase shift encountered on one round-trip transit of the resonator. When the reflectivity is high, the shape of the interferometer transmission peaks can be obtained by using the approximation $\sin(\Delta\phi/2) \approx \Delta\phi/2$ where $\Delta\phi$ is the small difference between ϕ and a multiple of 2π, and ϕ is nearly equal to a multiple of 2π. In this approximation, the full width at half maximum of the resonance is given by

$$\phi_{FWHM} = 2(1 - R_m)/\sqrt{R_m}. \tag{A2.1.78}$$

The finesse of the resonator is the ratio of the free spectral range divided by the resonance width:

$$\text{finesse} = 2\pi/\phi_{FWHM} = \pi\sqrt{R_m}/(1 - R_m). \tag{A2.1.79}$$

For less than ideal cavity mirrors, the finesse can be stated as a function of total resonator loss

$$\text{finesse} \approx \pi/(1 - \text{loss}/2). \tag{A2.1.80}$$

Such resonators, when illuminated by a single-frequency laser, can provide accurate measurements of low levels of loss in components inserted in the resonator.

A2.1.10 Frequency selection and frequency stability

The spectral properties of laser resonators are determined by the laser gain medium and the temporal character of the laser as well as by the resonator. The gain bandwidth of the active laser medium can be many times wider than the free spectral range of the resonator and support many longitudinal modes. As the laser oscillations build up, there is a frequency narrowing or frequency selection similar to transverse mode selection. High-gain pulsed lasers, however, typically have insufficient time to resolve single-mode operation. Effects such as spectral and spatial 'hole burning' can also limit frequency selection.

In mode-locked lasers many modes are locked in phase to synthesize a single short pulse (see section A3.6.3). The time–bandwidth limit specifies the minimum pulsewidth attainable for a given bandwidth or the minimum bandwidth required to support a given pulse duration. The minimum time–bandwidth product is obtained for pulses that have a Gaussian shape in time with no further amplitude or phase modulation. Frequently, the time–bandwidth product is given in terms of pulsewidth in seconds, Δt_{FWHM}, and spectral width, $\Delta \upsilon_{\mathrm{FWHM}}$, in hertz:

$$(\Delta t_{\mathrm{FWHM}} \Delta \upsilon_{\mathrm{FWHM}})_{\min} = 0.44. \qquad \text{(A2.1.81)}$$

In terms of wavenumber, it is $(\Delta t_{\mathrm{FWHM}} \Delta \tilde{\nu}_{\mathrm{FWHM}})_{\min} = 14.72$ ps cm^{-1} and in terms of wavelength,

$$(\Delta t_{\mathrm{FWHM}} \Delta \lambda_{\mathrm{FWHM}} / \lambda^2)_{\min} = 1.472 \times 10^{-6} \text{ ps nm}^{-1}$$

where λ is the central wavelength. The time–bandwidth product can also be stated as the uncertainty principle $(\Delta t_\sigma \Delta E_\sigma)_{\min} = \hbar/2$ given in terms of the variance.

The properties of the laser gain have an effect on frequency selection. When the gain is inhomogeneously broadened, gain in a narrow spectral region can be depleted without depleting the gain in neighbouring regions (see section A1.9.2). An example is a gas discharge where gain broadening is due to Doppler shifting in a distribution of velocities. Multiple modes can oscillate, each drawing gain from a different velocity population. The familiar helium–neon laser exhibits this type of behaviour with typically two or three modes oscillating simultaneously. The depletion of gain in a narrow spectral region of an inhomogeneously broadened laser is called spectral hole burning. It is also possible to have spatial hole burning in a standing-wave laser resonator. The gain is not depleted at the nodes of the standing-wave laser oscillation and remains to provide gain for modes of a different frequency with different node locations. Spatial hole burning can occur with either inhomogeneous or homogeneous gain broadening. When the laser gain is homogeneously broadened, energy extraction in a narrow spectral region will uniformly decrease gain over the entire gain bandwidth. Gain broadening by lifetime-limiting collisions or thermal vibrations in solids are examples of homogeneous broadening.

Techniques for frequency selection include the addition of dispersive elements such as prisms and gratings in the resonator (see section A3.3). Etalons within the resonator are frequently used for gain narrowing as are multiple-element birefringent filters. Individual elements of birefringent filters are wavelength-dependent high-order waveplates providing a retardation between orthogonal polarizations. These elements are placed between polarizers or Brewster-angle surfaces to provide favoured transmission for wavelengths that have integer orders of retardation. The order or retardation is adjusted by rotating the plate: using plates of different thickness can provide a single region of high transmission in a wider spectral range. Etalons can be as simple as an uncoated plane-parallel plate of a transmitting material or a precise assembly of two closely spaced surfaces with multilayer-dielectric reflective coatings to provide higher finesse. It is more difficult to control multiple etalons and it is common to use only a single etalon. It is also common to use combinations of these techniques.

To obtain frequency-stable output from a laser, it is first necessary to make the laser as stable as possible before proceeding to active control techniques. Mechanical stability and the reduction of mechanical vibrations are a first essential step. Temperature stability must be considered next. At this point it is necessary to consider the stability of the pump source. This is true even for a cw laser pumped by another cw laser. Active control of laser frequency requires an external reference and feedback control of the resonator. This is usually done with piezoelectric control of the cavity length (see section A3.3.3). Frequency reference standards can be a stable external etalon for relative frequency stabilization or an atomic transition reference for absolute stabilization. Sub-Doppler spectroscopy techniques may be appropriate for the absolute frequency standard. Even greater absolute frequency accuracy is obtained by reference to cryogenically cooled atoms held in a trap or transiting through an atomic fountain (see chapters D6.1 and D6.2).

Stable single-frequency operation is significantly easier to attain in low-power lasers. Injection locking is a technique for transferring the frequency characteristics of the low-power laser to a higher-power laser. To

accomplish this, the resonances of the locking laser and locked laser must be actively controlled and locked together. One technique for accomplishing this is called Pound–Driever locking (see section C3.3.6). The seeding laser beam is modulated to produce fm sidebands along with the central frequency. The sidebands are outside the resonance of the higher-power laser and are reflected. The central-frequency portion of the beam is partially transmitted into the second laser and drives the oscillation of that laser. The phase of the combined reflected locking beam and transmitted oscillation of the second laser change slightly if the two resonances are not exactly matched in frequency. This produces an amplitude modulation in the combined beam with the fm sidebands. The phase and amplitude of this amplitude modulation provide the error signal used to control the cavity length of the second laser. Ring resonators are suited to this type of locking. A unidirectional oscillation is established in the second laser. The combined transmitted and reflected beams from the second laser are directed away at an angle from the incident locking beam. This greatly helps in the isolation of the laser generating the locking radiation.

A2.1.11 Temporal resonator characteristics

Mode locking many cavity modes requires some modulation technique that couples the modes in phase. Acousto-optic modulators are used for active mode locking of cw laser systems. An acoustic modulation driven at a frequency that matches the resonator mode spacing is established in a transparent material. This modulation produces sidebands on the modes, which couple to the adjacent modes, locking many modes in phase. Saturable absorbers are normally used with higher-peak-power pulsed mode-locked lasers. Saturable absorbers appropriate for mode locking must have short relaxation times much less than the round-trip cavity transit times. As the laser oscillation builds up with random fluctuations, the strongest fluctuation will begin to saturate the absorption. This fluctuation will build up more rapidly and come to dominate the laser oscillation. Gain will be depleted by the strong fluctuation reducing the amplitude of secondary fluctuations. The rapid saturation of the absorption on each cavity transit is a modulation that couples additional modes. The depth of modulation, gain, gain bandwidth and dispersion in the resonator determine the width on the mode-locked pulses. Kerr-lens mode locking is an additional technique for producing very short mode-locked pulses. This technique uses the optical Kerr effect or change in refraction index that is produced by very high intensity pulses.

In a Q-switched laser, oscillation is held off by introducing high losses in the resonator while an inverted population is built up by some means of laser pumping. When the gain has reached a high level, the loss is abruptly removed (see section A3.6.1). Laser oscillation develops in an intense short pulse that may be as short as a few cavity transits of the resonator. Electro-optic and acousto-optic Q-switches are frequently used to control cavity losses for the purpose of Q-switched operation. An acoustic wave deflects a beam by Bragg diffraction in the acousto-optic modulator. When the rf signal to a piezo-electric transducer is turned off, the resonator returns to the low-loss condition. Pockels cells, waveplates and polarizers are used for electro-optic modulation. A double pass through a quarter-wave-retardation waveplate rotates the polarization of the resonator beam and the beam is deflected out of the resonator at the polarizer. When voltage is applied to the Pockels cell, a polarization retardation is produced that cancels the retardation of the waveplate, loss is minimized and a Q-switched laser oscillation rapidly develops. These topics are discussed in detail in chapter A3 and in the references.

A2.1.12 Fibre laser resonators

Brief mention is made here of fibre laser resonators. In fibre lasers, the resonator beam is primarily a guided wave inside a fibre. Pump radiation, often generated by semiconductor diode lasers, is coupled into the cladding of the optical fibre. The pump radiation is absorbed over a relatively great length in a fibre core that is doped with the active laser ion. The fibre core is manufactured with a higher refractive index

than that of the cladding and the core is usually chosen to be of a size that will support only a single-fibre mode. Bragg reflectors can be established in the fibre by techniques such as processing with ultraviolet radiation. It is necessary to engineer the coupling of the pump radiation into the fibre carefully, predict the free-space propagation of the fibre laser radiation outside the fibre and to optimize the placement of discrete elements of the fibre laser resonator not incorporated in the fibre itself. Gaussian optics is useful in each of these areas. Fibre lasers have remarkable properties such as high efficiency and simplicity of operation. Optical properties are also remarkable, with high powers exceeding 100 W having been demonstrated. Broad wavelength availability and high coherence are also being achieved.

A2.1.13 Conclusion

The goal of this chapter has been to present an overview of laser resonators and the techniques of resonator analysis and design. Most of the detail of this discussion has focused on the fundamental aspects of Gaussian optics and stable resonators. Unstable resonators were discussed with illustrative examples. Other topics were briefly mentioned. It is hoped that the information presented is sufficient to address basic issues in resonator design and the management of laser beams. There has been a substantial amount of work performed on the topics of optical resonators since the first demonstrations of lasers in the early 1960s, and investigation of optical resonators and beam propagation still continues. Several of the references have more extensive presentations and detailed bibliographies.

References

[1] Fox A G and Li T 1960 Resonant modes in an optical maser *Proc. IRE* **48** 18 904–5
Fox A G and Li T 1961 Resonant modes in maser interferometers *Bell Syst. Tech. J.* **40** 453–8

[2] Boyd G D and Gordon J P 1961 Confocal multi-mode resonator for millimeter through optical-wavelength masers *Bell Syst. Tech. J.* **40** 489–508

[3] Siegman A E 1986 *Lasers* (Mill Valley, CA: University Science Books)

[4] Hall D R and Jackson P E (ed) 1990 *The Physics and Technology of Laser Resonators* (Philadelphia, PA: Institute of Physics Publishing)

[5] Hodgson and Weber 1997 *Optical Resonators, Fundamentals, Advanced Concepts and Applications* (Berlin: Springer)

[6] Selveto O 1998 *Principles of Lasers* 4th edn (New York: Plenum)

[7] Kochner W 1976 *Solid-State Laser Engineering* (New York: Springer)

[8] Yariv A 1989 *Quantum Electronics* 3rd edn (New York: Wiley)

[9] Siegman A E 2000 Laser beams and resonators: the 1960s *IEEE J. Selected Topics Quantum Electron.* **6** 1380–8

[10] Siegman A E 2000 Laser beams and resonators: beyond the 1960s *IEEE J. Selected Topics Quantum Electron.* **6** 1389–99

[11] Kogelnik H and Li T 1966 Laser beams and resonators *Appl. Opt.* **5** 1550–67

[12] Abramowitz M and Stegun I A 1972 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Washington, DC: National Bureau of Standards)

[13] Siegman A E 1990 New developments in laser resonators *Proc. SPIE* **1224** 2–14

[14] Siegman A E and Townsend S W 1993 Output beam propagation and beam quality from a multimode stable-cavity laser *IEEE J. Quantum Electron.* **29** 1212–17

[15] Goodman J W 1968 *Introduction to Fourier Optics* (San Francisco, CA: McGraw-Hill)

[16] Oughstun E E 1987 *Unstable Resonator Modes (Progress in Optics 24)* ed E Wolf (Amsterdam: North-Holland) pp 165–387

[17] Siegman A E 1991 Defining the effective radius of curvature for a non-ideal optical beam *IEEE J. Quantum Electron.* **27** 1146–8

[18] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes in C: The Art of Scientific Computing* 2nd edn (New York: Cambridge University Press)

[19] Frigo M and Johnson S G *FFTW, A C Subroutine Library for Computing the Discrete Fourier Transform* (Cambridge, MA: MIT) ⟨http://www.fftw.org⟩

[20] Horwitz 1974 Asymptotic analysis of unstable resonator modes *J. Opt. Soc. Am.* **63** 1528–43

[21] Krupke W F and Sooy W R 1969 Properties of an unstable confocal resonator CO_2 laser system *IEEE J. Quantum Electron.* **QE-5** 575–86

A2.2
Waveguide laser resonators

Chris Hill

A2.2.1 Introduction

This chapter links the two topics of optical waveguide propagation (chapter A6) and optical resonator physics (chapter A2.1). An optical resonator can contain any sort of aperture, lens, mirror, prism and obstruction. The total resonator thus formed will have its own self-repeating field patterns. The principle that a 'resonator mode' is self-repeating in phase and amplitude (or that it is an *eigenvector* for the resonator round trip) is rather general and powerful, extending even to apparently complicated structures. The presence of a *waveguide*, as either a minor or a major 'obstruction', does not affect this principle and need not cause alarm. Waveguides are, from one popular and useful viewpoint, only 'apertures'—albeit extended 3D ones. This chapter tries to make readers familiar with their presence and effects.

The introduction to optical resonator properties in chapter A2.1 deliberately ignores any interaction of light with the laser-cavity side walls. Apertures and mirror edges may be used for transverse-mode control but otherwise the light is assumed to propagate freely between the resonator mirrors. We now consider a class of resonator with intentional wall effects: the resonator path includes a *waveguide*, for example a hollow dielectric tube in which light is guided by a series of Fresnel reflections from the tube walls. We will see how some resonator properties of waveguide lasers may be modelled theoretically and how such models compare with actual laser devices.

A general waveguide resonator is shown in figure A2.2.1. The total influence of all the optical elements, including the guide, will determine what self-consistent field patterns may exist. Our modelling problem is to find these *resonator modes*, with their round-trip losses and resonant frequencies. The important point here is that, inside a sufficiently narrow guide, the Gaussian-beam modes and propagation equations of chapter A2.1 are inappropriate. Instead, we must recognize guide wall effects by using another set of functions (modes) which obey boundary conditions at those walls. We must also treat the coupling of optical radiation at the guide apertures and the free-space propagation to and from the mirrors.

Now, if we properly account for the coupling in, the propagation through and the coupling out again from the guide, so that the waveguide output for *any* Gaussian-beam mode (at the input plane) is again expressed in terms of Gaussian-beam modes (at the output plane), we have as full a description as can be desired. The mathematical quantities involved may be called waveguide coupling matrices, sets of overlap integrals or 'transfer functions' (in terms of transverse modes of different spatial frequencies, rather than the more usual electrical waves of different radian frequencies).

Early workers on waveguide lasers sought definite *gain/bandwidth* advantages. In particular, because of favourable wall interactions, narrow-bore waveguide lasers with CO_2 as the active medium could run at much higher pressures than conventional Gaussian-resonator lasers, and thus offered increased frequency tunability. Chapter B3.1 says more about CO_2 waveguide laser device characteristics and technology. Here we concentrate on the optical properties of a 'waveguide resonator', where *the presence of the waveguide*

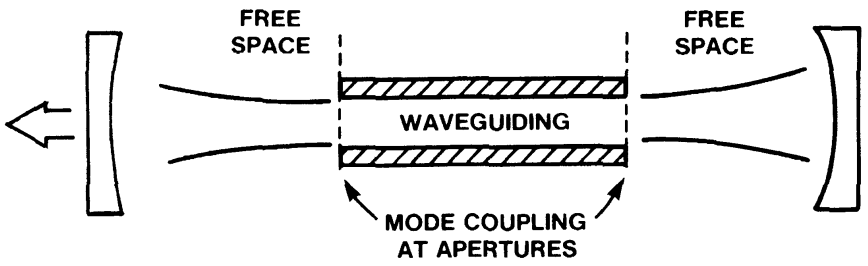


Figure A2.2.1. Sketch of a general linear waveguide resonator.

significantly affects the mode structure of the optical resonator. This seems the only reasonable answer to the question, When does a resonator become a waveguide resonator?

We restrict our discussions (although this is not essential) to infrared hollow dielectric guides. These are hollow guides as opposed to solid or clad optical fibre guides. They are assumed to be uniform, straight and of rectangular or circular cross section, with half-widths or radii much greater than the radiation wavelength ($a \gg \lambda$). Such structures possess sets of *waveguide modes.* These are field patterns which can propagate along the guides, with tolerably low loss, maintaining their transverse shape. Formally, they are solutions of Maxwell's equations obeying the appropriate boundary conditions at the guide walls (as discussed in chapter A6). With our assumptions, and freely using $a \gg \lambda$ to discard terms, we can derive fairly simple mode fields and propagation constants (and in any case we omit most of the mathematical details here). The lowest-order linearly polarized mode of a square-bore guide, for instance, is called EH_{11} and looks rather like the lowest-order linearly polarized free-space mode TEM_{00}. As we shall see later, the fact that guide attenuation losses are strongly mode-dependent may be used to our advantage in controlling the transverse-mode pattern actually emitted by the laser. When talking of mode fields, coupling losses and resonators, we will not use specific guide dimensions or wavelengths, since the theory is fairly general. The most common real devices, and most of our examples in this chapter, involve square guides and the familiar CO_2 laser medium ($\lambda \simeq 10$ μm). There is no direct treatment here of laser media, the effects of guide walls on discharges or their consequences for laser design and engineering. Some of this material is covered in chapter B3.1.

This chapter has four further sections. In section A2.2.2 we briefly discuss light propagation and modes in hollow dielectric waveguide structures. Section A2.2.3 is a rather general account of *waveguide resonator analysis.* Section A2.2.4 presents some detail of first-order (single-mode) waveguide resonator theory. Finally, section A2.2.5 is concerned with the resonator properties of some real waveguide lasers and how we interpret them in terms of available resonator theory.

A2.2.2 Propagation in hollow dielectric waveguides

In 1964, Marcatili and Schmeltzer proposed the use of hollow dielectric waveguides as low-loss components for laser amplifiers [1]. They calculated the allowed *modes of propagation* of hollow dielectric waveguides by solving the standard wave equation in free space but with dielectric boundary conditions; field components were matched at the guide walls instead of falling away to zero at infinity. These modes are field patterns which keep their shape, but experience a mode-dependent phase shift, as they propagate along the guide. Hollow dielectric waveguides used in typical gas waveguide lasers have half-widths or radii $a \sim 1$ mm, so the assumption $a \gg \lambda$ is reasonable. They cannot be made very much smaller without severe attenuation ($\propto a^{-3}$), and are inherently *multi-mode* or *over-moded.* When terms of first or higher order in λ/a are removed, the rather complex expressions for the field amplitudes of the waveguide modes reduce to simple 'first-order' ones. But the corresponding simplified propagation constants include essential terms of order λ/a and $(\lambda/a)^2$. For each waveguide mode the propagation constant can be written in the form $k = \beta + i\alpha$,

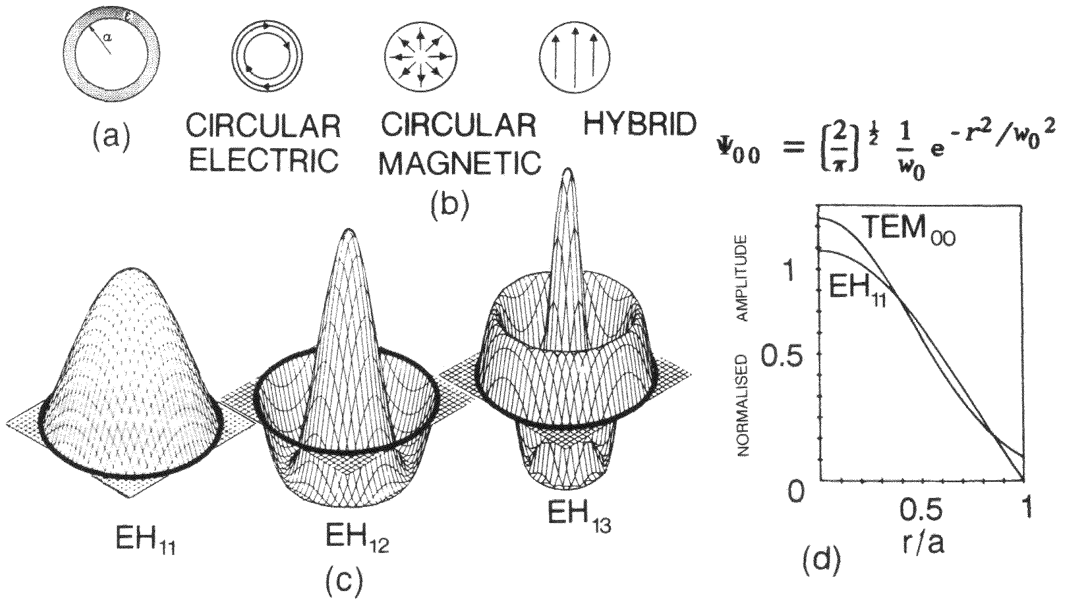


Figure A2.2.2. Characteristics of circular waveguide: (*a*) circular waveguide geometry, (*b*) electric field patterns of circular waveguide modes, (*c*) amplitude profiles if low-order EH_{1m} modes and (*d*) amplitude profile of EH_{11} and its best fit Gaussian Ψ_{00} ($w0 = 0.64a$). EH_{11} is normalized according to equations (A2.2.2) and (A2.2.3) with $a = 1$.

where β is the phase constant and α the amplitude attenuation constant. Propagation along the long axis of the waveguide is described by a term $\exp(ikz)$.

A2.2.2.1 Waveguide mode expressions

In this section we describe the simplified fields and propagation constants for the square and circular waveguide geometries. Inconsistent notation for waveguide modes is usually the first thing that irritates workers in this subject. A brief study of Degnan [2] or other reviews may help.

For circular cross section dielectric guides (radius a, polar coordinates r and φ as in figure A2.2.2(*a*), we usually have a single dielectric constant $\varepsilon = (n_a - ik_a)^2$ for the wall material. Circular guides may support transverse circular electric modes (TE_{0m}), transverse circular magnetic modes (TM_{0m}), and sets of higher-order hybrid modes (EH_{nm}) which have both radial and tangential electric and magnetic fields. Figure A2.2.2(*b*) shows the first-order field amplitudes and transverse patterns for the lowest-order mode of each type. In practice, linear polarization with a fixed axis is usually desired, and often enforced with a Brewster window or diffraction grating. We can express an arbitrary linearly polarized transverse field pattern within a circular guide as a combination of LP_{nm} modes, where n is the azimuthal or angular mode number and m is the radial mode number. The LP_{0m} modes are identical with the EH_{1m} modes of Marcatili and Schmeltzer [1] or the HE_{1m} modes of Snitzer [3]. This important set of zero-angular-order modes is sufficient to describe any circularly symmetric field pattern and to model many experiments. Figure A2.2.2(*c*) shows the amplitude distribution of the fundamental hybrid mode EH_{11} (or HE_{11} or LP_{01}). The higher-order modes must be included for a full description of even well-aligned lasers, and are essential for any account of misalignment effects.

The first-order fields are given by

$$E_{nm} = f_{nm} J_n(\rho_{nm} r/a) \cos(n\varphi) \qquad 0 \leq r \leq a \tag{A2.2.1}$$

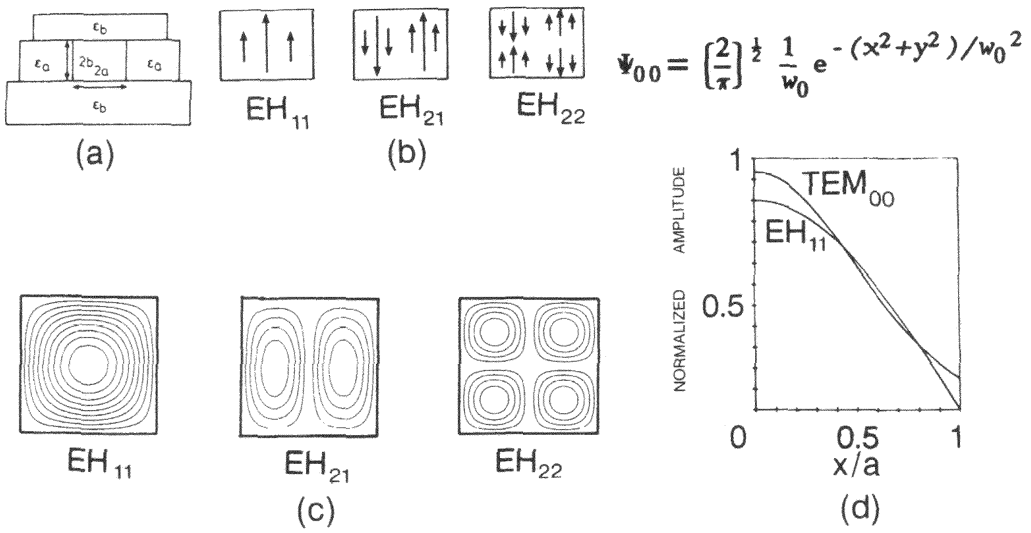


Figure A2.2.3. (*a*) Rectangular waveguide geometry, (*b*) electric field lines of low-order EH_{mn} modes (*c*) low-order amplitude contours for a square-bore guide and (*d*) amplitude profile of EH_{11} and its best fit Gaussian $\Psi_{00}(w_0 = 0.70\,a)$. EH_{11} is normalized according to equation (A2.2.7) with $a = b = 1$.

where ρ_{nm} is the m^{th} root of the nth-order Bessel function, i.e. $J_n(\rho_{nm}) = 0$, and

$$f_{nm} = \begin{cases} \sqrt{2}[\sqrt{\pi}\,a\,J_{n+1}(\rho_{nm})]^{-1} & \text{if } n = 2, 3, 4, \ldots \\ [\sqrt{\pi}\,a\,J_1(\rho_{0m})]^{-1} & \text{if } n = 0 \end{cases} \tag{A2.2.2}$$

for the even LP_{nm} ($n \neq 1$) modes. The axis of polarization is arbitrary: neither the field strengths nor the propagation constants depend on it. The field strengths for the even linearly polarized ($n = 1$) modes obey the same form:

$$E_{1m} = \sqrt{2}[\sqrt{\pi}\,a\,J_2(\rho_{1m})]^{-1} J_1(\rho_{1m} r/a)\cos\varphi \tag{A2.2.3}$$

but the axis of polarization is not arbitrary. This choice of even (cosine) modes forces the choice of circular magnetic TM_{0m} modes (plus HE_{2m}) for x-polarized light, and circular electric TE_{0m} modes (plus HE_{2m}) for y-polarized light. The normalization and the propagation constants are given by

$$\int_0^{2\pi} \int_0^a E_{nm}(r, \varphi) E_{n'm'}(r, \varphi) r\, \mathrm{d}r\, \mathrm{d}\varphi = \delta_{nn'}\delta_{mm'} \tag{A2.2.4}$$

$$\beta_{nm} \simeq \frac{2\pi}{\lambda}\left[1 - \frac{1}{2}\left[\frac{\lambda\rho_{nm}}{2\pi a}\right]^2\right] \tag{A2.2.5}$$

$$\alpha_{nm} \simeq \frac{1}{a}\left[\frac{\lambda\rho_{nm}}{2\pi a}\right]^2 \mathrm{Re}\left[\frac{1}{2}(\varepsilon + 1)(\varepsilon - 1)^{-\frac{1}{2}}\right] \tag{A2.2.6}$$

Figure A2.2.3 shows a general rectangular cross-section dielectric guide (width $2a$ and height $2b$) with two different dielectric materials. In practice, many guides use a single material but many others are 'hybrid' with one or two metal walls (which may be electrodes for transverse rf excitation). Common guide materials are aluminium, alumina (Al_2O_3) and beryllia (BeO). Figure A2.2.3 shows two complex relative dielectric constants:

$$\varepsilon_a = (n_a - \mathrm{i}k_a)^2 \qquad \varepsilon_b = (n_b - \mathrm{i}k_b)^2$$

where n_a and n_b are refractive indices and k_a and k_b are extinction coefficients; all four are positive real numbers. These rectangular guides support two sets of linearly polarized EH_{mn} hybrid modes but do not

support circularly polarized modes. In a ceramic–metal hybrid guide with one or two horizontal metal walls, the horizontally polarized E^xH_{mn} modes are usually favoured over the vertically polarized E^yH_{mn} modes, chiefly because of the loss factors associated with the metal.

The first-order fields for the linearly polarized EH_{mn} modes are

$$E_{m,n}(x, y) = ab^{-\frac{1}{2}} \begin{bmatrix} \cos \\ \sin \end{bmatrix} \left(\frac{m\pi x}{2a}\right) \begin{bmatrix} \cos \\ \sin \end{bmatrix} \left(\frac{n\pi y}{2b}\right); \qquad m, n = \begin{matrix} \text{odd} \\ \text{even} \end{matrix} \tag{A2.2.7}$$

with normalization

$$\int_{-a}^{a} \int_{-b}^{b} E_{mn}(x, y) E_{m'n'}(x, y)\, dx\, dy = \delta_{mm'}\delta_{nn'}. \tag{A2.2.8}$$

The propagation constants $k_{mn} = \beta_{mn} + i\alpha_{mn}$ are given by

$$\beta_{mn} \simeq \frac{2\pi}{\lambda}\left[1 - \frac{1}{2}\left[\frac{m\lambda}{4a}\right]^2 - \frac{1}{2}\left[\frac{n\lambda}{4b}\right]^2\right] \tag{A2.2.9}$$

and (for x-polarized modes)

$$\alpha_{mn} \simeq \frac{m^2}{a}\left[\frac{\lambda}{4a}\right]^2 \mathrm{Re}(\varepsilon_a(\varepsilon_a - 1)^{-\frac{1}{2}}) + \frac{n^2}{b}\left[\frac{\lambda}{4b}\right]^2 \mathrm{Re}((\varepsilon_b - 1)^{-\frac{1}{2}}) \tag{A2.2.10}$$

Figure A2.2.3 also has some contour maps for the lowest-order modes in a square waveguide, and a comparison of the EH_{11} field amplitude with that of the fundamental Gaussian beam which best approximates it (see section A2.2.4.1).

Remember the benefits of this modal approach to propagating light. The appropriate (Maxwell) equations are solved once for all, and thereafter little effort is needed to find the field (at any point in the guide) due to a specified initial field (at any previous point). Instead of solving a fresh set of wave equations, we need only perform a set of multiplications by the complex numbers $\exp(ik_{mn}z)$. This is a great benefit and immediately invites a *matrix* treatment. Any possible field (consistent with our simplified model) will be represented by a linear combination of EH_{mn} modes; and any possible propagation, reflection, scattering etc will produce another linear combination of the same modes. We can calculate once for all the self-coupling and cross-coupling coefficients to describe this, arrange them in suitably ordered propagation and coupling matrices, and manipulate them easily in one of several popular matrix-based software packages.

A2.2.3 Waveguide resonator analysis

This section gives a rather general account of resonator analysis as applied to waveguide lasers. A waveguide resonator (figure A2.2.1) may be viewed as a free-space Gaussian resonator perturbed by an aperturing waveguide tube, or as a tube resonator perturbed by the addition of free-space sections. We summarize the usual methods of calculating the properties of waveguide resonators and later examine some of the geometric constraints and principles in the design of a resonator. Having read chapter A2.1, readers should be familiar with the main features of *optical resonators* and free-space Gaussian beams, so that concepts such as round-trip eigenvalues and their associated *resonator modes* may be introduced and examined without difficulty. Even so, large amounts of time and effort can be saved if such terms as 'mode', 'resonator mode', 'multi-mode' and 'mode loss' are clearly understood. At the risk of some duplication, we discuss these key concepts as they arise.

A2.2.3.1 The concept of resonator modes

The real waveguide lasers used in the factory or on the bench are more or less complicated assemblies with electrical, mechanical and optical elements including active media. We wish to model important laser properties such as output power, beam shape, frequency tunability and frequency purity. We discard all specific real-life elements and consider the very simple, and fairly abstract, idea of an optical resonator which possesses some *resonator modes*. The link with real-life devices will reappear shortly. A resonator mode is a field distribution which repeats itself in shape and in phase after one round trip of the resonator.

Let us start with a well-aligned Gaussian-beam resonator with two large-aperture, spherically curved mirrors, and recall parts of chapter A2.1. We know that a standard free-space scalar wave equation, when we require the light to stay near the z-axis with the dominant variation being simply the axial propagation term $\exp(ikz)$, yields Gaussian beam functions. These are the TEM_{pq} free-space *modes* of propagation: their properties change very slowly with wavelength and their transverse shape is given by a simple Gaussian $\exp(-r^2/w^2)$, multiplied by Laguerre polynomials (in cylindrical coordinates) or Hermite polynomials (in Cartesian coordinates). We can define a common beam-waist position z_0 and a common beam-waist radius w_0, for the whole orthonormal set of beams. Then, formally, we can express the free-space field E_{fs} anywhere along the resonator axis as a linear combination of Gaussian beams:

$$E_{\mathrm{fs}} = \Sigma b_{pq} \Psi_{pq}(z - z_0, w_0) \tag{A2.2.11}$$

$$b_{pq} = \int E_{\mathrm{fs}} \Psi_{pq}^* \, \mathrm{d}A \tag{A2.2.12}$$

where the Ψ_{pq} are the TEM_{pq} Hermite–Gaussian or Laguerre–Gaussian functions *including* the z-dependent amplitude and phase factors given in chapter A2.1. The '*' indicates complex conjugation and the integral is performed over the infinite cross section (though only the area near the z-axis contributes significantly). The b_{pq} are complex coefficients which, during lossless free-space propagation, do not depend on axial position z. Again we see great benefits from the modal approach, as apparently complicated wave equations or diffraction integrals are replaced by the simple Gaussian beam transformations.

Thus equations (A2.2.11) and (A2.2.12) show a 'decomposition' of some arbitrary function E_{fs} into an orthonormal set of functions Ψ_{pq} representing various spatial frequencies, just as in Fourier analysis we split some function $f(x)$ into sines and cosines. If the original field E_{fs} is associated with a precise temporal frequency, then so is each term of the sum: the precise frequency information is extra and not contained in the spatial form of the modes of propagation.

By contrast, resonator modes are *modes of oscillation*. This concept has two essentials: the mode shape must be one of the self-repeating *transverse modes* of the cavity, and the mode frequency must yield a precise *axial* resonance. Conventionally, we assign two transverse-mode integers r and s, and one longitudinal-mode integer j, where j is the number of full 2π propagation phase shifts per round trip. Thus, a general resonator mode frequency is $\nu_{j,rs}$. For most real lasers, $j > 10\,000$ and $r, s < 10$, with r and s referring to the transverse-mode order in the x and y (or radial and azimuthal) directions. Usually a change of 1 part in 10^4 in λ will make no important difference to the transverse part (the shapes and amplitudes of the Gaussian beams), whereas the change $j \rightarrow j + 1$ (or a single-pass phase change $\sim\pi$) is crucial in laser frequency studies. This means that laser axial modes and transverse modes are physically decoupled and can be treated separately. They are both present in the idea of an oscillating or resonating mode. It often helps to imagine a laser emitting some beam of arbitrary transverse shape with a single well-defined frequency; this shape, and this frequency $\nu_{j,rs}$, constitute a resonator mode by definition.

Formally, the self-consistency condition translates into the complex eigenvalue equation:

$$M E_{\mathrm{fs}} = \gamma E_{\mathrm{fs}} \tag{A2.2.13}$$

where each self-consistent field E has its complex eigenvalue γ. We will define a *round-trip loss* $1-|\gamma|^2$ and *relative phase shift* $\arg(\gamma)$. Since the *round-trip matrix* M is complex and may be large, we may be no further forward. However, we see that if enough of M can be specified with reasonable accuracy, equation (A2.2.13) can be solved in principle. For stable open resonators, the resonator transverse modes are accurately given (see chapter A2.1) by pure Gaussian beams with appropriate phase shifts: that is, by choosing w_0 and z_0 correctly we can make M diagonal. Each eigenvector can then be written as a single term:

$$E_{\mathrm{fs}} = \Psi_{pq}(z - z_0, w_0). \tag{A2.2.14}$$

The Ψ_{pq} relative phases are given by $2(p+q+1)\cos^{-1}[\pm(g_1g_2)^{1/2}]$, where the 2 refers to a *round* trip or double pass between the mirrors, and the parameters $g_{1,2} = 1 - L/R_{1,2}$ were introduced in chapter A2.1. But in general, with misalignments, obstructions or perturbations, M is not diagonal and each eigenvector is a mixture or linear combination as in our equation (A2.2.11); the b_{pq} may now depend on z. We are still free to choose any w_0 and z_0 but no choice will yield pure Ψ_{pq} as the self-consistent fields, and equation(A2.2.13) will generally be tedious to solve without a computer. Applying this empty resonator theory to active lasers, we *assume* that the various γ and E are unaffected by the laser medium—except that the overall cavity loss $1-|\gamma|^2$ is exactly balanced by the round-trip gain. This neglect of the active medium is convenient but (not surprisingly) may lead to inaccurate results (see section A8.3).

If we understand all this, then the 'free-space' part of our problem is solved. To analyse a given resonator we break one round trip into a sequence of well-defined segments, such as free-space propagation between the mirrors or reflection from a mirror. We choose a base set of free-space Gaussian beams Ψ_{pq}. In practice, with limited computing resources, p and q cannot take infinitely many values, so our base set is judiciously *truncated*. By ABCD matrices or by explicit Huygens–Fresnel diffraction calculations, we mathematically describe the effect of each segment on our arbitrary field (= linear combination of Ψ_{pq}) by a matrix. The ordered product $M = M_n \ldots M_3 M_2 M_1$ of all these matrices is the grand round-trip matrix, whose eigenvectors are self-consistent linear combinations and whose eigenvalues define the round-trip losses and phase shifts.

We often hope that the largest eigenvalue (lowest loss) refers to some field:

$$E = b_{00}\Psi_{00} + b_{01}\Psi_{01} + b_{10}\Psi_{10} + \cdots \tag{A2.2.15}$$

(truncated at some $p_{\max}$, $q_{\max}$) which is nearly pure TEM_{00}; that is if we normalize to $\Sigma|b_{pq}|^2 = 1$, then we want $1 - |b_{00}|^2 \ll 1$. This seems less likely as larger perturbations are introduced, and is usually untrue of unstable resonators. *The perturbation that most concerns us here is a hollow dielectric waveguide, lying along the z-axis and acting as a 'three-dimensional aperture'.* From our present point of view, some obvious questions are: How do we form a matrix to describe the effect of this guide on our Gaussian beams? That is, how does radiation couple into the guide, propagate along it and couple out again? Is it easy to decide whether a given waveguide will significantly perturb the first few resonator modes? (If it will not, we may ignore it for most purposes.) And, importantly, does the eventual waveguide resonator matrix theory show any agreement with experiment?

A2.2.3.2 Waveguide modes

Suppose, by analogy with equation (A2.2.11), we have an orthonormal set of guide modes, and can express the field E_{wg} anywhere inside the guide as

$$E_{\mathrm{wg}} = \Sigma a_{mn} E_{mn} \tag{A2.2.16}$$

$$a_{mn} = \int E_{\mathrm{wg}} E^*_{mn}\, \mathrm{d}A \tag{A2.2.17}$$

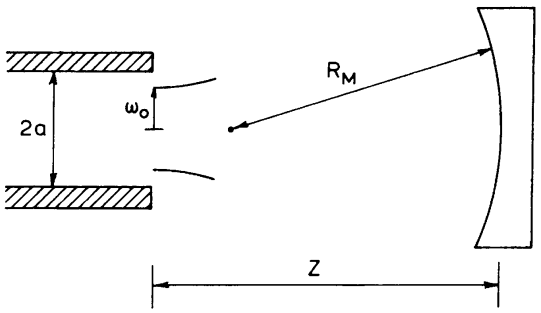


Figure A2.2.4. Sketch of general curved mirror–waveguide combination.

where we integrate over the guide cross section. These functions E_{mn} will usually stand for the linearly polarized EH_{mn} modes, just as the Ψ_{pq} functions stand for the TEM_{pq} modes. Because waveguides are lossy, the complex coefficients a_{mn} depend on z. There is no analogous mode waist radius w_0 or waist plane at some z_0; the guide mode has the same spatial shape at every z within the guide.

For Gaussian beams, an 'ideal' reflector is a large-aperture mirror whose curvature exactly matches that of the incident beams. Each beam is reflected back on itself, with no coupling to any other beam: that is, the Gaussian beam coupling matrix of the mirror is diagonal. For waveguide modes, the corresponding reflector is a plane mirror aligned perpendicular to the guide axis and placed immediately against the guide end. Here again the modes are reflected back along the z-axis without cross-coupling, so we have amplitude self-consistency. It is then fairly clear that, with two such mirrors, we have a simple waveguide resonator whose modes of oscillation, or resonator modes, are pure waveguide modes—with, again, the important extra condition of phase self-consistency. The resonator matrix M can be written diagonally; its non-zero elements are the propagation constants $\exp(i2k_{mn}L)$. This diagonal feature is unique to this (in principle) simple tube resonator, called dual case I (see section A2.2.4.1).

Later, we will meet other designs which offer nearly pure modes. We will also see that, in practice, even this simplest resonator does not always follow 'first-order' theory. In general, with perturbations, M must be non-diagonal and the eigenvectors will be linear combinations of the E_{mn} with several significant non-zero terms a_{mn}. Note that we can order the values of m and n so that $\Sigma a_{mn}E_{mn}$ is a column vector and M is a square matrix. If we understand this, the waveguide part of our problem will be solved also.

A2.2.3.3 Mode coupling, coupling losses and mode losses

The missing step so far is the coupling between free space and the waveguide. In mathematical terms this is a change of basis; given the two orthonormal sets of field patterns, we express the coupling coefficients as overlap integrals across the guide aperture and form these coefficients into a coupling matrix. When describing mode-coupling loss, we must specify the modes from which and into which radiation is coupled. For example, let us consider the fundamental waveguide mode EH_{11}, together with a general curved-mirror reflector (figure A2.2.4).

The *amplitude coupling coefficient* between EH_{11} and TEM_{pq} is given by $\int E_{11}\Psi^*_{pq}\,dA$, that is the integral of the product of the waveguide mode field and the complex conjugate of the free-space field. The modulus squared of this is the *coupling efficiency*, which differs from unity by the *coupling loss*. Similar coefficients, efficiencies and losses could be defined for any EH_{mn} mode or for any sum of modes. However, of immediate interest is the amplitude *self*-coupling coefficient of the fundamental mode, that is the amount of EH_{11} which reappears as EH_{11} after propagation to and from the reflector. This is defined as the overlap integral of EH_{11} with the returned (conjugate) field due to a 'launched' EH_{11}. The modulus squared will be

the EH_{11} self-coupling efficiency $|c_{11,11}|^2$ and the self-coupling *loss* will be

$$\Gamma_{11} = 1 - |c_{11,11}|^2 \tag{A2.2.18}$$

Γ_{11} is commonly called the 'EH_{11} coupling loss'. We can find these fields and integrals by diffraction theory or by assembling and tracking our linear combination of Ψ_{pq} but the task may not be easy. There is much more on this subject in the literature, and section A2.2.4 summarizes some results.

A particular combination of waveguide modes $\Sigma a_{mn} E_{mn}$ at the guide exit, associated with a given resonator mode, will experience its own self-coupling efficiency and loss. But this is a coherent assembly of guide modes, and the resonator mode coupling efficiency is not generally the weighted sum of the individual EH_{mn} coupling efficiencies. Instead it depends crucially on the relative phases of the waveguide modes. This is the simple but vital reason why a multi-mode approach is needed for all but the simplest resonators or all but the crudest estimates of laser loss and transverse-mode quality. Unless great pains are taken to force single-mode operation, significant higher-order mode amplitudes (a_{mn}) may creep in and cause unpredictable spatial interference; and small changes of only 1–2% in round-trip loss can significantly alter the available laser power. The relative higher–order mode powers $|a_{mn}|^2$ may seem very small, but interference involves complex amplitudes.

Similarly, the power lost per unit length through guide attenuation of waveguide mode combinations is not generally $\Sigma 2\alpha_{mn}|a_{mn}|^2$. In first-order theory (section A2.2.4), where terms of order λ/a are neglected, the EH_{mn} mode fields form a complete orthonormal set across the guide aperture and all the fields vanish at the walls. But to describe attenuation we must treat the small non-zero fields at the walls; and, when two or more modes are present, these fields include interference terms. Many resonator models tend to neglect this fact, either ignoring it totally or assuming that, with several modes travelling up and down the guide, the interference terms will rapidly 'wash out' to leave a smooth average loss. We should be aware of such neat but inaccurate assumptions, and realize that the distribution of loss over the round-trip path in real lasers is seldom easy to calculate.

A2.2.3.4 Single-mode, few-mode and multi-mode theory

Once our method of resonator analysis is settled, and we can describe each resonator segment, the question usually arises: How many modes should, or must, we include? Broadly speaking, there have been three answers: one, a few and many; that is single-mode theory, few-mode theory and multi-mode theory.

In *single-mode* theory we assume that the lasing resonator mode is pure EH_{11} (or, rarely, another pure EH_{mn} mode); therefore, the resonator loss is formed by pure EH_{11} attenuation and coupling losses and no interference effects are considered. This may be a good approximation for some lasers with well-aligned near-field plane mirrors or strong inbuilt mode selection; it is usually hopeless for lasers with gratings or misaligned mirrors or almost any perturbation. It requires all higher-order modes to be either very faintly excited or very strongly damped. But higher-order modes tend not to be suppressed very effectively by attenuation in typical guides: if we begin with equal amounts of EH_{11} and EH_{12}, it may take many metres of guide propagation before the ratio is even 2:1 in favour of EH_{11}. (To check this we evaluate $L \sim 0.5(\alpha_{12} - \alpha_{11})^{-1} \ln 2$). Also, by making guides with low EH_{11} losses, we inevitably reduce any mode discrimination due to guide loss: the inference is that an 'ideal' laser of minimum EH_{11} loss, with an excellent straight smooth guide and plane mirrors near the guide ends, has negligible inbuilt transverse-mode discrimination and may emit beams whose transverse shapes are quite unlike E_{11} or Ψ_{00}.

To model real lasers in any detail, we must consider at least the *first few* modes. It is generally considered that about five or ten will do: the lowest-loss resonator mode and its loss will be modelled with fair accuracy, so that adding extra modes does not greatly change the results; beyond five or ten modes the first-order theory must be suspect; and, since with N modes we generally have an $N \times N$ complex matrix, the computing load for $N > 10$ may be too heavy.

Multi-mode resonator models, when compared with real lasers (see section A2.2.5), tend to give encouraging but not *very* accurate results. They are far better than single-mode models, and considerably better than three- or five-mode models, but yield diminishing returns as our initial approximations become strained and then broken.

In fact, these approximations are likely to collapse before we achieve numerical convergence to 'final' resonator predictions. This is, again, simply because typical waveguides are short and fat. Although the capillary waveguides used in lasers appear relatively long and slender, in terms of transverse-mode discrimination they are not. The influence of interfering high-order modes can be complicated and resonator losses can seem to depend (in theory) very sharply on precise guide geometry (see section A2.2.5). Some caution is needed when we interpret rapidly varying loss curves, derived from idealized matrix equations, in terms of real laser behaviour. Real lasers have imperfections—roughnesses, random mirror defects, scattering centres—which may be very difficult to model accurately but which, on the whole, tend to smooth out such 'ideal' sharp variations or spikes.

In practice, we must draw a line at some reasonable level of theoretical complexity; it becomes unproductive to consider hundreds of modes in search of a small increase in numerical accuracy. Extra caution is perhaps needed at this point, because the number-crunching power of cheap computers continues to advance. It becomes ever easier to evaluate hundreds or thousands of matrix coefficients by numerical integration or FFT-based diffraction integrals, pass them through matrix inversion routines, and plot the 'results' cheerfully. Section A2.2.4 offers some rough limits to common waveguide laser assumptions. The only general lesson is that we should keep clearly in mind the type of results and the degree of accuracy that will satisfy us; use the necessary number of modes to that end; and lower our expectations if we find that number too high.

A2.2.4 First order theory and its limits

This section discusses the *single-mode* theory of waveguide resonators. This theory, built on the waveguide mode expressions in section A2.2.2 and the coupling coefficients in section A2.2.3, gives our first view of the impact of a waveguiding region within the resonator. We quote some important results from the literature on coupling theory, the dual case I resonator and Rigrod modelling, but we stress that the resonator properties of most waveguide lasers are poorly described by single-mode theory.

A2.2.4.1 Coupling loss theory of single-mode waveguide resonators

This topic introduces the essential concepts in resonator analysis, and remains interesting for three main reasons. First, if we trust that our resonator transverse mode is almost a pure waveguide mode, we want to estimate coupling losses and judge their effect on laser power (see section A2.2.5 where the Rigrod equation is discussed). Second, in resonator design, we want to exploit mode-dependent coupling losses and force near-single-mode operation; here the relative losses of several modes are important. The third reason is that coupling losses are important in experiments and applications involving waveguide transmission, even if there is no resonator.

We consider the general reflector (waveguide plus spherical mirror) of figure A2.2.4. A guide mode is launched from the aperture and, after reflection, couples to itself and to other modes. Researchers have found the returned field distribution in two main ways. One is to express the initial field as a linear combination of Gaussian modes and propagate it to and from the mirror according to the Gaussian beam equations of Kogelnik and Li [4], assuming that the mirror aperture is effectively infinite. The other is to use scalar diffraction integrals. The two approaches give identical results [5]. Early work [6–8] predicted the existence of three reflector configurations which result in low EH_{11} coupling loss. For a circular guide of radius a, with

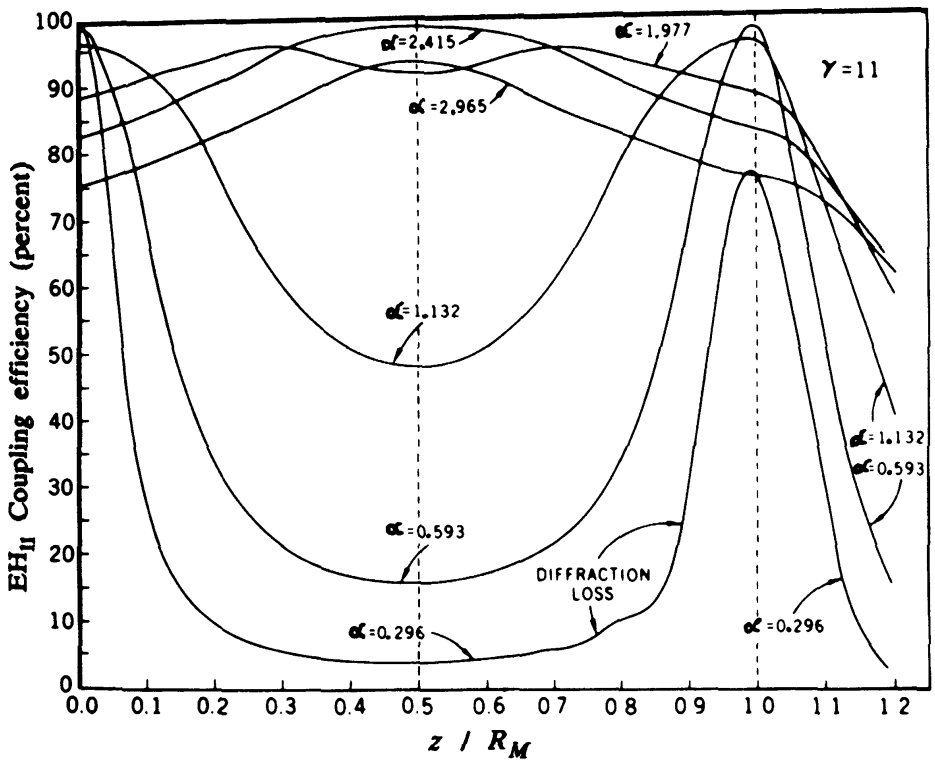


Figure A2.2.5. EH_{11} coupling losses for circular guides. Curves are shown for various values of $\alpha = (ka^2/R_M)$ and a constant mirror half width of $11a$. (From Degnan and Hall [7].)

$E(r)$ the launched field and $E'(r)$ the returned field, equation (A2.2.18) may be written:

$$\Gamma_{11} = 1 - \left| \int_0^a E(r)E'(r)2\pi r \, dr \right|^2 . \tag{A2.2.19}$$

Evaluating this for a range of mirror-guide distances and mirror curvatures yields the set of curves in figure A2.2.5. The three low-loss reflector geometries (figure A2.2.6) were named by Degnan and Hall [7]:

Case I: large R mirrors very near the guide ($z \simeq 0$, $z/R \simeq 0$),

Case II: large R mirrors centred near the guide entrance ($z \simeq R$),

Case III: mirrors with $R \sim 2b$ and $z \sim b$, where $b = \pi w_0^2/\lambda$.

Here w_0 and b are the beam-waist radius and Rayleigh range of the EH_{11} mode's *approximating Gaussian*, i.e. the TEM_{00} beam having maximum overlap with EH_{11} across the guide aperture ($z = 0$). This occurs for $w_0/a \sim 0.64$ (circular) or $w_0/a \sim 0.70$ (square), and the power overlap is ~98% in both geometries. We expect that any reflector which efficiently recouples this TEM_{00} to itself will do the same for EH_{11}. Thus, the simplest possible model of a general waveguide resonator, with only EH_{11} inside the guide and only TEM_{00} outside, suggests the choice of *phase-matched mirrors*, whose curved surfaces coincide with the phase fronts of this special TEM_{00} at z_1 and z_2, according to $R = z + b^2/z$. For such mirrors, Abrams [9] found that Γ_{11} is ~1.48% for case III and always less than 7% (circular bore). For case III, Degnan and Hall [7] revised this estimate to ~1.38% and found that $\Gamma_{12} \sim 78\%$. A reasonably well-built device with one or two case III mirrors usually offers the best chance of guaranteed TEM_{00}-like operation. On the whole, the approximating Gaussian concept is neat and intuitive and predicts Γ_{11} minima with fair accuracy, but with short fat guides its practical value has not always been realized. Note: this definition $b = \pi w_0^2/\lambda = z_R$ is common in the waveguide laser literature. It is half the 'b' of chapter A2.1 and Kogelnik and Li [4].

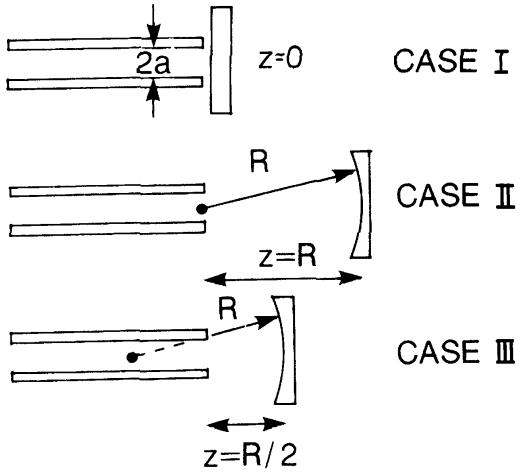


Figure A2.2.6. Low-loss coupling configurations for a fundamental waveguide mode.

In the far field, any launched field pattern will revert to a roughly spherical wave centred near the guide aperture. Case II is a far-field phase-matched reflector, for which any guide mode has low coupling loss. The poor mode discrimination and large z make it unpopular, unless the extra space is needed for intracavity elements.

Case I is used in most of today's commercial and scientific gas waveguide lasers: just a plane mirror at each end of the guide, placed against, or within a couple of millimetres of, the exit. There has been much study of the EH_{11} coupling loss for small (realistic) departures from perfect case I ($z = 0$, $R = \infty$, zero tilt), with surprising disagreements. Some simple and useful results are:

EH_{mn} coupling losses Γ_{11} ($z \ll b$). Degnan and Hall [7] give 38.4 $N^{-1.5}$% for circular guides, where $N = a^2/\lambda z$ is the reflector Fresnel number (numerical approximation to diffraction integral results). Boulnois and Agrawal [10] give = 33 $N^{-1.5}$% for square guides (asymptotic approximation); the same answer is obtained from a Hermite–Gaussian expansion [11].

EH_{11} self-coupling efficiencies for square guides ($N > 1$). Calculations by Boulnois and Agrawal [10] yield the 1-D result:

$$|c_{mm}|^2 \simeq \left(1 - \frac{m^2}{6N^{1.5}} - \frac{\pi m^4}{240N^{2.5}} + \frac{m^4}{72N^3}\right)^2 \tag{A2.2.20}$$

from which (keeping two terms) $\Gamma_{mn} = (1/6)(m^2 + n^2)N^{-1.5}$. Their analytical expressions, not given here, greatly reduce the computer time needed for square-bore coupling-loss calculations, if misalignment is not involved, and coupling amplitudes and cross-coupling coefficients are not required.

A2.2.4.2 Dual case I waveguide lasers

We recall that in a dual case I resonator a single round trip is very nearly equivalent to propagation along an undistorted guide of length $2L$, and the resonator transverse modes are essentially the transverse modes of the waveguide itself. The mode frequencies are found from a standard phase equation:

$$2\beta_{mn} n(\nu_{j,mn}) L = 2j\pi \tag{A2.2.21}$$

where the laser medium refractive index $n(\nu)$ includes a small anomalous dispersion term. By taking the difference between two mode frequencies, we can remove the axial mode integer $j \sim 10^5$. In the absence of

mode pulling ($n(\nu) = 1$) we have:

$$\nu_{j,mn} \simeq \frac{jc}{2L} + \frac{c\lambda}{32}\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) \qquad \text{(rectangular)} \tag{A2.2.22}$$

$$\nu_{j,mn} \simeq \frac{jc}{2L} + \frac{c\lambda\rho_{nm}^2}{8\pi^2 a^2} \qquad \text{(circular).} \tag{A2.2.23}$$

Equations (A2.2.22) and (A2.2.23) show roughly how the mode frequencies depend on waveguide geometry, although, in practice, the exact frequencies will depend on many other small corrections. In particular, the difference in frequency between two different transverse modes depends strongly on waveguide bore but weakly on waveguide length, in terms of typical manufacturing tolerances; the derivatives $\partial\nu/\partial a$ vary as a^{-3}.

By contrast with stable open resonators (chapter A2.1), these waveguide laser transverse-mode spacings vary as m^2 and n^2 or ρ_{nm}^2 and can exceed the axial mode spacing $c/2L$. Homogeneously broadened lasers tend, with some mostly unwelcome exceptions, to operate on the one mode (of all the available modes) with the highest ratio of small-signal gain to threshold loss. Now, if we fix $|\nu_{j,11} - \nu_{j,12}|$ at an integer multiple of $c/2L$, then the EH_{12} mode always coincides with a lower-loss EH_{11} mode and, *other things being equal*, EH_{11} must be preferred. This 'coincidence' of transverse-mode frequencies, if it can be achieved, will hold very well for different axial modes under the same line, and fairly well for different CO_2 lines. Formally, this implies:

$$L = 16sa^2/\lambda\Delta \qquad s = 1, 2, 3, 4, \ldots \qquad \text{(rectangular)} \tag{A2.2.24}$$

where $\Delta = |m^2 - m'^2 + n^2 - n'^2|$. This is for a square-bore dual case I laser and any two modes EH_{mn} and $EH_{m'n'}$. For the circular-bore dual case I device we obtain:

$$L = 4\pi^2 sa^2/\lambda(|\rho_{nm}^2 - \rho_{n'm'}^2|) \qquad \text{(circular)} \tag{A2.2.25}$$

for any two modes LP_{nm} and $LP_{n'm'}$. There is some evidence that such choices of guide geometry can improve the average laser mode quality by increasing the proportion of the active signature over which the fundamental mode is preferred.

It may be asked 'When is a case I ?' We can show from equations (A2.2.10) and (A2.2.20) that at $\lambda = 10\ \mu$m the plane-mirror coupling loss $\Gamma_{mn}(N \gg 1)$ becomes comparable with the first-order single-pass alumina guide loss $2\alpha_{mn}L$ when $z \simeq 0.18L^{-2/3}$ (z and L in mm), independent of guide width and nearly independent of mode number [12]. It seems reasonable to propose, as a working definition of a case I mirror, that z should be considerably less than this value, so that the coupling loss can be neglected. However, this typically requires $z < 1$ mm, and rules out many existing case I mirrors. Another reasonable definition is that a dual case I laser, where the mirror-guide distances z are gradually increased, ceases to be dual case I when the fundamental resonator mode has appreciable non-EH_{11} content.

There are no universally accepted definitions of mode quality. Formally, we would wish to know all the significant terms in the decompositions $\Sigma a_{mn}E_{mn}$ or $\Sigma b_{pq}\Psi_{pq}$, over the full laser signature and for a range of cavity perturbations. This is a most unrealistic goal. Many users are satisfied if a laser, after a brief warm-up, yields a 'reasonable' single-dot transverse intensity pattern with long-term power variations of a few per cent. In this case, extensive but simple beam measurements are adequate (see chapter A2.1 and section C5). Other users may set fierce limits on short-term frequency fluctuations, unwanted mode beating, and departure from pure EH_{11} or pure TEM_{00}. They may also demand wide tunability and no line-hopping or multi-lining.

A2.2.4.3 Rigrod analysis for waveguide lasers

A useful and deliberately simple model for the output power of a homogeneously broadened gas laser has been widely adopted in the guise of one or other form of the Rigrod equation. It allows comparisons of

waveguide resonator theory with experiment by expressing a measurable quantity, the laser output power, in terms of resonator losses. The gain is assumed constant along the length L of the laser and does not depend on direction; when homogeneous broadening is dominant it is fully described by

$$g(z) = \frac{g_0(\nu_0)}{1 + I_+ + I_-} - 2\alpha = \frac{1}{I_+}\frac{\mathrm{d}I_+}{\mathrm{d}z} = -\frac{1}{I_-}\frac{\mathrm{d}I_-}{\mathrm{d}z}. \tag{A2.2.26}$$

This $g(z)$ is the saturated gain in intensity per unit length experienced by both the *forward* and *backward* travelling plane waves in the amplifying section of the resonator. These plane-wave intensities are normalized to the line-centre saturation intensity $I_s(\nu_0)$, the intensity in whose presence the available gain drops to $g_0(\nu_0)/2$ (half the small-signal value). Thus:

$$I_+ \equiv I(\text{forward})/I_s(\nu_0) \qquad I_- \equiv I(\text{backward})/I_s(\nu_0).$$

There is a uniform loss constant 2α per unit length. By considering the boundary conditions defined by the mirror reflectances $R_1 = 1 - A_1 - T_1$ and $R_2 = 1 - A_2 - T_2$ (where A is the mirror loss and T the transmittance) and manipulating this gain equation, Rigrod obtained essentially this expression for line-centre output power when $\alpha = 0$:

$$P = \frac{I_s(\nu_0)A_b\sqrt{R_1T_2}}{(\sqrt{R_1} + \sqrt{R_2})(1 - \sqrt{R_1R_2})}[g_0(\nu_0)L - \ln(R_1R_2)^{-\frac{1}{2}}]. \tag{A2.2.27}$$

This represents the normalized intensity incident on the outcoupler, multiplied by (i) the normalization factor $I_s(\nu_0)$, (ii) the area of the incident beam A_b and (iii) the outcoupler transmittance T_2. Often there is only one outcoupler ($T_1 = 0$) and all the dissipative loss is lumped into one term A; then $R_1 = 1 - A$ and $R_2 = 1 - T$. To account for a small uniform guide loss, we can put $R_1 = 1 - A - 4\alpha L$. The full equation (A2.2.26) cannot be integrated explicitly for a resonator but Rigrod [13] derived a useful approximate form. For a homogeneously broadened laser with a Lorentzian linewidth $\Delta\nu$ and $f \equiv (\nu - \nu_0)/\Delta\nu$:

$$P_\nu = \frac{I_s(\nu_0)A_b\sqrt{R_1T_2}[(g_0(\nu_0) - 2\alpha(1 + 4f^2))L + (1 + 4f^2)\ln(R_1R_2)^{\frac{1}{2}}]}{(\sqrt{R_1} + \sqrt{R_2})(1 - \sqrt{R_1R_2})(1 - 2\alpha L/\ln(R_1R_2)^{\frac{1}{2}})} \tag{A2.2.28}$$

The waveguide beam area is usually defined as $A_b = \pi w_e^2$, where $w_e \equiv w_0/\sqrt{2}$ is the $1/e$-power radius of the EH_{11} approximating Gaussian (see section A2.2.4.1). For a square guide $A_b \simeq \pi(0.70/\sqrt{2})^2 \simeq 0.78a^2$. If the resonator mode *is* pure EH_{11}, the assumption of uniform A_b (and, hence, uniform active medium 'filling factor') will hold well—better than for open resonators with Gaussian beams of varying $w(z)$.

The waveguide distributed loss is 2α, not α as sometimes assumed in the literature. Unluckily, if the guide is narrow enough to force EH_{11}-like operation, then $2\alpha_{11}$ is often not greatly less than the *saturated* gain; but if it is wide enough to make $2\alpha_{11}$ very small, perturbations and lack of discrimination may introduce higher-order modes whose losses are not very small. Common dual case I CO_2 waveguide lasers tend to have lumped losses A of 1–3%; higher values naturally occur in more complex resonators. The best obtainable alumina waveguides seem to have losses of $2\alpha_{11} \geq 1\%\ \mathrm{m}^{-1}$ for $2a = 1.5$ mm, and the actual laser mode is never pure EH_{11}. Thus, it is usually unwise to ignore guide losses. Equation (A2.2.28) is simple and useful; it pays at least lip service to distributed losses; and in many cases, it can be fitted fairly well to measured power.

In a single-mode model, the coupling loss is considered as part of the lumped loss A. In a multi-mode model, or in non-trivial resonators with several sources of loss, the easiest solution is often to insert an average distributed loss $-(1/L)\ln|\gamma|$ in equation (A2.2.28). As mentioned in section A2.2.3.3, we hope that loss variations will 'wash out' along the resonator path, but few or no real lasers have had their dissipative loss distributions examined in detail.

Rigrod analysis is often used to extract best-fit values of laser parameters, such as the g_0, I_s and A. One parameter, such as T_2 or active gain length, is varied while the others are kept constant; and output power is recorded for various pressures (or gas mixes, input powers and so on). Even after much curve-fitting, the error bars are almost always quite large ($\sim \pm 10\%$ or worse). The equations are so simple that this is not surprising. What use is the idea of constant A_b, when we know that changing the pressure or overpumping the discharge will seriously alter the output mode (and hence A_b, and hence 2α, and probably the effective g_0 and I_s too) ? Why suppose that the loss is uniform in a waveguide laser, when we know that injecting an EH_{11} matched beam into a typical guide gives clear periodic variations in loss? The answer is that Rigrod analysis is *simple* and, though obviously deficient, *accurate enough to be useful* (see section A2.2.5) in understanding and designing waveguide laser resonators.

A2.2.5 Real waveguide resonators: experiment and theory

A2.2.5.1 Distant mirrors

From time to time, the versatile toolbox of waveguide resonator theory has been refined on real-life lasers. Some problems, such as a spatially resolved treatment of gain and saturation, or accurate determination of waveguide E_{mn} and k_{mn}, are very hard. A few are both interesting and not too difficult, at least initially, and are summarized here: plane mirrors, then tilted plane mirrors; intraline tunability and line selection; finally, unpleasant behaviour such as line hopping and resonator mode 'hooting'. Multi-mode predictions and detailed power/mode/frequency studies have rarely been made for real devices.

If we begin with a square-bore dual case I resonator, in theory the fundamental transverse mode is pure EH_{11} and the others are pure higher-order EH_{mn} (which can be ranked in order of increasing loss). As one mirror is moved away, the round-trip efficiency for a pure EH_{mn} mode is given by the product of guide transmission and mirror coupling efficiencies; if these are both near unity, we can write the round-trip loss as the sum of guide loss and coupling loss, or roughly $4\alpha_{mn}L + (1/6)(m^2+n^2)N^{-1.5}$ (see the similar small-term approximations, such as $R_1 = 1 - A - 4\alpha L$, introduced for the Rigrod treatment). But, however slightly at first, modes of the same parity now begin to couple among themselves. The fundamental resonator mode at a fixed wavelength is now whatever field minimizes the sum of guide loss and coupling loss. Thus, the resonator may do better than $4\alpha_{11}L + \Gamma_{11}$ by mixing in (say) some EH_{13}, EH_{15}, and so on. But it may also do worse, and the fundamental mode loss may be higher or lower than the pure EH_{11} loss.

This is an important illustration of the central issue. For given wavelength and geometry, the resonator can 'choose' only from the available set of self-consistent fields. The resonator cannot in this case choose pure EH_{11}, because pure EH_{11} is no longer a resonator mode—it ceased to be one when the mirror was moved.

If we fix L and vary z, the fundamental loss curve often shows non-monotonic behaviour (wiggles). This is confirmed by experiments where laser power fades when z is increased initially but revives when the plane mirror is pushed yet further away. If we fix z and vary L (easier to model than do), the fundamental loss curve often shows rather sharp periodic 'spiky' behaviour. The periods correspond with the 'beat wavelengths' over which pairs of important modes (notably EH_{11} and EH_{13}) change their relative guide propagation phase shifts by 2π. The published predictions of loss spikes [11, 14] are no doubt unrealistically sharp (see section A2.2.3.4) but experiments confirm that phase-period effects are important in real lasers—often more important than attenuation effects. This is further proof of shortness and fatness; if guides were very long and thin, high-order modes would be damped out efficiently.

There is another large literature based on beat wavelengths or phase periods in multi-mode waveguide structures. Simple relationships between lengths and mode-dependent phase shifts as in equation (A2.2.24), if they can be assumed to hold for multiple modes in long, straight and low-loss rectangular guides, lead to 'degenerate' total phase shifts (modulo 2π). For propagation, the spatial pattern formed from a coherent

sum of modes repeats itself at regular intervals; for resonators, the different transverse-mode frequencies coincide. There are also many sub-interval effects that scale with the guide Fresnel number $a^2/\lambda L$. These 'Talbot' or 'kaleidoscope' relations were (re)invented in the 1980s and applied to infrared hollow dielectric waveguides and slabs. They are probably more important nowadays in lasers, amplifiers, compact splitters and recombiners at shorter wavelengths around 1–2 μm, where the investments in technology are greater and the length scales $L \simeq a^2/\lambda$ are more convenient. The repetitions and coincidences, which would be exact in ideal guides that were lossless and otherwise obeyed first-order theory, are still impressive in imperfect real guides.

A2.2.5.2 Tilted mirrors and folded lasers

Some people, presented with a 'well-aligned' waveguide laser and an EH_{11}-like output, cannot resist 'tweaking' a mirror to be sure that it is well aligned. The immediate result is familiar: the power drops sharply, the mode becomes asymmetric or multi-dot, and things are never quite as good again. Tilt experiments are interesting for several reasons. First, of course, they are relatively easy, often easier than varying z or L. Figure A2.2.7 shows a typical arrangement. A tilted plane mirror is a fair first approximation to a laser diffraction grating (see A2.2.5.3). Interesting things tend to happen with tilts of order λ/a, which is typically several milliradians for CO_2 guides: microradian precision is not needed, and the important dispersive tilts in CO_2 lasers happen to be 2–8 mrad. A curved guide can readily be modelled as a sequence of short segments with small tilts between them. Moreover, tilting is a simple and controllable way to introduce and study non-EH_{11} guide modes.

Suppose we tilt an ideal case I plane mirror ($z = 0$) around its vertical diameter: call this a pure x-tilt φ_x. To first order, we get a linear phase shift $\exp(i2\pi\varphi_x/\lambda)$ across the aperture width. The EH_{mn} fields are sine and cosine functions and their tilt-coupling coefficients emerge as sums of sinc functions [5]. There is no 'y-coupling' between modes of different n. With circular guides (Bessel functions and polar coordinates), any tilt couples together modes with different n and different m.

For non-zero z, the tilt causes both a linear phase shift and a displacement Δx of the recoupling field, and numerical solution is generally required. As φ_x and Δx grow to (typically) several mrad and a significant fraction of the guide width, the number of modes needed for a given accuracy tends to increase sharply. The theoretical dependence of resonator loss on z or L, for modest and relevant tilt values such as 6–8 mrad, can be surprisingly acute, and 'case I' assumptions may be improper even for $z = 1$–2 mm [12]. Results for a small square-bore rf-excited laser are also shown in figure A2.2.7. The measured and predicted powers $P(\varphi_x)$ agree very well and the Rigrod values for g_0 and I_s fall within the usual (rather wide) limits.

Folded-waveguide CO_2 designs are popular because they squeeze long active lengths into compact devices and are well suited to slab-like rf excitation structures [15]. A V-fold or U-fold with plane mirrors is no harder to model than a distant plane mirror in a linear resonator. Curved folding mirrors can improve output mode quality but they cause astigmatism and the mode-coupling coefficients need heavy computation [16]. Folds with 'part waveguiding' or Brewster-cut prisms reduce the losses and increase the theoretical difficulty. In one example [17], the CO_2 resonator had two waveguides, a fold prism, a plane mirror and a curved mirror, a Brewster window *and* an electro-optic modulator. Measured and predicted powers are in fair agreement but this is not even a true multi-transverse-mode model, simply an iterative lumped-element approach to a real device where no closed-form solution exists.

A2.2.5.3 Tunability and line selection

For each available laser transition, a resonator has a set (perhaps small, perhaps very large) of potentially oscillating resonator modes. Many CO_2 devices can choose from several tens of distinct lines, each with 1–10 potential resonator modes. Well-behaved lasers ought to select the one mode/line combination offering the

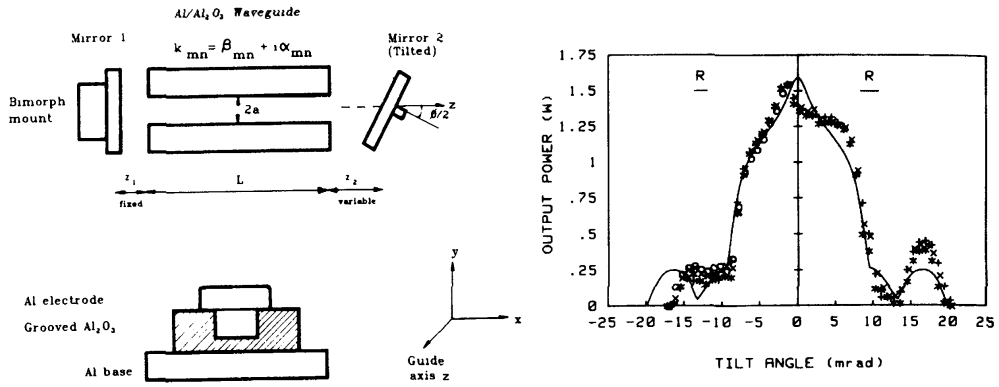


Figure A2.2.7. Square-bore rf-excited CO_2 waveguide laser with adjustable mirror mounts. (From Hill and Colley [12].)

highest ratio of small-signal gain to threshold loss; the competitive processes characteristic of homogeneous broadening should suppress all other laser frequencies. This is often a reasonable description of real lasers. The main problem of CO_2 tunability theory is then to calculate the relevant gains and losses and find how far away from line centre we may move our chosen mode (usually the fundamental mode) before it slips below threshold or another mode takes over. (There is a separate literature asking whether one can *formally* show that self-consistent fields exist, and that if they exist they are orthogonal, but we need not explore this here.)

Setting $P(\nu) = 0$ in equation (A2.2.28) gives the maximum possible continuous tuning range:

$$2|\nu - \nu_0|_{\text{zero power}} = \Delta\nu \left(\frac{g_0(\nu_0)L}{2\alpha L + \ln(R_1 R_2)^{-\frac{1}{2}}} - 1 \right)^{\frac{1}{2}}. \tag{A2.2.29}$$

In these simple terms, the modelling problem is just to find the average loss $2\alpha = -(1/L)\ln|\gamma|$. However, no systematic check of laser tuning ranges against multi-mode theory seems to have been performed. To model 'hopping' to other resonator modes under the same CO_2 line, we need to know their losses and relative frequencies. These depend on guide material constants (poorly known) and precise guide geometry (hard to measure accurately and, if distorted, hard to model). To model hopping to another line, we need to know the relative gains of the two lines and the precise positions of the resonator modes within the lineshapes.

In principle, we can probe the laser cavity for modes which, though still below threshold, are 'bubbling under' and threatening to lase; and this can be done neatly by reinjecting some of the laser's own output ([18, 19]; note that UK authors, but not usually US authors, will call this an 'autodyne' configuration). The total problem of mode selection and tunability still looks very difficult. Nevertheless it is important to see, at least qualitatively, why waveguide CO_2 lasers are especially prone to line hopping. Their ability to run cw at high pressures (200–400 torr) offers a linewidth $\Delta\nu$ of 1–2 GHz. It is possible, but rather difficult, to obtain reliable dual case I single-mode CO_2 tuning ranges of $\sim$1 GHz [20, 21]. Complex cavities can raise this to $\sim$1.5 GHz. Usually, unwanted lines or modes break in after a few hundred MHz. With narrow guides only 100–300 grating lines are typically illuminated but 'loss of resolution' is an incomplete explanation.

Lasing on a desired line, on the fundamental mode in a dual case I grating-tuned laser, will cease for one of three main reasons. We cannot exceed $|\nu - \nu_0| = c/4L$ because an adjacent axial mode will take over. We may fall below threshold according to equation (A2.2.29). Or another mode may achieve a higher gain/loss ratio. If this is a mode under the same line, we can infer the mode frequency separation from the laser signature or from heterodyne measurements, estimate the mode losses, and hope that the experimental mode-hopping point coincides with the theoretical point where the gain/loss ratios are equal. This seems a straightforward test of theory. It is not, and is seldom if ever tried. Real lasers do not obey ideal theory and unwanted modes tend to be seen as nuisances to be eliminated, not as resonator features to be studied.

A2.2.5.4 Resonator mode degeneracies: hopping and hooting

So far we have accepted that a cw CO_2 waveguide laser chooses, from instant to instant, the one resonator mode with the highest ratio of small-signal gain to threshold loss. All other potential modes are frozen out by rapid collisional-broadening effects in the high-pressure active medium. At many times and for many devices this is not true. Real gas discharge waveguide lasers are imperfect and cannot suppress other modes unless the gain/loss ratios are *clearly* larger than that of the fundamental mode; but what *clearly* means for a specific laser depends on molecular excitation rates and transverse mode spatial overlaps, and is very hard to quantify. The rate equations can be readily solved only for cases well removed from real life. In very simple terms, the laser cannot be expected to run reliably single-frequency when two equal-loss modes lie within a few MHz; or when mode hopping is imminent and the (gain/loss) ratios are very nearly equal; or, similarly, when line hopping is imminent. Thus we may see more than one frequency emitted on one line, or more than one line emitted, or both; and this is highly irritating to many users.

If, with a multi-mode resonator model, we plot the losses for the first few resonator modes as functions of z or φ_x, comparing curves for several known CO_2 wavelengths, it is tempting to identify regions of curve-crossing or near-coincidence with observations of multi-frequency output. This cannot be pushed too far, because even the multi-mode approaches outlined here are highly idealized and restricted to empty resonators, whereas multi-frequency waveguide CO_2 lasers are (by our definitions) non-ideal and badly behaved. For instance, we have not tried to model any effect that the two or more resonating modes have on each other; that is a nonlinear problem needing much more work. Despite these reservations, studies of cw mode beating ('hooting') and line-hopping provide encouraging support for multi-mode methods.

A2.2.6 Summary

The general aim of this chapter has been to offer readers a brief and not mathematically detailed review of waveguide laser resonators, so that the literature can be approached with a knowledge of the theory's present limitations, and without a few common misconceptions. We began by introducing *modes of propagation* (both free-space and waveguide). Resonator modes are self-consistent field patterns, with certain definite round-trip phase shifts, which are potentially available to the laser cavity as *modes of oscillation*; and each is associated with a *single* frequency. The presence in the laser field of several guide modes does not imply several different oscillation frequencies.

From our point of view there is nothing special about *waveguide* resonators as such. It is possible, and may sometimes help, to see them as conventional open resonators with more or less strongly perturbing 3D apertures. The transition from free-space propagation to waveguiding, as the perturbation strengthens, is not necessarily sharp. And we stress that many existing commercial and scientific 'waveguide' devices are not safely clear of this transition region; still less are they safely within the domain of single-mode theory. They are *short and fat*. This forces us towards a multi-mode treatment if we want reasonably accurate performance comparisons and predictions.

Current theory cannot explain all our experimental observations. Nevertheless, the steady expansion of waveguide laser technology in industrial, medical and remote-sensing applications should prompt continuing improvements in the depth and accuracy of waveguide resonator theory.

References

[1] Marcatili E A J and Schmeltzer R A 1964 Hollow metallic and dielectric waveguides for long distance optical transmission and lasers *Bell Syst. Tech. J.* **43** 1783–809
[2] Degnan J J 1976 The waveguide laser: a review *Appl. Phys.* **11** 1–33
[3] Snitzer E 1961 Cylindrical dielectric waveguide modes *J. Opt. Soc. Am.* **51** 491–8
[4] Kogelnik H and Li T 1966 Laser beams and resonators *Appl. Opt.* **15** 1550–67
[5] Hill C A and Hall D R 1985 Coupling loss theory of single-mode waveguide resonators *Appl. Opt.* **24** 1283–90

[6] Abrams R L and Chester A N 1974 Resonator theory for hollow waveguide lasers *Appl. Opt.* **13** 2117–25
[7] Degnan J J and Hall D R 1973 Finite-aperture waveguide-laser resonators *IEEE J. Quantum Electron.* **QE-9** 901–10
[8] Abrams R L and Bridges W B 1973 Characteristics of sealed-off waveguide CO_2 lasers *IEEE J. Quantum Electron.* **QE-9** 940–6
[9] Abrams R L 1972 Coupling losses in a hollow waveguide laser resonator *IEEE J. Quantum Electron.* **QE-8** 838–43
[10] Boulnois J-L and Agrawal G P 1982 Mode discrimination and coupling losses in rectangular waveguide resonators with conventional and phase-conjugate mirrors *J. Opt. Soc. Am.* **72** 853–60
[11] Hill C A 1988 Transverse modes of plane-mirror waveguide lasers *IEEE J. Quantum Electron.* **QE-24** 1936–46
[12] Hill C A and Colley A D 1990 Misalignment effects in a CO_2 waveguide laser *IEEE J. Quantum Electron.* **QE-26** 323–8
[13] Rigrod W W 1978 Homogeneously broadened cw lasers with uniform distributed loss *IEEE J. Quantum Electron.* **QE-14** 377–81
[14] Gerlach R, Wei D and Amer N M 1984 Coupling efficiency of waveguide laser resonators formed by flat mirrors: analysis and experiment *IEEE J. Quantum Electron.* **QE-20** 948–63
[15] Newman L A and Hart R A 1987 Recent R&D advances in sealed-off CO_2 lasers *Laser Focus/Electro-opt.* 80–96
[16] Banerji J, Davies A R, Hill C A, Jenkins R M and Redding J R 1995 Effects of curved mirrors in waveguide resonators *Appl. Opt.* **34** 3000–8
[17] Hill C A, Pearson G N, Tapster P, Vaughan J M and Miller G M 1996 Polarization states and output powers of a CO_2 laser with an electro-optic phase retarder *Appl. Opt.* **35** 5381–5
[18] Pearson G N, Harris M, Hill C A, Vaughan J M and Homby A M 1995 Inter-transverse-mode injection locking and subthreshold gain measurements in a CO_2 waveguide laser *IEEE J. Quantum Electron.* **QE-31** 1064–8
[19] Shackleton C J, Loudon R, Hill C A, Shepherd T J, Harris M and Vaughan J M 1995 Transverse modes above and below threshold in a single-frequency laser *Phys. Rev.* A **52** 4908–20
[20] Abrams R L 1974 Gigahertz tunable waveguide CO_2 laser *Appl. Phys. Lett.* **25** 304–6
[21] Gonchukov S A, Kornilov S T and Protsenko E D 1978 Tunable waveguide laser *Sov. Phys.—Tech. Phys.* **23** 1084–6

Reviews

Abrams R L 1979 *Laser Handbook* vol 3, ed M L Stitch (Amsterdam: North-Holland) pp 41–88
Hall D R and Hill C A 1987 *Handbook of Molecular Lasers* ed P K Cheo (New York and London: Marcel Dekker) pp 165–258
Hill C A 1989 Theory of waveguide laser resonators *The Physics and Technology of Laser Resonators* ed D R Hall and P E Jackson (Bristol: Adam Hilger) (This forms the basis of the present chapter.)
Smith P W, Wood O R II, Maloney P J and Adams C R 1981 Transversely excited waveguide gas lasers *IEEE J. Quantum Electron.* **17** 1166–81

Other reading

Bel'tyugov V N, Gracheva E V, Kuznetsov A A, Ochkin V N, Sobolev N N, Troitskii Yu V and Udalov Yu B 1988 Frequency selectivity of a multimode waveguide gas laser with a diffraction grating *Sov. J. Quantum Electron.* **18** 599–604
Chester A N and Abrams R L 1972 Mode losses in hollow waveguide lasers *Appl. Phys. Lett.* **21** 576–8
Degnan J J 1973 Waveguide laser mode patterns in the near and far field *Appl. Opt.* **12** 1026–30
Henderson D M 1976 Waveguide lasers with intracavity electrooptic modulators: misalignment loss *Appl. Opt.* **15** 1066–70
Hill C A and Hall D R 1986 Waveguide resonators with a tilted mirror *IEEE J. Quantum Electron.* **QE-22** 1078–87
Hill C A, Redding J R and Colley A D 1990 Multimode treatment of misaligned CO_2 waveguide lasers *J. Mod. Opt.* **37** 473–81
Laakmann K D and Steier W H 1976 Waveguides: characteristic modes of hollow rectangular dielectric waveguides *Appl. Opt.* **15** 1334–40
Merkle G and Heppner J 1983 CO_2 waveguide laser with Fox-Smith mode selector *IEEE J. Quantum Electron.* **QE-19** 1663–7
Roullard F P III and Bass M 1977 Transverse mode control in high gain, millimeter bore, waveguide lasers *IEEE J. Quantum Electron.* **QE-13** 813–19
Smith P W 1971 A waveguide gas laser *Appl. Phys. Lett.* **19** 132–4
Tang F and Henningsen J O 1987 Conditions for single-line and single-mode tuning of a CO_2 waveguide laser *Appl. Phys.* B **44** 93–8

A3
Laser beam control

Jacky Byatt

In most laser applications it is necessary to focus, modify or shape the laser beam by using lenses and other optical elements. In general, laser-beam propagation can be approximated by assuming that the laser beam has an ideal Gaussian intensity profile, corresponding to the theoretical TEM_{00} mode. Coherent Gaussian beams have transformation properties that require special consideration. The output from single-mode (TEM_{00}) lasers is highly Gaussian (helium–neon lasers and argon–ion lasers are very good examples). In contrast, high-power lasers often operate with many modes and are highly non-Gaussian. The propagation of non-Gaussian lasers is discussed later in this chapter.

The basic properties of laser resonators and Gaussian laser beams are discussed in chapter A2. In the fundamental TEM_{00} mode, the beam emitted from a laser forms as a Gaussian transverse irradiance profile as shown in figure A3.1. The Gaussian shape is truncated at some diameter either by the internal dimensions of the laser or by some limiting aperture in the optical train.

A3.1 Transforming a Gaussian beam with simple lenses

It is clear from chapter A2 that Gaussian beams transform differently to non-Gaussian beams. Siegman [1] uses matrix transformations to treat the general problem of Gaussian beam propagation with lenses and mirrors. A less rigorous, but in many ways more insightful, approach to this problem has been developed by Self [2]. Self shows a method to model transformations of a laser beam through simple optics, under paraxial conditions, by calculating the Rayleigh range and beam waist location following each individual optical element. These parameters are calculated using a formula analogous to the well-known standard lens formula.

The standard lens equation is written as

$$\frac{1}{s} + \frac{1}{s''} = \frac{1}{f} \tag{A3.1}$$

where s is the object distance, s'' is the image distance and f is the focal length of the lens. For Gaussian beams, Self has derived an analogous formula by assuming that the waist of the input beam represents the object, and the waist of the output beam represents the image. The formula is expressed in terms of the Rayleigh range of the input beam.

In the regular form

$$\frac{1}{s + z_R^2/(s - f)} + \frac{1}{s''} = \frac{1}{f} \tag{A3.2}$$

or, in dimensionless form

$$\frac{1}{(s/f) + (z_R/f)^2/(s/f - 1)} + \frac{1}{s''/f} = 1. \tag{A3.3}$$

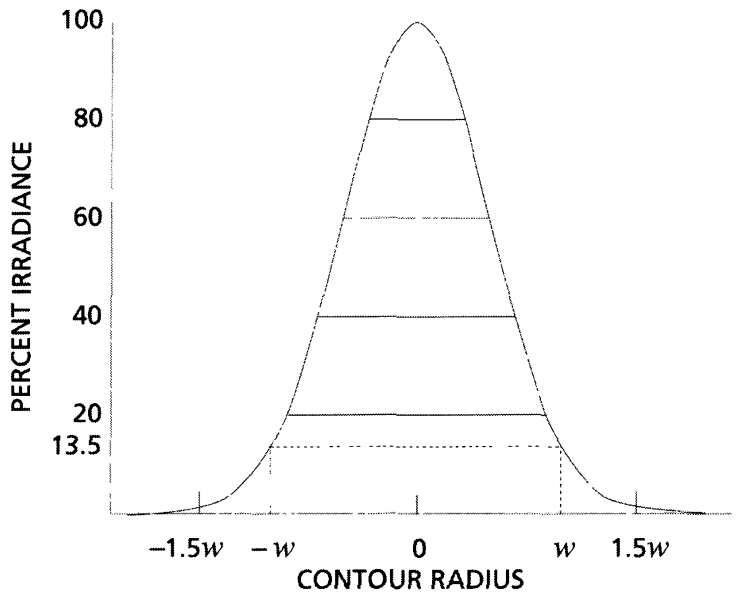


Figure A3.1. Irradiance profile of the fundamental TEM_{00} mode.

In the far-field limit as $z_R \rightarrow 0$ this reduces to the geometric optics equation. A plot of s/f *versus* s''/f for various values of z_R/f is shown in figure A3.2. For a positive thin lens, the three distinct regions of interest correspond to real object and real image, real object and virtual image, and virtual object and real image.

The main differences between Gaussian beam optics and geometric optics, highlighted in such a plot, can be summarized as follows.

- There are both a maximum and a minimum image distance for Gaussian beams.
- The maximum image distance occurs at $s = f + z_R$, rather than at $s = f$.
- There is a common point in the Gaussian beam expression at $s/f = s''/f = 1$. For a simple positive lens, this is the point at which the incident beam has a waist at the front focus and the emerging beam has a waist at the rear focus.
- A lens appears to have a shorter focal length as z_R/f increases from zero (i.e. there is a Gaussian focal shift).

Self recommends calculating z_R, w_0 and the position of w_0 for each optical element in the system in turn so that the overall transformation of the beam can be calculated. To carry this out, it is also necessary to consider magnification: w_0''/w_0. The magnification is given by

$$m = \frac{w_0''}{w_0} = \frac{1}{\sqrt{\{[1-(s/f)]^2 + (z_R/f)^2\}}}. \tag{A3.4}$$

The Rayleigh range of the output beam is then given by

$$z_R'' = m^2 z_R. \tag{A3.5}$$

All the formulae are written in terms of the Rayleigh range of the input beam. Unlike the geometric case, the formulae are not symmetric with respect to input and output beam parameters. For backtracing beams, it is useful to know the Gaussian beam formula in terms of the Rayleigh range of the output beam:

$$\frac{1}{s} + \frac{1}{s'' + z_R''^2/(s''-f)} = \frac{1}{f}. \tag{A3.6}$$

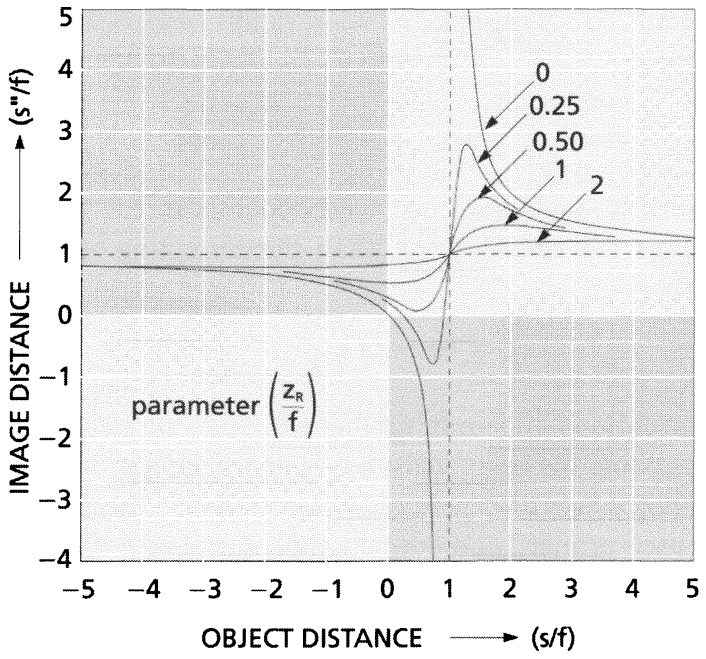


Figure A3.2. Plot of lens formula for Gaussian beams with normalized Rayleigh range of the input beam as the parameter.

A3.1.1 Beam concentration

The spot size and focal position of a Gaussian beam can be determined from the previous equations. Two cases of particular interest occur when $s = 0$ (the input waist is at the first principal surface of the lens system) and $s = f$ (the input waist is at the front focal point of the optical system). For $s = 0$, we get

$$s'' = \frac{f}{1 + (\lambda f/\pi w_0^2)^2} \tag{A3.7}$$

and

$$w = \frac{\lambda f/w_0}{[1 + (\lambda f/\pi w_0^2)]^{1/2}}. \tag{A3.8}$$

For the case of $s = f$, the equations for image distance and waist size reduce to the following:

$$s'' = f$$

and

$$w = \lambda f/\pi w_0.$$

Substituting typical values into these equations yields nearly identical results and, for most applications, the simpler, second set of equations can be used.

A3.1.1.1 Calculating a correcting surface

In cases where the laser uses a partially-transmitting output mirror, the first lens seen by the laser beam is the output mirror itself. The beam is refracted as it passes through the second surface of the output mirror. If the mirror has a flat second surface, the apparent beam waist moves closer to the mirror and the divergence is increased. To counteract this, laser manufacturers often put a radius on the second surface to collimate the beam by making a waist at the output mirror, as shown in the case of a typical helium–neon laser cavity consisting of a flat high reflector and an output mirror with a radius of curvature of 20 cm separated by 15 cm.

If the laser is operating at 633 nm, the intrinsic beam waist radius (w_0), beam output radius (w_{200}), and beam half-angle divergence θ are

$$w_0 = 0.13 \text{ mm} \qquad \omega_{200} = 0.26 \text{ mm} \qquad \text{and} \qquad \theta = 1.5 \text{ mrad}$$

however, with a flat second surface, the divergence nearly doubles to 2.8 mrad. By solving equation (A3.7) for f, with $s'' = 15$ cm, we see that the focal length of the correcting output coupler should be 15.1 cm. Using the lens-makers formula

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{A3.9}$$

with the appropriate sign convention and assuming that $n = 1.5$, we get a convex correcting curvature of approximately 5.5 cm. At this point, the beam waist has been transferred to the output coupler, with a radius of 0.26 mm, and the far-field half-angle divergence is reduced to 0.76 mrad, a factor of nearly four.

Correcting surfaces are used primarily on output couplers whose radius of curvature is a metre or less. For longer radius output couplers, the effects are much less dramatic.

A3.1.1.2 Depth of focus

Depth of focus ($\pm\Delta z$), that is, the range in image space over which the focused spot diameter remains below an arbitrary limit, can be derived using the following equation

$$w(z) = w_0\left[1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2\right]^{1/2}. \tag{A3.10}$$

The first step in performing a depth-of-focus calculation is to set the allowable degree of spot size variation. If we choose a typical value of 5%, or $w(z) = 1.05w_0$, and solve for $z = \Delta z$, the result is

$$\Delta z = \pm\frac{0.32\pi w_0^2}{\lambda}.$$

Since the depth of focus is proportional to the square of focal spot size, and focal spot size is directly related to f-number, the depth of focus is proportional to the square of the f-number of the focusing system.

A3.1.2 Truncation

In a diffraction-limited lens, the diameter of the image spot is

$$d = K \times \lambda \times f/\# \tag{A3.11}$$

where K is a constant dependent on truncation ratio (the ratio of the Gaussian beam diameter to the limiting aperture diameter) and pupil illumination, λ is the wavelength of light, and $f/\#$ is the speed of the lens at truncation. The intensity profile of the spot is strongly dependent on the intensity profile of the radiation filling the entrance pupil of the lens. For uniform pupil illumination, the image spot takes on an Airy disc intensity profile as shown in figure A3.3. If the pupil illumination is Gaussian in profile, an image spot of Gaussian profile results as shown in figure A3.4. When the pupil illumination is between these two extremes, a hybrid intensity profile results.

In the case of the Airy disc, the intensity falls to zero at the point $d_{\text{zero}} = 2.44 \times \lambda \times f/\#$, defining the diameter of the spot (see figure A3.3). When the pupil illumination is not uniform, the image spot intensity

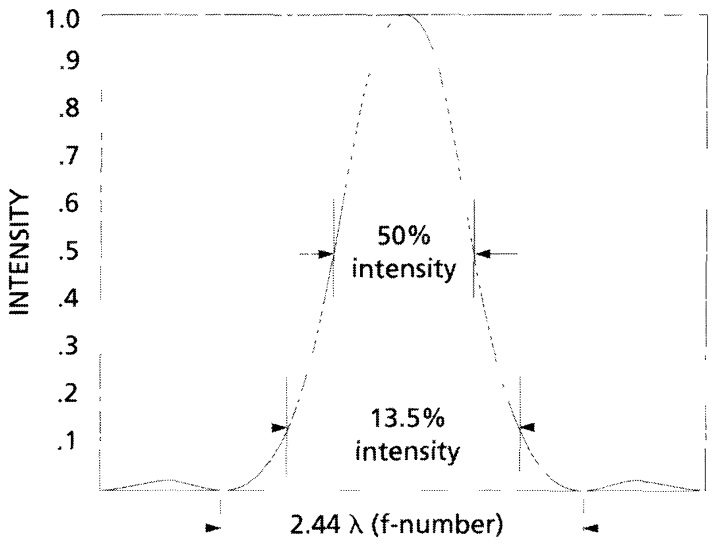


Figure A3.3. Airy disc intensity distribution at the image plane.

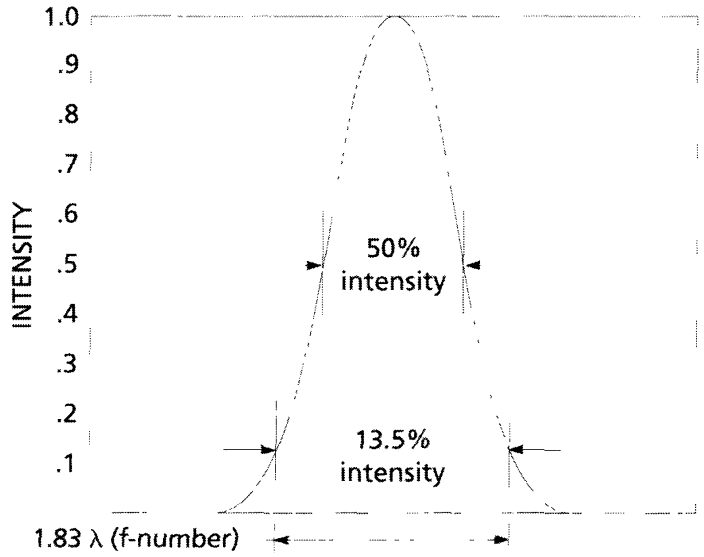


Figure A3.4. Gaussian intensity distribution at the image plane.

never falls to zero, making it necessary to define the diameter at some other point. This is commonly done for two points:

$$d_{\mathrm{FWHM}} = 50\% \text{ intensity point}$$

and

$$d_{1/e^2} = 13.5\% \text{ intensity point.}$$

It is helpful to introduce the truncation ratio

$$T = \frac{D_{\mathrm{b}}}{D_{\mathrm{t}}} \tag{A3.12}$$

where D_{b} is the Gaussian beam diameter measured at the $1/e^2$ intensity point and D_{t} is the limiting aperture diameter of the lens. If $T = 2$, which approximates uniform illumination, the image spot intensity profile approaches that of the classic Airy disc. When $T = 1$, the Gaussian profile is truncated at the $1/e^2$ diameter,

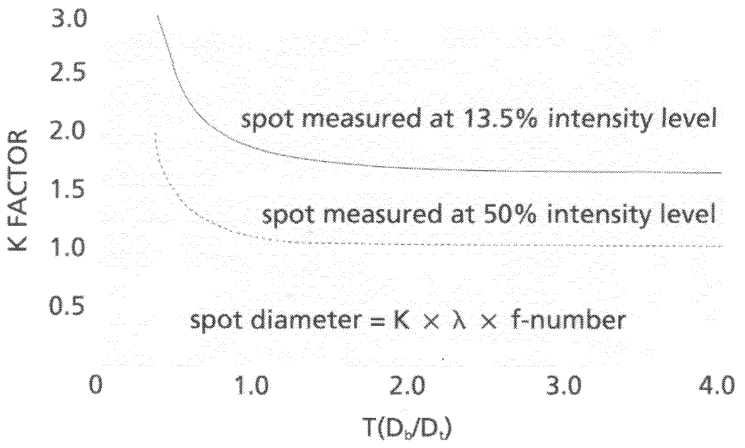


Figure A3.5. *K* factors as a function of truncation ratio.

and the spot profile is clearly a hybrid between an Airy pattern and a Gaussian distribution. When $T = 0.5$, which approximates the case for an untruncated Gaussian input beam, the spot intensity profile approaches a Gaussian distribution.

Calculation of spot diameter for these or other truncation ratios requires that K is evaluated. This is accomplished with the formulae

$$K_{\mathrm{FWHM}} = 1.029 + \frac{0.7125}{(T - 0.2161)^{2.179}} - \frac{0.6445}{(T - 0.2161)^{2.221}} \tag{A3.13}$$

and

$$K_{1/e^2} = 1.6449 + \frac{0.6460}{(T - 0.2816)^{1.821}} - \frac{0.6445}{(T - 0.2816)^{1.891}}. \tag{A3.14}$$

The K function, plotted in figure A3.5, permits calculation of on-axis spot diameter for any beam truncation ratio.

The optimal choice for truncation ratio depends on the relative importance of spot size, peak spot intensity, and total power in the spot as demonstrated in table A3.1. The total power loss, P_{L}, from the spot after propagating through an aperture can be calculated by using

$$P_{\mathrm{L}} = \mathrm{e}^{-2(D_{\mathrm{t}}/D_{\mathrm{b}})^2} \tag{A3.15}$$

for a truncated Gaussian beam. A good compromise between power loss and spot size is often a truncation ratio of one. When $T = 2$ (approximately uniform illumination), fractional power loss is 60%. When $T = 1$, d_{1/e^2} is just 8.0% larger than when $T = 2$, while fractional power loss is down to 13.5%. Because of this large saving in power with relatively little growth in the spot diameter, truncation ratios of 0.7 to 1.0 are typically used. Ratios as low as 0.5 might be employed when laser power must be conserved. However, this low value often wastes too much of the available clear aperture of the lens.

The mathematics of the effects of truncation on a real-world laser beam are beyond the scope of this chapter. For an in-depth treatment of this problem, please refer to [3].

A3.1.3 Non-Gaussian laser beams

In the real world, perfectly Gaussian laser beams are very hard to find. Low-power beams from helium–neon lasers can be a close approximation, but the higher the power of the laser and the more complex the excitation mechanism (e.g. transverse discharges, flash-lamp pumping), or the higher the order of the mode, the more the beam deviates from the ideal.

Table A3.1. Spot diameters and fractional power loss for three values of truncation.

Truncation ratio	d_{FWHM}	d_{1/e^2}	d_0	P_{L} (%)
∞	1.03	1.64	2.44	100
2.0	1.05	1.69	—	60
1.0	1.13	1.83	—	13.5
0.5	1.54	2.51	—	0.03

As discussed in section A2.1.4.3, the M^2 factor has come into general use to address the issue of non-Gaussian beams. For a theoretical Gaussian beam, the value of the radius-divergence product is

$$w_0\theta = \lambda/\pi. \tag{A3.16}$$

For a real laser beam, we have

$$w_{0M}\theta_M = M^2\lambda/\pi > \lambda/\pi. \tag{A3.17}$$

where w_{0M} and θ_M are the $1/e^2$ intensity waist radius and the far-field half-divergence angle of the real laser beam, respectively. For a typical helium–neon laser operating in TEM_{00} mode, $M^2 < 1.1$. Ion lasers typically have an M^2 factor ranging from 1.1 to 1.7. For high-energy multimode lasers, the M^2 factor can be as high as 20 or 30. In all cases, the M^2 factor affects the characteristics of a laser beam and cannot be neglected in optical designs, and truncation, in general, increases the M^2 factor of the beam. M^2 factors into the equations for beam diameter and wavefront radius as follows:

$$w_M(z) = w_{0M}[1 + (z\lambda M^2/\pi w_{0M}^2)^2]^{1/2} \tag{A3.18}$$

and

$$R_M(z) = z\lfloor 1 + (\pi w_{0M}^2/z\lambda M^2)^2\rfloor. \tag{A3.19}$$

The definition for the Rayleigh range remains the same for a real laser beam and becomes

$$z_{\mathrm{R}} = \pi w_{0M}^2/\lambda \tag{A3.20}$$

and the lens equation (A3.6) becomes

$$\frac{1}{s + (z_{\mathrm{R}}/M^2)/(s - f)} + \frac{1}{s''} = \frac{1}{f} \tag{A3.21}$$

or, in normalized fashion [4]

$$\frac{1}{(s/f) + (z_{\mathrm{R}}/M^2 f)^2/(s/f - 1)} + \frac{1}{s''/f} = 1. \tag{A3.22}$$

A3.2 Transverse modes and mode control

The fundamental TEM_{00} mode is only one of many transverse modes that satisfy the round-trip propagation criteria shown in equation (A3.2) (see section A2.1.4). Figure A3.6 shows examples of the primary lower-order Hermite–Gaussian (rectangular) modes. Note that the subscripts m and n in the eigenmode TEM_{mn} are correlated to the number of nodes in the x and y directions. In each case, adjacent lobes of the mode are 180° out of phase.

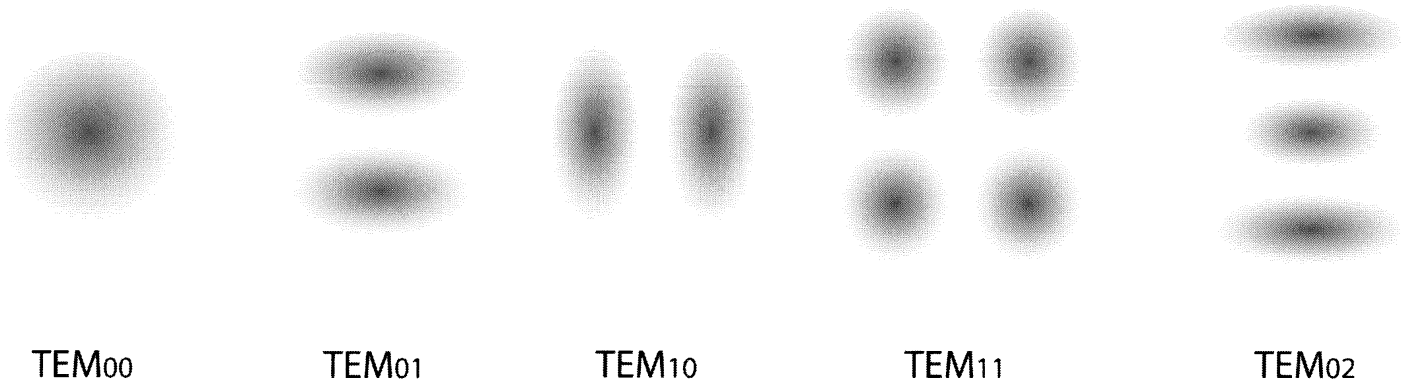


Figure A3.6. Low-order Hermite–Gaussian resonator modes.

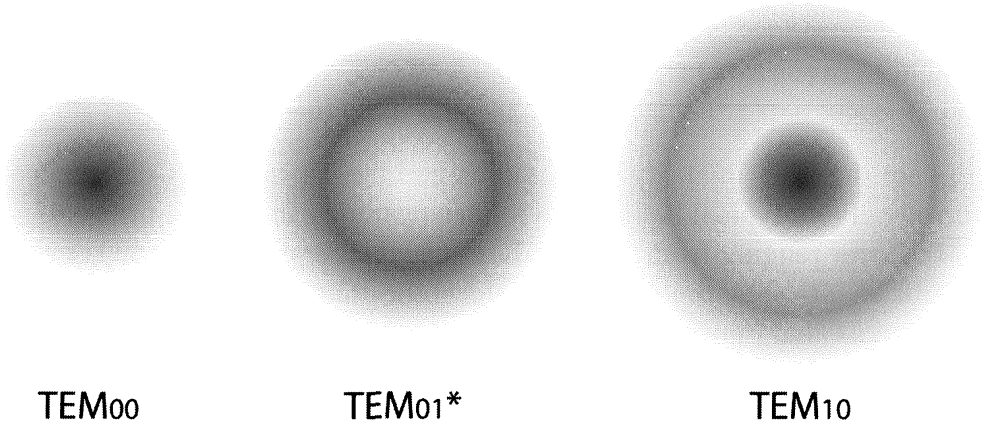


Figure A3.7. Low-order axisymmetric Laguerre–Gaussian resonator modes.

The propagation equation can also be written in cylindrical form in terms of radius (ρ) and angle (ϕ). The eigenmodes ($E_{\rho\phi}$) for this equation are a series of axially symmetric modes, which, for stable resonators, are closely approximated by Laguerre–Gaussian functions, denoted by $\mathrm{TEM}_{\rho\phi}$. For the lowest order mode, TEM_{00}, the Hermite–Gaussian and Laguerre–Gaussian functions are identical, but for higher order modes, they differ significantly, as shown in figure A3.7.

The mode, TEM^*_{01}, also known as the 'bagel' or 'doughnut' mode, is considered to be a superposition of the Hermite–Gaussian TEM_{10} and TEM_{01} modes, locked in phase quadrature [5].

In real-world lasers, the Hermite–Gaussian modes predominate since strain, slight misalignment or contamination on the optics tends to drive the system toward rectangular coordinates. Nonetheless, the Laguerre–Gaussian TEM_{10} 'target' or 'bulls-eye' mode is clearly observed in well-aligned gas-ion and helium–neon lasers with the appropriate limiting apertures.

A3.2.1 Mode control

The transverse modes for a given stable resonator each have different beam diameters and divergences. The lower order the mode, the smaller the beam diameter, the lower the M^2 value and, although it is not intuitively obvious, the narrower the far-field divergence angle (the divergence is determined by the size of the uniphase lobes making up the higher-order mode, not the size of the mode itself). For example, the TEM^*_{01} doughnut mode is approximately 1.5 times greater than the diameter of the fundamental TEM_{00} mode, and the Laguerre TEM_{10} target mode is twice the diameter of the TEM_{00} mode. The theoretical M^2 values for the TEM_{00},

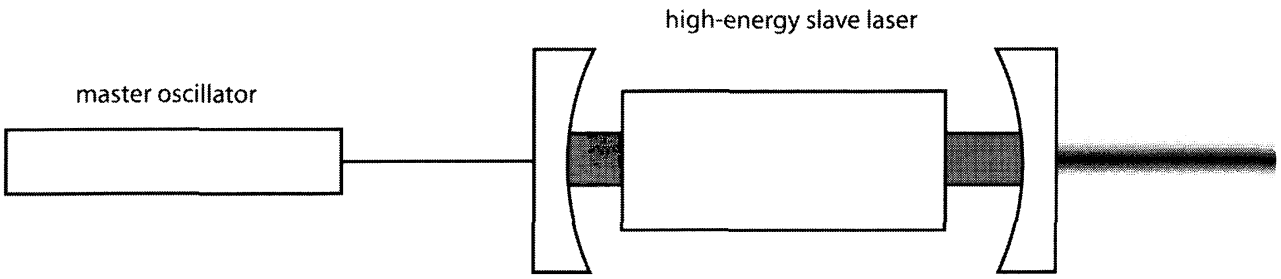


Figure A3.8. Configuration for laser injection locking.

TEM^*_{01} and TEM_{10} modes are 1.0, 2.3 and 3.6, respectively [6]. Because of its smooth intensity profile, low divergence and ability to be focused to a diffraction-limited spot, it is usually desirable to operate in the lowest order mode possible, TEM_{00}. Lasers, however, tend to operate at the highest order mode possible, either in addition to, or instead of TEM_{00}, because the larger beam diameter may allow them to extract more energy from the lasing medium.

The primary method for reducing the order of the lasing mode is to add sufficient loss to the higher-order modes so that they cannot oscillate, without significantly increasing the losses at the desired, lower-order mode. In most lasers this is accomplished by placing a fixed or variable aperture inside the laser cavity. Because of the significant differences in beam diameter, the aperture can cause significant diffraction losses for the higher-order modes without impacting the lower-order modes. As an example, consider the case of a typical argon–ion laser with a long-radius cavity and a variable mode-selecting aperture.

When the aperture is fully open, the laser oscillates in the TEM_{10} target mode. As the aperture is slowly reduced, the output changes smoothly to the TEM^*_{01} doughnut mode and, finally, to the TEM_{00} fundamental mode. (Many gas lasers will support only one mode at a time due to the nature of the discharge. Other lasers can have several modes operating simultaneously.)

In many lasers, the limiting aperture is provided by the geometry of the laser itself. For example, by designing the cavity of a helium–neon laser so that the diameter of the fundamental mode at the end of the laser bore is approximately 60% of the bore diameter, the laser will naturally operate in the TEM_{00} mode.

A3.2.2 Injection locking

In high-power high-gain lasers, suppressing higher-order modes with an aperture can be very difficult, if not impossible. One technique developed to address this problem is 'injection locking'.

An injection-locking scheme is shown in the figure A3.8. The relatively weak output of a well-controlled 'master' laser, operating at frequency ω_1 and intensity I_1, is fed into the cavity of a freely oscillating, higher-power, uncontrolled 'slave' laser, operating at frequency ω_0 and intensity I_0. If the frequency ω_1 is tuned very close to ω_0, the coherent photons from the master laser will compete with and overwhelm the ω_0 'noise build up' in the slave for the oscillator gain, 'locking' the output of the slave to that of the master, with output at frequency ω_1 and intensity I_0.

According to Siegman, the full locking range for the oscillator is given by

$$\Delta\omega_{\text{lock}} = 2(\omega_1 - \omega_0) \approx \frac{2\omega_0}{Q_e}\sqrt{\frac{I_1}{I_0}} \tag{A3.23}$$

where Q_e is a quality factor describing the losses from the resonator mirror and cavity configuration (for more information see Q-switching, later). As can be seen from the equation, the higher the intensity of the master laser beam, the wider the locking range.

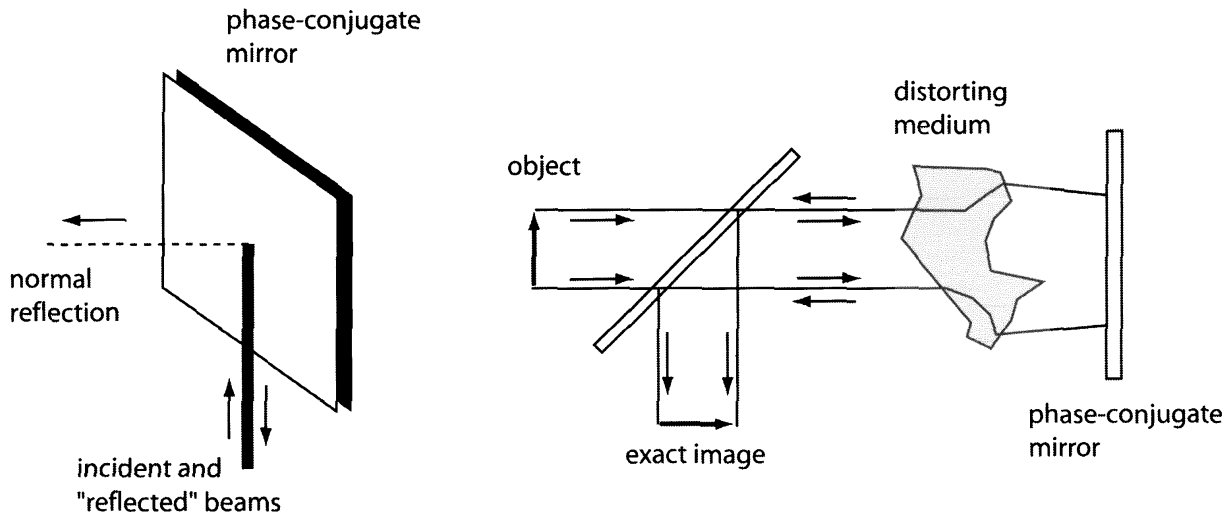


Figure A3.9. 'Reflection' properties of a phase-conjugate mirror.

A similar technique, called injection seeding, is used with high-energy pulsed lasers. In this case, the beam from a well-controlled cw laser replaces the 'noise' around which the pulsed laser builds up its output.

A3.2.3 Mode control with phase-conjugate mirrors

A major problem encountered in high-power and high-energy lasers and laser amplifiers is the wavefront distortion caused by inhomogeneities in the laser medium and optical cavity. These inhomogeneities can result from high-speed turbulence, naturally occurring impurities in the lasing medium, or thermal effects caused by the excitation method or the laser beam itself. Wavefront distortion and mode purity can be dramatically reduced by using phase-conjugate mirrors.

Optical phase conjugation is the generic name for a variety of nonlinear optical processes that are capable of 'reflecting' waves in such a manner that both the direction of propagation and the phase for each component of the wave is exactly reversed (see section A4.6). This is illustrated in figure A3.9. Whereas, with a regular mirror, only a ray normal to the mirror is reflected back upon itself, with a phase-conjugate mirror, *all* rays are reflected back upon themselves. Furthermore, if the wave passes through a distorting element on its way to the phase-conjugate mirror, the 'reflected' wave, after passing back through the distorting element, has exactly the same phase characteristics as the incident wave.

This has profound implications for high-energy laser applications, because, by using a phase-conjugate mirror as part of the laser cavity, or as a reflector in a laser amplifier, the distorting effects of the laser media can be effectively eliminated (see chapter C1.3).

Two common techniques for generating a phase-conjugate mirror are stimulated Brillouin scattering (SBS) (see section A4.11.2), which uses acoustic waves to generate the reflection, and degenerate four-wave mixing, wherein two counter-propagating probe beams interact with the incident and reflected main beams to generate the reflection [7].

A3.3 Single axial mode operation

A3.3.1 Theory of longitudinal modes

In a laser cavity, the requirement that the field exactly reproduce itself in relative amplitude and phase means that the only allowable laser wavelengths or frequencies are given by the formulae

$$\lambda = \frac{P}{N} \qquad \text{or} \qquad \nu = \frac{Nc}{P} \tag{A3.24}$$

where λ is the laser wavelength, ν is the laser frequency, c is the speed of light in a vacuum, N is an integer whose value is determined by the lasing wavelength and P is the effective perimeter of the beam as it makes one round trip, taking into account the effects of the index of refraction, *etc.* For a conventional two-mirror cavity where the mirrors are separated by optical length L, these formulae revert to the familiar

$$\lambda = \frac{2L}{N} \qquad \text{or} \qquad \nu = \frac{Nc}{2L}.$$

These allowable frequencies are referred to as longitudinal modes. The spacing between the modes, in terms of frequency, is given by

$$\Delta\nu = \frac{c}{P}. \tag{A3.25}$$

As can be seen from equation (A3.25), the shorter the laser cavity, the greater the cavity mode spacing. By differentiating the expression for ν with respect to P we arrive at

$$\delta\nu = -\frac{Nc}{P^2}\delta P \qquad \text{or} \qquad \delta\nu = -\frac{Nc}{2L^2}\delta L. \tag{A3.26}$$

Consequently, for a helium–neon laser operating at 632.8 nm, with an effective cavity length of 25 cm, the mode spacing is approximately 600 MHz, and a 100 nm change in cavity length will cause a given longitudinal mode to shift by approximately 190 MHz.

The number of longitudinal laser modes that are present in a laser depends primarily on two factors: the length of the laser cavity and the width of the gain envelope of the lasing medium. For example, the gain of the red helium–neon laser is centred at 632.8 nm and has a width (FWHM) of approximately 1.4 GHz, meaning that, with a 25 cm laser cavity length, only two or three longitudinal modes can be present simultaneously, and a change in cavity length of less than one micron will cause a given mode to 'sweep' completely through the gain. By doubling the cavity length, the number of oscillating longitudinal modes that can fit under the gain curve doubles.

The gain of a gas-ion laser (e.g. argon or krypton) is approximately five times broader than that of a helium–neon laser, and the cavity spacing is typically much greater, allowing many more modes to oscillate simultaneously.

In most cases involving conventional lasers, only the portion of energy in the laser gain that is very close in frequency to that of a given longitudinal mode can contribute to the output energy in that mode; consequently, the greater the number of longitudinal modes present, the higher the total output energy. Likewise, a mode oscillating at a frequency near the peak of the gain will have higher energy than one oscillating at the fringes. This has a significant impact on the performance of a laser system because, as vibration and temperature changes cause small changes in the cavity length, modes sweep back and forth through the gain. A laser operating with only two or three longitudinal modes can experience power fluctuations of 10% or more, whereas a laser with ten or more longitudinal modes will see mode-sweeping fluctuations of 2% or less.

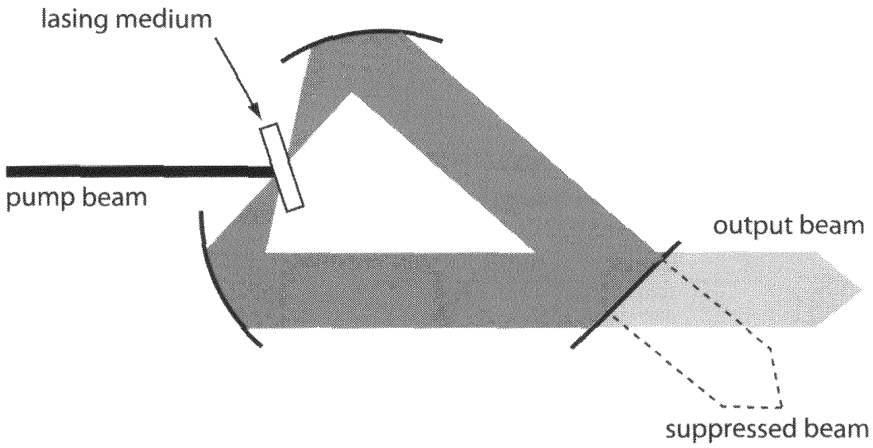


Figure A3.10. Ring laser cavity.

A3.3.2 Selecting a single longitudinal mode

A laser that operates with a single longitudinal mode is called a single-frequency laser. There are two ways to force a conventional two-mirror laser to operate with a single longitudinal mode. The first is to design the laser with a short enough cavity that only a single mode can be sustained. For example, in the helium–neon laser described before, a 10 cm cavity would allow only one mode to oscillate. This is not a practical approach for most gas lasers because, by the time the cavity is short enough to suppress additional modes, there is insufficient energy in the lasing medium to sustain any lasing action at all.

The second method is to introduce an element, typically a low-finesse Fabry–Pérot etalon, into the laser cavity. The free spectral range of the etalon should be several times the width of the gain curve, and the reflectivity of the surfaces should be sufficient to provide 10% or greater loss at frequencies half a laser mode spacing away from the etalon peak. The etalon is mounted at a slight angle to the optical axis of the laser to prevent parasitic oscillations between the etalon surfaces and the laser cavity.

Once the mode is selected, the challenge is to optimize and maintain its output power. Since the laser mode moves if the cavity length changes slightly, and the etalon pass band shifts if the etalon spacing varies slightly, it important that both be stabilized. Various mechanisms are used. Etalons can be passively stabilized by using zero-expansion spacers and athermalized designs, or they can be thermally stabilized by placing the etalon in a precisely controlled oven. Likewise, the overall laser cavity can be passively stabilized, or alternatively, the laser cavity can be actively stabilized by providing a mechanism to control cavity length.

The ring laser: Lasers with more than two resonator mirrors, as shown in figure A3.10, are commonly used to produce a single longitudinal mode. These lasers have a mode spacing of c/P where P is the length of the perimeter of the laser cavity. Normally, these lasers produce counterpropagating standing waves, but by introducing a device such as a Faraday rotator into the laser cavity, or injecting leakage from a partially-reflecting mirror back into the laser cavity, the laser beam can be constrained to propagate in a single direction as a travelling wave. The travelling wave sweeps through the laser gain, eliminating the spatial hole burning found with standing wave lasers, effectively increasing the homogeneity of the laser gain. When an etalon (or other frequency-selecting element) is introduced into the laser beam, more of the available energy is coupled into the single longitudinal mode than with a standing-wave laser. In addition, since adjacent modes are suppressed, the laser can be pumped with higher energy, further increasing single-mode output. For example, a ring dye laser can produce ten times the single frequency output of an equivalent standing-wave dye laser.

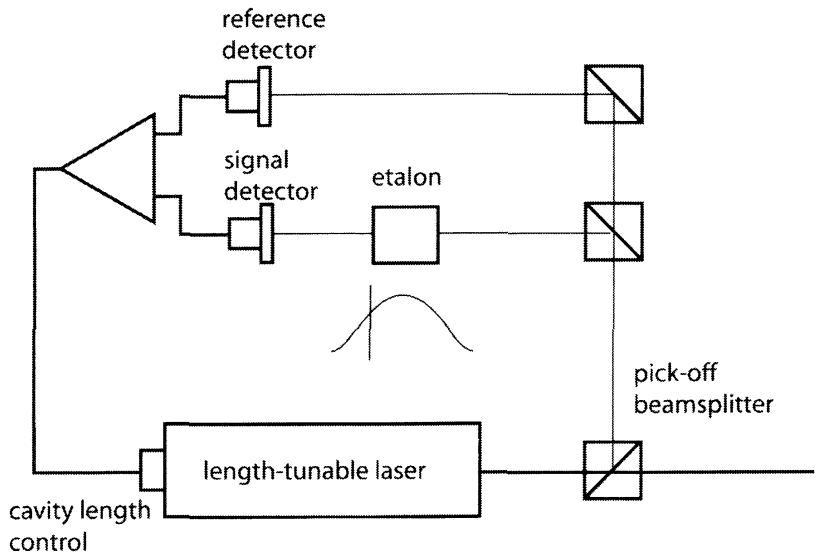


Figure A3.11. Laser frequency stabilization scheme.

A3.3.3 Frequency stabilization

The frequency output of a single-longitudinal-mode laser can be stabilized by precisely controlling the laser cavity length. A moderate level of stability can be accomplished passively by building an athermalized resonator structure and carefully controlling the laser environment to eliminate expansion, contraction and vibration; or actively, by using a mechanism to determine the frequency (either relatively or absolutely) and quickly adjusting the laser cavity length to maintain the frequency within the desired parameters.

A typical stabilization scheme is shown in figure A3.11. A portion of the laser output beam is directed into a low-finesse Fabry–Pérot etalon, and tuned to the side of the transmission band. The throughput is compared to an unattenuated reference beam, as shown in the figure. If the laser frequency increases, the ratio of attenuated power to unattenuated power increases. If the laser frequency decreases, the ratio decreases. In other words, the etalon is used to create a frequency discriminant that converts changes in frequency to changes in power. By 'locking' the discriminant ratio at a specific value (e.g. 50%) and providing negative feedback to the device used to control cavity length, output frequency can be controlled. If the frequency increases from the preset value, the length of the laser cavity is increased to drive the frequency back to the set point. If the frequency decreases, the cavity length is decreased. The response time of the control electronics is determined by the characteristics of the laser system being stabilized. This system can, in general, be made as stable as the reference etalon. Mechanical, acoustic and thermal stabilization of the etalon can improve the absolute stability of this setup.

Other techniques can be used to provide a discriminant. One common method used to provide an ultrastable, long-term reference is to replace the etalon or add to the etalon setup, an absorption cell or atomic beam with a spectral absorption line within the laser's tuning band. The stabilization system can then be used to maintain the laser frequency at the centre of the appropriate transition. This method is the basis for modern atomic clocks (see chapters C3.3, D6.1 and D6.2).

A major problem encountered with tightly controlled stabilized lasers is unwanted frequency shifts caused by cavity length perturbations that are beyond the handling capability of the control electronics. This is a particular problem with dye lasers, where bubbles or thickness changes in the dye jet can change the effective cavity length sufficiently to move the frequency completely off of the discriminant. This is handled in commercial systems by providing two correction loops: a narrow-range high-frequency correction loop (typically with a piezoelectric actuator) that controls the cavity length when the system is in lock, and a

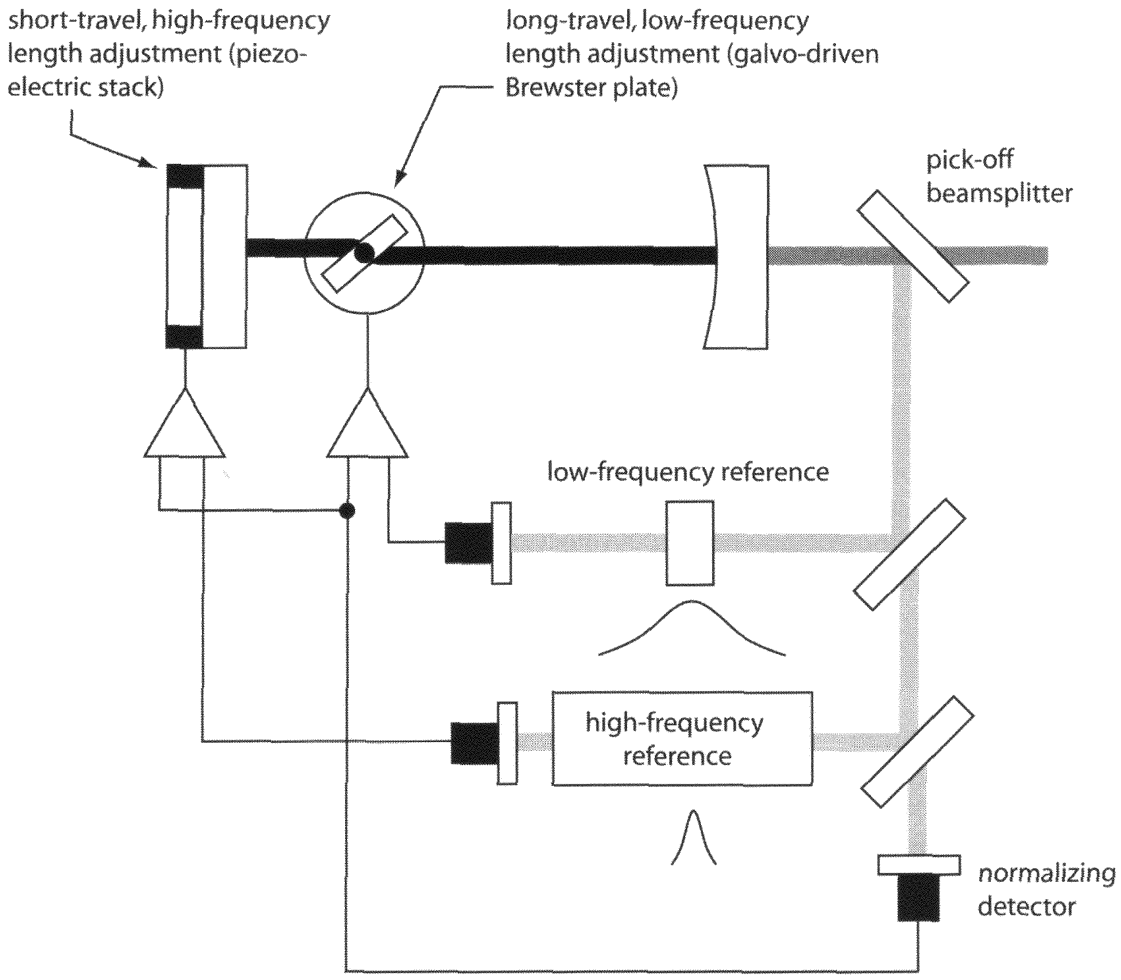


Figure A3.12. Correcting for long-range perturbations.

lower-frequency, longer-range correction loop (for example using a pair of galvanometer-driven tilting plates) to control the overall cavity length to within the range of the piezoelectric actuator. The discriminant must also be capable of providing the appropriate error signal over the maximum possible frequency jump. One method is to use two control etalons, one to provide a narrow discriminant for small perturbations, the second to provide a broader discriminant for long-range perturbations (see figure A3.12).

Another stabilization method, shown in figure A3.13, is used with commercial helium–neon lasers. It takes advantage of the fact that, for an internal mirror tube, the adjacent modes are orthogonally polarized. The cavity length is designed so that two modes can oscillate under the gain curve, as shown in figure A3.14. The two modes are separated outside the laser by a polarization-sensitive beamsplitter. Stabilizing the relative amplitude of the two beams stabilizes the frequency of both beams (see section B3.6.7).

The cavity length changes needed to stabilize the laser cavity are very small. In principle, the maximum adjustment needed is that required to sweep the frequency through one free spectral range of the laser cavity (the cavity mode spacing). For the helium–neon laser cavity described earlier, the required change is only 320 nm, well within the capability of piezoelectric actuators.

Commercially available systems can stabilize frequency output to 1 MHz or less. Laboratory systems that stabilize the frequency to within tens of kilohertz have been developed.

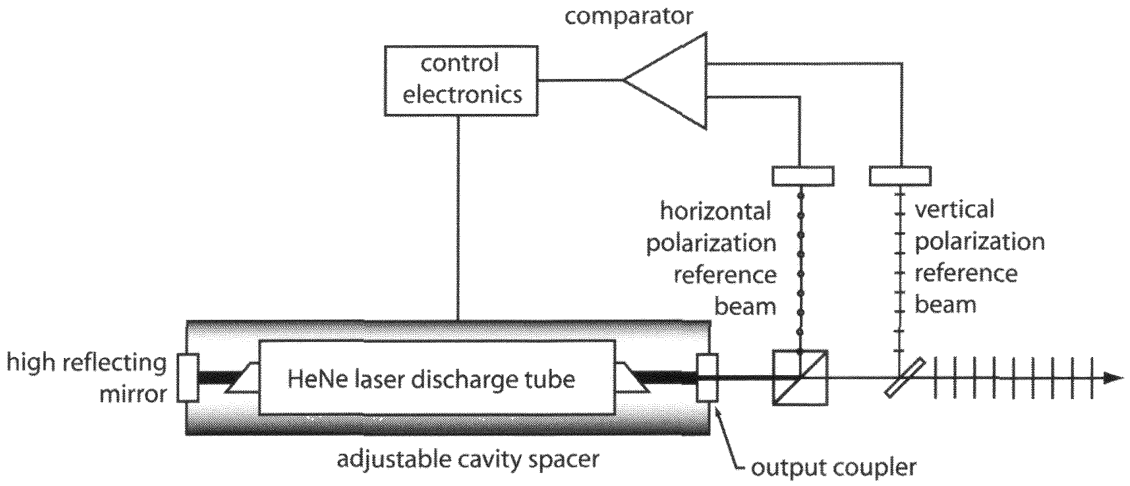


Figure A3.13. Frequency stabilization for a helium–neon laser.

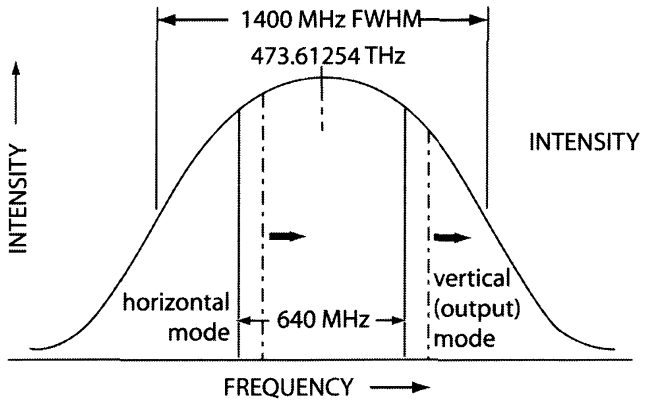


Figure A3.14. Two orthogonal modes oscillating in a helium–neon gain envelope.

A3.4 Tunable operation

Many lasers can operate at more than one wavelength. Argon and krypton lasers can operate at discrete wavelengths ranging from the ultraviolet to the near infrared. Dye lasers can be continuously tuned over a spectrum of wavelengths determined by the fluorescence bandwidths of the specific dyes (typically about 150 nm). Alexandrite and titanium sapphire lasers can be tuned continuously over specific spectral regions.

To create a tunable laser, the coatings on the cavity optics must be sufficiently broadband to accommodate the entire tuning range, and a variable-wavelength tuning element must be introduced into the cavity, either between the cavity optics or replacing the high-reflecting optic, to introduce loss at undesired wavelengths.

Three tuning mechanisms are in general use: Littrow prisms, diffraction gratings and birefringent filters. Littrow prisms (see figure A3.15) and their close relative, the full-dispersing prism, are used extensively with gas lasers that operate at discrete wavelengths. In its simplest form, the Littrow prism is a 30°–60°–90° prism with the surface opposite the 60° coated with a broadband high-reflecting coating. The prism,which replaces the end mirror in the laser, is oriented so that the desired wavelength is reflected back along the optical axis and the other wavelengths are dispersed off axis. By rotating the prism, individual lines can be chosen. To improve performance, the prism's angles can be modified so that the beam enters the prism exactly at Brewster's angle, thereby reducing intracavity losses. For higher power lasers that require greater dispersion to separate closely spaced lines, the Littrow prism can be replaced by a full-dispersing prism coupled with a high reflecting mirror.

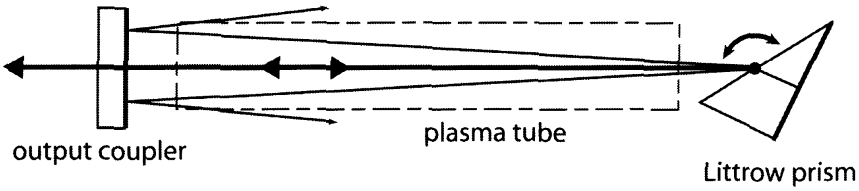


Figure A3.15. Littrow prism used to select a single wavelength.

Gratings are used for laser systems that require a higher degree of dispersion than that of a full-dispersing prism. They are usually placed inside the laser cavity.

Birefringent filters have come into general use for continuously tunable dye and Ti:sapphire lasers, since they introduce significantly lower loss than do gratings. The filter is made from a thin, crystalline-quartz plate with its fast axis oriented in the plane of the plate. The filter, placed at Brewster's angle in the laser beam, acts like a weak etalon with a free spectral range wider than the gain curve of the lasing medium. Rotating the filter around the normal to its face shifts the transmission bands, tuning the laser. Since there are no coatings and the filter is at Brewster's angle (thereby polarizing the laser), there are no inherent cavity losses at the peak of the transmission band. A single filter does not have as significant a line-narrowing effect as does a grating, but this can be overcome by stacking multiple filter plates together, with each successive plate having a smaller free spectral range.

A3.5 Beam shape and astigmatism in diode lasers

Two characteristics of the output beams of semiconductor (diode) lasers are astigmatism (caused when the vertical and horizontal beam waists are not coincident) and ellipticity (caused when the vertical and horizontal beam waists are different sizes). In applications requiring collimation or transformation of the diode laser's beam, these characteristics must be considered, and often corrected, if the wavefront is to approach the diffraction limit.

A3.5.1 Correcting astigmatism in collimators

Index-guided diode lasers typically exhibit a small amount of astigmatism (between 2 μm and 8 μm). Gain-guided diode lasers usually have between 30 μm and 60 μm of astigmatism.

In many applications, it is important to have as little wavefront distortion as possible in the final collimated or focused beam. The amount of astigmatic wavefront distortion, W, caused by the axial astigmatism of the diode laser can be given by the following expression:

$$W \approx (NA^2 \times z)/2\lambda \tag{A3.27}$$

where NA is the numerical aperture of the diode, λ is the laser wavelength and z is the axial astigmatism. The divergence can be expressed as a function of the wavefront or the axial astigmatism:

$$\text{Divergence} = 8W/\varphi \tag{A3.28}$$

where φ is the diameter of the clear aperture and W is expressed in the same units as φ.

The simplest way to correct astigmatism in a diode laser collimator is to place a plano-concave cylinder lens in front of the collimator. If the collimating lens is focused on the front facet of the diode, the cylinder should have negative refracting power with the power oriented parallel to the junction. If the collimating lens is focused on the waist behind the exit facet of the diode laser, then the cylinder lens should have positive

power and be oriented perpendicular to the plane of the junction. If the cylinder lens is placed before the collimating lens, the cylinder radius R is given (approximately) by:

$$R = \frac{\varphi^2}{8nD(1 - \cos u)} \tag{A3.29}$$

where φ is the clear aperture of the lens, n is the refractive index of the lens and u is the smaller of the half acceptance angle of the collimator or the divergence of the laser [8].

If the cylinder is placed after the collimating lens and we assume that the two lenses are thin, then we can treat the astigmatism as defocus, and calculate the necessary focal lengths using

$$\frac{1}{f_{\text{cyl}}} = \frac{1}{f_{\text{coll}}} - \frac{1}{f_{\text{ast}}} \tag{A3.30}$$

where f_{ast} is the focal length of the collimator plus the amount of astigmatism present.

A tilted plate, placed between the negative and positive elements of a beam expander, will add longitudinal astigmatism The amount of astigmatism added depends on the angle of tilt and the thickness of the plate. The added astigmatism, D, can be calculated using the following expression:

$$D = \frac{1}{\sqrt{n^2 - \sin^2(U_{\text{p}})}} \left[\frac{n^2 \cos^2(U_{\text{p}})}{n^2 - \sin^2(U_{\text{p}})} - 1 \right] \tag{A3.31}$$

where t is the thickness of the plate, U_{p} is the tilt angle and n is the index of refraction of the tilt plate.

A3.5.2 Circularizing a diode laser

The elliptical output of a diode laser can be circularized in a variety of ways. The most common technique is to use an anamorphic beam expander. There are two general classes of anamorphic beam expanders currently used with diode lasers: beam expanders that use cylinder lenses and beam expanders that use prisms.

Beam expanders that use cylinder lenses to circularize the beam usually do so in the form of a Galilean beam-expanding telescope. This has two advantages over a prism telescope: the beam is not displaced from the original centreline and the cylindrical elements that make up the beam expander can be adjusted to correct for any natural astigmatism in the diode output. Unfortunately, for large magnifications, the length of the telescope becomes excessively long. Furthermore, the telescope is not easily adjustable—an important factor since the ellipticity ratio can vary dramatically from diode to diode.

Because of the inherent disadvantages of the cylindrical beam expander, most designers opt to go with prism magnification for beam circularization. The prisms are relatively easy to manufacture with good transmitted wavefront, and they are easy to align. The most common configuration is the Brewster telescope shown in figure A3.16. It is compact and it produces an exit beam parallel to the incoming beam. Differences in elliptical ratios can be accommodated by rotating the prism pair to different angles. The chief disadvantage is that the exit beam is displaced. This can be corrected by adding additional prisms, as shown in figure A3.17.

A3.6 Q-switching, mode-locking and cavity dumping

Various techniques are used to generate laser pulses with much higher peak powers than for the same laser operating continuously. The most common are Q-switching, mode-locking and cavity dumping.

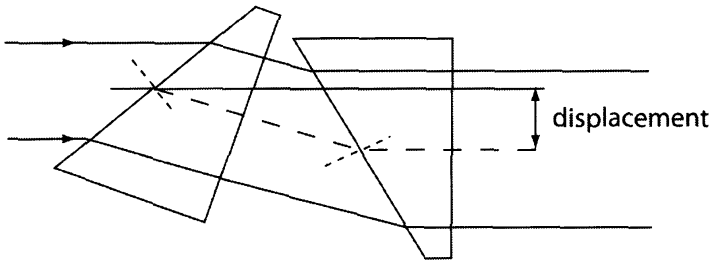
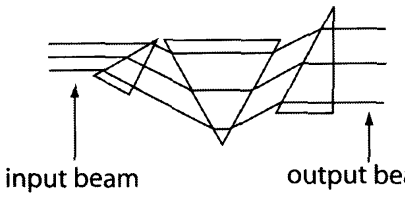


Figure A3.16. The Brewster telescope. Polarization must be in the p-plane for low reflective loss if the Brewster surface is not AR coated.



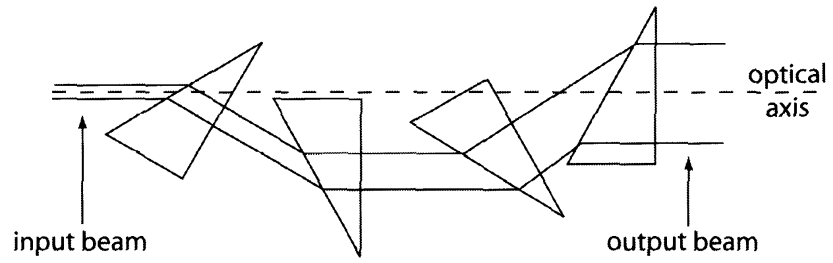


Figure A3.17. Anamorphic beam expanders that eliminate beam offset.

A3.6.1 Q-switching

In a lasing system, the output power is proportional to the population inversion between the upper and lower energy levels of the lasing transition. The greater the inversion, the higher the output power. In an oscillating laser, the upper laser level is constantly being depopulated by emission stimulated by the photons circulating in the laser cavity, as well as by spontaneous emission and by non-radiating relaxation mechanisms. At equilibrium (i.e., cw operation) the excitation mechanism is repopulating the upper laser level at the same rate as the other mechanisms are depopulating it.

The Q (or quality factor) of a laser cavity is defined by the equation

$$Q = \frac{2\pi P}{\lambda \delta} \tag{A3.32}$$

where P is the perimeter of the laser cavity ($2L$ for a linear laser) and δ is the round-trip cavity loss. Increasing the round-trip cavity loss (i.e. 'spoiling' the Q of the cavity) reduces, or effectively eliminates, the main source of stimulated emission depopulation—photons directed back into the lasing medium by the cavity mirrors. Assuming that the rate of population of the upper lasing level remains constant, the population inversion will then increase dramatically until a new equilibrium is reached, essentially storing up energy in the lasing medium. If the cavity losses are suddenly reduced (Q is 'switched' to its normal value), the round-trip gain will be much larger than the cavity loss, and the energy within the laser cavity will build up at an unusually

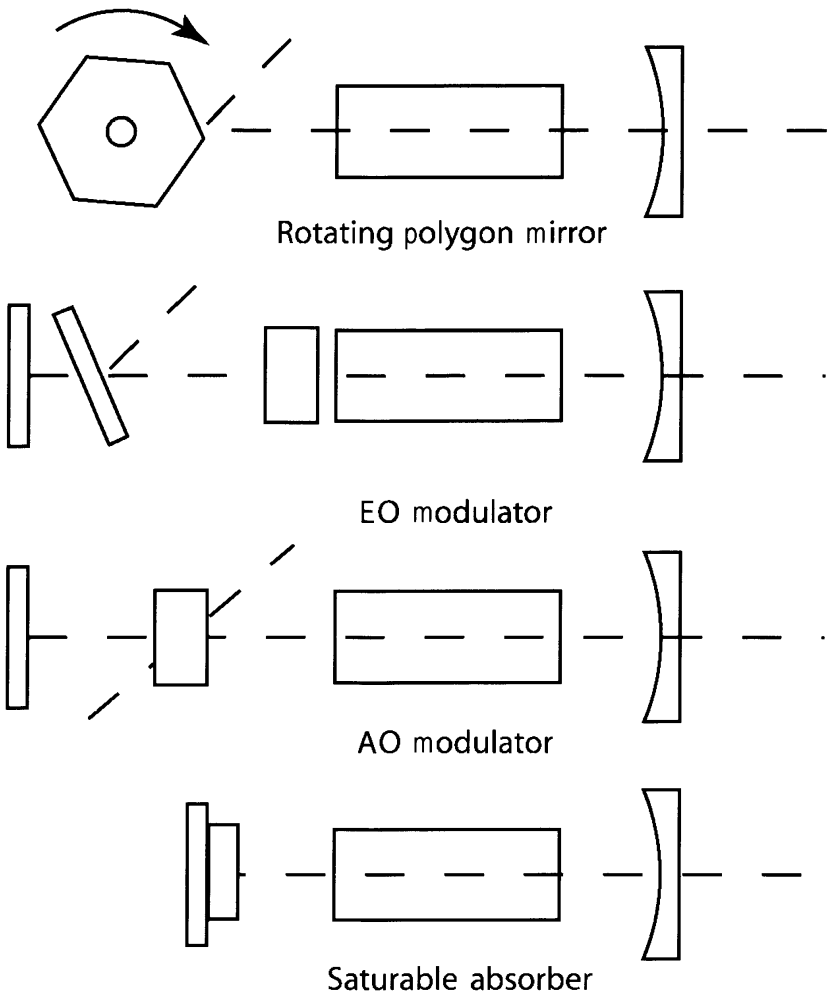


Figure A3.18. Techniques used to Q-switch a laser.

rapid rate, resulting in a 'giant pulse' (see section C2.2.2). The peak power in the Q-switched pulse can be three to four orders of magnitude higher than the power obtained from a non-switched laser using the same excitation mechanism.

For Q-switching to be most effective, the time needed to depopulate the upper laser level *via* spontaneous emission and non-radiating mechanisms must be much greater than the time needed to populate the level, as is the case for many solid-state materials (e.g. ruby, Nd:YAG). It is not true, however, for many gas-discharge lasers (e.g. helium–neon, argon–ion); Q-switching these lasers typically increases peak power by only 50%. Consequently, Q-switching is rarely used with gas lasers based on atomic transitions. Q-switching is, however, used on some molecular gas lasers (e.g. CO_2 lasers), because the vibrational and rotational transitions typically have a longer relaxation time than atomic transitions.

Four techniques routinely used to Q-switch a laser are shown in figure A3.18.

Rotating mirrors. One of the earliest methods was to use a rotating high-reflecting mirror or multifaceted prism as the high reflector. As the mirror spins, it alternately goes in and out of alignment changing the cavity Q. This technique has several weaknesses: slow switching speed, mechanical complexity, vibration and alignment problems, high motor wear and the inability to precisely time or trigger the pulses.

Electro-optic and acousto-optic Q-switching. A second method is to insert an electro-optic (EO) or acousto-optic (AO) shutter inside the laser cavity. The electro-optic shutter operates by rotating the polarization of

the recirculating light in the cavity by 90° and directing it out of the cavity with a polarization-selecting optical element. The acousto-optic shutter creates an rf-induced Bragg grating which deflects light out of the cavity. The electro-optic shutter is faster, with high hold-off (insertion loss in the low-Q state), but it is more expensive, requires a fast-switching power supply and has, in general, a higher insertion loss. The acousto-optic shutter is simpler and has very low insertion loss, but it is slower switching and should not be used with high-gain lasers. Both techniques can provide precise triggering and synchronization of the laser pulses.

Passive Q-switching. Lasers can also be Q-switched by placing an easily saturable absorbing medium inside the laser cavity. The absorber lowers the cavity gain but not sufficiently to eliminate totally the ability of the laser to oscillate. When the population inversion builds up sufficiently to overcome the additional cavity loss, the laser begins to oscillate weakly. The absorber quickly saturates and becomes transparent, restoring the cavity Q. Saturable absorbers are widely used in commercial lasers because they are simple and require no electronics. The main drawback of passive Q-switching is the inability to trigger pulses externally and the systems exhibit much greater pulse-to-pulse timing jitter than is typically observed with other techniques.

A3.6.2 Cavity dumping

In a cw laser, the power circulating inside the laser cavity is much greater than the power escaping through the output coupler. For example, the circulating power in a 10 mW helium–neon laser cavity is approximately 2 W; for a 20 W ion laser cavity, approximately 400 W. By replacing the output couplers on these lasers with non-transmitting, high-reflecting mirrors, this circulating power can be increased by nearly an order of magnitude. Cavity dumping is used to access (dump) the circulating energy inside the laser cavity (see section C2.2.3). The effect is the same as if the output coupler were suddenly pulled away, letting all of the circulating power escape in a pulse whose width, essentially, is the round-trip time of the laser cavity.

Cavity-dumping schemes employing acousto-optic and electro-optic modulators are shown in figure A3.19. As was the case with acousto-optic Q-switching, an rf signal creates a Bragg diffraction grating in the modulator, which deflects the beam out of the cavity, except that in this case, the laser is oscillating fully when the deflection occurs (Q-switching in reverse). The switching speed is determined by the time it takes the acoustic signal to travel across the beam; consequently, the beam is often focused at the modulator, as shown in the figure, to speed up the process. Electro-optic cavity dumping occurs when the Pockels cell rotates the polarization.

A3.6.3 Mode-locking

Mode-locking is a method of producing a train of very narrow, extremely high-peak-power pulses from a cw or long-pulse laser. A complete understanding of mode-locking is beyond the scope of this article, and for detailed information the author recommends the text by Anthony Siegman [1] (see also chapter C2.3).

As with most other processes involving laser manipulation, mode-locking begins with the introduction of a selective loss into the laser cavity, forcing the laser to react to the effects of that loss. In the case of mode-locking, a periodic loss is introduced, which is timed to coincide with the round-trip time of the laser cavity. In one approach, the aperture at the laser mirror is widened and narrowed periodically at a frequency equal to the cavity round-trip time ($c/2L$). While the aperture is narrowed, fewer photons are transmitted to the gain medium. After the first round trip, there will be a non-uniform spatial distribution of photons throughout the cavity, with a higher density of photons in the group that passed through the open aperture (the favoured group). Consequently, as the favoured group passes through the gain medium, it adds more stimulated emission photons to its group than the other photons in the cavity, accentuating the non-uniformity. The process is repeated, with the round-trip time and the opening and closing of the aperture

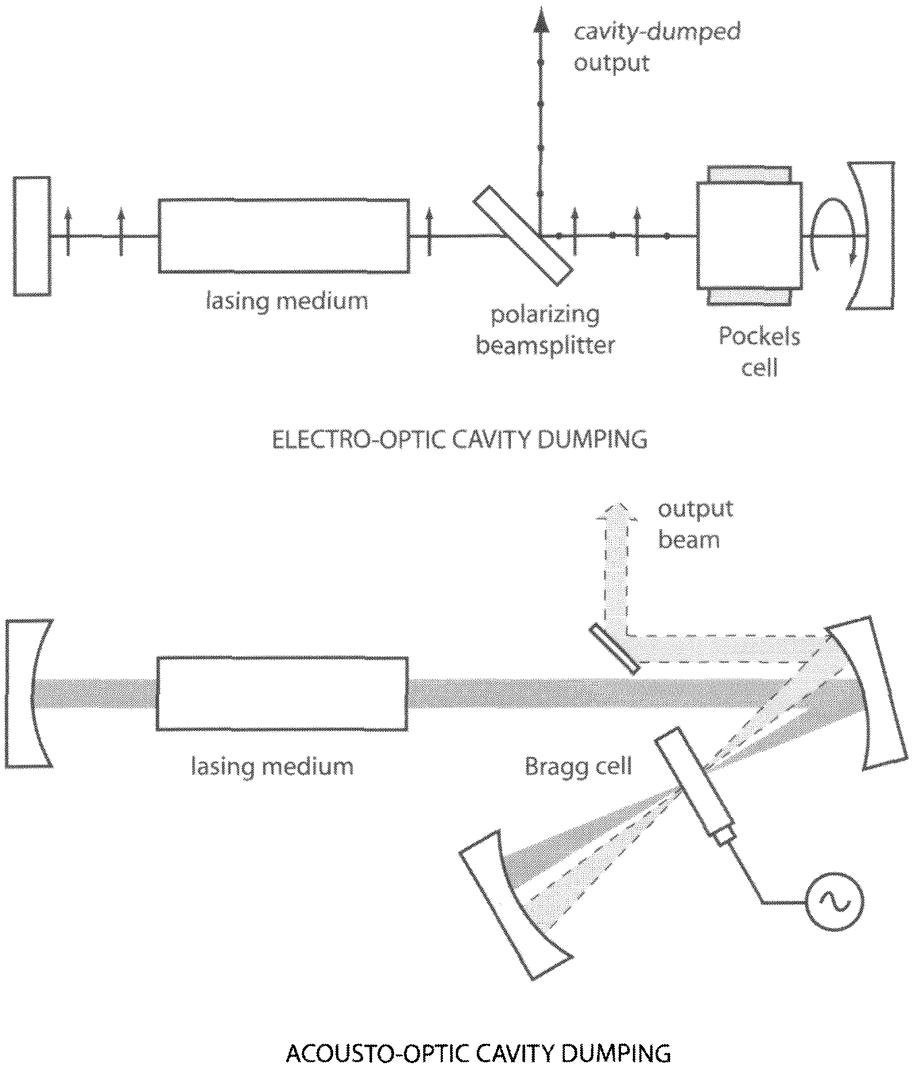


Figure A3.19. Electro-optic and acousto-optic cavity dumping configurations.

perfectly synchronized. The favoured group always sees the open aperture, while the rest of the circulating photons experience losses. As the energy in the favoured group increases, the width of the group narrows because there is somewhat more loss in the wings of the group than in the centre. Ultimately, the energy in the favoured group is so high that it absorbs all the remaining gain in the lasing medium, and the output becomes a train of very narrow pulses exactly spaced by the round-trip time of the laser cavity.

Three main techniques are used to mode-lock a laser: active mode-locking, passive mode-locking and synchronous pumping.

Active mode-locking. Active mode-locking is achieved by placing a modulation element inside the laser cavity and operating it at a frequency very close to the round-trip time of the laser cavity. Because lasers can be effectively mode-locked with only 20% to 30% amplitude modulation, hold-off is not an issue, and acousto-optic modulators are typically used because of their low insertion loss. This technique is used to mode-lock a wide variety of gas and solid-state lasers.

Passive mode-locking. It is possible to mode-lock a cw laser by placing a saturable absorber in the laser cavity. On the surface, this setup is very similar to the passive Q-switching scheme discussed before, but in practice, the parameters of the absorber, laser cavity, gain medium and excitation mechanism are substantially different.

As a population inversion builds up in the lasing medium, the precursor of laser oscillation is noise in the form of stimulated and spontaneous emission. As the noise builds up, there will be one noise spike with sufficient energy to saturate the absorber and make its way around the resonator and back through the gain medium where it can be amplified. Although eventually other noise spikes will have sufficient energy to saturate the absorber, this initial spike will be preferentially amplified until it absorbs all the available gain. This technique has been used to create the shortest optical pulses with durations of less than 30 fs.

Synchronous pumping. It is possible to mode-lock a laser by modulating the gain medium instead of modulating the cavity losses. This is done extensively with laser-pumped laser systems, such as dye lasers, by mode-locking the pump source (e.g. an argon laser) with an acousto-optic modulator and then pumping the dye laser with the mode-locked train of pulses. The key is to have the round-trip time of the dye laser cavity closely match that of the pump laser. This can lead to some very large systems. For optimal mode-locking, it is critical that the cavity lengths of both lasers, as well as the modulation frequency of the pump laser, be maintained precisely.

A3.7 Beam quality—limits and measurement

Defining the quality of a laser beam is often a difficult proposition because most lasers do not operate in a pure Gaussian mode. Indeed, in many cases, a Gaussian mode is undesirable (e.g. for photolithography applications, a 'top-hat' beam with uniform intensity over the transverse area is most desirable). The main parameters that go into determining beam quality are beam diameter and divergence, beam intensity profile, beam wavefront, mode uniformity, noise and frequency stability. No one instrument can measure all of these parameters, but there are instruments available that can measure each of the parameters.

Power and energy meters can measure peak power, average power, pulse energy and, in many cases, beam noise (rapid intensity fluctuations). Slit, knife-edge and pinhole scanning profilometers can provide accurate two-dimensional intensity profiles for cw laser beams as small as a few microns in diameter; CCD beam profilometers can give accurate two- and three-dimensional intensity profiles of both pulsed and cw beams, but with reduced resolution. Scanning interferometers are used to detect longitudinal modes and the presence of multiple transverses modes. They can also be used to determine the characteristics of the wavefront at a given point. Wavemeters can compare the laser output to a stabilized source and measure both relative and absolute frequency drift. Finally, the M^2 meter automatically measures beam diameter at multiple points on a focused beam and compares the observed beam waist and far-field divergence with that of a theoretical Gaussian, providing important information about how the beam propagates through an optical system (see section A2.1.4.3).

A3.7.1 Frequency and amplitude stabilization

The output of a freely oscillating laser will fluctuate in both amplitude and frequency. Fluctuations of less than 0.1 Hz are commonly referred to as 'drift'; faster fluctuations are termed 'noise' or, when talking about sudden frequency shifts, 'jitter'.

The major sources of noise in a laser are fluctuations in the pumping source and changes in length or alignment caused by vibration, stress and changes in temperature. For example, unfiltered line ripple can cause output fluctuations of 5–10% or more.

Likewise, a 10 μrad change in alignment can cause a 10% variation in output power, and, depending upon the laser, a 1 μm change in length can cause amplitude fluctuations of up to 50% (or more) and frequency fluctuations of several gigahertz.

High-frequency noise (>1 MHz) is caused primarily by 'mode beating'. For example, in a helium–neon laser, transverse Laguerre–Gaussian modes of adjacent order are separated by a few tens of MHz. If multiple

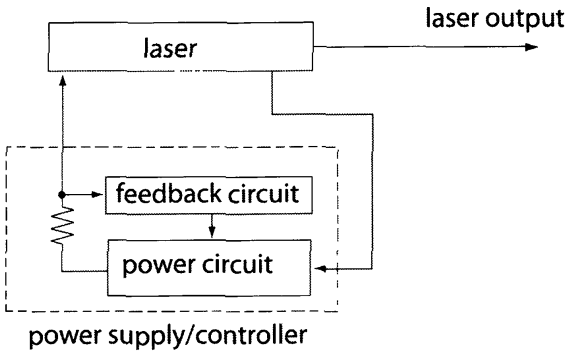


Figure A3.20. Automatic current control schematic diagram.

transverse modes oscillate simultaneously, heterodyne interference effects, or 'beats', will be observed at the difference frequencies. Likewise, mode beating can occur between longitudinal modes at frequencies of

$$\Delta\nu_{\text{longitudinal}} = \frac{c}{2L} \text{ or } \frac{c}{P} \tag{A3.33}$$

where L is the mirror separation of a linear laser cavity and P is the perimeter of a ring laser cavity. Mode beating can cause peak-to-peak power fluctuations of several percent. The only way to eliminate this noise component is to limit the laser output to a single transverse and single longitudinal mode.

Finally, when all other sources of noise have been eliminated, we are left with quantum noise, the noise generated by the spontaneous emission of photons from the upper laser level in the lasing medium. The randomness of this emission is proportional to the square root of the number of photons being detected. Thus the signal-to-noise ratio is proportional to the square root of the number of photons being detected. The laser beam signal quality increases as the laser intensity increases. It is impossible to suppress spontaneous noise, but, in most applications, it is inconsequential.

A3.7.2 Methods for suppressing amplitude noise and drift

Two primary methods are used to stabilize amplitude fluctuations in commercial lasers: automatic current control (ACC), also known as current regulation, and automatic power control (APC), also known as light regulation. In ACC, the current driving the pumping process passes through a stable sensing resistor, as shown in figure A3.20, and the voltage across this resistor is monitored. If the current through the resistor increases, the voltage drop across the resistor increases proportionately. Sensing circuitry compares this voltage to a reference and generates an error signal that causes the power supply to reduce the output current appropriately. If the current decreases, the inverse process occurs. ACC is an effective way to reduce noise generated by the power supply, including line ripple and fluctuations.

With APC, instead of monitoring the voltage across a sensing resistor, a small portion of the output power in the beam is diverted to a photodetector, as shown in figure A3.21, and the voltage generated by the detector circuitry is compared to a reference. As output power fluctuates, the sensing circuitry generates an error signal that is used to make the appropriate corrections to maintain constant output.

Automatic current control effectively reduces amplitude fluctuations caused by the driving electronics, but it has no effect on amplitude fluctuations caused by vibration or misalignment. Automatic power control can effectively reduce power fluctuations from all sources. Neither of these control mechanisms has any impact on frequency stability.

Not all cw lasers are amenable to APC as described. For the technique to be effective, there must be a monotonic relationship between output power and a controllable parameter (typically current or voltage). For example, throughout the typical operating range of a gas-ion laser, an increase in current will increase the

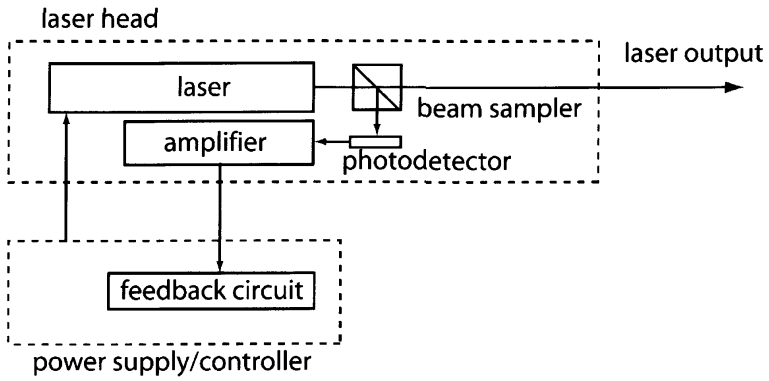


Figure A3.21. Automatic power control schematic diagram.

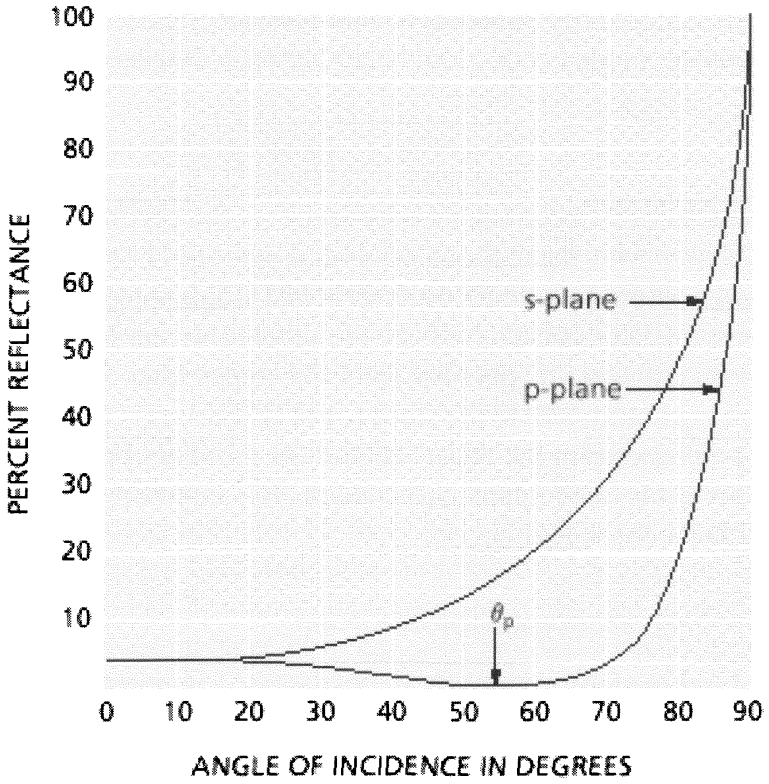


Figure A3.22. Reflectivity of the glass surface as a function of incidence angle.

output power and *vice versa*. This is not the case for some lasers. The output of a helium–neon laser is very insensitive to discharge current, and an increase in current may increase or decrease laser output. In a helium cadmium laser, where electrophoresis determines the density and uniformity of cadmium ions throughout the discharge, a slight change in discharge current in either direction can effectively kill lasing action.

If traditional means of APC are not suitable, the same result can be obtained by placing an acousto-optic modulator inside the laser cavity and using the error signal to control the amount of circulating power ejected from the cavity.

One consideration that is often overlooked in an APC system is the geometry of the light pickoff mechanism itself. One's first instinct is to insert the pickoff optic into the main beam at a 45° angle, so that the reference beam exits at a 90° angle. However, as shown in figure A3.22, for uncoated glass, there is almost a 10% difference in reflectivity for s and p polarization. In a randomly polarized laser, the ratio of the s and p components is not necessarily stable, and using a 90° reference beam can actually increase amplitude

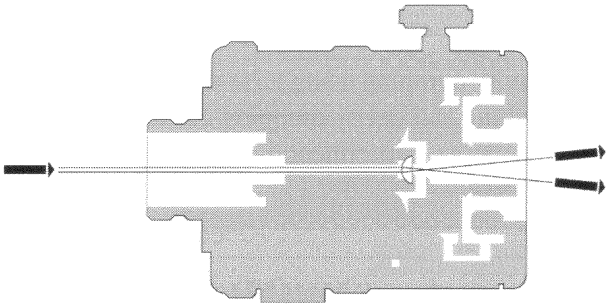


Figure A3.23. Spatial filtering.

fluctuations. This is of much less concern in a laser that has a high degree of linear polarization (e.g., 500:1 or better), but even then there is a slight presence of the orthogonal polarization. Good practice dictates that the pickoff element is inserted at an angle of 25° or less.

Never use the reflection off a Brewster window, used to polarize a laser, as the reference beam, because this is, in general, not representative of the actual laser output.

A3.8 Spatial filtering

Laser light scattered from dust particles residing on optical surfaces may produce interference patterns resembling holographic zone planes. Such patterns can cause difficulties in interferometric and holographic applications where they form a highly detailed, contrasting and confusing background which interferes with desired information. In other cases, stray coherent light from satellite beams (beams generated by reflections from the front and back surfaces of the output coupler) or incoherent light from a laser discharge may interfere with a measurement. Spatial filtering is a simple way of suppressing this interference and maintaining a very smooth beam irradiance distribution. The scattered light propagates in different directions from the laser light and hence is spatially separated at a lens focal plane. By centring a small aperture around the focal spot of the direct beam, as shown in figure A3.23, it is possible to block scattered light while allowing the direct beam to pass unscathed. The result is a cone of light with a smooth irradiance distribution which can be refocused to form a collimated beam that is almost equally smooth.

As a compromise between ease of alignment and complete spatial filtering, it is best that the aperture diameter is about two times the $1/e^2$ beam contour at the focus, or about 1.33 times the 99% throughput contour diameter.

References

[1] Siegman A 1986 *Lasers* (Sausalito, CA: University Science Books)
[2] Self S A 1983 Focusing of spherical Gaussian beams *Appl. Opt.* **22** 658
[3] Belland P and Crenn J 1982 Changes in characteristics of a Gaussian beam weakly diffracted by a circular aperture *Appl. Opto.* **21** 522–7
[4] Sun H 1998 Thin lens equation for a real laser beam with weak lens aperture truncation *Opt. Eng.* **37** 2906–13
[5] Freiberg R J and Halsted A S 1969 Properties of low order transverse modes in argon ion lasers *Appl. Opt.* **8** 355–62
[6] Rigrod W W 1963 Isolation of axi-symmetric optical-resonator modes *Appl. Phys. Lett.* **2** 51–3
[7] Kuttner 1985 Laser beam scanning *Optical Engineering* vol 8 (New York: Dekker) p 352
[8] Unsbo P 1995 Phase conjugation and four-wave mixing *Doctoral Thesis* Royal Institute of Technology, Stockholm

Further reading

Born M and Wolf E 1999 *Principles of Optics* 7th edn (Cambridge: Cambridge University Press)
Koechner W 1999 *Solid-State Laser Engineering* 5th edn (Berlin: Springer)

Q-Switching and Cavity Dumping:

Truntna R and Siegman A E 1977 Laser cavity dumping using an antiresonant ring *IEEE J. Quantum Electron.* **13** 955–62
Chesler R B and Maydan D 1971 Q-Switching and cavity dumping of Nd:YAG lasers *J. Appl. Phys.* **42** 1028–34
Chesler R B, Karr M A and Geusic J E 1970 An experimental and theoretical study of high repetition rate Q-switched Nd:YAG lasers *Proc. IEEE* **58** 1899–914

Phase conjugation:

Hellwarth R W 1977 Generation of time-reversed wave fronts by nonlinear refraction *J. Opt. Soc. Am.* **67** 1–3
Rockwell D A 1988 A review of phase-conjugate solid-state lasers *IEEE J. Quantum Electron.* **24** 1124–40
Bloom D M and Bjorklund G C 1977 Conjugate wave-front generation and image reconstruction by four-wave mixing *Appl. Phys. Lett.* **31** 592–4
Lera G and Nieto-Vesperinas M 1990 Phase conjugation by four-wave mixing of statistical beams *Phys. Rev.* A **41** 6400–5

Frequency stabilization:

Hall J L, Hils D, Salomon C and Chartier J-M 1987 Towards the ultimate laser resolution *Laser Spectroscopy* vol VIII, ed W Persson and S Svanberg (Heidelberg: Springer) pp 376–80
Hall J L 1986 Stabilizing lasers for applications in quantum optics *Quantum Optics IV, Proc. 4th Int. Symp. (February 10–15, Hamilton, New Zealand)* ed J D Harvey and D F Walls (Berlin: Springer) pp 273–84
Hils D and Hall J L 1989 Ultra-stable cavity-stabilized lasers with subhertz linewidth *Proc. 4th Int. Symp. on Frequency Standards and Metrology* ed A De Marchi (Heidelberg: Springer) pp 162–73

A4
Nonlinear optics

Robert W Boyd

A4.1 Basic concepts

Nonlinear optics is the study of the interaction of light with matter under conditions such that the linear superposition principle is not valid. The origin of this breakdown of the linear superposition principle can usually be traced to a modification of the optical properties of the material medium induced by the presence of an intense optical field. With a few important exceptions [1], only laser light is sufficiently strong to lead to an appreciable modification of the nonlinear optical properties of a material system and, for this reason, the field of nonlinear optics is basically the study of the interaction of laser light with matter. In this context it is important to distinguish two different sorts of nonlinear optical effects: (1) effects associated with the nonlinear optical response of the material contained within the laser cavity itself; and (2) effects induced by a prescribed laser beam outside of the laser cavity. In this chapter, we are concerned primarily with the second possibility, which constitutes the traditional field of nonlinear optics. Nonlinear optical processes occurring within the laser cavity itself constitute a central aspect of laser physics, as described in chapter A1, and lead to important effects such as laser instabilities and chaos [2] and to self mode locking of lasers [3]. The treatment of nonlinear optics presented in this chapter is necessarily limited in scope. More detailed treatments can be found in various monographs on the subject [4–10] as well as in the research literature. The present treatment follows most closely the notational conventions of [5].

Nonlinear optical effects can often be described by assuming that the response of the material system, as described by the induced dipole moment per unit volume, that is the polarization $\tilde{P}(t)$ can be expressed as a power series expansion in the strength $\tilde{E}(t)$ of the applied laser field as

$$\begin{aligned}\tilde{P}(t) &= \chi^{(1)}\tilde{E}(t) + \chi^{(2)}\tilde{E}^2(t) + \chi^{(3)}\tilde{E}^3(t) + \cdots \\ &\equiv \tilde{P}^{(1)}(t) + \tilde{P}^{(2)}(t) + \tilde{P}^{(3)}(t) + \cdots .\end{aligned} \tag{A4.1}$$

Here the first term describes ordinary linear optics, the second term describes second-order nonlinear optical effects, etc. The quantity $\chi^{(1)}$ is the linear susceptibility, the quantity $\chi^{(2)}$ is the second-order susceptibility, etc. We shall see later that there is a significant qualitative difference between second-order and third-order nonlinear optical effects. To summarize these differences briefly, we note that second-order nonlinear optical effects involve processes involving the simultaneous interaction of three photons, whereas third-order processes involve the interaction of four photons. Thus second-order nonlinear optics includes processes such as second-harmonic generation, sum- and difference-frequency generation and optical rectification; in contrast, third-order nonlinear optical effects include processes such as third-harmonic generation, the intensity dependence of the refractive index and four-wave mixing processes.

Equation (A4.1) has been written in a highly simplified form. In general, the relation between the polarization and the applied laser field must treat the tensor nature of the nonlinear coupling and any possible

frequency dependence of the nonlinear susceptibility elements. One particularly useful way of generalizing equation (A4.1) to deal with such issues is to express $\tilde{P}(t)$ and $\tilde{E}(t)$ in terms of their frequency components as

$$\tilde{\boldsymbol{P}}(\boldsymbol{r},t)=\sum_n \boldsymbol{P}(\omega_n)\mathrm{e}^{-\mathrm{i}\omega_n t} \qquad \tilde{\boldsymbol{E}}(\boldsymbol{r},t)=\sum_n \boldsymbol{E}(\omega_n)\mathrm{e}^{-\mathrm{i}\omega_n t} \tag{A4.2}$$

where the summation extends over all positive and negative frequency components of the field. We then define the second-order susceptibility to be the coefficient relating the amplitude of the nonlinear polarization to the product of two field amplitudes according to

$$P_i(\omega_n+\omega_m)=\sum_{jk}\sum_{(nm)}\chi^{(2)}_{ijk}(\omega_n+\omega_m,\omega_n,\omega_m)E_j(\omega_n)E_k(\omega_m). \tag{A4.3}$$

Here i, j and k refer to the Cartesian components of the fields and the notation (nm) indicates that we are to sum over n and m while holding the sum $\omega_n+\omega_m$ fixed. By way of illustration, second-harmonic generation is described using these conventions by the susceptibility $\chi^{(2)}_{ijk}(2\omega,\omega,\omega)$, sum-frequency generation by the susceptibility $\chi^{(2)}_{ijk}(\omega_1+\omega_2,\omega_1,\omega_2)$ and difference-frequency generation by the susceptibility $\chi^{(2)}_{ijk}(\omega_1-\omega_2,\omega_1,-\omega_2)$. We similarly define the third-order susceptibility through the relation

$$\begin{aligned}P_i(\omega_o+\omega_n+\omega_m)=&\sum_{jkl}\sum_{(mno)}\chi^{(3)}_{ijkl}(\omega_0+\omega_n+\omega_m,\omega_o,\omega_n,\omega_m)\\&\times E_j(\omega_o)E_k(\omega_n)E_l(\omega_m).\end{aligned} \tag{A4.4}$$

Third-harmonic generation is then described by the susceptibility $\chi^{(3)}_{ijkl}(3\omega,\omega,\omega,\omega)$ and the intensity-dependent refractive index is described by $\chi^{(3)}_{ijkl}(\omega,\omega,\omega,-\omega)$. The intensity dependence of the refractive index is alternatively described in terms of the nonlinear refractive index coefficient n_2, defined by the relation

$$n=n_0+n_2I, \tag{A4.5}$$

where I is the laser intensity, which is related to the nonlinear susceptibility through

$$n_2=\frac{12\pi^2}{n_0^2c}\chi^{(3)}. \tag{A4.6}$$

It is often convenient to measure I in units of W cm^{-2}, in which case n_2 is measured in units of cm^2 W^{-1}. We then find that numerically

$$n_2(\mathrm{cm}^2\ W^{-1})=\frac{12\pi^2}{n_0^2c}10^7\chi^{(3)}(\mathrm{esu})=\frac{0.0395}{n_0^2}\chi^{(3)}(\mathrm{esu}). \tag{A4.7}$$

A4.2 Mechanisms of optical nonlinearity

In this section we present a brief summary of the various physical mechanisms that can lead to a nonlinear optical response of a material system. We first make some general comments regarding the conditions under which various types of optical nonlinearities can occur.

A4.2.1 Influence of inversion symmetry on second-order nonlinear optical processes

A well-known result states that the second-order susceptibility $\chi^{(2)}$ necessarily vanishes for a material possessing inversion symmetry. Thus, second-order effects cannot occur in liquids, gases or glasses, nor can they occur in any of the crystal classes that possess inversion symmetry.

A4.2.2 Influence of time response on nonlinear optical processes

It should be noted that only very fast physical mechanisms can lead to an appreciable response for processes in which the output frequency is different from the input frequencies, because, in order for such processes to occur, the material has to be able to respond at the difference frequencies of the various interacting fields. In contrast, processes such as the intensity-dependent refractive index can occur even as the consequence of sluggish mechanisms, because in this case the average intensity of the incident light field can lead to a change of the refractive index. We thus conclude that only very fast processes can lead to processes such as harmonic generation.

A4.2.3 Non-resonant electronic response

Perhaps the most important source of optical nonlinearity is the response of bound electrons to an applied laser field. The electronic response can lead to both second- and third-order nonlinear optical processes. For the important case of non-resonant excitation, this mechanism has a very short response time. This response time can be estimated as the time required for the electron cloud surrounding the atomic nucleus to move in response to an applied laser field; this time is of the order of magnitude of the period associated with the motion of an electron in a Bohr orbit about the nucleus, which is of the order of magnitude of 10^{-16} s.

Nonresonant electronic response can be described theoretically in one of several ways. One is to solve Schrödinger's equation for an atom in the presence of an intense laser field and extract that part of the induced response which is second or third order in the amplitude of the applied field. Another is to develop a totally classical model of the optical response based, for instance, on adding nonlinear contributions to the restoring force introduced into the equation of motion used in the Lorentz model of the atom. These approaches lead to consistent predictions in relevant limits. At an even more elementary level, one can make order-of-magnitude estimates [4, 11, 12] of the size of the nonlinear optical response by arguing that the ratio of linear to nonlinear optical response will be of the order of $(E/E_{\text{at}})^n$, where n is the order of the nonlinearity. Since $E_{\text{at}} = m^2e^5/\hbar^4 = 1.9 \times 10^7$ statvolt cm^{-1}, this argument leads to the prediction that

$$\chi^{(2)} = \hbar^4/8m^2e^5 = 7.29 \times 10^{-9} \text{ cm (statvolt)}^{-1} \tag{A4.8}$$

$$\chi^{(3)} = \hbar^8/8m^4e^{10} = 4.25 \times 10^{-16} \text{ cm}^2 \text{ (statvolt)}^{-2}. \tag{A4.9}$$

These values are in good order-of-magnitude agreement with the measured values for typical nonlinear optical materials.

A4.2.4 Molecular orientation

The molecular orientation effect occurs in anisotropic molecules and leads to a nonlinear optical response as a consequence of the tendency of molecules to become aligned along the electric field vector of the incident laser field. This process is illustrated in figure A4.1. This alignment tends to increase the refractive index of the material, that is it leads to a positive value of $\chi^{(3)}$. This process typically has a response time of the order of 1 ps and produces a nonlinear optical response of the order of 10^{-12} esu. The contribution to the third-order susceptibility resulting from molecular orientation can be expressed as

$$\chi^{(3)} = \frac{4N(\alpha_3 - \alpha_1)^2}{75kT} \tag{A4.10}$$

where N is the number density of molecules and $(\alpha_3 - \alpha_1)$ is the difference in polarizabilities along the principal dielectric axes of the molecule.

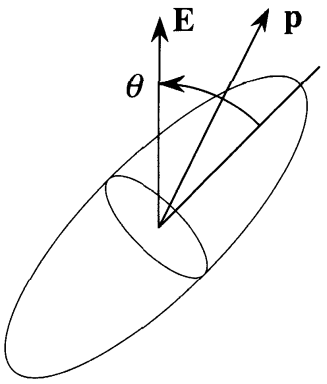


Figure A4.1. Origin of the molecular orientation effect, illustrating the tendency of an anisotropic molecule to become oriented in a electric field.

A4.2.5 Electrostriction

Electrostriction is the tendency of materials to become compressed in the presence of a static or oscillating electric field. Since for most materials the refractive index increases with material density, this process leads to a positive value of $\chi^{(3)}$, typically of the order of 10^{-13} esu. The response time of electrostriction is typically of the order of 1 ns. The contribution to the third-order susceptibility resulting from electrostriction can be expressed as

$$\chi^{(3)} = \frac{1}{48\pi^2} C_T \gamma_e^2 \tag{A4.11}$$

where C_T is the isothermal compressibility and where $\gamma_e \equiv \rho \partial \epsilon / \partial \rho$ is the electrostrictive constant.

A4.2.6 Photorefractive effect

The photorefractive effect [13, 14] leads to a large nonlinear optical response but one that cannot usually be described in terms of a third-order (or any order) nonlinear susceptibility. The photorefractive effect occurs as a consequence of the tendency of weakly bound electric charges within an optical material to migrate from regions of high intensity to regions of low intensity. This charge imbalance leads to the establishment of an electric field within the material, which modifies the refractive index of the material by means of the linear electrooptic (Pockels) effect. This basic process is illustrated in figure A4.2. The photorefractive effect cannot be described in terms of a nonlinear susceptibility because the resulting change in refractive index tends to be independent of the strength of the incident laser field. Stronger laser fields tend to speed up the process of charge redistribution but do not change the final charge distribution. Typically a laser beam of intensity 1 W cm^{-2} will produce a photorefractive response with a response time of the order of 1 s.

A4.3 Nonlinear optical materials

The development of applications of nonlinear optics has historically been limited by the availability of materials with the required optical and environmental properties and much effort has gone into the development of superior materials for use in nonlinear optics [15–17]. A brief representative sample of some materials of interest in second- and third-order nonlinear optics are given in tables A4.1 and A4.2. More complete listings of materials properties are to be found various references [9, 18–20]. A particularly useful approach toward the development of superior materials for nonlinear optics has been the development of nanocomposite materials [21–23].

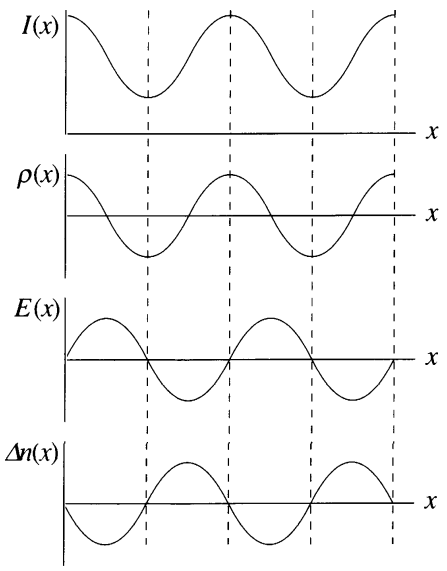


Figure A4.2. Origin of the photorefractive effect. $I(x)$ represents the spatially modulated laser intensity, $\rho(x)$ represents the free-charge distribution, $E(x)$ the static electric field created by this charge distribution and $\Delta n(x)$ the resulting change in refractive index.

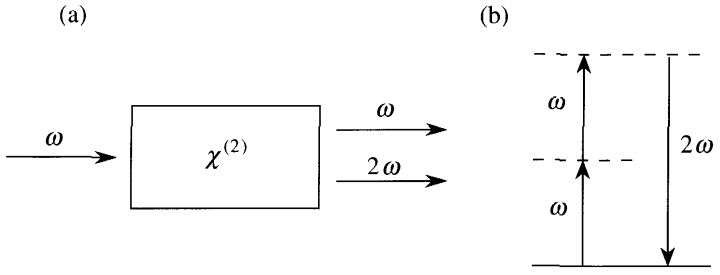


Figure A4.3. (*a*) The geometry of second-harmonic generation and (*b*) its description in terms of an energy level diagram.

A4.4 Second- and third-harmonic generation

Second-harmonic generation is the process in which an incident field at frequency ω_1 is converted to an output field at frequency $\omega_2 = 2\omega_1$ by means of the second-order response of the material system. This was, in fact, one of the first nonlinear optical processes to be studied in detail [37] and was discovered shortly after the invention of the laser. Second-harmonic generation can be a very efficient process, leading to conversion efficiencies approaching 100%. This process is described pictorially in figure A4.3.

Second-harmonic generation can be described mathematically by introducing coupled-amplitude equations that describe the propagation of the fundamental and second-harmonic waves. We take the fundamental wave to have amplitude $A_1(z)\exp(ik_1z)$, where $k_1 = n_1\omega_1/c$ is its wavevector magnitude, and take the second-harmonic wave to have amplitude $A_2(z)\exp(ik_2z)$, where $k_2 = n_2\omega_2/c$ is its wavevector magnitude. The coupled amplitude equations are derived by introducing the nonlinear polarizations $P(2\omega) = \chi^{(2)}A_1^2\exp(2ik_1)$ and $P(\omega) = 2\chi^{(2)}A_2A_1^*\exp[i(k_2 - k_1)z]$ of equation (A4.3) into the driven wave equation and then making the slowly varying amplitude approximation. The resulting equations have the form

$$\frac{dA_1}{dz} = \frac{4\pi i\omega_1^2\chi^{(2)}}{k_1c^2}A_2A_1^*e^{-i\Delta kz} \tag{A4.12}$$

Table A4.1. Properties of several second-order nonlinear optical materials

Crystal (class)	Transmission range (μm)	Refractive index (at 1.06 μm)	Nonlinear coefficient (pm V^{-1})
Silver gallium selenide $AgGaSe_2$ ($\bar{4}2m$)	0.78–18	$n_o = 2.7010$ $n_e = 2.6792$	$d_{36} = 33$ (at 10.6 μm)
β-barium borate BBO ($3m$)	0.21–2.1	$n_o = 1.6551$ $n_e = 1.5425$	$d_{22} = 2.3$ $d_{24} = d_{15} \leq 0.1$
Lithium iodate $LiIO_3$ (6)	0.31 – 5	$n_o = 1.8517$ $n_e = 1.7168$	$d_{31} = -7.11$ $d_{33} = -7.02$ $d_{14} = 0.31$
Lithium niobate $LiNbO_3$ (3m)		$n_o = 2.234$ $n_e = 2.155$	$d_{31} = -5.95$ $d_{33} = -34.4$
Potassium dihydrogen phosphate KH_2PO_4 (KDP)	0.18 – 1.55	$n_o = 1.4944$ $n_e = 1.4604$	$d_{36} = 0.63$
$KTiOPO_4$ KTP ($mm2$)	0.35–4.5	$n_x = 1.7367$ $n_y = 1.7395$ $n_z = 1.8305$	$d_{31} = 6.5$ $d_{32} = 5.0$ $d_{33} = 13.7$ $d_{24} = 6.6$ $d_{15} = 6.1$

From a variety of sources. By convention, $d = \frac{1}{2}\chi^{(2)}$. The tensor nature of the nonlinear coefficients is expressed in contracted notation, in which the first index of d_{il} represents any of the three Cartesian indices and the second index l represents the product of two Cartesian indices acording to the rule $l = 1$ implies xx, 2 implies yy, 3 implies zz, 4 implies yz or zy, 5 implies zx or xz, and 6 implies xy or yx. To convert d_{il} to the Gaussian cgs units of cm statvolt^{-1}, each entry should be divided by 4.189×10^{-4}.

and

$$\frac{dA_2}{dz} = \frac{2\pi i\omega_2^2\chi^{(2)}}{k_2c^2}A_1^2e^{i\Delta kz} \tag{A4.13}$$

where $\Delta k = 2k_1 - k_2$. These equations express the fact that the amplitude of the second-harmonic wave is driven by the A_1^2 and that the generated second-harmonic wave acts back on the fundamental wave through the factor $A_2A_1^*$. Coupled amplitudes for other nonlinear optical processes are derived using analogous procedures.

Second-harmonic generation can occur with good efficiency only if the wavevector mismatch factor Δk that appears in equations (A4.12) and (A4.13) is much smaller than the inverse of the length L of the

Table A4.2. Third-order nonlinear optical coefficients of various materials.

Material	n_0	$\chi^{(3)}$ (esu)	n_2 (cm^2 W^{-1})	Comments
Crystals				
Al_2O_3	1.8	2.2×10^{-14}	2.9×10^{-16}	
CdS	2.34	7.0×10^{-12}	5.1×10^{-14}	1.06 μm
diamond	2.42	1.8×10^{-13}	1.3×10^{-15}	
GaAs	3.47	1.0×10^{-10}	3.3×10^{-13}	1, 1.06 μm
Ge	4.0	4.0×10^{-11}	9.9×10^{-14}	THG $\lvert\chi^{(3)}\rvert$
LiF	1.4	4.4×10^{-15}	9.0×10^{-17}	
Si	3.4	2.0×10^{-10}	2.7×10^{-14}	THG $\lvert\chi^{(3)}\rvert$
TiO_2	2.48	1.5×10^{-12}	9.4×10^{-15}	
ZnSe	2.7	4.4×10^{-12}	3.0×10^{-14}	1.06 μm
Glasses				
fused silica	1.47	1.8×10^{-14}	3.2×10^{-16}	
As_2S_3 glass	2.4	2.9×10^{-11}	2.0×10^{-13}	
BK-7	1.52	2.0×10^{-14}	3.4×10^{-16}	
BSC	1.51	3.6×10^{-14}	6.4×10^{-16}	
Pb Bi gallate	2.3	1.6×10^{-12}	1.3×10^{-14}	
SF-55	1.73	1.5×10^{-13}	2.0×10^{-15}	
SF-59	1.953	3.1×10^{-13}	3.3×10^{-15}	
Nanoparticles				
CdSSe in glass	1.5	1.0×10^{-12}	1.8×10^{-14}	non-res.
CS 3-68 glass	1.5	1.3×10^{-8}	2.3×10^{-10}	res.
gold in glass	1.5	1.5×10^{-8}	2.6×10^{-10}	res.
Polymers				
polydiacetylenes				
PTS		6×10^{-10}	$3. \times 10^{-12}$	non-res.
PTS		-4×10^{-8}	-2×10^{-10}	res.
9BCMU			1.9×10^{-10}	$\lvert n_2\rvert$, res.
4BCMU	1.56	-9.2×10^{-12}	-1.5×10^{-13}	non-res, β = 0.01 cm MW^{-1}
Liquids				
acetone	1.36	1.1×10^{-13}	2.4×10^{-15}	
benzene	1.5	6.8×10^{-14}	1.2×10^{-15}	
carbon disulphide	1.63	2.2×10^{-12}	3.2×10^{-14}	$\tau = 2$ ps
CCl_4	1.45	8.0×10^{-14}	1.5×10^{-15}	
diiodmethane	1.69	1.1×10^{-12}	1.5×10^{-14}	
ethanol	1.36	3.6×10^{-14}	7.7×10^{-16}	
methanol	1.33	3.1×10^{-14}	6.9×10^{-16}	
nitrobenzene	1.56	4.1×10^{-12}	6.7×10^{-14}	
water	1.33	1.8×10^{-14}	4.1×10^{-16}	

Table A4.2. (Continued.)

Material	n_0	$\chi^{(3)}$ (esu)	n_2 (cm^2 W^{-1})	Comments
Other materials				
air	1.0003	1.2×10^{-17}	5.0×10^{-19}	
Ag		2.0×10^{-11}		THG $\lvert\chi^{(3)}\rvert$
Au		5.4×10^{-11}		THG $\lvert\chi^{(3)}\rvert$
vacuum	1	2.4×10^{-33}	1.0×10^{-34}	
cold atoms	1.0	5.1	0.2	(EIT BEC)
fluorescein dye in glass	1.5	$2 + 2i$	$0.035(1 + i)$	$\tau = 0.1$ s

Here n_0 is the linear refractive index. The third-order susceptibility $\chi^{(3)}$ is defined by equation (A4.1). This definition is consistent with that introduced by Bloembergen [4]. Some workers use an alternative definition which renders their values four times smaller. In compiling this table we have converted the literature values when necessary to the definition of equation (A4.1). The quantity β is the coefficient describing two-photon absorption. Reference [24] provides an extensive tabulation of third-order nonlinear optical susceptibilities. The values of $\chi^{(3)}$ given in this reference need to be multiplied by a factor of four to conform with the standard convention of Bloembergen, which is the convention used in the present article. Other references used are [25–36].

interaction region. When this condition is met, the interaction is said to be phase matched. Phase matching is typically achieved by making use of the natural birefringence of standard second-order nonlinear optical crystals and propagating the fundamental and second-order fields with orthogonal polarizations [38]. For $\Delta k = 0$, and assuming that only a fundamental field is present at the input to the medium, these equations can be solved exactly to find that

$$A_2(z) = \sqrt{n_1/n_2} A_1(0) \tanh(z/l) \tag{A4.14}$$

where

$$l = \frac{(n_1 n_2)^{1/2} c}{4\pi\omega_1 \chi^{(2)} |A_1(0)|} \tag{A4.15}$$

gives the characteristic distance scale over which the interaction occurs. Note that this model predicts that asymptotically the conversion efficiency can approach 100%. Second-harmonic generation in the plane-wave limit has been described more completely by Armstrong *et al* [39] and the effects of laser-beam focusing on this process have been described by Boyd and Kleinman [40].

Radiation at the third-harmonic frequency can be created in one of two ways. One procedure is to create the third harmonic directly by means of a third-order interaction in which the amplitude of the nonlinear polarization is given by

$$P(3\omega) = \chi^{(3)}.E^3 \tag{A4.16}$$

Third-order interactions of this sort (and higher-order interactions, which give rise, for instance, to fifth- and seventh-harmonic generation) tend to be less efficient than second-order interactions but have the advantage that they can be used even at short wavelengths where standard nonlinear optical crystals are not transmitting. This approach to the generation of third-harmonic radiation has been described in detail by Ward and New [41] and by Miles and Harris [42].

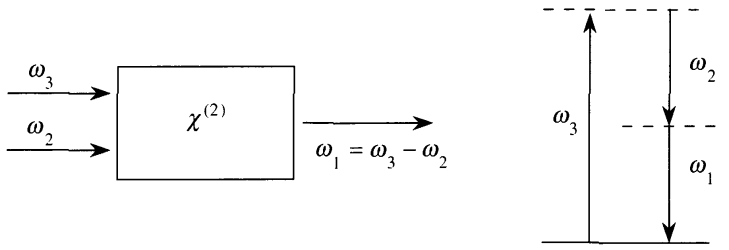


Figure A4.4. Illustration of the relation between difference frequency generation and optical parametric amplification. Note that amplification of the lower frequency input field ω_2 accompanies the creation of the difference frequency field ω_1.

The other approach to the generation of radiation at the third-harmonic frequency is to first generate a field at frequency 2ω through the process of second-harmonic generation followed by sum-frequency generation of the fields at frequencies ω and 2ω to produce an output at frequency 3ω. This approach can often be considerably more efficient than direct third-harmonic generation because lower-order processes tend to be stronger than third-order processes. In fact, through a judicious choice of experimental conditions it is possible to produce radiation at the third-harmonic frequencies with efficiency exceeding 80% [43].

A4.5 Optical parametric oscillation

An important technological application of nonlinear optics is the construction of parametric oscillators, which can produce tunable radiation over broad spectral regions spanning the infrared, visible and ultraviolet.

To understand the operation of an optical parametric oscillator (OPO), let us first examine the nature of the amplification that accompanies the process of difference frequency generation, which is illustrated in figure A4.4. The left-hand side of this figure shows input waves at frequencies ω_3 and ω_2 with $\omega_3 > \omega_2$ incident on a second-order nonlinear optical material, within which the difference frequency wave at frequency $\omega_3 - \omega_2$ is generated. The energy-level diagram on the right-hand side of this figure reveals that one photon must be added to the field at frequency ω_2 for every photon that is created at frequency ω_1. The process of difference-frequency generation thus automatically leads to amplification of the lower-frequency input field.

This conclusion can be reached more rigorously by considering the coupled-waves equations describing the interaction of the two low-frequency waves,

$$\frac{dA_1}{dz} = \frac{4\pi i\omega_1^2\chi^{(2)}}{k_1c^2}A_3A_2^*e^{i\Delta kz} \tag{A4.17}$$

$$\frac{dA_2}{dz} = \frac{4\pi i\omega_2^2\chi^{(2)}}{k_2c^2}A_3A_1^*e^{i\Delta kz} \tag{A4.18}$$

where

$$\Delta k = k_3 - k_1 - k_2. \tag{A4.19}$$

These equations can readily be solved for arbitrary boundary conditions. The solution for the special case of perfect phase matching ($\Delta k = 0$) and for no input at one of the lower frequencies (i.e. $A_2(0) = 0$) is given by

$$A_1(z) = A_1(0)\cosh\kappa z \Longrightarrow \tfrac{1}{2}A_1(0)\exp(gz) \tag{A4.20}$$

$$A_2(z) = i\left(\frac{n_1\omega_2}{n_2\omega_1}\right)^{1/2}\frac{A_3}{|A_3|}A_1^*(0)\sinh\kappa z \Longrightarrow O(1)A_1^*(0)\exp(gz) \tag{A4.21}$$

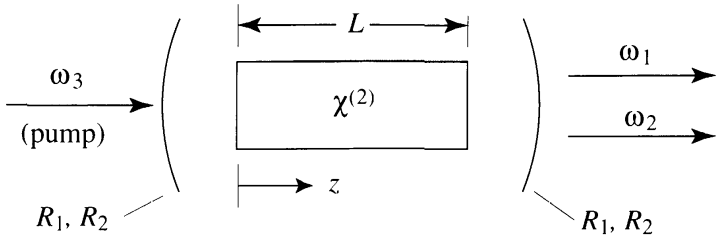


Figure A4.5. Layout of the optical parametric oscillator.

where

$$g = (\kappa_1 \kappa_2^*)^{1/2} \qquad \kappa_j = \frac{4\pi i \omega_j^2 \chi^{(2)} A_3}{k_j c^2}. \tag{A4.22}$$

In these equations the symbol $\Longrightarrow$ denotes the asymptotic behavior at large z and the symbol $O(1)$ denotes a number of the order of unity. Clearly, both waves asymptotically experience exponential growth.

The optical layout of an optical parametric oscillator is shown in figure A4.5. Here a pump wave of frequency ω_3 is incident on a second-order nonlinear optical crystal located inside an optical resonator. The end mirrors are assumed to be identical and to have reflectivities R_1 and R_2 at frequencies ω_1 and ω_2 respectively. The oscillator is said to be singly resonant if the end mirror reflectivity is large at either ω_1 or ω_2 and is said to be doubly resonant if the end mirror reflectivity is large at both ω_1 and ω_2. Generally speaking, doubly resonant oscillators have lower threshold pump intensities but singly resonant oscillators are more readily operated in a stable manner because they do not require the independent establishment of a cavity resonance condition for the two separate frequencies ω_1 and ω_2.

Let us next consider the threshold condition for the establishment of oscillation in an OPO. For simplicity, we consider a simple model that applies to the doubly resonant oscillator. We assume that $R_1 = R_2 \approx 1$, that $\Delta k = 0$ and that the frequencies exactly meet the cavity resonance condition. The threshold condition can then be expressed as

$$(e^{2gL} - 1) = 2(1 - R). \tag{A4.23}$$

Here the left-hand side of the equation can be interpreted as the fractional energy gain per pass and the right-hand side of the equation can be interpreted as the fractional energy loss per pass. The factor of two appears in the exponential because g is defined to be the amplitude gain coefficient. By expanding the exponential on the left-hand side to first order in gL, we find that the threshold condition can be expressed [44] as

$$gL = (1 - R). \tag{A4.24}$$

Through use of equation (A4.22), we can use this result to determine the laser intensity required to reach the threshold for parametric oscillation.

The output frequencies of an OPO are usually controlled by adjusting the orientation of the nonlinear mixing crystal to determine which set of frequencies ω_1 and ω_2 (with $\omega_1 + \omega_2 = \omega_3$) satisfy the phase-matching condition $\Delta k = 0$. OPOs tend to be broadly tunable because the tuning range is limited only by the limits of transparency of the crystal and by the limits over which the phase-matching relation can be established. Optical parametric oscillation was first observed experimentally by Giordmaine and Miller [44]. Continuous-wave OPO operation was first achieved by Smith *et al* [45]. Early work on OPOs has been reviewed by Byer and Herbst [47]. Recent work has stressed newer materials such as beta-barium borate [46].

A4.6 Optical phase conjugation

Optical phase conjugation [48–51] is a nonlinear optical process that holds considerable promise for applications such as aberration correction, image processing and novel forms of interferometry [52]. The name

phase conjugation derives from the fact that certain nonlinear optical processes have the ability to transform a field of the form

$$\tilde{E}(r,t) = A(r)\mathrm{e}^{\mathrm{i}kz-\mathrm{i}\omega t} + \text{c.c.} \tag{A4.25}$$

into the form

$$\tilde{E}_{\mathrm{pc}}(r,t) = A^*(r)\mathrm{e}^{-\mathrm{i}kz-\mathrm{i}\omega t} + \text{c.c.} \tag{A4.26}$$

In addition to propagating in a direction opposite to that of the incident field, the wavefront of the phase-conjugate wave is changed from A to A^*. The nature of the phase-conjugation process is illustrated in figure A4.6(*a*), which shows an optical field falling onto a phase conjugating device which is often referred to as a phase conjugate mirror (PCM). The 'phase-conjugate' nature of this reversed wavefront allows it to remove, in double pass, the influence of aberrations in optical systems. The novel quantum statistical properties of the phase conjugation process have been described by Gaeta and Boyd [53].

The two primary means for forming a phase conjugate wavefront are degenerate four-wave mixing, which is illustrated in figure A4.6(*b*), and stimulated Brillouin scattering, which is illustrated in figure A4.6(*c*). In the four-wave mixing interaction, a signal beam of amplitude A_3 interacts with two counterpropagating plane-wave pump beams of amplitudes A_1 and A_2 in a third-order nonlinear optical medium. Under these conditions, the dominant phase-matched contribution to the nonlinear optical susceptibility is of the form

$$P_{\mathrm{NL}} = 6\chi^{(3)}A_1A_2A_3^*. \tag{A4.27}$$

Since the nonlinear polarization is proportional to A_3^*, it will generate an output field that is the phase conjugate of the input field, that is a field proportional to A_3^*. The mutual interaction of the signal and conjugate beams can be described by the coupled amplitude equations

$$\frac{\mathrm{d}A_3}{\mathrm{d}z} = \mathrm{i}\kappa A_4^* \qquad \frac{\mathrm{d}A_4}{\mathrm{d}z} = \mathrm{i}\kappa A_3^* \tag{A4.28}$$

where $\kappa = (12\pi\omega/nc)\chi^{(3)}A_1A_2$. The solution to these equations for the boundary conditions appropriate to the situation illustrated shows that the amplitude of the generated conjugate field is given by

$$A_4(0) = A_3^*(0)\frac{\mathrm{i}\kappa}{|\kappa|}\tan|\kappa|L. \tag{A4.29}$$

We see that the generated field is indeed proportional to the complex conjugate of the input field. We also see that a phase-conjugate mirror can have a reflectivity of greater than 100%, because the pump waves provide energy to the phase-conjugate wave.

The other standard configuration for forming a phase-conjugate wavefront is through stimulated Brillouin scattering (SBS), as illustrated in figure A4.6(*c*). SBS is described in more detail in section A4.11 of this chapter. This process leads to phase conjugation because an aberrated input wave will produce a highly non-uniform volume intensity distribution in the focal region. The gain coefficient of the SBS process is proportional to the laser intensity and the resulting nonuniform gain distribution will tend to generate an output wave whose wavefronts match those of the input wave.

A4.7 Self-focusing of light

Self-focusing is an example of a self-action effect of light. Other examples of self-action effects include self-trapping of light and the break-up of a beam of light into multiple filaments. These effects are illustrated schematically in figure A4.7. These particular self-action effects can occur only if the nonlinear refractive index coefficient n_2 is positive. Self-focusing (figure A4.7(*a*)) occurs because the refractive index at the centre of the laser beam is larger than in the wings of the laser beam. This effect causes the material medium to act

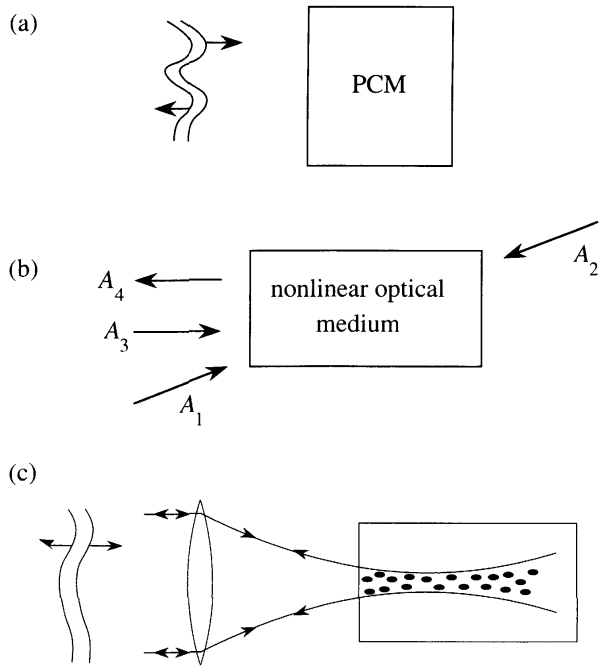


Figure A4.6. (*a*) Illustration of the nature of the phase-conjugation process, (*b*) phase conjugation by degenerate four-wave mixing and (*c*) phase conjugation by stimulated Brillouin scattering.

like a positive lens, bringing the light to a focus within the material medium. Self-trapping (figure A4.7(*b*)) occurs when the tendency of a beam to converge because of self-focusing precisely compensates for the tendency of the beam to diverge due to diffraction effects. Simple arguments [5, 54] show that this balance can occur only if the laser beam carries the critical power

$$P_{\mathrm{cr}} = \lambda^2/8n_0n_2. \tag{A4.30}$$

As a point of reference, P_{cr} has the value 30 kW for carbon disulphide at a wavelength of 700 nm. Self-trapped filaments are also known as spatial solitons. The use of spatial solitons has been proposed for optical switching applications [55]. According to the simple model leading to equation (A4.30), self-trapped filaments can have any diameter d, as long as the total power contained in the beam has the value P_{cr}. Only if the power of the laser beam exceeds P_{cr} can self-focusing occur. It is readily shown, on the basis of Fermat's principle, that the distance from the entrance face of the nonlinear material to the self-focus is given by

$$z_{\mathrm{f}} = \frac{2n_0}{0.61}\frac{w_0^2}{\lambda}\frac{1}{(P/P_{\mathrm{cr}}-1)^{1/2}} \tag{A4.31}$$

where w_0 is the beam diameter. If the laser power is much greater than P_{cr}, another process known as filamentation (figure A4.7(*c*)) can occur. In this process, the beam breaks up into multiple small filaments, each of which carries power P_{cr}. The origin of this process is that small perturbations on the incident laser wavefront experience exponential spatial growth as the consequence of near-forward four-wave mixing processes [56]. The maximum value of this growth rate is is given by $g = (\omega/c)n_2I$ and it occurs at the characteristic filamentation angle $\theta_{\mathrm{max}} = \sqrt{2}g/k$. Filamentation is an undesirable process and methods to suppress filamentation include the use of spatial filtering to remove aberrations from the laser wavefront, the use of specially structured beams [57] and the use of quantum interference effects [58] to eliminate the nonlinear response leading to filamentation.

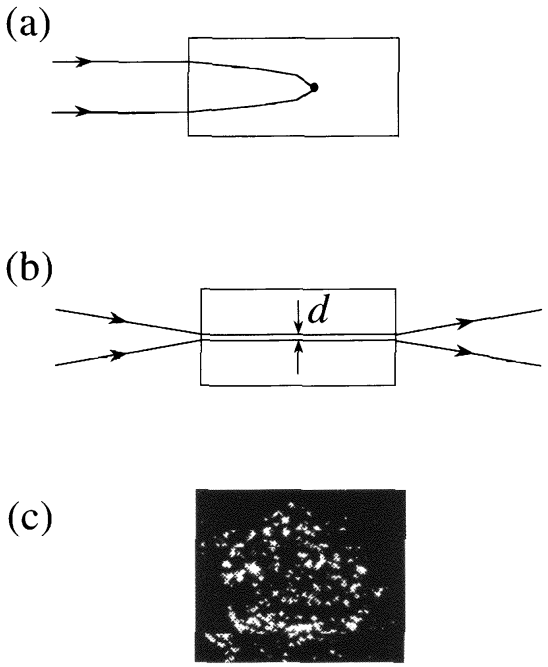


Figure A4.7. Several self-action effects of light are illustrated: (*a*) self-focusing, (*b*) self-trapping of light and (*c*) the break-up of a beam of light into multiple filaments.

A4.8 Optical solitons

Optical solitons are beams of light that propagate without changing their form, that is they propagate as a self-similar solution to the wave equation [59]. By a *temporal soliton*, one means a pulse of light that propagates through a dispersive medium with no change in shape because of a balancing of dispersive and nonlinear effects. By a *spatial soliton*, one means a beam of light that propagates with a constant transverse profile because of a balance between diffraction and self-focusing effects. Similarly, a *spatio-temporal soliton* is a pulse that propagates without spreading in time or in the transverse directions.

The basic equation describing the propagation of optical solitons is a generalization of the so-called nonlinear Schrödinger equation and has the form

$$\frac{\partial A}{\partial z} + \frac{1}{v_g}\frac{\partial A}{\partial t} + \frac{1}{2}\mathrm{i}\beta_2\frac{\partial^2 A}{\partial t^2} + \frac{1}{2\mathrm{i}k}\nabla_\mathrm{T}^2 = \frac{\mathrm{i}n_0 n_2 \omega_0}{2\pi}|A|^2 A \tag{A4.32}$$

where $v_g = (\partial k/\partial\omega)^{-1}$ is the group velocity, $\beta_2 = \partial^2 k/\partial\omega^2$ is a measure of the dispersion in the group velocity, ∇_T^2 is the transverse Laplacian and ω_0 is the central frequency of the pulse. The term involving β_2 describes the tendency of the pulse to spread in time due to dispersive effects, the term involving ∇_T^2 describes diffraction and the term involving n_2 describes self-phase modulation and self-focusing effects. The propagation of temporal solitons can be described by this equation by discarding the transverse Laplacian and the propagation of spatial solitons can be described by this equation by discarding the time derivatives. This equation possesses solutions relevant to many different physical situations. For instance, it possesses solutions in the form of bright solitons (e.g. a bright pulse on a zero background) or dark solitons (a decrease in intensity in an otherwise uniform non-zero background). Optical solitons can occur only for certain values of the material parameters. For instance, bright temporal solitons can occur only if n_2 and β_2 have opposite signs and dark temporal solitons can occur only if these quantities have the same sign. Similarly, bright spatial solitons can occur only if n_2 is positive, whereas dark spatial solitons can occur only if n_2 is negative. Only certain solutions to equation (A4.32) are stable to small perturbations. For instance, bright spatial solitons

are stable in one transverse dimension but are unstable in two transverse dimensions.

A particularly important example of optical solitons are bright temporal solitons propagating through an optical fibre. The solution to equation (A4.32), with the transverse terms discarded, that describes this occurrence is given by

$$A_s(z, \tau) = A_s^0 \operatorname{sech}(\tau/\tau_0)e^{i\kappa z} \tag{A4.33}$$

where $\tau = t - z/v_g$ and where the pulse amplitude A_s^0 and pulsewidth τ_0 must be related according to

$$|A_s^0|^2 = \frac{-2\pi k_2}{n_0 n_2 \omega_0 \tau_0^2} \tag{A4.34}$$

and where

$$\kappa = -k_2/2\tau_0^2 \tag{A4.35}$$

represents the phaseshift experienced by the pulse upon propagation. Note that the condition (A4.34) shows that β_2 and n_2 must have opposite signs in order for equation (A4.33) to represent a physical pulse in which the intensity $|A_s^0|^2$ and the square of the pulsewidth τ_0^2 are both positive. We can see from equation (A4.32) that, in fact, β_2 and γ must have opposite signs in order for group velocity dispersion to compensate for self-phase modulation.

For the case of an optical fibre, n_2 is positive with a value of approximately 3.2×10^{-16} cm^2 W^{-1}. Bright optical solitons can then occur only if β_2 is negative and consideration of the dispersion of the refractive index of silica glass shows that β_2 is negative only at wavelengths longer than 1.3 μm. Optical solitons of this sort have been observed experimentally [60].

A4.9 Optical bistability

Optical bistability refers to the possibility that a given optical system may possess two (or more) outputs for a given input. This possibility was first described theoretically by Szöke *et al* [64] and first observed experimentally by Gibbs *et al* [62]. Extensive treatments of bistability can be found in [61,63]. The realization that optical bistability can occur is important because it suggests that nonlinear optical techniques can be used to perform logical operations similar to those of electronic digital computers.

A standard design for a bistable optical device and its typical operating characteristics are shown in figure A4.8. Here a wave of amplitude A_1 is shown falling onto a device in the form of a Fabry–Pérot interferometer filled with a third-order nonlinear optical material. Such a device can be bistable in the sense that, under certain situations, there can be more than one output intensity for a given input intensity. The theoretical analysis of such a device proceeds by deriving a relation between the input amplitude A_1 and the output amplitude A_3. The result is the well-known Airy equation

$$A_3 = \frac{\tau^2 A_1}{1 - \rho^2 e^{2ikl - \alpha l}} \tag{A4.36}$$

where τ is the amplitude transmittance of either end mirror and ρ is the amplitude reflectivity. Here k is the total propagation constant and α is the total intensity absorption coefficient of the material within the resonator, including both their linear and nonlinear contributions. A nonlinear contribution to k or α or both can lead to bistable behaviour. As an illustration, if we assume that the material within the resonator is a saturable absorber, for which the absorption coefficient α changes with intensity according to

$$\alpha = \frac{\alpha_0}{1 + I/I_s} \tag{A4.37}$$

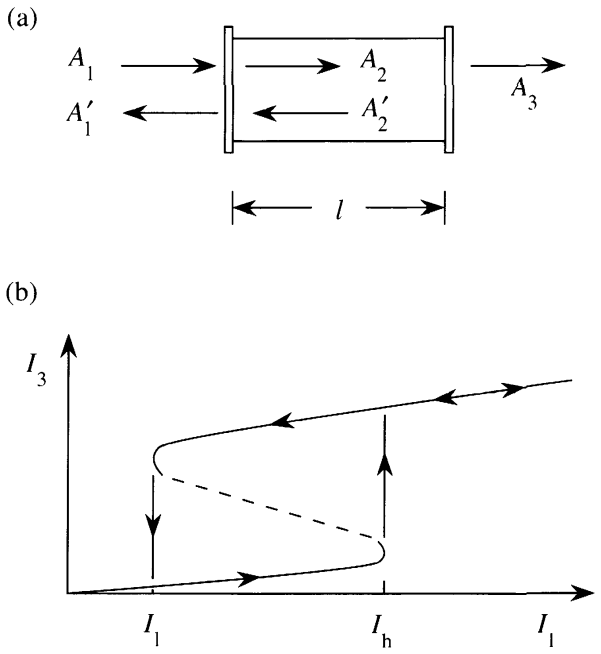


Figure A4.8. (*a*) Standard design for a bistable optical device and (*b*) typical operating characteristics.

where I_s is the saturation intensity, we find that the output intensity $I_3 = |A_3|^2$ is related to the input intensity $I_1 = |A_1|^2$ according to

$$I_1 = I_3\left(1 + \frac{C_0}{1 + 2I_3/TI_s}\right)^2 \tag{A4.38}$$

where $C_0 = R\alpha_0 l/(1 - R)$. Under certain conditions (in particular, for $C_0 > 8$), this equation predicts the occurrence of optical bistability. The input–output characteristics of the device under these conditions are illustrated schematically in figure A4.8(*b*). This curve has the form of a standard hysteresis loop and shows that, over a considerable range of input intensities, more than one output intensity can occur.

A4.10 Optical switching

Optical switching refers to the use of nonlinear optical methods to control one beam of light by a second beam of light. Extensive discussions of optical switching can be found in [65] and [66].

A prototypical design for an all-optical switch is shown in figure A4.9. In this design, a third-order nonlinear optical material is placed in one arm of a Mach–Zender interferometer (see section A5.2.3). Let us first analyse the operation of such a device in the absence of an applied control field. The relative phase of the two pathways through the interferometer changes with signal wave input intensity according to $\phi_{NL} = n_2(\omega/c)Il$ and thus the input can be directed either toward output port 1 or port 2 depending upon the input intensity. The threshold intensity for switching from one output port to the other is given by the condition that the nonlinear phase shift ϕ_{NL} is equal to π radians. In this configuration, the nonlinear interferometer could be used to separate low intensity pulses from high intensity pulses in a pulse train of variable intensity. More sophisticated types of switching behaviour can be obtained by applying an additional input field to the control port of the interferometer. For instance, the presence or absence of a control field can be used to direct the signal field either to output port 1 or port 2, so that the device would operate as a router.

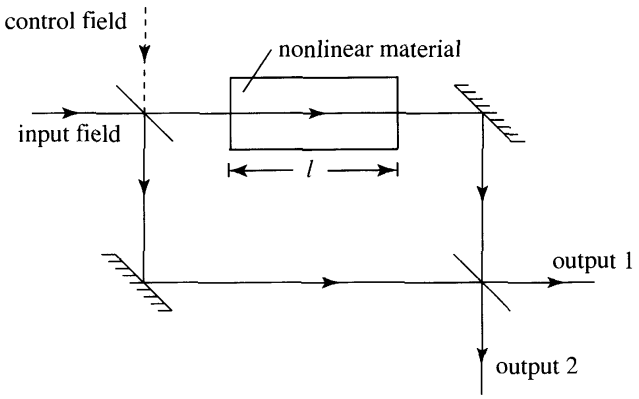


Figure A4.9. Typical design of an all-optical switching device, in the form of a nonlinear Mach–Zehnder interferometer.

A4.11 Stimulated light scattering

The scattering of light can occur either as a consequence of spontaneous or stimulated processes. The distinction between these two types of light-scattering processes can be understood by noting [67] that all light scattering occurs as a consequence of localized fluctuations in the optical properties of the material medium. From this perspective, spontaneous light scattering occurs as a consequence of fluctuations induced by thermal or by quantum mechanical zero-point fluctuations and stimulated light scattering occurs as a consequence of fluctuations that are induced by the presence of the laser field.

It is believed that all spontaneous light-scattering processes possess a stimulated analogue, and quantitative models have been presented that relate the optical coefficients of these two types of processes [68]. Some of the important light-scattering processes are as follows:

- Rayleigh scattering is the scattering of light from non-propagating density fluctuations. Rayleigh scattering does not produce a shift in the line-centre of the scattered light but does lead to a linewidth of the order of 5×10^{-4} cm^{-1}. The gain of the stimulated analogue of this process is of the order of 10^{-4} cm MW^{-1}.
- Rayleigh-wing scattering is the scattering of light from fluctuations in the orientation of anisotropic molecules. Rayleigh-wing scattering does not produce a shift in the line-centre of the scattered light, but does lead to a linewidth of the order of 10 cm^{-1}. The gain of the stimulated analogue of this process is of the order of 10^{-3} cm MW^{-1}.
- Brillouin scattering is the scattering of light from propagating sound waves. Brillouin scattering produces a shift of the scattered light of the order of 0.3 cm^{-1} with a linewidth of the order of 5×10^{-3} cm^{-1}. The gain of the stimulated analogue of this process is of the order of 10^{-2} cm MW^{-1}.
- Raman scattering is the scattering of light from vibrational modes of the molecules that constitute the scattering medium. Raman scattering produces a shift of the scattered light of the order of 1000 cm^{-1} with a linewidth of the order of 5 cm^{-1}. The gain of the stimulated analogue of this process is of the order of 5×10^{-3} cm MW^{-1}.

A4.11.1 Stimulated Raman scattering

Stimulated Raman scattering (SRS) [69] is characterized by exponential growth of a light wave at the Stokes sideband of the laser field. In particular, in this process the intensity $I_S(t)$ of a beam of light at the Stokes frequency $\omega_S = \omega_L - \omega_V$, where ω_L is the laser frequency and ω_V is the vibrational frequency of the molecule, increases with propagation distance z according to

$$I_S(z) = I_S(0)e^{gI_L z} \tag{A4.39}$$

Table A4.3. Properties of stimulated Raman scattering for several materials.

Substance	Frequency shift (cm^{-1})	Gain factor g ($cm\ GW^{-1}$)
Liquids		
benzene	992	3
water	3290	0.14
nitrogen	2326	17
oxygen	1555	16
Gases		
methane	2916	0.66 at 10 atm
hydrogen	4155 (vibrational)	1.5 (10 atm and above)
	450 (rotational)	0.5 (0.5 atm and above)
deuterium	2991 (vibrational)	1.1 (10 atm and above)
nitrogen	2326	0.071 (at 10 atm)
oxygen	1555	0.016 (at 10 atm)

where g is the Raman gain factor and I_L is the laser intensity. The Raman gain coefficient for various materials is given in table A4.3. The Stokes shift for SRS is sufficiently large that SRS is an important technique for laser frequency shifting [70].

SRS can be understood by assuming that the optical polarizability of a molecule depends on the interatomic separation $q(t)$ according to

$$\alpha(t) = \alpha_0 + (\partial\alpha/\partial q)_0[q(t) - q_0] \tag{A4.40}$$

where q_0 is the equilibrium value of the inter-atomic separation [71]. Note that a periodic oscillation of $q(t)$ will produce a periodic oscillation of $\alpha(t)$ and consequently of the refractive index $n(t)$. Such a periodic modulation of $n(t)$ will tend to amplify light at a frequency detuned from the laser frequency by the vibrational frequency. Moreover, the simultaneous presence of the laser and Stokes beams will tend to reinforce the molecular vibration at the beat frequency of the waves. This process leads to exponential growth of the Stokes sideband, with a gain factor given by

$$g = \frac{i\pi N\omega_S}{m\omega_v n_S c} \frac{(\partial\alpha/\partial q)_0^2}{[\omega_S - (\omega_L - \omega_v)] + i\gamma} \tag{A4.41}$$

where m is the reduced nuclear mass, γ is the vibrational damping rate and n_S is the refractive index at the Stokes frequency.

A4.11.2 Stimulated Brillouin scattering

The analysis of stimulated Brillouin scattering (SBS) shares many features in common with that of SRS but differs in the significant manner that SBS involves a collective excitation of the material medium. Consequently, the properties of SBS can be quite different in different directions. Our analysis here is restricted to geometries in which the laser and Stokes fields propagate in opposite directions.

The nature of the gain of the SBS process can be understood from the following perspective. The laser and counterpropagating Stokes waves beat together and, by means of the electrostrictive response of the material, produce a sound wave at the beat frequency which travels in the direction of the laser field. Some of

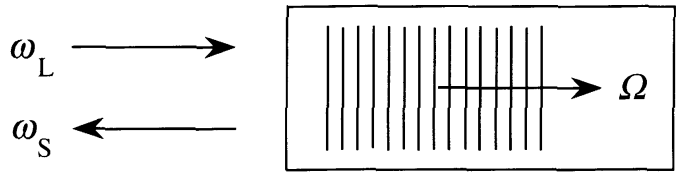


Figure A4.10. Illustration of the nature of stimulated Brillouin scattering.

Table A4.4. Properties of stimulated Brillouin scattering for a variety of materials.[a]

Substance	$\Omega_B/2\pi$ (MHz)	$\Gamma/2\pi$ (MHz)	g_0 (cm MW^{-1})
CS_2	5850	52.3	0.15
Acetone	4600	224	0.02
Toluene	5910	579	0.013
CCl_4	4390	520	0.006
Methanol	4250	250	0.013
Ethanol	4550	353	0.012
Benzene	6470	289	0.018
H_2O	5690	317	0.0048
Cyclohexane	5550	774	0.0068
CH_4 (1400 atm)	150	10	0.1
Optical glasses	15 000-26 000	10–106	0.004–0.025
SiO_2	25 800	78	0.0045

[a] Values are quoted for a wavelength of 0.694 μm. To convert to other laser frequencies ω, recall that Ω_B is proportional to ω, Γ is proportional to ω^2 and g_0 is independent of ω.

the laser light then scatters from this sound wave and, in doing so, becomes Stokes-shifted and consequently reinforces the Stokes wave. But since the Stokes wave is now stronger it tends to produce a stronger sound wave and, in this manner, the growth of the sound and Stokes wave mutually reinforce each other. These phenomena are illustrated in figure A4.10. A consistent analysis of this situation shows that the Stokes wave experiences exponential amplification according to

$$I_S(z) = I_S(0)e^{g_0 I_L} \tag{A4.42}$$

where the Brillouin gain factor is given by

$$g_0 = \frac{\gamma_e^2 \omega^2}{nvc^3 \rho_0 \Gamma_B} \tag{A4.43}$$

and where γ_e is the electrostrictive constant introduced earlier in equation (A4.11), ω is the laser frequency, n is the refractive index, v is the velocity of sound, ρ_0 is the mean material density and Γ_B is the phonon-damping rate. The properties of SBS for several materials are summarized in table A4.4.

A4.12 Multi-photon absorption

Multi-photon absorption refers to optical processes in which more than one photon is removed from the optical field in a single optical transition. Some typical multi-photon absorption processes as well as normal

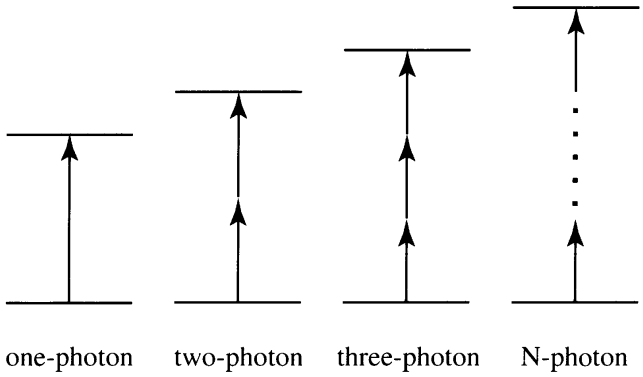


Figure A4.11. Illustration of one-photon and multi-photon absorption processes.

one-photon absorption are shown in figure A4.11. Multiphoton absorption processes are important for a number of reasons, including the fact that multi-photon absorption constitutes a nonlinear loss mechanism that can limit the efficiency of certain optical interactions and also because of applications of multi-photon absorption such as the use of two-photon microscopy for biological applications [72].

Multiphoton absorption can be described theoretically by calculating the transition rate from the ground state to the final state through use of time-dependent quantum-mechanical perturbation theory. The resulting expression is often referred to as Fermi's golden rule. This method applied to the case of two-photon absorption leads to a prediction for the value of the two-photon absorption cross section $\sigma_{ng}^{(2)}(\omega)$, which is defined such that the transition rate for transitions from level g to level n is given by

$$R_{ng}^{(2)} = \sigma_{ng}^{(2)}(\omega) I^2. \tag{A4.44}$$

One finds that the two-photon cross section has the form

$$\sigma_{ng}^{(2)}(\omega) = \frac{8\pi^3}{n^2 c^2} \left| \sum_m \frac{\mu_{nm}\mu_{mg}}{\hbar^2(\omega_{mg} - \omega)} \right|^2 \rho_f(\omega_{ng} = 2\omega) \tag{A4.45}$$

In this expression, μ_{nm} is the electric-dipole matrix element connecting levels n and m and $\rho_f(\omega_{ng} = 2\omega)$ is the density of states for the g to n transition evaluated at the laser frequency ω. Experimentally, two-photon cross sections are often quoted with intensities measured in units of photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. One finds, either from laboratory measurement [73] or from evaluation of equation (A4.45), that a typical value of the two-photon cross section is

$$\overline{\sigma}_{ng}^{(2)} \approx 10^{-49} \frac{\mathrm{cm}^4\,\mathrm{s}}{\mathrm{photon}^2}. \tag{A4.46}$$

These predictions can readily be extended to higher-order multi-photon transition rates.

A4.13 Optically induced damage

A topic of great practical importance is laser-induced damage of optical materials. Laser-induced damage is important because this process limits the maximum amount of optical power that can be transmitted through a given material and, consequently, limits the efficiency of many nonlinear optical interactions. The field of optical damage has been described in several review articles [74–77]; several investigations of optical damage include the following references [78,79]. There are several different mechanisms that lead to optical damage. In brief summary, these mechanisms are as follows:

Table A4.5. Optical damage threshold of fused silica.[a]

Pulse duration	Threshold fluence (J cm^{-2})	Threshold intensity (GW cm^{-2})	Comments
1 ps	1.3	1300	Deviation from $\tau^{1/2}$ scaling
10 ps	4.1	410	
100 ps	13	130	
1 ns	41	41	
10 ns	130	13	
100 ns	410	4.1	

[a] From Stuart *et al* [78] and other sources.

- Linear absorption, leading to localized heating and cracking of the optical material. This is the dominant damage mechanism for continuous-wave and long-pulse ($\gtrsim 1$ μs) laser beams.
- Avalanche breakdown, which is the dominant mechanism for pulsed lasers (shorter then $\lesssim 1$ μs) for intensities in the range 10^9–10^{12}/ W cm^{-1}.
- Multiphoton ionization or dissociation of the optical material, which is the dominant mechanism for intensities in the range 10^{12}–10^{16} W cm^{-2}.
- Direct (single cycle) field ionization, which is the dominant mechanism for intensities $>10^{10}$ W cm^{-2}.

This summary suggests that the avalanche breakdown mechanism is the dominant optical damage mechanism for laser pulses most often encountered in the laboratory and in applications. The nature of this mechanism is that a small number of free electrons initially present within the optical material are accelerated to high energies through their interaction with the laser field. These electrons can then impact-ionize other atoms within the material, thereby producing additional electrons which are subsequently accelerated by the laser field and eventually producing still more electrons. Some fraction of the energy imparted to each electron will lead to a localized heating of the material, which can eventually lead to damage of the material due to cracking or melting. The small number of electrons initially present within the material are created by one of several processes, including thermal excitation multi-photon excitation, or free electrons resulting from crystal defects.

Empirical evidence shows that, for many materials and for laser pulse lengths τ in the range of 10 ps to 100 ns, the threshold fluence for laser damage increases with pulse duration as $\tau^{1/2}$ and, consequently, the threshold intensity decreases with pulse duration as $\tau^{-1/2}$. Results for the case of fused silica are shown in table A4.5.

A4.14 Strong-field effects and high-order harmonic generation

Recent advances have led to the development of lasers that can produce pulses of only a few femtoseconds duration. New nonlinear optical phenomena become accessible with these short laser pulses for two different reasons: (1) ultrashort laser pulses of only modest energy can produce super-intense fields; and (2) nonlinear optical self-action effects are qualitatively different when excited by such short pulses, because of the dominance of dispersive and space-time coupling effects. Some of these new features are reviewed in the present section.

Let us first consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects have historically been modelled using the power-series expansion of equation (A4.1),

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{at} = e/a_0^2 = 2 \times 10^7$ statvolt cm^{-1} $= 6 \times 10^9$ V cm^{-1}. This field strength corresponds to a laser intensity of $I_{at} = 4 \times 10^{16}$ W cm^{-2}, which constitutes the threshold field-strength for exciting non-perturbative nonlinear optical response.

One of the dramatic consequences of excitation with intensities comparable to the atomic unit of intensity I_{at} is the occurrence of high-harmonic generation [80, 81]. In brief, if an atomic gas jet is irradiated by high-intensity laser radiation, it is observed that all odd harmonics of the laser frequency, up to some maximum value N_{max}, are emitted. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with N_{max} as large as 341.

This phenomenon can be understood in terms of a simple physical model [82]. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse and, each time the electron passes near the nucleus, it undergoes violent acceleration and emits a short pulse of radiation. This radiation is in the form of a train of short pulses; the spectrum of the radiation is the square of the Fourier transform of this pulse train, which contains the harmonics of the oscillation period up to some maximum frequency that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made qualitative to show that the maximum harmonic number is given by

$$N_{max}\hbar\omega = 3.17K + U_p \tag{A4.47}$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ioization energy of the atom.

Nonlinear optical self-action effects are also profoundly modified through excitation with ultrashort laser pulses. New phenomena come into play, including self-steepening of the laser pulse and space time focusing effects, in which different spectral components of the laser pulse undergo differing amounts of self-focusing. These effects have been described by a nonlinear envelope equation [83] which is a generalization of equation (A4.32) and have been studied extensively by Gaeta [84] in the context of supercontinuum generation.

References

[1] Lewis G N, Lipkin D and Magel T T 1941 *J. Am. Chem. Soc.* **63** 3005
[2] Boyd R W, Raymer M G and Narducci L M (ed) 1986 *Optical Instabilities* (Cambridge: Cambridge University Press)
[3] Spence D E, Kean P N and Sibbett W 1991 *Opt. Lett.* **16** 42
Sibbett W, Gran R S and Spence D E 1994 *Appl. Phys.—Lasers Opt.* B **58** 171
[4] Bloembergen N 1964 *Nonlinear Optics* (New York: Benjamin)
[5] Boyd R W 1992 *Nonlinear Optics* (San Diego, CA: Academic)
[6] Butcher P N and Cotter D 1990 *The Elements of Nonlinear Optics* (Cambridge: Cambridge University Press)
[7] Hannah D C, Yuratich M A and Cotter D 1979 *Nonlinear Optics of Free Atoms and Molecules* (Berlin: Springer)
[8] Shen Y R 1984 *The Principles of Nonlinear Optics* (New York: Wiley)
[9] Sutherland R L 1996 *Handbook of Nonlinear Optics* (New York: Dekker)
[10] Agrawal G P and Boyd R W (ed) 1992 *Contemporary Nonlinear Optics* (Boston, MA: Academic)
[11] Boyd R W 1996 *Laser Sources and Applications* ed A Miller and D M Finlayson (Scottish Universities Summer School in Physics) (Bristol: IOPP)
[12] Boyd R W 1999 *J. Mod. Opt.* **46** 367
[13] Günter P and Huignard J-P (ed) 1988 *Photorefractive Materials and Their Applications Vols 1 and 2 (Topics in Applied Physics 61 and 62)* (New York: Springer)
[14] D D Nolte (ed) 1995 *Photorefractive Effects and Materials* (Dordrecht: Kluwer)
[15] Prasad P N and Williams D J 1991 *Introduction of Nonlinear Opitcal Effects in Molecules and Polymers* (New York: Wiley)
[16] Chemla D S and Zyss J 1987 *Nonlinear Optical Properties of Organic Molecules and Crystals* vols 1 and 2 (Orlando, FL: Academic)

[17] Boyd R W and Fischer G L 2001 Nonlinear optical materials *Encyclopedia of Materials, Science and Technology* (Amsterdam: Pergamon) 6237-44
[18] Cleveland Crystals, Inc, 19306 Redwood Road, Cleveland, Ohio 44110 USA provides a large number of useful data sheets which may also be obtained at http://www.clevelandcrystals.com.
[19] Smith A V maintains a public domain nonlinear optics data base SNLO, which can be obtained at http://www.sandia.gov/imrl/XWEB1128/xxtal.htm
[20] Chase L L and van Stryland E W 1995 *CRC Handbook of Laser Science and Technology* (Boca Raton, FL: Chemical Rubber Company) section 8.1 (This reference provides an extensive tabulation of third-order nonlinear optical susceptibilities. The values of $\chi^{(3)}$ given in this reference need to be multiplied by a factor of four to conform with the standard convention of Bloembergen, which is the convention used in the present article.)
[21] Hache F, Ricard D, Flytzanis C and Kreibig U 1988 *Appl. Phys.* A **47** 347
[22] Sipe J E and Boyd R W 1992 *Phys. Rev.* A **46** 1614
Boyd R W and Sipe J E 1994 *J. Opt. Soc. Am.* B **11** 297
Fischer G L, Boyd R W, Gehr R J, Jenekhe S A, Osaheni J A, Sipe J E and Weller-Brophy L A 1995 *Phys. Rev. Lett.* **74** 1871
[23] Nelson R L and Boyd R W 1999 *Appl. Phys. Lett.* **74** 2417
[24] Chase L L and van Stryland E W 1995 *CRC Handbook of Laser Science and Technology* (Boca Raton, FL: Chemical Rubber Company) section 8.1
[25] Bloembergen N *et al* 1969 *Opt. Commun.* **1** 195
[26] Vogel E M *et al* 1991 *Phys. Chem. Glasses* **32** 231
[27] Hall D W *et al* 1989 *Appl. Phys. Lett.* **54** 1293
[28] Lawrence B L *et al* 1994 *Electron. Lett.* **30** 447
[29] Carter G M *et al* 1985 *Appl. Phys. Lett.* **47** 457
[30] Molyneux *et al* S 1993 *Opt. Lett.* **18** 2093
[31] Erlich J E *et al* 1993 *J. Mod. Opt.* **40** 2151
[32] Sutherland R L 1996 *Handbook of Nonlinear Optics* (New York: Dekker) ch 8
[33] Pennington D M *et al* 1989 *Phys. Rev.* A **39** 3003
[34] Euler H and Kockel B 1935 *Naturwiss.* **23** 246
[35] Hau L V *et al* 1999 *Nature* **397** 594
[36] Kramer M A, Tompkin W R and Boyd R W 1986 *Phys. Rev.* A **34** 2026
[37] Franken P A, Hill A E, Peters C W and Weinreich G 1961 *Phys. Rev. Lett.* **7** 118
[38] Midwinter J E and Warner J 1965 *Brit. J. Appl. Phys.* **16** 1135
[39] Armstrong J A, Bloembergen N, Ducuing J and Pershan P S *Phys. Rev.* **127** 1918
[40] Boyd G D and Kleinman D A 1968 *J. Appl. Phys.* **39** 3597
[41] Ward J F and New G H C 1969 *Phys. Rev.* **185** 57
[42] Miles R B and Harris S E 1973 *IEEE J. Quantum Electron.* **QE- 9** 470
[43] Craxton R S 1980 *Opt. Commun.* **34** 474
Seka W, Jacobs S D, Rizzo J E, Boni R and Craxton R S 1980 *Opt. Commun.* **34** 469
[44] Giordmaine J A and Miller R C 1965 *Phys. Rev. Lett.* **14** 973
Giordmaine J A and Miller R C 1966 *Appl. Phys. Lett.* **9** 298
[45] Smith R G *et al* 1968 *Appl. Phys. Lett.* **12** 308
[46] Bosenberg W R, Pelouch W S and Tang C L 1989 *Appl. Phys. Lett.* **55** 1952
[47] Byer R L and Herbst R L 1977 *Tunable Infrared Generation* ed Y R Shen (Berlin: Springer)
[48] Zel'dovich B Ya, Popovichev V I, Ragulsky V V and Faizullov F S 1972 *JETP Lett.* **15** 109
[49] Zel'dovich B Ya, Pilipetsky N F and Shkunov V V 1985 *Principles of Phase Conjugation* (Berlin: Springer)
[50] Fisher R A 1983 (ed) *Optical Phase Conjugation* (Orlando, FL: Academic)
[51] Boyd R W and Grynberg G 1992 *Contemporary Nonlinear Optics* ed G P Agrawal and R W Boyd (Boston, MA: Academic)
[52] Gauthier D J, Boyd R W, Jungquist R K, Lisson J B and Voci L L 1989 *Opt. Lett.* **14** 325
[53] Gaeta A L and Boyd R W 1988 *Phys. Rev. Lett.* **60** 2618
[54] Svelto O 1974 *Progress in Optics* vol XII, ed E Wolf (Amsterdam: North-Holland)
[55] Blair S, Wagner K and McLeod R 1994 *Opt. Lett.* **19** 1943
[56] Bespalov V I and Talanov V I 1966 *JETP Lett.* **3** 307
[57] Maillotte H, Monneret J and Froehly C 1990 *Opt. Commun.* **77** 241
[58] Jain M, Xia H, Yin G Y, Merriam A J and Harris S E 1996 *Phys. Rev. Lett.* **77** 4326
[59] Zakharov V E and Shabat A B 1972 *Sov. Phys.–JETP* **34** 62
[60] Mollenauer L F, Stolen R H and Gordon J P 1980 *Phys. Rev. Lett.* **45** 1095
[61] Gibbs H M 1985 *Optical Bistability* (Orlando, FL: Academic)
[62] Gibbs H M, McCall S L and Venkatesan T N 1976 *Phys. Rev. Lett.* **36** 113
[63] Lugiato L A 1984 'Theory of optical bistability' *Progress in Optics* vol XXI, ed E Wolf (Amsterdam: North-Holland)
[64] Szöke A, Daneu V, Goldhar J and Kurnit N A 1969 *Appl. Phys. Lett.* **15** 376
[65] Stegeman G I and Miller A 1993 *Photonics in Switching* (San Diego, CA: Academic)

[66] Gibbs H M, Khitrova G and Peyghambarian N (ed) 1990 *Nonlinear Photonics* (Berlin: Springer)
[67] Fabelinskii I L 1986 *Molecular Scattering of Light* (New York: Plenum)
[68] Hellwarth R W 1963 *Phys. Rev.* **130** 1850
[69] Kaiser W and Maier M 1972 *Laser Handbook* ed F T Arechi and E O Schulz-DuBois (Amsterdam: North-Holland)
[70] Simon U and Tittel F K 1994 *Methods of Experimental Physics* vol III, ed R G Hulet and F B Dunning (Orlando, FL: Academic)
[71] Garmire E, Pandarese F and Townes C H 1963 *Phys. Rev. Lett.* **11** 160
[72] Denk W, Strickler J H and Webb W W 1990 *Science* **248** 73
Xu C and Webb W W 1997 *Topics in Fluorescence Spectroscopy, Volume 5: Nonlinear and Two-Photon-Induced Fluorescence* ed J Lakowicz (New York: Plenum) ch 11
[73] Xu C and Webb W W 1996 *J. Opt. Soc. Am.* B **13** 481
[74] Bloembergen N 1974 *IEEE J. Quantum Electron.* **10** 375
[75] Lowdermilk W H and Milam D 1981 *IEEE J. Quantum Electron.* **17** 1888
[76] Raizer Y P 1965 *Sov. Phys.–JETP* **21** 1009
[77] Manenkov A A and Prokhorov A M 1986 *Sov. Phys.–Usp.* **29** 104
[78] Stuart B C *et al* 1995 *Phys. Rev. Lett.* **74** 2248
Stuart B C *et al* 1996 *Phys. Rev.* B **53** 1749
[79] Du D *et al* 1994 *Appl. Phys. Lett.* **64** 3071
[80] Ferray M *et al* 1988 *J. Phys. B: At. Mol. Opt. Phys.* **21** L31
[81] Chang Z *et al* 1997 *Phys. Rev. Lett.* **79** 2967
[82] Corkum P B 1993 *Phys. Rev. Lett.* **71** 1994
[83] Brabec T and Krausz F 1997 *Phys. Rev. Lett.* **78** 3282
[84] Gaeta A L 2000 *Phys. Rev. Lett.* **84** 3583

A5
Interference and polarization

Alan Rogers

A5.1 Introduction

Light is a form of electromagnetic radiation. It is distinguished from other forms only by its particular wavelength range: 0.4–0.7 μm for the visible range, with regions below 0.4 μm (ultraviolet) and above 0.7 μm (infrared) also conventionally classified as lying within the 'optical' range.

Electromagnetic radiation exhibits both wave and particle (i.e. photon) properties. Wave properties are usually more appropriate for description of behaviour at the longer wavelengths (radio and microwave wavelengths for example) where photon energies are small and, therefore, the number of photons per energy range is large; particle properties are more appropriate at the shorter wavelengths (x-rays, γ-rays, for example), where there are very few photons per energy range and the discrete, particulate nature of the radiation is more evident in the, now, relatively rare occurrence of photon arrival at a detector. The optical range is intermediate between these two regimes, so that it is sometimes more useful to work with the wave description and sometimes with the photon description. Examples of the former are the subject of the present chapter, interference and polarization; examples of the latter are photodetection and photoemission processes.

Interference and polarization are both concerned with wave interaction. Interference is concerned with the interaction between waves, whilst polarization is concerned with the interaction of a wave within itself. They are not independent: the interference between waves depends upon their relative polarization states, and both interference and polarization phenomena depend upon the 'purity' of the wave—on its 'coherence'. We shall, however, deal with each topic separately and in turn, cross-linking the two where appropriate.

A5.2 Interference

A5.2.1 Wave coherence

We shall begin by considering the interference between 'pure' waves. 'Pure' in this sense means that the wave is idealized as a sinusoid in all time and space, so that the way in which such waves interact is constant for all time and space. This is a useful idealization in that many of the light sources with which we deal in practice (including most lasers) are 'pure' for the observation times and experimental volumes which are used. A stricter definition of this 'purity' is that the wave within itself, or two (or more) interfering waves, maintain a constant phase relationship over these times and spaces—they are then said to be self- or mutually-'coherent', respectively. We shall deal more fully with the concept of coherence and its quantitative effect on interference phenomena in section A5.2.4.

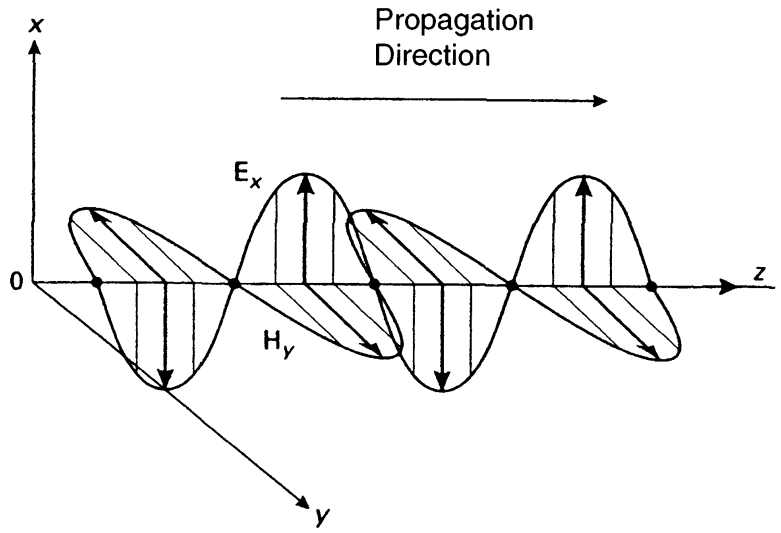


Figure A5.1. Electromagnetic wave. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

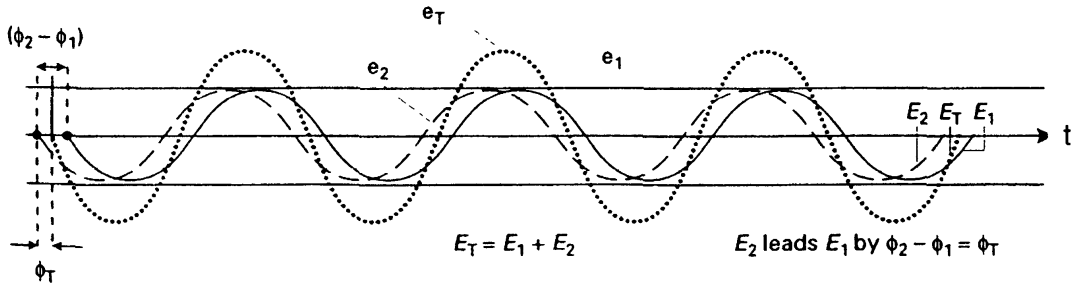


Figure A5.2. Addition of two waves of the same frequency. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

A5.2.2 Coherent-wave interference

Light, in the wave description, comprises electric and magnetic fields oscillating in phase and mutually at right angles in space (figure A5.1). We know that these fields are vector fields since they represent forces (on unit charge and unit magnetic pole, respectively). The fields will thus add vectorially. Consequently, when two light waves are superimposed on each other we obtain the resultant by constructing their vector sum at each point in time and space.

It is clear that the amplitude of such a wave in a given direction can only be affected by another if they both have components in that direction. Two waves oscillating mutually at right angles cannot, therefore, influence each other and, to find the effect at other angles, one wave must be resolved in the direction of the other. The simplest case to deal with is that where the two waves of the same frequency are both oscillating in the same direction and propagating in the same (orthogonal) direction, so that their mutual interaction is maximized.

If two such sinusoids are added, the result is another sinusoid. Suppose that two such light waves given, via their electric fields, as:

$$e_1 = E_1 \cos(\omega t + \phi_1)$$
$$e_2 = E_2 \cos(\omega t + \phi_2)$$

have the same oscillation direction and are superimposed at a point in space (figure A5.2). We know that the

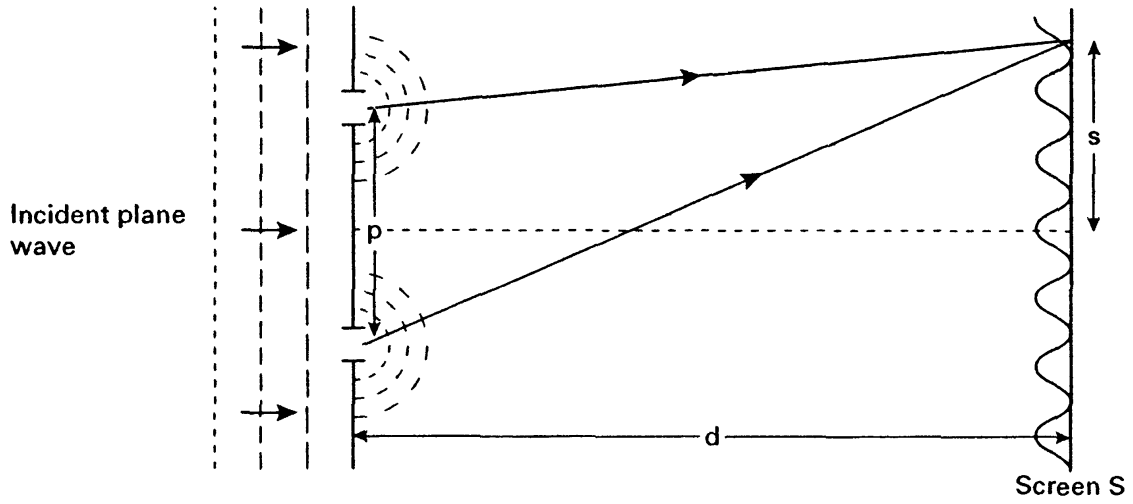


Figure A5.3. Two-slit interference. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

resultant field at the point will be given, using elementary trigonometry, by:

$$e_t = E_T \cos(\omega t + \phi_T)$$

where:

$$E_T^2 = E_1^2 + E_2^2 + 2E_1E_2\cos(\phi_2 - \phi_1)$$

and

$$\tan\phi_T = \frac{(E_1 \sin\phi_1 + E_2 \sin\phi_2)}{(E_1 \cos\phi_1 + E_2 \cos\phi_2)}.$$

For the important case where $E_1 = E_2 = E$, say, we have:

$$E_T^2 = 4E^2 \cos^2 \frac{(\phi_2 - \phi_1)}{2} \tag{A5.1}$$

and

$$\tan\phi_T = \tan\frac{(\phi_2 + \phi_1)}{2}.$$

The intensity of the wave will be proportional to E_T^2 so that, from (A5.1) it can be seen to vary from $4E^2$ to 0, as $(\phi_2 - \phi_1)/2$ varies from 0 to $\pi/2$.

Consider now the arrangement shown in figure (A5.3). Here two slits, separated by a distance p, are illuminated by a plane wave. The portions of the wave which pass through the slits will interfere on the screen S, a distance d away. Now each of the slits will act as a source of cylindrical waves, from Huygens' principle. Moreover, since they originate from the same plane wave, they will start in phase. On a line displaced a distance s from the line of symmetry on the screen the waves from the two slits will differ in phase by:

$$\delta = \frac{2\pi}{\lambda} \cdot \frac{sp}{d} = k \cdot \frac{sp}{d} (d \gg s, p) \qquad \text{with } k = \frac{2\pi}{\lambda}.$$

Thus, as s increases, the intensity will vary between a maximum and zero, in accordance with equation (A5.1). These variations will be viewed as fringes, i.e. lines of constant intensity parallel with slits. They are known as Young's fringes, after their discoverer, and are the simplest example of light interference. We shall now consider some important measuring instruments which depend upon these interference principles.

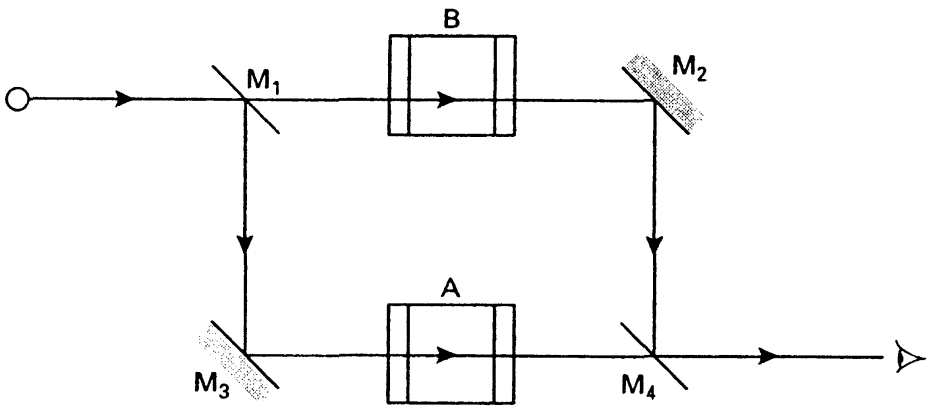


Figure A5.4. Basic Mach–Zehnder interferometer. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

A5.2.3 Interferometers

In section A5.2.2 the essentials of dual-beam interference were discussed. Although very simple in concept, the phenomenon is extremely useful in practice. The reason for this is that the maxima of the resulting fringe pattern appear where the phase difference between the interfering light beams is a multiple of 2π. Any quite small perturbation in the phase of one of the beams will thus cause a transverse shift in the position of the fringe pattern which, using opto-electronic techniques, is readily observed to about 10^{-4} of the fringe spacing. Such a shift is caused by, for example, an increase in path length of one of the beams by one hundredth of a wavelength, or about 5×10^{-9} m for visible light. This means that differential distances of this order can be measured, leading to obvious applications in, for example, sensitive strain monitoring on mechanical structures.

Another example of a dual-beam interferometer is shown in figure A5.4. Here the two beams are produced from the partial reflection and transmission at a dielectric, or partially silvered, mirror M_1. Another such mirror, M_4, recombines the two beams after their separate passages. Such an arrangement is known as a Mach–Zehnder interferometer and is used extensively to monitor changes in the phase differences between two optical paths. An optical-fibre version of a Mach–Zehnder interferometer is shown in figure A5.5. In this case the 'mirrors' are optical couplings between the cores of the two fibres. The 'fringe pattern' consists effectively of just one fringe, since the fibre core acts as an efficient spatial filter. However, the light which emerges from the fibre end (E) clearly will depend on the phase relationship between the two optical paths when the light beams recombine at R and, thus, it will depend critically on propagation conditions within the two arms. If one of the arms varies in temperature, strain, density, etc compared with the other, then the light output will also vary. Hence the latter can be used as a sensitive measure of any physical parameters which are capable of modifying the phase propagation properties of the fibre.

Finally, figure A5.6(a) shows another, rather more sophisticated variation of the Mach–Zehnder idea. In this case the beams are again separated by means of a beam-splitting mirror, but are returned to the same point by fully silvered mirrors placed at the ends of the two respective optical paths. (The plate P is necessary to provide equal optical paths for the two beams in the absence of any perturbation). This arrangement is called the Michelson interferometer after the experimenter who, in the late 19th century, used optical interferometry with great skill to make many physical advances [1]. His interferometer (not to be confused with his 'stellar' interferometer, of which more later) allows for a greater accuracy of fine adjustment by control of the reflecting mirrors, but uses, of course, just the same basic interferometric principles as before. The optical-fibre version of this device is shown in figure A5.6(b).

For completeness and because of its historical importance, mention must be made of the use of Michel-

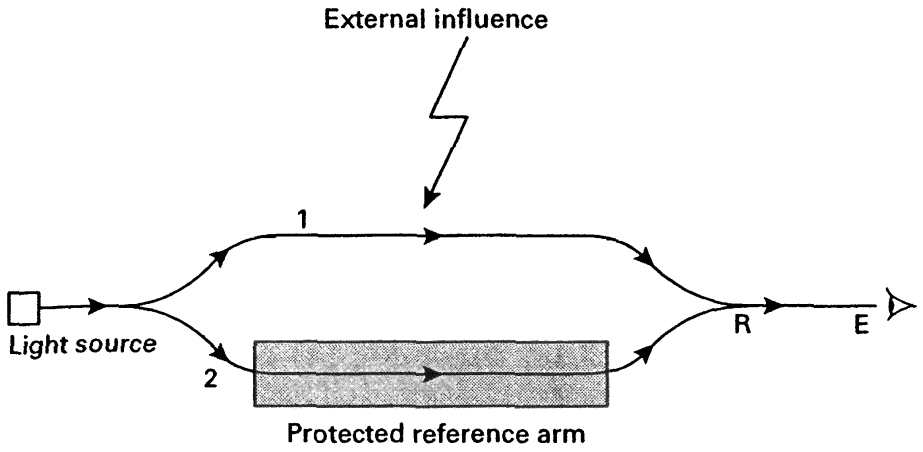


Figure A5.5. An optical-fibre Mach–Zehnder interferometer. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

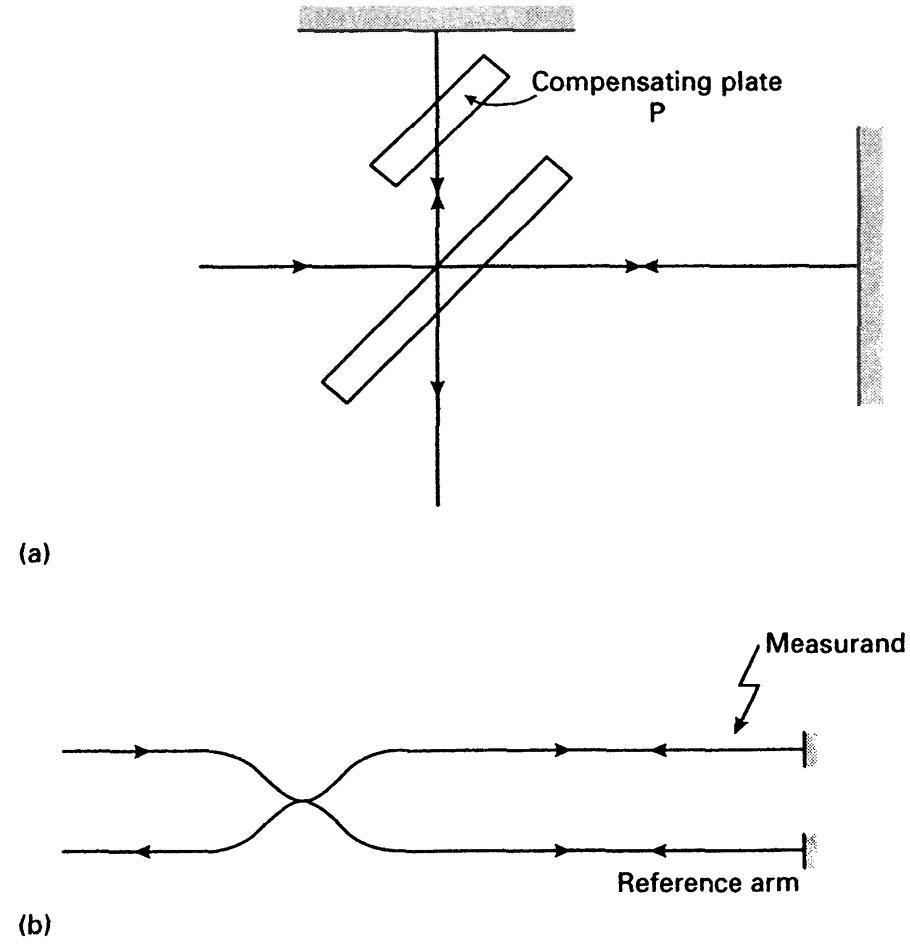


Figure A5.6. Michelson interferometers. (*a*) Bulk version; (*b*) optical-fibre version. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

son's interferometer in the famous Michelson–Morley experiment of 1887 [2]. This demonstrated that light travelled with the same velocity in each of two orthogonal paths, no matter what was the orientation of the interferometer with respect to the earth's 'proper' motion through space. This result was crucial to Einstein's formulation of special relativity in 1905 and is thus certainly one of the most important results in the history

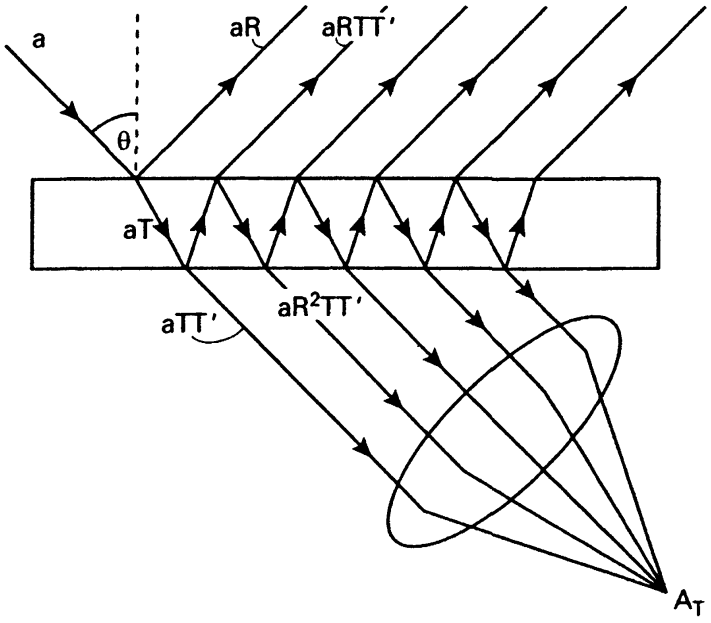


Figure A5.7. Multiple wave interference. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

of experimental physics.

Valuable as dual-beam interferometry is, it suffers from the limitation that its accuracy depends upon the location of the maxima (or minima) of a sinusoidal variation. For very accurate work, such as precision spectroscopy, this limitation is severe. By using the interference amongst many beams, rather than just two, we find that we can improve the accuracy very considerably. We can see this by considering the arrangement of figure A5.7. Light from a single source gives a large number of phase-related, separate beams by means of multiple reflections and transmissions within a dielectric (e.g. glass) plate. For a given angle of incidence (θ) there will be fixed values for the amplitude transmission (T, T') and amplitude reflection (R) coefficients, as shown. If we start with a wave of amplitude 'a', the waves on successive reflections will suffer attenuation by a constant factor and will increase in phase by a constant amount. If we consider the transmitted light only, then the total amplitude which arrives at the focus of the lens L is given by the sum:

$$A_T = aTT' \exp(\mathrm{i}\omega t) + aTT'R^2 \exp \mathrm{i}(\omega t - ks) + aTT'R^4 \exp \mathrm{i}(\omega t - 2ks) + \cdots$$

where, again,

$$k = \frac{2\pi}{\lambda}$$

and where s is the optical path difference between successive reflections at the lower surface (including the phase changes on reflection and transmission). The sum can be expressed as:

$$A_T = aTT' \sum_{p=0}^{\infty} R^{2p} \exp \mathrm{i}(\omega \mathrm{t} - \mathrm{pks})$$

which is a geometric series whose sum value is:

$$A_T = \frac{aTT' \exp(\mathrm{i}\omega \mathrm{t})}{(1 - R^2 \exp(-\mathrm{iks}))}.$$

Hence the intensity (I) of the light is given by:

$$I \propto |A_T|^2 = \frac{(aTT')^2}{(1 + R^4 - 2R^2 \cos ks)}. \tag{A5.2}$$

We note from this equation the ratio of maximum and minimum intensities:

$$\frac{I_{\max}}{I_{\min}} = \frac{(1 + R^2)^2}{(1 - R^2)^2}$$

so that the fringe contrast increases with R. However, as R increases so does the attenuation between the successive reflections. Hence, the total transmitted light power will fall. Figure A5.8 shows how I varies with ks for different values of R. We note that the fringes become very sharp for large values of R. Hence the position of the maxima may now be accurately determined. Further, since the spacing of the maxima specifies ks, this information can be used to determine either k or s, if the other is known. Consequently, multiple interference may be used either to select (or measure) a very specific wavelength, or to measure very small changes in optical path length.

The physical reason for the sharpening of the fringes as the reflectivity increases is indicated in figure A5.9. The addition of the multiplicity of waves is equivalent to the addition of vectors with progressively decreasing amplitude and increasing relative phase. For small reflectivity (figure A5.9(a)) the wave amplitudes decrease rapidly, so that the phase increase has a relatively small effect on the resultant wave amplitude. In the case of high reflectivity (figure A5.9(b)), the reverse is the case and a small successive phase change rapidly reduces the resultant.

Two important devices based on these ideas of multiple reflection are the Fabry–Pérot interferometer and the Fabry–Pérot etalon [3]. In the former case the distance between the two surfaces is finely variable for fringe control; in the case of the etalon the surfaces are fixed. In both cases the flatness and parallelism of the surfaces must be accurate to $\sim\lambda/100$ for good quality fringes. This is difficult to achieve in a variable device, and the etalon is preferred for most practical purposes.

The Fabry–Pérot interferometer is extremely important in opto-electronics. We have already noted its wavelength selectivity but we should also note its ability to store optical energy by continually bouncing light between two parallel mirrors. For this reason it is often called a Fabry–Pérot 'cavity' and is, roughly speaking, the optical equivalent of an electronic oscillator. The optical term is 'resonator', and it is this property which makes it an integral feature in all lasers. Because of its importance it is useful to be aware of the parameters which characterize the performance of the Fabry–Pérot resonator. There are three main parameters: (i) finesse; (ii) resolving power; (iii) free spectral range.

These parameters relate, as is to be expected, to the instrument's ability to separate closely-spaced optical wavelengths. The first is a measure of the sharpness of the fringes. This measure is normalized to the separation of the fringes for a single wavelength, since, clearly, there is no advantage in having narrow fringes if they are all crowded together, so that the orders of different wavelengths overlap. We hence define a quantity:

$$\Phi = \frac{\text{separation of successive fringes}}{\text{width at half maximum of a single fringe}}$$

Φ is called the 'finesse' and is roughly equivalent to the 'Q' ('quality' factor measuring the sharpness of the resonance) of an electronic oscillator.

It is easy to derive an expression for Φ from equation (A5.2) as follows. Equation (A5.2) may be written in the form:

$$I = \frac{I_{\max}}{1 + F \sin^2(\frac{1}{2}\Psi)}$$

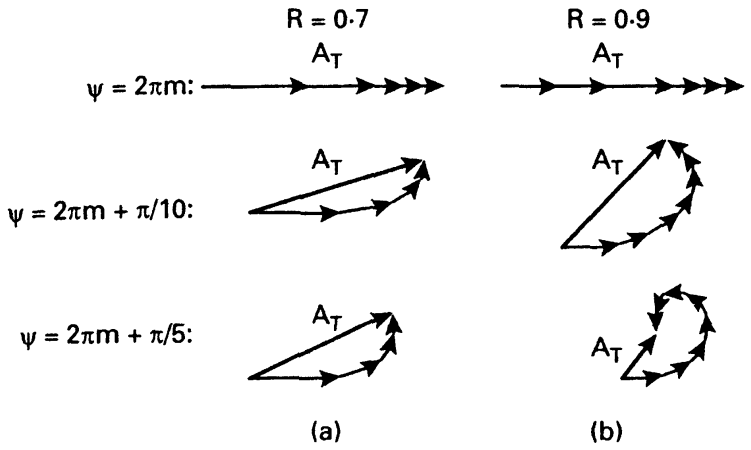


Figure A5.8. Variation of intensity with optical path, for various reflectivities, in a multiple interference plate. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

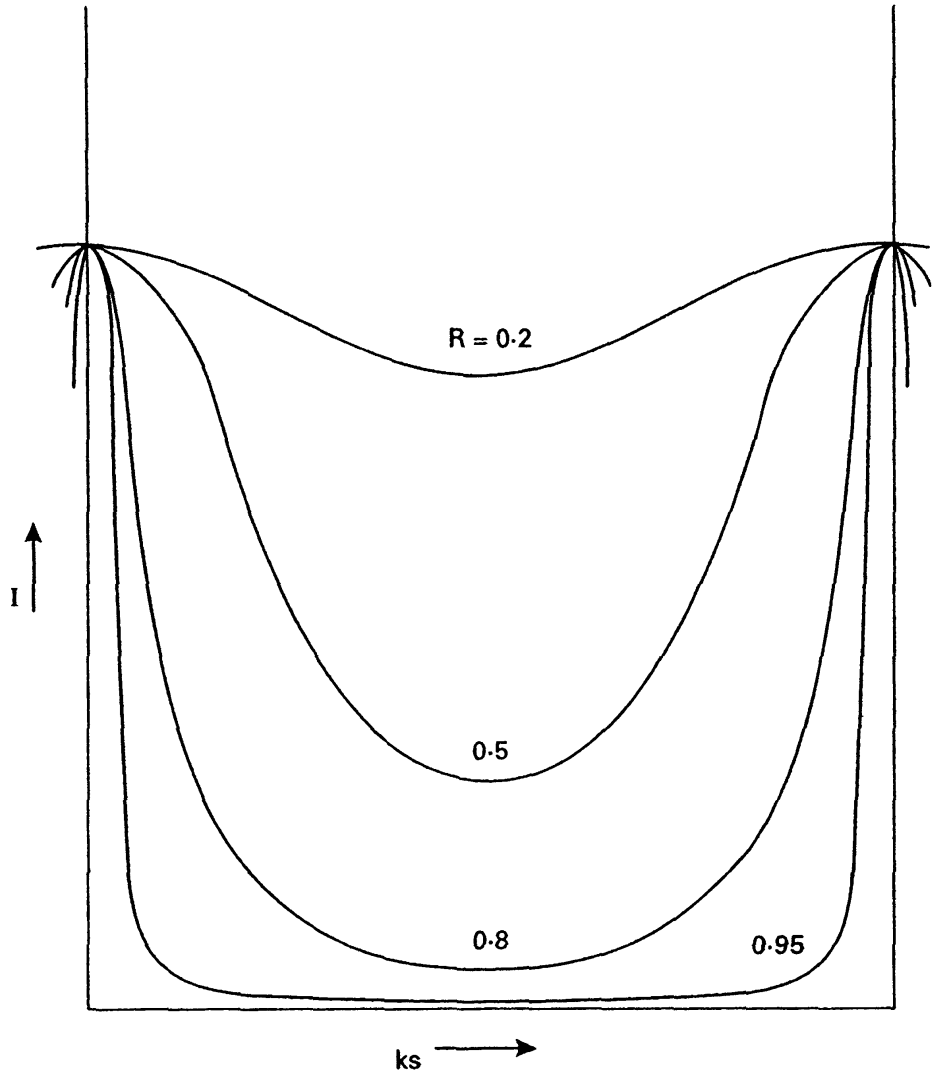


Figure A5.9. Graphical illustration of the dependence of fringe sharpness on reflectivity (R). (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

where

$$F = \frac{4R^2}{(1-R^2)^2}$$

and

$$\Psi = ks.$$

F is sometimes known as the 'coefficient of finesse'. From this it is clear that $I = I_{\max}/2$ when:

$$\Psi_h = \frac{2}{\sqrt{F}}.$$

Hence the width at half maximum $= 2\Psi_h = 4/\sqrt{F}$. The 'Ψ distance' between successive maxima is just 2π and, thus, the finesse is given by:

$$\Phi = \frac{2\pi}{2\Psi_h} = \frac{\pi\sqrt{F}}{2} = \frac{\pi R}{(1-R^2)}.$$

This quantity has a value of two for a dual beam interferometer. For a Fabry–Pérot etalon with $R = 0.9$ its value is 15. Clearly, the higher the value of R the sharper are the fringes for a given fringe separation and the more wavelength-selective is the device.

The next quantity we need to look at is the resolving power. This is a measure of the smallest detectable wavelength separation ($\delta\lambda$) at a given wavelength (λ) and is defined as:

$$\rho = \frac{\lambda}{\delta\lambda}.$$

If we take λ to be that which corresponds to a ψ difference equal to the width of the half maximum, we find that:

$$\rho = \frac{\lambda}{\delta\lambda} = p \times \text{finesse}$$

i.e. $\rho = p\Phi$ where p is the 'order' of the maximum. If the etalon is being viewed close to normal incidence, then p will be effectively just the number of wavelengths in a double passage across the etalon. If the etalon has optical thickness t we have $p = 2t/\lambda$ and with:

$$F = \frac{4R^2}{(1-R^2)^2}$$

we have:

$$\rho = \pi t \frac{\sqrt{F}}{\lambda}.$$

Resolving power, ρ, is typically of the order of 10^6, compared with a figure $\sim 10^4$ for a dual beam interferometer such as the Michelson (see section A5.2.5(i)). The ratio of these figures thus represents the improvement in accuracy afforded by multiple beam interferometry over dual beam techniques.

Finally, we define a quantity concerned with the overlapping of orders. If the range of wavelengths ($\Delta\lambda$) under investigation is such that the $(p+1)$th maximum of λ is to coincide with the pth maximum of $(\lambda + \Delta\lambda)$, then, clearly, there is an unresolvable confusion. For this to be so:

$$(p+1)k = p(k + \Delta k)$$

so that

$$\frac{\Delta k}{k} = \frac{\Delta\lambda}{\lambda} = \frac{1}{p}.$$

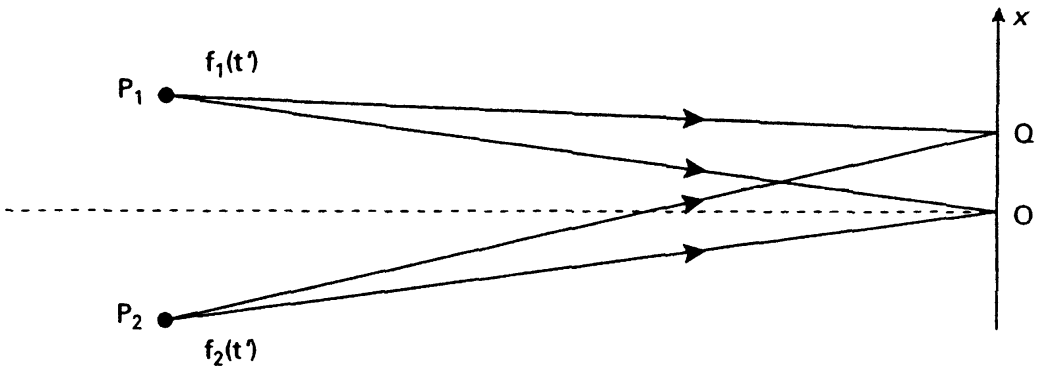


Figure A5.10. Interference between partially-coherent sources. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

Again, close to normal incidence we may write, with $p = 2t/\lambda$:

$$\Delta\lambda = \frac{\lambda}{p} = \frac{\lambda^2}{2t}.$$

$\Delta\lambda$ is called the 'free spectral range' of the etalon and represents the maximum usable wavelength range without recourse to prior separation of the confusable wavelengths.

For a more detailed discussion of the Fabry–Pérot interferometer see [3].

A5.2.4 Interference between partially-coherent waves

In dealing with the subjects of interference it has been assumed that each of the interfering waves bears a constant phase relationship to the others in both time and space. Such an assumption cannot be valid for all time and space intervals since the atomic emission processes which give rise to light are largely uncorrelated, except for the special case of laser emission. In this section we shall look at interference between waves which are not fully coherent, but only 'partially coherent'. These are waves which are mutually related in phase to only a limited extent, which must be quantified. This can conveniently be done by considering again the problem of dual beam interference, but this time with partially-coherent beams.

Consider the two-beam interference diagram of figure A5.10. It is clear, from our previous look at this topic, that interference fringes will be formed if the two waves bear a constant phase relationship to each other, but we must now consider the form of the interference pattern for varying degrees of mutual coherence. In particular, we must consider the 'visibility' of the pattern; in other words the extent to which it contains measurable structure and contrast.

At the point 0 (in figure A5.10) the (complex) amplitude resulting from the two sources P_1 and P_2 is given by:

$$A = f_1(t'') + f_2(t'')$$

where $t'' = t' + \tau_0$, τ_0 is the time taken for light to travel from P_1 or P_2 to 0. If f_1, f_2 represent the electric field amplitudes of the waves, the observed intensity at 0 will be given by the square of the modulus of this complex number. Hence, in this case the optical intensity is given by:

$$I_0 = \langle AA^* \rangle = \langle (f_1(t'') + f_2(t''))(f_1^*(t'') + f_2^*(t'')) \rangle$$

where the triangular brackets indicate an average taken over the response time of the detector (e.g. the human eye) and we assume that f_1 and f_2 contain the required constant of proportionality ($K^{1/2}$) to relate optical intensity with electric field strength, i.e. $I = KE^2$.

At point Q the amplitudes will be:

$$f_1(t'' - \tau/2),\ f_2(t'' + \tau/2)$$

τ being the time difference between paths P_2Q and P_1Q. Writing $t = t'' - \tau/2$, we have the intensity at Q:

$$I_Q = \langle (f_1(t) + f_2(t+\tau))(f_1^*(t) + f_2^*(t+\tau)) \rangle$$

i.e.

$$\begin{aligned} I_Q = {} & \langle f_1(t) f_1^*(t) \rangle + \langle f_2(t) f_2^*(t) \rangle \\ & + \langle f_2(t+\tau) f_1^*(t) \rangle + \langle f_1(t) f_2^*(t+\tau) \rangle. \end{aligned} \tag{A5.3}$$

The first two terms are clearly the independent intensities, I_1, I_2, of the two sources at Q. The second two terms will have values that depend upon the extent to which f_1 and f_2 correlate in phase and amplitude when displaced in time by τ. We may, in fact, define a 'mutual correlation function', $c_{12}(t)$:

$$\begin{aligned} c_{12}(t) &= \langle f_1(t) f_2^*(t+\tau) \rangle \\ c_{12}^*(t) &= \langle f_1^*(t) f_2(t+\tau) \rangle. \end{aligned}$$

We may note, in passing, that each of these terms will be zero if f_1 and f_2 have orthogonal polarizations, since, in that case, neither field amplitude has a component in the direction of the other, there can be no superposition, and the two cannot interfere. Hence, the average value of their product is again just the product of their averages, each of which is zero, being a sinusoid.

If $c_{12}(\tau)$ is now written in the form:

$$c_{12}(\tau) = |c_{12}(\tau)| \exp(i\omega\tau)$$

(which is valid provided that f_1 and f_2 are sinusoids in ωt) we have:

$$c_{12}(\tau) + c_{12}^*(\tau) = 2|c_{12}(\tau)| \cos \omega\tau.$$

Hence, provided that we observe the light intensity at Q with a detector which has a response time very much greater than the coherence times (self and mutual) of the sources (so that the time averages are valid) then we may write the intensity at Q as (equation (A5.3)):

$$I_Q = I_1 + I_2 + 2|c_{12}(\tau)| \cos \omega\tau. \tag{A5.4}$$

As we move along x we shall effectively increase τ, so we shall see a variation in intensity whose amplitude will be $2|c_{12}(\tau)|$ (i.e. twice the modulus of the mutual correlation function) and which varies about a mean value equal to the sum of the two intensities (figure A5.11). Thus we have an experimental method by which the mutual correlation of the sources, $c_{12}(\tau)$, can be measured.

If we now define a fringe visibility for this interference pattern by:

$$V = \frac{(I_{\max} - I_{\min})}{(I_{\max} + I_{\min})}$$

which quantifies the contrast in the pattern, i.e. the difference between maxima and minima as a fraction of the mean level, then, from equation (A5.4):

$$V(\tau) = \frac{2|c_{12}(\tau)|}{(I_1 + I_2)}$$

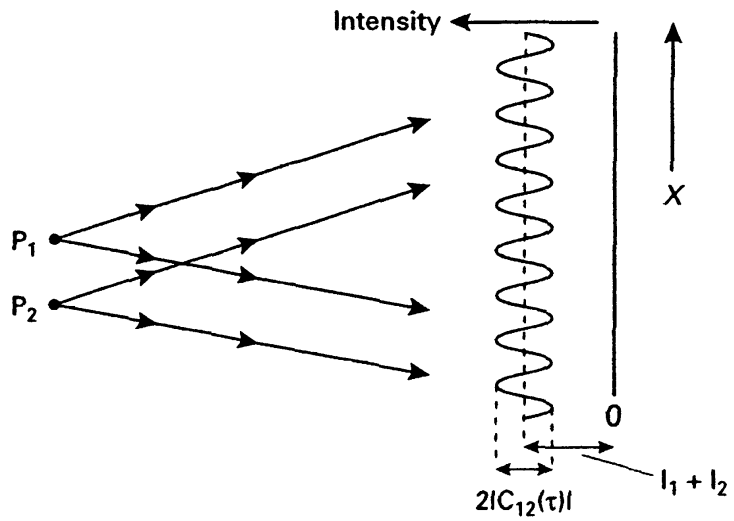


Figure A5.11. Mutual coherence function ($|C_{12}(\tau)|$) from the two-source interference pattern. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

so that the visibility of the fringes is seen to be directly related to the mutual correlation of the sources. We further define a 'coherence' function, $\gamma(t)$, which is just the mutual correlation function normalized to its value when $\tau = 0$, so that it now only depends upon differences between the phases at τ. The value of the mutual correlation function at $\tau = 0$ is given by:

$$c_{12}(0) = \langle f_1(t) f_2^*(t) \rangle = K\langle E_1 E_2 \rangle = (I_1 I_2)^{1/2}$$

so with:

$$\gamma(\tau) = \frac{|c_{12}(\tau)|}{|c_{12}(0)|}$$

we have:

$$\gamma(\tau) = \frac{|c_{12}(\tau)|}{(I_1 I_2)^{1/2}}.$$

Hence the visibility function $V(\tau)$ is related to the coherence function $\gamma(\tau)$ by:

$$V(\tau) = \frac{2(I_1 I_2)^{1/2}}{(I_1 + I_2)} \cdot \gamma(\tau)$$

and, if the two intensities are equal, we have:

$$V(\tau) = \gamma(\tau)$$

i.e. the visibility and coherence functions are identical.

From this we may conclude that, for equal intensity coherent sources, the visibility is 100% ($\gamma = 1$); for incoherent sources it is zero; and for partially coherent sources the visibility gives a direct measure of the actual coherence.

If we arrange that the points P_1 and P_2 are pinholes equidistant from and illuminated by a single source S, then the visibility function clearly measures the self-coherence of S. Suppose now that the two holes are placed in front of an extended source, S, as shown in figure A5.12, and that their separation is variable.

The interference pattern produced by these sources of light now measures the correlation between the two corresponding points on the extended source. If the separation is initially zero and is increased until

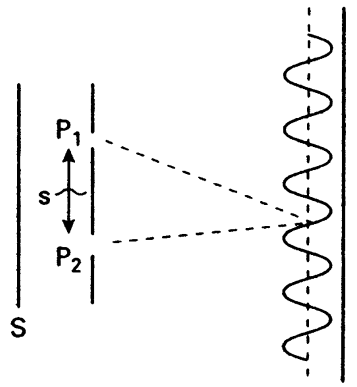


Figure A5.12. Extended-source interference. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

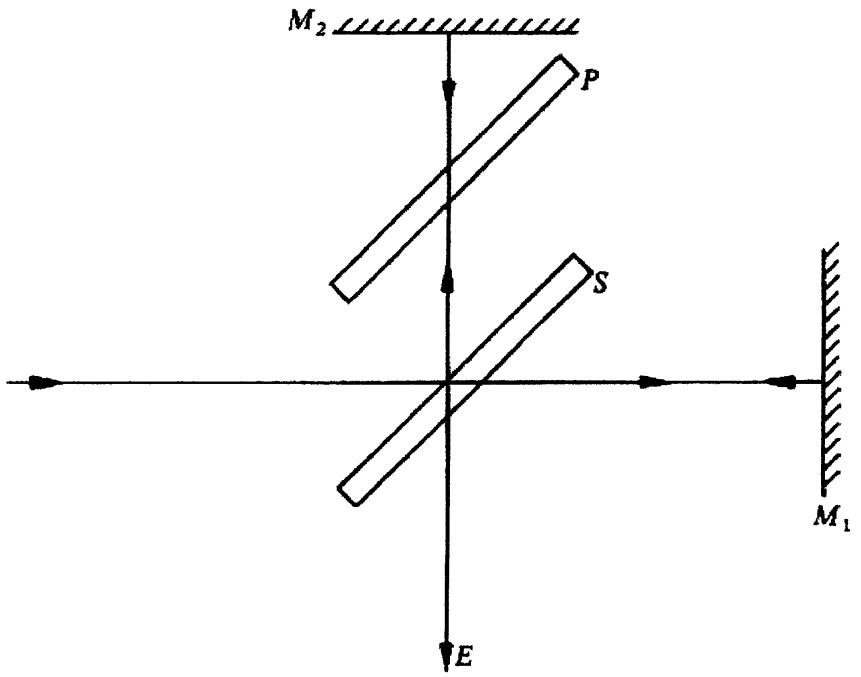


Figure A5.13. Arrangement for the Michelson interferometer.

the visibility first falls to zero, the value of the separation at which this occurs defines a spatial coherence dimension for the extended source. Also, if the source is isotropic, a coherence area is correspondingly defined. In other words, in this case, any given source point has no phase correlation with any point that lies outside the circular area of which it is the centre point.

A5.2.5 Practical examples

In order to appreciate fully the practical importance of the concept of optical coherence we shall conclude with four examples of the concept in action.

(i) The Michelson and Twyman–Green interferometers. The Michelson [1] interferometer is a very sensitive device for measuring optical path length differences. It is also important for its crucial role in the original

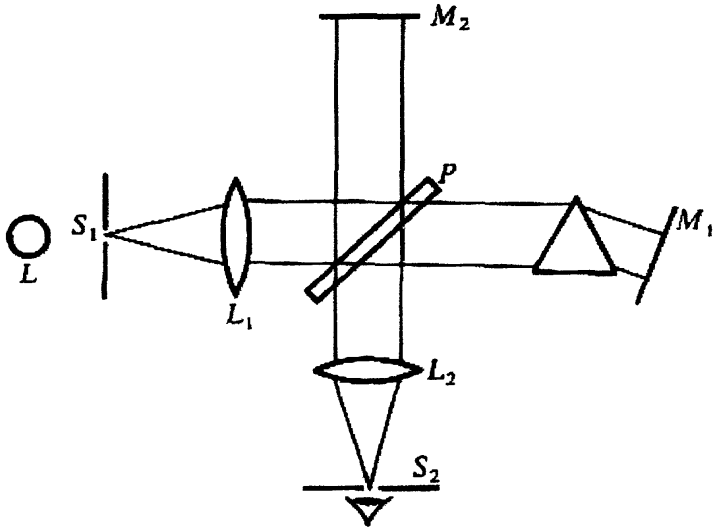


Figure A5.14. The Twyman–Green interferometer.

formulation of special relativity. The basic arrangement is shown in figure A5.13. Light from a collimated, extended source is split into two beams by a partial mirror S. The two light signals are then returned to S by the two plane mirrors M_1 and M_2. The partial mirror then returns parts of each light signal in the direction E, to a screen where they are allowed to interfere. The plate P is included to compensate that path (i.e. via mirror 2) for the extra distance travelled along the other path (via mirror 1) as a result of the triple passage of the partial mirror S. Hence, in this pristine state, the two signals in direction E are precisely matched in path of travel. (This is important to ensure that high-visibility interference is obtained over a broad range of wavelengths, in the face of dispersion in the mirror material). With an extended source the field, looking into the direction E, will appear as a set of interference rings corresponding to the rays from different positions on the source aperture. The interference pattern can be made to consist of straight lines by setting M_1 and M_2 at a small angle to each other, thus creating a linear phase difference (which then dominates) across the field. If an optical component is now introduced into one of the arms, its added optical path length will displace the fringes. By measuring this displacement, accurate measurements can be made of, for example, length and refractive index. This interferometer was used to measure the number of wavelengths of cadmium light in 1 m, thus leading to a new definition of that length unit.

The Michelson interferometer's contribution to the formulation of special relativity derives from the fact that, if the interferometer is moving physically in the direction of one of the light paths, the double passage of the light in that direction should be shorter, according to classical electromagnetic theory, than in the orthogonal direction, thus resulting in a fringe displacement. It thus should have been possible to detect the motion of the earth through the aether in its passage around the sun. The fact that Michelson and Morley, who performed this experiment in 1887, could detect no such motion, led Einstein (in 1905) to abandon altogether the notion of the aether, and to formulate entirely new ideas about the nature of space and time, in his Special Theory of Relativity.

An important variation of the Michelson interferometer idea is shown in figure A5.14. It is known as the Twyman–Green interferometer [4]. In this case the illuminating light originates from a point source and is rendered plane by a lens, so that a uniform phase front results. If an optical component is now inserted into one of the arms and care is taken to ensure that the light returning to P is again plane, after having twice traversed the inserted component, the interference pattern at S will be distorted by any imperfections in the component, such as variations in refractive index. This allows the imperfections to be accurately quantified. This interferometer is thus a powerful tool in the preparation of high quality optical elements, even to the

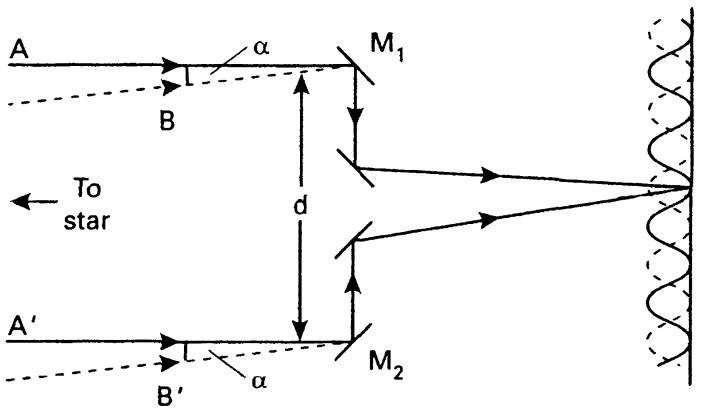


Figure A5.15. Michelson's stellar interferometer. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

extent of arranging for its output to control polishing functions, in order to equalize path lengths to less than a fraction of a wavelength.

(ii) Michelson's stellar interferometer. The concept of the spatial coherence of a light source is used in an instrument known as Michelson's stellar interferometer to measure the angular diameter of (nearer) stars (figure A5.15). If the star subtends an angle α at the two mirrors, spaced at distance d, then the two monochromatic (with the aid of an optical filter) rays A, A' (essentially parallel due to the very large distance of the star), from one point at the edge of the star, will be coherent and will produce an interference pattern with visibility = 1. Similarly, so also will the two rays B, B' from a diametrically opposite point at the other edge of the star. If the distance d between the mirrors is such that the ray B' is just one wavelength closer to M_2 than B is to M_1 then the second interference pattern, from BB', will coincide with the first from AA'. All the intermediate points across the star produce interference patterns between these two to give a total resultant visibility of zero. Hence, the value of d for which the fringe visibility first disappears provides the angular diameter of the star as λ/d (in fact, due to the circular rather than rectangular area, it is $1.22\ \lambda/d$). This method was first used by Michelson in 1920 [5] to determine the angular diameter of the star Betelgeuse as 0.047 seconds of arc. Distances between mirrors (d) of up to 10 m have been used. (Betelgeuse is a large star in the constellation of Orion and is quite close to Earth (~4 light years).) The vast majority of stars are too distant even for this very sensitive method to be of any use.

(iii) The Mach–Zehnder interferometer. Consider now the two-arm, optical-fibre, Mach–Zehnder arrangement (see section 2.3) of figure A5.16. A measurand (quantity to be measured) M in a Mach–Zehnder interferometer causes a phase change in arm 1 which is detected by means of a change in the position of the interference pattern resulting from the recombination of the light at point R. Interference can only occur if the recombining beams have components of the same polarization and if the difference in path length between the two arms is less than the source coherence length. This is not practicable with an LED, which has a coherence length ~0.02 mm, but even a modest semiconductor laser has a coherence length ~1 m (coherence time ~5 ns) and can easily be used in this application. A single mode He–Ne laser has a coherence length of several kilometres. It is clear that, in order to make an accurate measurement of M, it is necessary to choose a source with a fairly large coherence length. However, if the coherence length is too large, every reflection in the system interferes with every other, and an unwanted interference 'noise' results. This is an important problem for the opto-electronic designer: the coherence of the source must be optimized for the system in question.

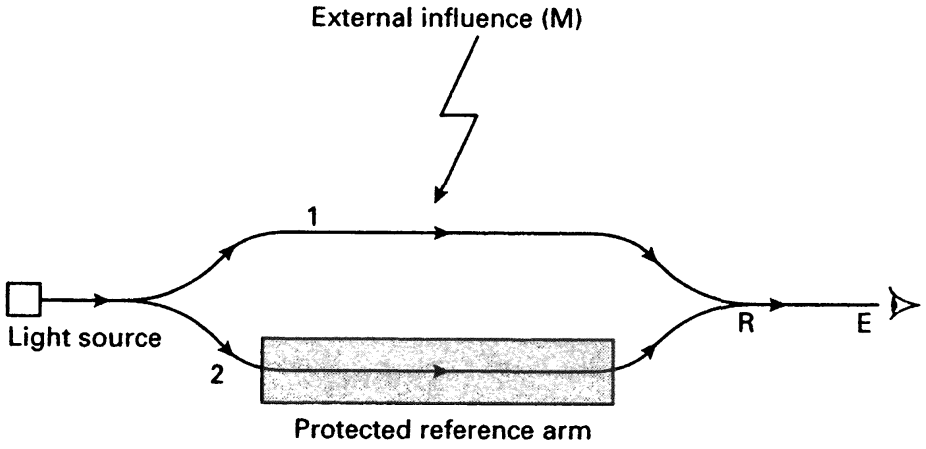


Figure A5.16. An optical-fibre Mach–Zehnder interferometer. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

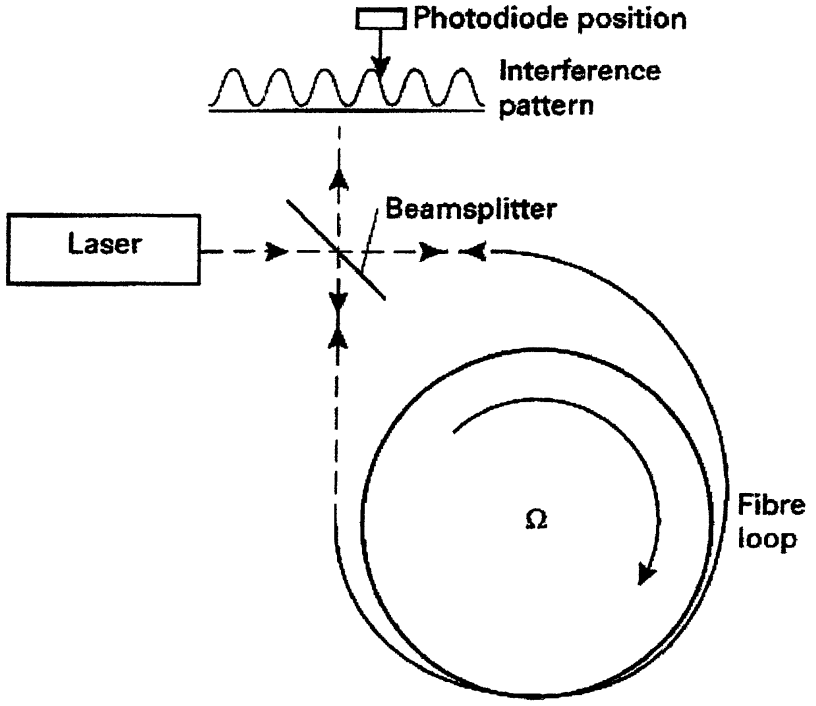


Figure A5.17. A Sagnac interferometer: the optical-fibre gyroscope.

This Mach–Zehnder arrangement is widely used in optical measurement technology. A special case of the Mach–Zehnder interferometer is described in the next section.

(iv) The optical-fibre gyroscope. A rather more sophisticated example of the effect of coherence occurs in the optical-fibre gyroscope (figure A5.17) [6]. The principle of the gyroscope is essentially the same as for the Mach–Zehnder interferometer we have just considered, the only important difference being that the two interfering arms now exist in the same fibre; the distinction between them lies in the fact that the light propagates in opposite directions in the two arms. The single source launches light in each of the two directions around a fibre loop and the light which emerges from the two ends of the loop interferes to produce a pattern on the screen. If the loop now rotates about an axis normal to its plane in, say, a clockwise direction, the light travelling in that direction will see its exit end receding from it, while the counter-propagating light

will see its end approaching. Consequently, the two components will traverse different optical paths and will emerge with a different relative phase relationship compared to that when the loop is stationary. The effect of the rotation is thus to cause a shift in the interference pattern, the magnitude of which is a measure of the rotation. Clearly, the greater the rotational velocity the greater will be the path length shift, and the coherence length of the source must be large enough to embrace the range of rotation which is to be measured. However, if the coherence length is too great another problem arises. Some of the light will be backscattered, by the fibre material, as it propagates, and light which is backscattered from a region around the half-way point will itself interfere for the two directions (figure A5.17) and construct its own interference pattern. This will generate a noise level which will degrade the device performance. Clearly, the greater the coherence length of the source the greater will be the region around the midpoint from which this can occur and thus the greater will be the noise level. Hence a compromise or 'trade-off' has to be struck, as it always does in device and system design. Gyroscopes are very important devices for navigation and automatic flight control. The conventional gyroscope based on the conservation of angular momentum in a spinning metal disc is highly developed, but contains parts which take time to be set in motion ('spin-up' time) and which wear. The device is also relatively expensive both to install and to maintain. The optical-fibre gyroscope overcomes all these problems (but, inevitably, has some of its own).

The phase difference between the counter-propagating signals is given by [7]:

$$\Phi = \frac{8\pi A\Omega}{c_0\lambda_0}$$

or

$$\Phi = \frac{2\pi LD\Omega}{c_0\lambda_0}$$

where A is the total effective area of the coil (i.e. the total area enclosed by N turns), L is the total length of the fibre , D is the diameter of the coil, c_0 the velocity of light in free space, λ_0 the wavelength of the light in free space and Ω is the rotation rate. Let us now insert some numbers into these expressions. Suppose that we use a wavelength of 1 μm with a coil of length 1 km and a diameter of 0.1 m. This gives:

$$\Phi = 2.1\Omega$$

For the earth's rotation of $15\,^\circ\,h^{-1}$ (7.3×10^{-5} radians s^{-1}) we must therefore be able to measure $\sim 1.5 \times 10^{-4}$ radian of phase shift. This can quite readily be done. In fact it is possible, using this device, to measure $\sim 10^{-6}$ radian of phase shift, corresponding to $\sim 5 \times 10^{-7}$ radians s^{-1} of rotation rate.

A5.3 Polarization

A5.3.1 Introduction

We know that the electric and magnetic fields, for a freely propagating light wave, lie transversely to the propagation direction and orthogonally to each other. Normally, when discussing polarization phenomena, we fix our attention on the electric field, since it is this that has the most direct effect when the wave interacts with matter. In saying that an optical wave is 'polarized' we are implying that the direction of the electric field is either constant or is changing in an ordered, prescribable manner. In general, the tip of the electric vector circumscribes an ellipse, performing a complete circuit in a time equal to the period of the wave, or in a distance of one wavelength. Clearly, the two parameters are equivalent in this respect.

As is well known, linearly polarized light can conveniently be produced by passing any light beam through a sheet of 'polaroid'. This is a material which absorbs light of one linear polarization (the 'acceptance' direction) to a much smaller extent ($\sim$1000 times) than the orthogonal polarization, thus effectively allowing just one linear polarization state to pass. The material's properties result from the fact that it consists of

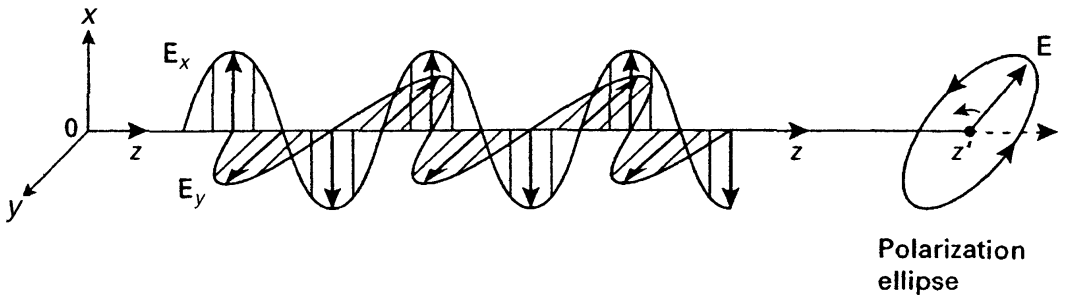


Figure A5.18. Electric field components for an elliptically-polarized wave. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

long-chain polymeric molecules aligned in one direction (the acceptance direction) by stretching a plastic, and then stabilizing it. Electrons can move more easily along the chains than transversely to them, and thus the optical wave transmits easily only when its electric field lies along this acceptance direction. The material is cheap and allows the use of large optical apertures. It thus provides a convenient means whereby, for example, a specific linear polarization state can be defined; this state then provides a ready polarization reference which can be used as a starting point for other manipulations.

In order to study these manipulations and other aspects of polarization optics, we shall begin by looking more closely at the polarization ellipse.

A5.3.2 The polarization ellipse

The most general form of polarized light wave propagating in the Oz direction is derived from the two linearly polarized components in the Ox and Oy directions (figure A5.18):

$$E_x = e_x \cos(\omega t - kz + \delta_x)$$
$$E_y = e_y \cos(\omega t - kz + \delta_y). \tag{A5.5a}$$

If we eliminate $(\omega t - kz)$ from these equations we obtain the expression:

$$\frac{E_x^2}{e_x^2} + \frac{E_y^2}{e_y^2} + \frac{2E_xE_y}{e_xe_y}\cos(\delta_y - \delta_x) = \sin^2(\delta_y - \delta_x) \tag{A5.5b}$$

which is the ellipse (in the variables E_x, E_y) circumscribed by the tip of the resultant electric vector at any one point in space over one period of the combined wave. This can only be true, however, if the phase difference $(\delta_y - \delta_x)$ is constant in time, or, at least, changes only slowly when compared with the speed of response of detector. In other words, we say that the two waves must have a large mutual 'coherence'. If this were not so then relative phases and, hence, resultant field vectors would vary randomly within the detector response time, giving no ordered pattern to the behaviour of the resultant field and thus presenting to the detector what would be, essentially, unpolarized light. Assuming that the mutual coherence is good, we may investigate further the properties of the polarization ellipse.

Note, first, that the ellipse always lies in the rectangle shown in figure A5.19, but that the axes of the ellipse are not parallel with the original x, y directions. The ellipse is specified as follows: with e_x, e_y, $\delta(= \delta_y - \delta_x)$ known, then we define $\tan\beta = e_y/e_x$. The orientation of the ellipse, α, is given by:

$$\tan 2\alpha = \tan 2\beta \cos\delta.$$

Semi-major and semi-minor axes a, b are given by:

$$e_x^2 + e_y^2 = a^2 + b^2 \sim I \qquad \text{(the wave intensity)}.$$

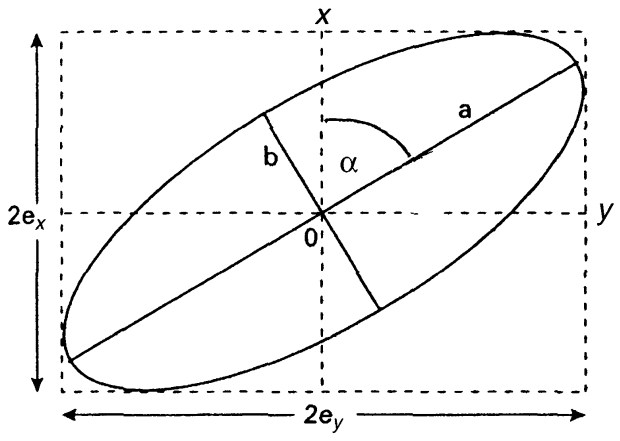


Figure A5.19. The polarization ellipse. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

Also, the ellipticity, e, is given by:

$$e = \tan\chi = \pm\frac{b}{a} \qquad \text{(the sign determines the sense of the rotation)}$$

where:

$$\sin 2\chi = -\sin 2\beta \sin\delta.$$

We should note also that the electric field components along the major and minor axes are always in quadrature (i.e. $\pi/2$ phase difference, the sign of the difference depending on the sense of the rotation).

Linear and circular states of polarization may be regarded as special cases where the polarization ellipse degenerates into a straight line or a circle, respectively. A linear state is obtained with the components in equations (A5.5(a)) when either:

$$\left.\begin{aligned} e_x &= 0 \\ e_y &\neq 0 \end{aligned}\right\} \qquad \text{linearly polarized in } Oy \text{ direction}$$

$$\left.\begin{aligned} e_x &\neq 0 \\ e_y &= 0 \end{aligned}\right\} \qquad \text{linearly polarized in } Ox \text{ direction}$$

or,

$$\delta_y - \delta_x = m\pi$$

where m is an integer. In this latter case the direction of polarization will be at an angle:

$$\begin{aligned} &+\tan^{-1}(e_y/e_x) \qquad m \text{ even} \\ &-\tan^{-1}(e_y/e_x) \qquad m \text{ odd} \end{aligned}$$

with respect to the Ox axis. A circular state is obtained when

$$e_x = e_y$$

and

$$(\delta_y - \delta_x) = (2m + 1)\pi/2$$

i.e. in this case the two waves have equal amplitudes and are in phase quadrature. The waves will be right-hand circularly polarized when m is even, and left-hand circularly polarized when m is odd.

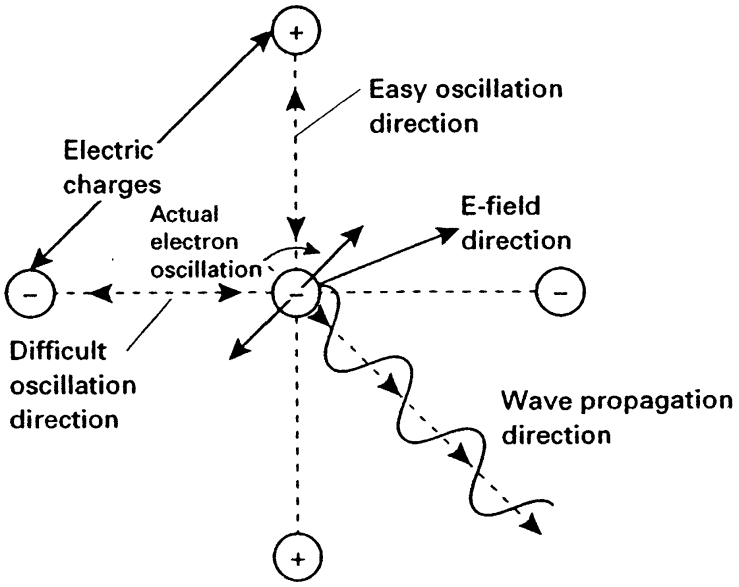


Figure A5.20. Electron response to electric field in an anisotropic medium. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

A5.3.3 Material interactions

Light can become polarized as a result of the intrinsic directional properties of matter: either the matter which is the original source of the light, or the matter through which the light passes. These intrinsic material directional properties are the result of directionality in the bonding which holds together the atoms of which the material is made. This directionality leads to variations in the response of the material according to the direction of an imposed force, be it electric, magnetic or mechanical. The best-known manifestation of directionality in solid materials is the crystal, with the large variety of crystallographic forms, some symmetrical, some asymmetrical. The characteristic shapes which we associate with certain crystals result from the fact that they tend to break preferentially along certain planes known as cleavage planes, which are those planes between which atomic forces are weakest.

It is not surprising, then, to find that directionality in a crystalline material is also evident in the light which it produces, or is impressed upon the light which passes through it.

In order to understand the ways in which we may produce polarized light, control it and use it, we must make a gentle incursion into the subject of crystal optics.

A5.3.4 Crystal optics

Light propagates through a material by stimulating the elementary atomic dipoles to oscillate and thus to radiate. In our previous discussions the forced oscillation was assumed to take place in the direction of the driving electric field but, in the case of a medium whose physical properties vary with direction, an anisotropic medium, this is not necessarily the case. If an electron in an atom or molecule can move more easily in one direction than another, then an electric field at some arbitrary angle to the preferred direction will move the electron in a direction which is not parallel with the field direction (figure A5.20). As a result, the direction in which the oscillating dipole's radiation is maximized (i.e. normal to its oscillation direction) is not the same as that of the driving wave.

The consequences of this simple piece of physics for the optics of anisotropic media are complex. One consequence is that the refractive index varies with the direction of the electric field of the wave, E. If we have a wave travelling in direction Oz, its velocity now will depend upon its polarization state: if the wave

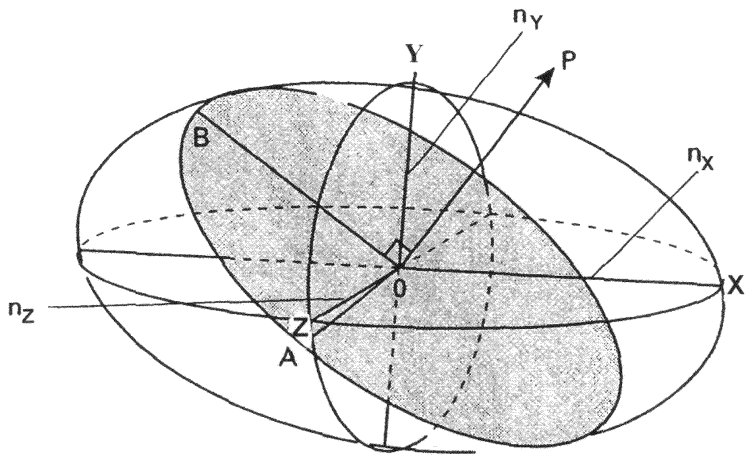


Figure A5.21. The index ellipsiod. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

is linearly polarized in the Ox direction it will travel with velocity c_0/n_x, while if it is linearly polarized in the Oy direction its velocity will be c_0/n_y. Hence, the medium is offering two refractive indices to the wave travelling in this direction: we have the phenomenon known as double refraction or 'birefringence'. A wave which is linearly polarized in a direction at 45° to Ox will split into two equal-amplitude components, linearly polarized in directions Ox and Oy, respectively, the two components travelling at different velocities. Hence, the phase difference between the two components will steadily increase and the composite polarization state of the wave will vary progressively from linear to circular back to linear again. The two special directions, Ox and Oy, for which there is no resolution of amplitude, are referred to as the birefringence axes.

This behaviour is, of course, a direct consequence of the basic physics which was discussed earlier: it is easier, in the anisotropic crystal, for the electric field to move the atomic electrons in one direction than in another. Hence, for the direction of easy movement, the light polarized in this direction can travel faster than when it is polarized in the direction for which the movement is more sluggish.

It follows from these discussions that an anisotropic medium may be characterized by means of three refractive indices, corresponding to polarization directions along Ox, Oy, Oz, and that these will have values n_x, n_y, n_z respectively. We can use this information to determine the refractive index (and thus the velocity) for a wave in any direction with any given linear polarization state. To do this we construct an 'index ellipsoid' or 'indicatrix', as it is sometimes called. This ellipsoid has the following important properties [8].

Suppose that we wish to investigate the propagation of light at an arbitrary angle to the crystal axes (polarization as yet unspecified). We draw a line, OP, corresponding to this direction within the index ellipsoid, passing through its centre O (figure A5.21). Now we construct the plane, also passing through O, for which OP is its normal. This plane will cut the ellipsoid in an ellipse. This ellipse has the property that the directions of its major and minor axes define the directions of the birefringence axes for this propagation direction, and the lengths of these axes OA and OB are equal to the refractive indices for these polarizations. Since these two linear polarization states are the only ones which propagate without change of polarization form for this crystal direction, they are sometimes referred to as the 'eigenstates' or 'polarization eigenmodes' for this direction, conforming to the matrix terminology of eigenvectors and eigenvalues.

The propagation direction we first considered, along Oz, corresponds, of course, to one of the axes of the ellipsoid, and the two refractive indices n_x, n_y, are the lengths of the other two axes in the central plane normal to Oz. The refractive indices n_x, n_y, n_z, are referred to as the principal refractive indices. Several

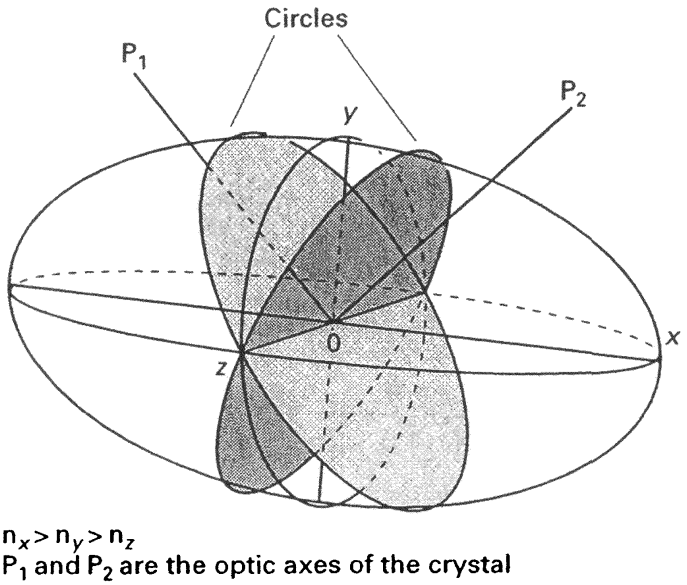


Figure A5.22. Ellipsiod for a biaxial crystal. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

other points are very well worth noting. Suppose, first, that

$$n_x > n_y > n_z$$

It follows that there will be a plane which contains Oz for which the two axes of interception with the ellipsoid are equal (figure A5.22). This plane will be at some angle to the $OyOz$ plane and will thus intersect the ellipsoid in a circle. This means that, for the light propagation direction corresponding to the normal to this plane, all polarization directions have the same velocity; there is no double refraction for this direction. This direction is an optic axis of the crystal and there will, in general, be two such axes, since there must also be such a plane at an equal angle to the $OyOz$ plane on the other side (see figure A5.22). Such a crystal, with two optic axes, is said to be biaxial. Suppose now that:

$$n_x = n_y = n_o \qquad \text{(say), the 'ordinary' index}$$

and

$$n_z = n_e \qquad \text{(say), the 'extraordinary' index.}$$

In this case one of the principal planes is a circle and it is the only circular section (containing the origin) which exists. Hence, in this case there is only one optic axis, along the Oz direction. Such crystals are said to be uniaxial. The crystal is said to be positive when $n_e > n_o$ and negative when $n_e < n_o$. For example, quartz is a positive uniaxial crystal, and calcite a negative uniaxial crystal. These features are, of course, determined by the crystal class to which these materials belong. It is clear that the index ellipsoid is a very useful device for determining the polarization behaviour of anisotropic media. Let us now consider some practical consequences of all of this.

A5.3.5 Retarding wave-plates

Consider a positive uniaxial crystal plate (e.g. quartz) cut in such a way (figure A5.23) as to set the optic axis parallel with one of the faces. Suppose a wave is incident normally on to this face. If the wave is linearly polarized with its electric field parallel with the optic axis, it will travel with refractive index n_e as we have described; if it has the orthogonal polarization, normal to the optic axis, it will travel with refractive index n_o.

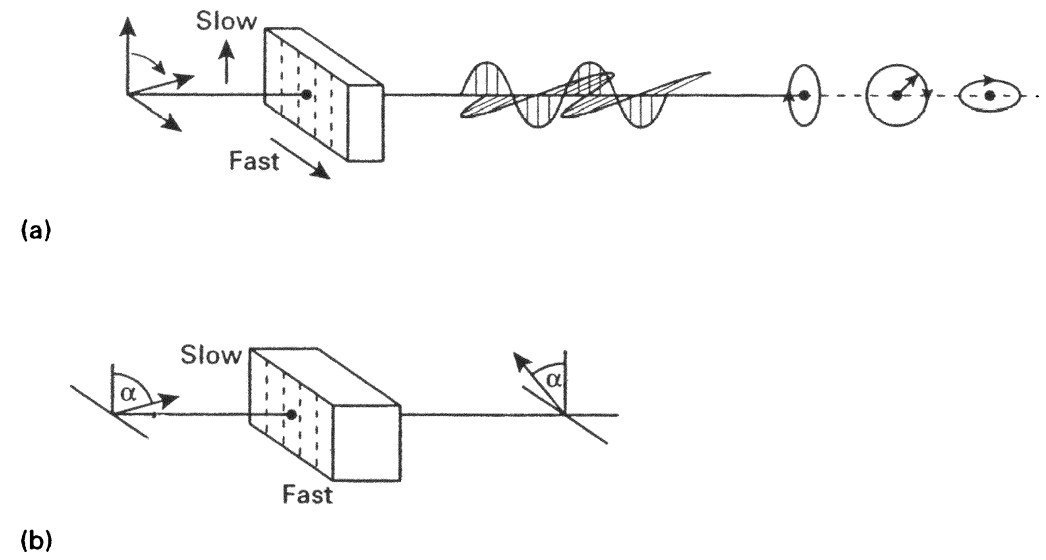


Figure A5.23. Polarization control with waveplates. (*a*) Quarter-wave plate; (*b*) half-wave plate. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

The two waves travel in the same direction through the crystal but with different velocities. For a positive uniaxial crystal $n_e > n_o$ and, thus, the light linearly polarized parallel with the optic axis will be a 'slow' wave, whilst the one at right angles to the axis will be 'fast'. For this reason the two crystal directions are often referred to as the 'slow' and 'fast' axes.

Suppose that the wave is linearly polarized at 45° to the optic axis. The phase difference between the components parallel with and orthogonal to the optic axis will now increase with distance, l, into the crystal according to:

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)l.$$

Hence, if, for a given wavelength λ,

$$l = \frac{\lambda}{4(n_e - n_o)}$$

then

$$\phi = \frac{\pi}{2}$$

and the light emerges from the plate circularly polarized. We have inserted a phase difference of $\pi/2$ between the components, equivalent to a distance shift of $\lambda/4$, and the crystal plate, when of this thickness, is called a 'quarter-wave' plate. It will (for an input polarization direction at 45° to the axes) convert linearly polarized light into circularly polarized light or *vice versa*. If the input linear polarization direction lies at some arbitrary angle α to the optic axis then the two components:

$$E\cos\alpha$$
$$E\sin\alpha$$

will emerge with a phase difference of $\pi/2$. We noted in section A5.3.3 that the electric field components along the two axes of a polarization ellipse were always in phase quadrature. It follows that these two components are now the major and minor axes of the elliptical polarization state which emerges from the plate. Thus, the ellipticity of the ellipse (i.e. the ratio of the major and minor axis) is just $\tan\alpha$ and, by varying the input polarization direction α, we have a means by which we can generate an ellipse of any

ellipticity. The orientation of the ellipse will be defined relative to the direction of the optic axis of the waveplate (figure A5.23(*a*)).

Suppose now that the crystal plate has twice the previous thickness and is used at the same wavelength. It becomes a 'half-wave' plate. A phase difference of π is inserted between the components (linear eigenstates). The result of this is that an input wave which is linearly polarized at angle α to the optic axis will emerge still linearly polarized but with its direction now at $-\alpha$ to the axis. The plate has rotated the polarization direction through an angle -2α. Indeed, any input polarization ellipse will emerge with the same ellipticity but with its orientation rotated through -2α (figure A5.23(*b*)).

It follows that, with the aid of these two simple plates, we can generate elliptical polarization of any prescribed ellipticity and orientation from linearly polarized light, which can itself be generated from any light source plus a simple polaroid sheet. Equally valuable is the reverse process: that of the analysis of an arbitrary elliptical polarization state or its conversion to a linear state. Suppose we have light of unknown elliptical polarization. By inserting an analyzing polarizer and rotating it around the axis parallel to the propagation direction, we shall find a position of maximum transmission and an orthogonal position of minimum transmission. These are the major and minor axes of the ellipse (respectively) and the ratio of the two intensities at these positions will give the square of the ellipticity of the ellipse, i.e.

$$e = \frac{b}{a} = \frac{E_b}{E_a} = \left(\frac{I_b}{I_a}\right)^{1/2}.$$

Clearly, the orientation of the ellipse is also known since this is, by definition, just the direction of the major axis and is given by the position at which the maximum occurs. In order to convert the elliptical state into a linear one, all we need is a quarter-wave plate (appropriate to the wavelength of the light used, of course). Since the components of electric field along the major and minor axis of the ellipse are always in phase quadrature (see section A5.3.2), the insertion of a quarter-wave plate with its axes aligned with the axes of the polarization ellipse brings the components into phase or into antiphase, and the light thus becomes linearly polarized. The quarter-wave plate is used in conjunction with a following polaroid sheet (or prism polarizer) and the two are rotated (independently) about the propagation axis until the light is extinguished. The quarter-wave plate must then have the required orientation in line with the ellipse axes, since only when the light has become linearly polarized can the polarizer extinguish it completely. (If there are no positions for which the light is extinguished, then it is not fully polarized.)

Such are the quite powerful manipulations and analyses which can be performed with very simple devices. However, manual human intervention for rotation of plates is not always convenient or even possible. In many cases polarization analysis and control must be done very quickly (perhaps in nanoseconds) and automatically, using electronic processing. For these cases more advanced polarization devices must be used and, in order to understand and use these, a more advanced theoretical framework is necessary. We shall introduce this in section A5.3.8. Before doing so, however, we shall first look at another very important component for polarization control and then at another crucial polarization parameter.

A5.3.6 Polarizing prisms

The same ideas as those just described are also useful in devices that produce linearly polarized light with a higher degree of polarization than a polaroid sheet is capable of and without its intrinsic loss (even for the 'acceptance' direction there is a significant loss). We shall look at just two of these devices, in order to illustrate the application of the ideas, but there are several others (these are described in most standard optics texts).

The first device is the Nicol prism, illustrated in figure A5.24. Two wedges of calcite crystal are cut as shown, with their optic axes in the same direction (in the plane of the page) and cemented together with 'Canada balsam', a material whose refractive index at visible wavelengths lies midway n_e between and n_o.

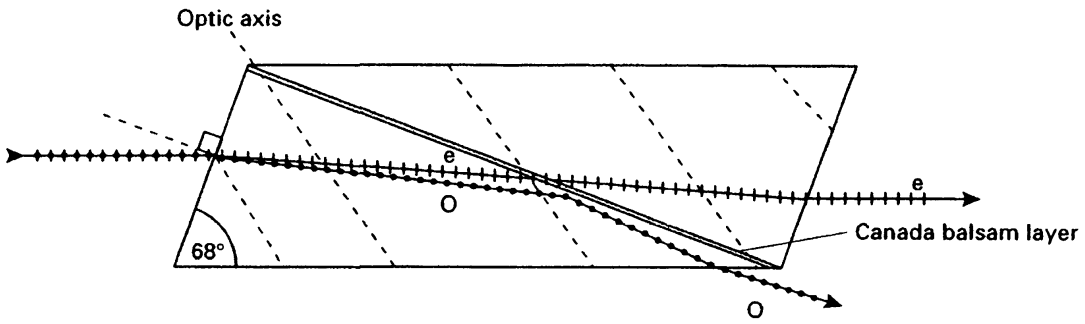


Figure A5.24. Action of the Nicol prism. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

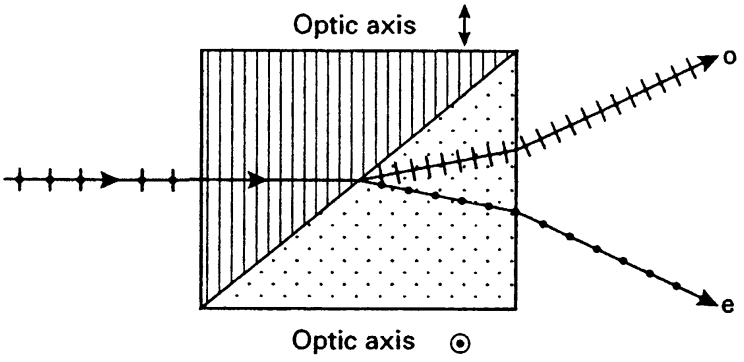


Figure A5.25. Action of the Wollaston prism. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

When unpolarized light enters parallel to the axis of the prism (as shown) and at an angle to the front face, it splits, as always, into the e and o components, each with its own refractive index, and thus each with its own refractive angle according to Snell's law. (Calcite is a negative uniaxial crystal so $n_o > n_e$.) When the light reaches the Canada balsam interface between the two wedges, it finds that the geometry and refractive indices have been arranged such that the ordinary (o) ray, with the larger deflection angle, strikes this interface at an angle greater than the total internal reflection (TIR) angle and is thus not passed into the second wedge, whereas the extra-ordinary (e) ray is so passed. Hence, only the e ray emerges from the prism and this is linearly polarized. Thus, we have an effective prism polarizer, albeit one of limited angular acceptance ($\sim 14°$) since the TIR condition is quite critical in respect of angle of incidence.

The second prism we shall discuss is widely used in practical polarization optics: it is called the Wollaston prism, and is shown in figure A5.25. Again we have two wedges of positive (say) uniaxial crystal. They are equal in size, placed together to form a rectangular block, (sometimes a cube) and have their optic axes orthogonal, as shown. Consider a wave entering normally from the left. The e and o waves travel with differing velocities and strike the boundary between the wedges at the same angle. On striking the boundary one of the waves sees a positive change in refractive index ($n_e - n_o$), the other a negative change ($n_e - n_o$), so that they are deflected, respectively, up and down (figure A5.25) through equal angles. The e and o rays thus diverge as they emerge from the prism allowing either to be isolated, or the two to be observed (or detected) simultaneously but separately. Also it is clear that, by rotating this prism around the propagation axis, we may reverse the positions of the two components.

It is extremely useful to be able to separate the two orthogonally polarized components in this controllable way. For example, consider the problem of the measurement of the rotation of the direction of a linearly-

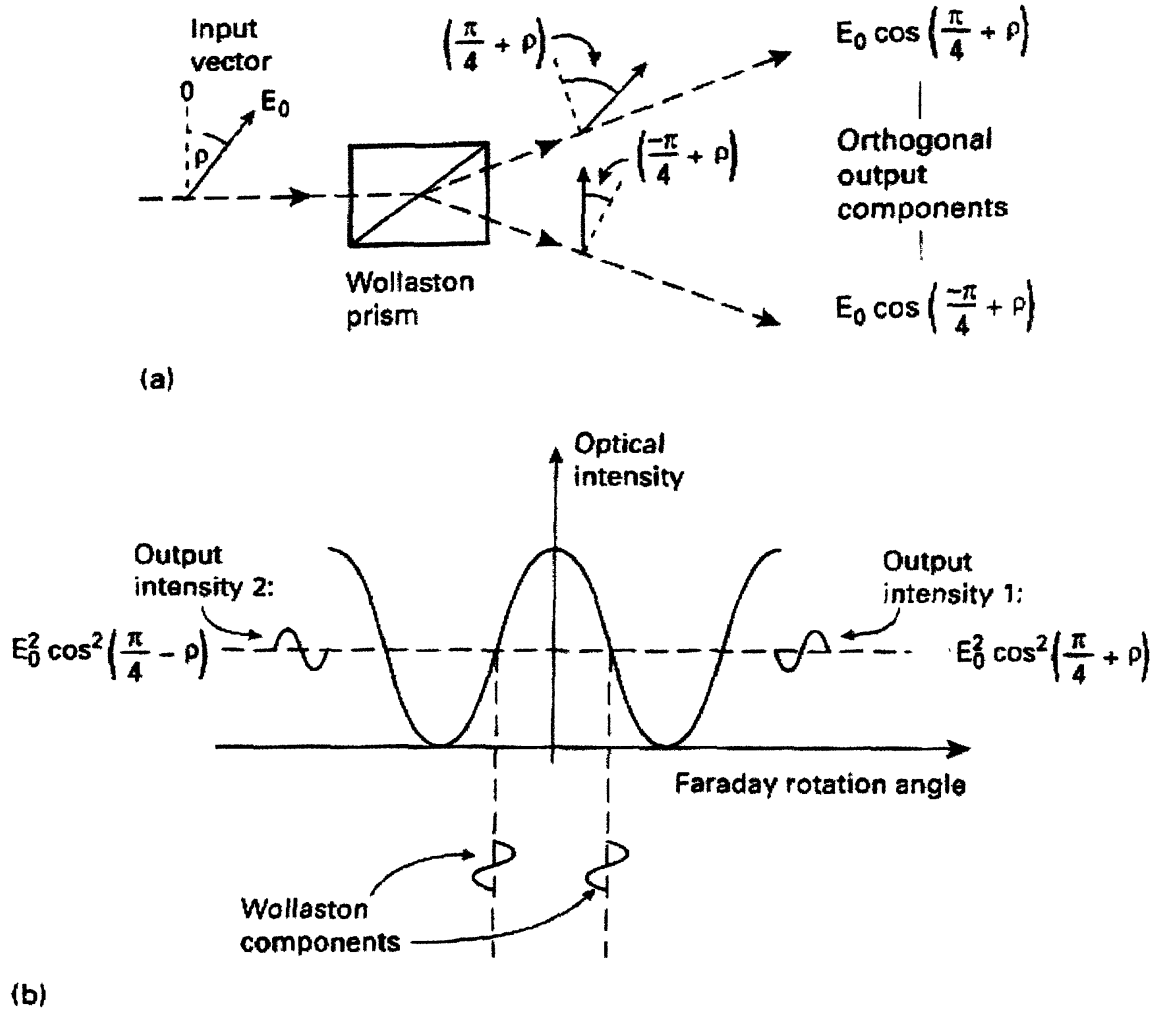


Figure A5.26. Measurement of polarization rotation. (*a*) Wollaston E-field components; (*b*) output intensities for Wollaston components.

polarized optical wave. Such a problem arises, for example, when measuring an electric current using the magneto-optic effect in single-mode optical fibre (see section A5.3.9(*b*)).

How can we actually measure such a rotation, ρ? Suppose that the emerging linearly polarized light falls on to a linear polarizer which is set with its polarization direction parallel with that of the light's *input* polarization direction (figure A5.26). In the absence of a magnetic field ($\rho = 0$) all the light will be passed by the polarizer (ignoring its intrinsic attenuation).

Let us assume that the electric field amplitude of the propagating light is E_0, so that an intensity proportional to E_0^2 is passed, in the absence of current. When current flows, the polarization is rotated through an angle ρ and only a field component $E_0 \cos\rho$ will now be passed by the polarizer, giving a measurable intensity proportional to $E_0^2 \cos^2\rho$. This intensity, in principle, allows ρ to be deduced.

However, there is a more convenient way to measure ρ. Suppose that, instead of a simple polarizer, we use a Wollaston prism, with its polarization axes set at $\pm 45°$ to the input polarization direction. We now have two intensity outputs from the Wollaston prism (see figure A5.26):

$$I_1 = KE_0^2 \cos^2(\tfrac{1}{4}\pi - \rho)$$
$$I_2 = KE_0^2 \cos^2(\tfrac{1}{4}\pi + \rho)$$

where K is the usual universal constant.

If we detect these two intensities separately (by measuring the optical powers falling on two separate photodiodes: remember that power = intensity × area), then we can readily arrange for the electronics to construct the function

$$S = \frac{I_1 - I_2}{I_1 + I_2} = \frac{\cos^2(\frac{1}{4}\pi - \rho) - \cos^2(\frac{1}{4}\pi + \rho)}{\cos^2(\frac{1}{4}\pi - \rho) + \cos^2(\frac{1}{4}\pi + \rho)}$$

which gives, on manipulation of the functions

$$S = \sin 2\rho$$

and, if 2ρ is small ($\ll /2$):

$$S \approx 2\rho$$

Hence, with the aid of the polarizing beam-splitter, we have succeeded in measuring the polarization rotation independently of the light intensity and, thus, free from any variations in the source output, or variations in the attenuation along the optical path.

A5.3.7 Circular birefringence

So far we have considered only linear birefringence, where two orthogonal linear polarization eigenstates propagate, each remaining linear, but with different velocities. Some crystals also exhibit circular birefringence. Quartz (again) is one such crystal and its circular birefringence derives from the fact that the crystal structure spirals around the optic axis in a right-handed (dextro-rotatory) or left-handed (laevo-rotatory) sense depending on the crystal specimen: both forms exist in nature.

It is not surprising to find, in view of this knowledge and our understanding of easy motions of electrons, that light which is right-hand circularly polarized (clockwise rotation of the tip of the electric vector as viewed by a receiver of the light) will travel faster down the axis of a matching right-hand spiralled crystal structure then left-hand circularly polarized light. We now have circular birefringence: the two circular polarization components propagate without change of form (i.e. they remain circularly polarized) but at different velocities. They are the circular polarization eigenstates for this case.

The term 'optical activity' has been traditionally applied to this phenomenon, and it is usually described in terms of the rotation of the polarization direction of a linearly polarized wave as it passes down the optic axis of an 'optically active' crystal. This fact is exactly equivalent to the interpretation in terms of circular birefringence, since a linear polarization state can be resolved into two oppositely rotating circular components (figure A5.27). If these travel at different velocities, a phase difference is inserted between them. As a result of this, when recombined, they again form a resultant which is linearly polarized but rotated with respect to the original direction (figure A5.27). Hence 'optical activity' is equivalent to circular birefringence.

In general, both linear and circular birefringence might be present simultaneously in a material (such as quartz). In this case the polarization eigenstates which propagate without change of form (and at different velocities) will be elliptical states, the ellipticity and orientation depending upon the ratio of the magnitudes of the linear and circular birefringences, and on the direction of the linear birefringence eigen-axes within the crystal.

It should, again, be emphasized that only the polarization eigenstates propagate without change of form. All other polarization states will be changed into different polarization states by the action of the polarization element (e.g. a crystal component). These changes of polarization state are very useful in opto-electronics. They allow us to control, analyse, modulate and demodulate polarization information impressed upon a light beam, and to measure important directional properties relating to the medium through which the light has passed. We must now explore a more rigorous formalism to handle these more general polarization processes.

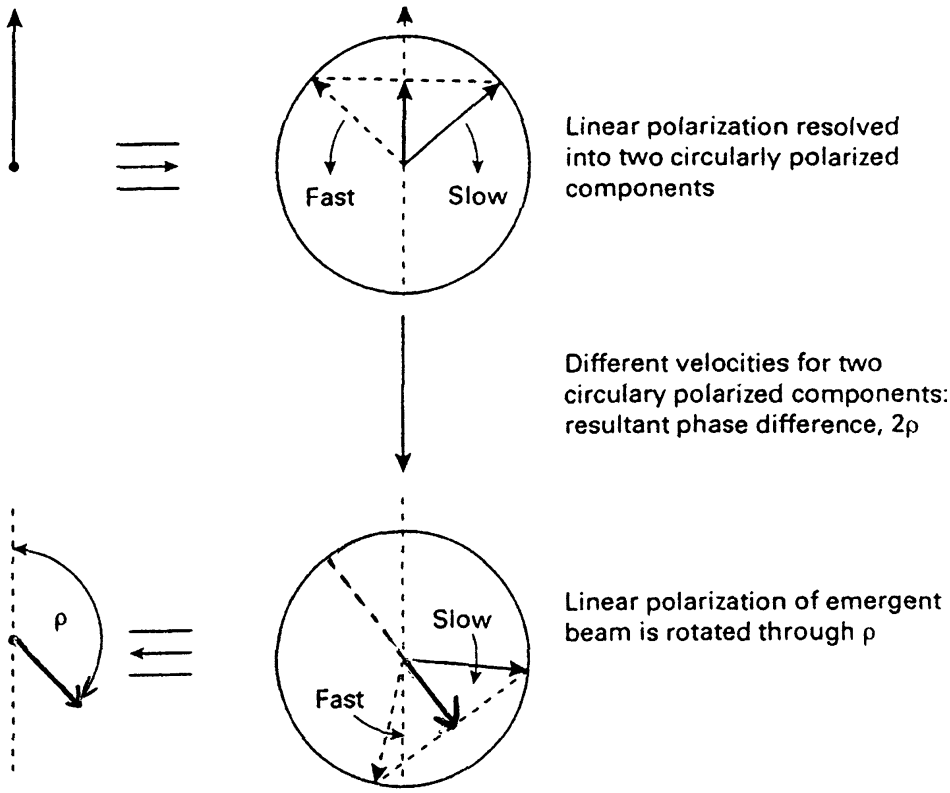


Figure A5.27. Resolution of linear polarization into circularly polarized components in circular birefringence (2ρ). (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

A5.3.8 Polarization analysis

As has been stated, with both linear and circular birefringence present, the polarization eigenstates (i.e. the states which propagate without change of form) for a given optical element are elliptical states, and the element is said to exhibit elliptical birefringence, since these eigenstates propagate with different velocities.

In general, if we have, as an input to a polarization-optical element, light of one elliptical polarization state, it will be converted, on emergence, into a different elliptical polarization state (the only exceptions being, of course, when the input state is itself an eigenstate). We know that any elliptical polarization state can always be expressed in terms of two orthogonal electric field components defined with respect to chosen axes Ox, Oy, i.e.

$$E_x = e_x \cos(\omega t - kz + \delta_z)$$
$$E_y = e_y \cos(\omega t - kz + \delta_y)$$

or, in complex exponential notation:

$$E_x = |E_x| \exp(\mathrm{i}\phi_x) \qquad (\phi_x = \omega t - kz + \delta_x)$$
$$E_y = |E_y| \exp(\mathrm{i}\phi_y) \qquad (\phi_y = \omega t - kz + \delta_y).$$

When this ellipse is converted into another by the action of a lossless polarization element, the new ellipse will be formed from components which are linear combinations of the old, since it results from directional resolutions and rotations of the original fields. Thus, these new components can be written:

$$E'_x = m_1 E_x + m_4 E_y$$
$$E'_y = m_3 E_x + m_2 E_y$$

or, in matrix notation:

$$\boldsymbol{E}' = \boldsymbol{M} \cdot \boldsymbol{E}$$

where

$$\mathbf{M} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix} \tag{A5.6}$$

and the m_n are, in general, complex numbers. $\mathbf{M}$ is known as a 'Jones' matrix after the mathematician who developed an extremely useful 'Jones calculus' for manipulations in polarization optics [9]. Now, in order to make measurements of the input and output states in practice we need a quick and convenient experimental method. In section A5.3.6 there was described a method for doing this which involved the manual rotation of a quarter wave plate and/or a polarizer, but the method we seek now must lend itself to automatic operation.

A convenient method for this practical determination is again to use the linear polarizer and the quarter-wave plate, but to measure the light intensities for a series of fixed orientations of these elements.

Suppose that $I\ (\theta, \varepsilon)$ denotes the intensity of the incident light passed by the linear polarizer set at angle θ to Ox, after the Oy component has been retarded by angle ε as a result of the insertion of the quarter-wave plate with its axes parallel with Ox, Oy. We measure what are called the four Stokes parameters, as follows:

$$\begin{aligned}
S_0 &= I(0°, 0) + I(90°, 0) = e_x^2 + e_y^2 \\
S_1 &= I(0°, 0) - I(90°, 0) = e_x^2 - e_y^2 \\
S_2 &= I(45°, 0) - I(135°, 0) = 2e_x e_y \cos\delta \\
S_3 &= I\left(45°, \frac{\pi}{2}\right) - I\left(135°, \frac{\pi}{2}\right) = 2e_x e_y \sin\delta \\
&(\delta = \delta_y - \delta_x).
\end{aligned}$$

These parameters can be measured directly, with the aid of a photodetector. If the light is 100% polarized, only three of these parameters are independent, since:

$$S_0^2 = S_1^2 + S_2^2 + S_3^2$$

S_0 being the total light intensity.

If the light is only partially polarized, the fraction:

$$\eta = \frac{(S_1^2 + S_2^2 + S_2^3)}{S_0^2}$$

defines the degree of polarization. In what follows we shall assume that the light is fully polarized ($\eta = 1$). It is easy to show that measurement of the S_n provides the ellipticity, e, and the orientation, α, of the polarization ellipse according to the relations:

$$\begin{aligned}
e &= \tan\chi \\
\sin 2\chi &= \frac{S_3}{S_0} \\
\tan 2\alpha &= \frac{S_2}{S_1}.
\end{aligned}$$

Now, these relations suggest a geometrical construction which provides a powerful and elegant means for description and analysis of polarization-optical phenomena. The Stokes parameters S_1, S_2, S_3 may be regarded as the Cartesian coordinates of a point referred to axes Ox_1, Ox_2, Ox_3. Thus, every elliptical polarization state corresponds to a unique point in three dimensional space. For a constant S_0 (lossless

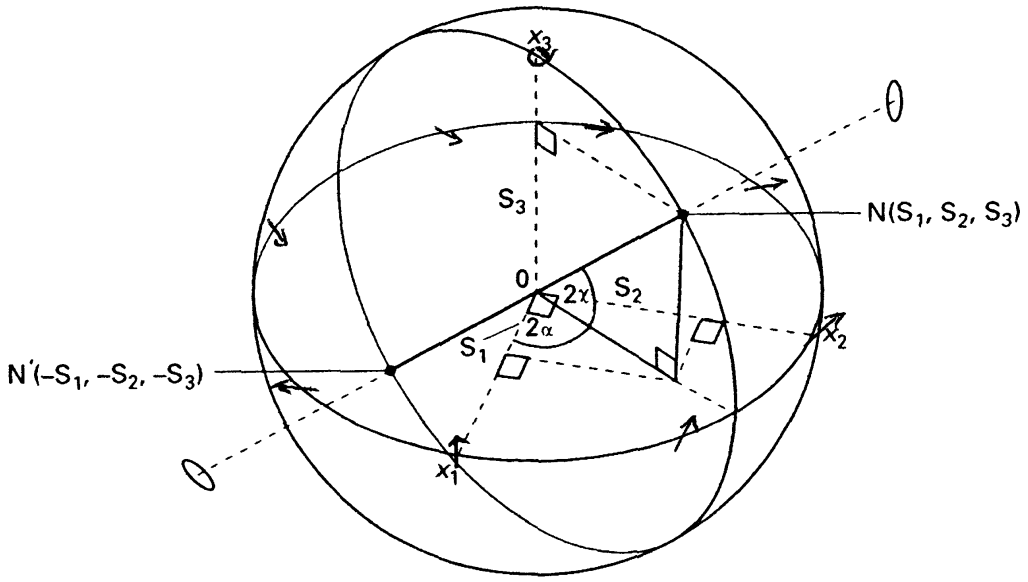


Figure A5.28. The Poincaré sphere: the eigenmode diameter (NN'). (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

medium) it follows that all such points lie on a sphere of radius S_0 — the Poincaré sphere (figure A5.28). The properties of the sphere are quite well known (see, for example [10]). We can see that the equator comprises the continuum of linearly polarized states whilst the two poles correspond to the two oppositely-handed states of circular polarization.

It is clear that any change resulting from the passage of light through a lossless element, from one polarization state to another, corresponds to a rotation of the sphere about a diameter. Now, any such rotation of the sphere may be expressed as a unitary 2 × 2 matrix, **M**. Thus, the conversion from one polarization state, $\boldsymbol{E}$, to another $\boldsymbol{E}'$, may also be expressed in the form:

$$\boldsymbol{E}' = \mathbf{M} \cdot \boldsymbol{E}$$

or

$$\begin{pmatrix} E'_x \\ E'_y \end{pmatrix} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}$$

$$E'_x = m_1 E_x + m_4 E_y$$
$$E'_y = m_3 E_x + m_2 E_y$$

where

$$\mathbf{M} = \begin{pmatrix} m_1 & m_4 \\ m_3 & m_2 \end{pmatrix}$$

and this **M** may be immediately identified with our previous **M**, in equation (A5.6). **M** is a Jones matrix [9] which completely characterizes the polarization action of the element and is also equivalent to a rotation of the Poincaré sphere. The two eigenvectors of the matrix correspond to the eigenmodes (or eigenstates) of the element (i.e. those polarization states which can propagate through the element without change of form). These two polarization eigenstates lie at opposite ends of a diameter of the Poincaré sphere and the polarization effect of the element is to rotate the sphere about this diameter (figure A5.29) through an angle which is equal to the phase which the polarization element inserts between its eigenstates.

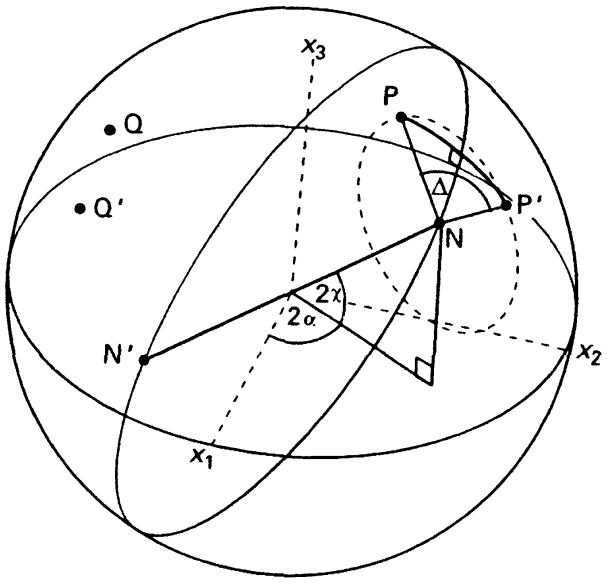


Figure A5.29. Rotation of the Poincaré sphere about the eigenmode diameter (NN'). (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

The polarization action of the element may thus be regarded as that of resolving the input polarization state into the two eigenstates with appropriate amplitudes, and then inserting a phase difference between them before recombining to obtain the emergent state. Thus, a pure rotator (e.g. optically-active crystal) is equivalent to a rotation about the polar axis, with the two oppositely-handed circular polarizations as eigenstates. The phase velocity difference between these two eigenstates is a measure of the circular birefringence. Analogously, a pure linear retarder (such as a wave plate) inserts a phase difference between orthogonal linear polarizations which measures the linear birefringence. The linear retarder's eigenstates lie at opposite ends of an equatorial diameter.

It is useful for many purposes to resolve the polarization action of any given element into its linear and circular birefringence components. The Poincaré sphere makes it clear that this may always be done, since any rotation of the sphere can always be resolved into two subrotations, one about the polar diameter and the other about an equatorial diameter.

From this brief discussion we can begin to understand the importance of the Poincaré sphere. It is a construction which converts all polarization actions into visualizable relationships in three dimensional space.

To illustrate this point graphically let us consider a particular problem. Suppose that we ask what is the smallest number of measurements necessary to define completely the polarization properties of a given lossless polarization element, about which we know nothing in advance. Clearly, we must provide known polarization input states and measure their corresponding output states, but how many input/output pairs are necessary: one, two, more?

The Poincaré sphere answers this question easily. The element in question will possess two polarization eigenmodes and these will be at opposite ends of a diameter. We need to identify this diameter. We know that the action of the element is equivalent to a rotation of the sphere about this diameter and through an angle equal to the phase difference which the element inserts between its eigenmodes. Hence, if we know one input/output pair of polarization states, we know that the rotation from the input to the output state must have taken place about a diameter lying in the plane which perpendicularly bisects the line joining the two states (see figure A5.29). Two other input/output states will similarly define another such plane and, thus, the required diameter is clearly seen as the common line of intersection of these planes.

Further, the phase difference inserted between the eigenstates (i.e. the sphere's rotation angle) is easily calculated from either pair of states, once the diameter is known.

Hence, the answer is that two pairs of input/output states will define completely the polarization properties of the element. Simple geometry has provided the answer. A good general approach is to use the Poincaré sphere to determine or visualize the nature of the solution to a problem, and then to revert to the Jones matrices to perform the precise calculations. Alternatively, some simple results in spherical trigonometry will usually suffice.

Having dealt with the theoretical tools by which polarization characteristics are manipulated and analysed, we shall now turn to some practical applications of the ideas. We shall look at ways in which directional properties within materials can impose polarizations on light waves passing through them.

A5.3.9 Applications of polarization optics

A polarized optical wave is essentially one which is asymmetrical with regard to its transverse vibrations. In other words, there is a preference for oscillation in some directions when compared with others.

When an optical wave passes through a material medium it does so by stimulating the elementary atomic dipoles to radiate. These secondary radiations combine vectorially with the primary wave to give rise to the resultant propagation through the medium, thus defining the latter's (complex) refractive index.

In such circumstances any directionality inherent in the medium itself, resulting either from its crystal structure or from externally applied asymmetrical forces, will be impressed also on the propagating wave. Consequently, carefully chosen materials can be used to control polarization state; and the polarization analysis of the resultant wave can be used sensitively to probe material structures.

Clearly then, polarization effects may arise naturally, or may be induced deliberately. Of those which occur naturally the most common are the ones which are a consequence of an anisotropic material, an asymmetrical material strain or asymmetrical waveguide geometries.

If an optical medium is compressed in a particular direction, there results the same kind of directional restriction on the atomic or molecular electrons as in the case of crystals and, hence, the optical polarization directions parallel and orthogonal to these imposed forces (for isotropic materials) will encounter different refractive indices.

Somewhat similarly, if an optical wave is being guided in a channel, or other type of guide, with a refractive index greater than its surroundings, we have to be aware of the effect of any asymmetry in the geometry of the guide's cross section. Clearly, if the cross section is a perfect circle, as in the case of an ideal optical fibre, all linear polarization directions must propagate with the same velocity. If, however, the cross section is elliptical, then it is not difficult to appreciate that a linear polarization direction parallel with the minor axis propagates at a different velocity from that parallel with the major axis.

The optical fibre is, in fact, a good medium for illustrating these passive polarization effects, since all real fibres possess the same directional asymmetry due to one or more of the following: non-circularity of core cross section; linear strain in the core; twist strain in the core. Bending will introduce linear strain and twisting will introduce circular strain (figure A5.30). Linear strain leads to linear birefringence, circular (twist) strain to circular birefringence.

The linear birefringence in 'standard' telecommunications optical fibre can be quite troublesome for high performance links since it introduces velocity differences between the two orthogonal linear polarization states, which lead to relative time lags of the order of 1–10 ps km^{-1}. Clearly, this distorts the modulating signal: a pulse in a digital system, for example, will be broadened, and thus degraded, by this amount (see section A6.5.2). This so-called 'polarization mode dispersion (PMD)' can be reduced by spinning the preform from which the fibre is being drawn, whilst it is being drawn, so as to average out the cross-sectional anisotropies. This 'spun preform' technique [11] reduces this form of dispersion to ~0.01 ps km^{-1}, i.e. by two orders of magnitude.

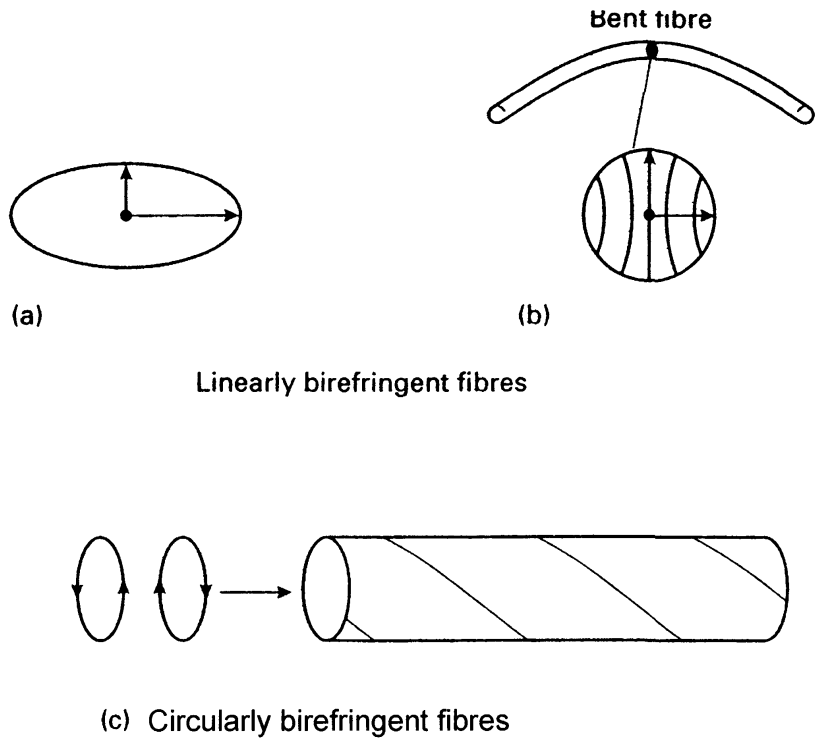


Figure A5.30. Birefringence in optical fibres. (*a*) Geometrical 'form'; (*b*) bending 'strain'; and (*c*) twist-strain circularly birefringent fibre. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

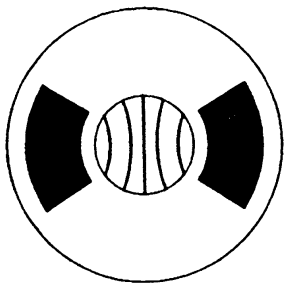


Figure A5.31. Asymmetrically doped, linearly birefringent optical fibre ('bow-tie'). (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

It is sometimes valuable deliberately to introduce linear or circular birefringence into a fibre. In order to introduce linear birefringence the fibre core may be made elliptical (with consequences previously discussed) or stress may be introduced by asymmetric doping of the cladding material which surrounds the core (figure A5.31) [12]. The stress results from asymmetric contraction as the fibre cools from the melt.

Circular birefringence may be introduced by twisting and then clamping the fibre or by spinning an asymmetric preform (from which the fibre is being pulled). One important application of fibre with a high value of linear birefringence ('hi-bi' fibre) is that linearly polarized light launched into one of the two linear eigenmodes will tend to remain in that state, thus providing a convenient means for conveying linearly polarized light between two points. The reason for this 'polarization holding' property is that light, when coupled (i.e. transferred) to the other eigenmode, will be coupled to a mode with a different velocity and will not, in general, be in phase with other previous light couplings into the mode; thus the various couplings will interfere destructively overall and only a small amplitude will result. There is said to be a 'phase mismatch'. (This is yet another example of wave interference!). Clearly, however, if a deliberate attempt is made to couple

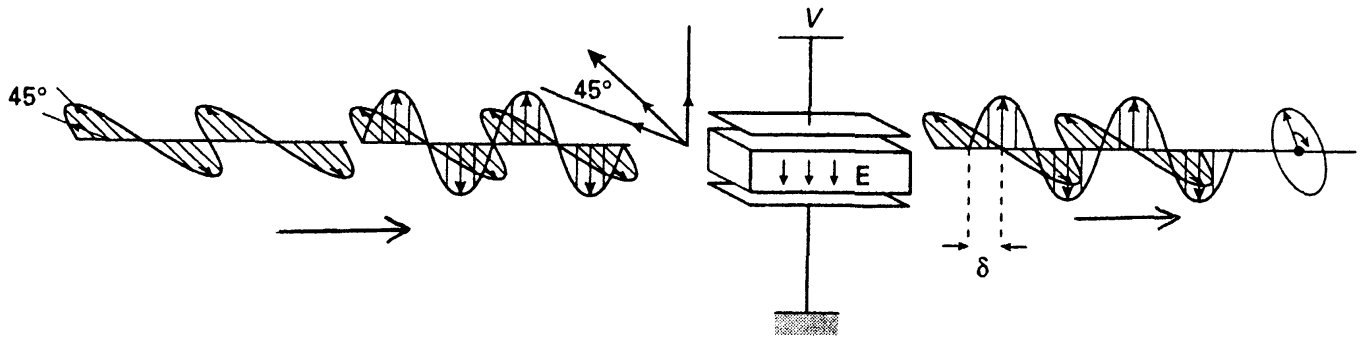


Figure A5.32. The electro-optic effect. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

light only at those points where the two modes are in phase, then constructive interference can occur and the coupling will be strong. This is known as 'resonant' coupling and has a number of important applications.

An extremely convenient way of inducing polarization anisotropies into materials is by subjecting them to electric and/or magnetic fields. As we know very well, these fields can exert forces on electrons, so it is not surprising to learn that, through their effects on atomic electrons, the fields can influence the polarization properties of media, just as the chemical bond restrictions on these electrons in crystals were able to do. The use of electric and magnetic fields thus allows us to build convenient polarization controllers and modulators. Some examples of the effects which can be used will help to establish these ideas.

(a) The electro-optic effect. When an electric field is applied to an optical medium the electrons suffer restricted motion in the direction of the field, when compared with that orthogonal to it. Thus, the material becomes linearly birefringent in response to the field. This is known as the electro-optic effect.

Consider the arrangement of figure A5.32. Here we have incident light which is linearly polarized at 45° to an electric field and the field acts on a medium transversely to the propagation direction of the light. The field-induced linear birefringence will cause a phase displacement between components of the incident light which lie, respectively, parallel and orthogonal to the field; hence the light will emerge elliptically polarized.

A (perfect) polarizer placed with its acceptance direction parallel with the input polarization direction will of course, pass all the light in the absence of a field. When the field is applied, the fraction of light power which is passed will depend upon the form of the ellipse, which in turn depends upon the phase delay introduced by the field. Consequently, the field can be used to modulate the intensity of the light, and the electro-optic effect is, indeed, very useful for the modulation of light.

The phase delay introduced may be proportional either to the field (Pockels effect) or to the square of the field (Kerr effect (see section A6.5.3)). All materials manifest a transverse Kerr effect. Only crystalline materials can manifest any kind of Pockels effect, or longitudinal (E field parallel with propagation direction) Kerr effect. The reason for this is physically quite clear. If a material is to respond linearly to an electric field, the effect of the field must change sign when the field changes sign. This means that the medium must be able to distinguish (for example) between 'up' (positive field) and 'down' (negative field). But it can only do this if it possesses some kind of directionality in itself, otherwise all field directions must be equivalent in their physical effects. Hence, in order to make the necessary distinction between up and down, the material must possess an intrinsic asymmetry and, hence, must be crystalline. By a similar argument a longitudinal E field can only produce a directional effect orthogonally to itself (i.e. in the direction of the optical electric field) if the medium is anisotropic (i.e. crystalline) for otherwise all transverse directions will be equivalent.

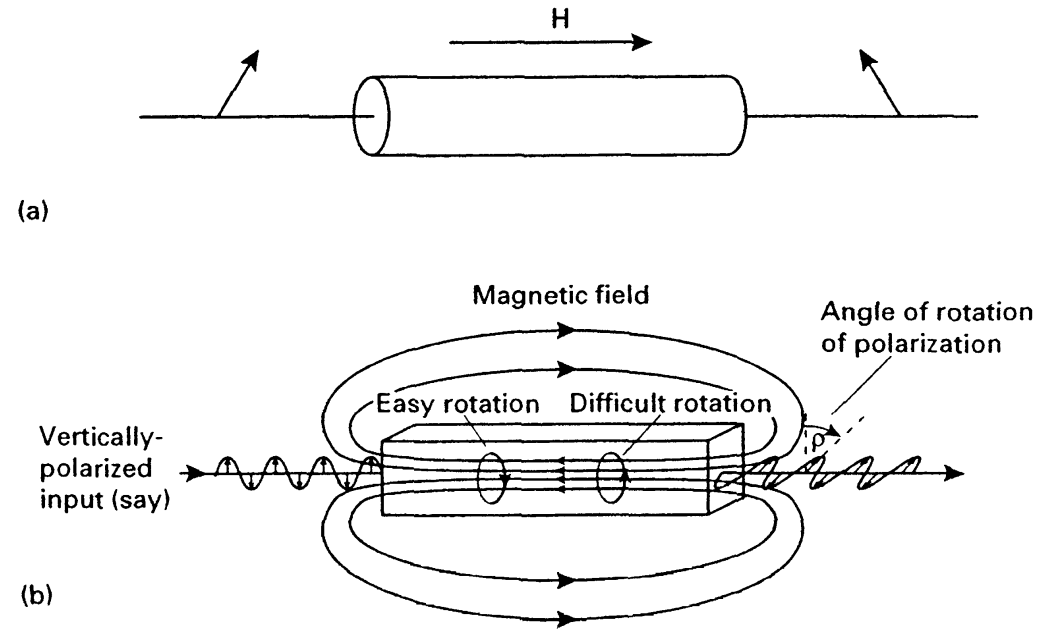


Figure A5.33. The Faraday magneto-optic effect. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

In addition to the modulation of light (phase or intensity/power) it is clear that the electro-optic effect could be used to measure an electric field and/or the voltage which gives rise to it. Several modulation and sensors are based on this idea.

(b) The magneto-optic effect. If a magnetic field is applied to a medium in a direction parallel with the direction in which light is passing through the medium, the result is a rotation of the polarization direction of whatever is the light's polarization state: in general, the polarization ellipse is rotated. The phenomenon, known as the Faraday (after its discoverer, in 1845) magneto-optic effect, is normally used with a linearly polarized input, so that there is a straightforward rotation of a single polarization direction (figure A5.33). The magnitude of the rotation due to a field, H, over a path length, L, is given by:

$$\rho = V \int_0^L H \cdot \mathrm{d}l$$

where V is a constant known as the Verdet constant: V is a constant for any given material, but is wavelength dependent. Clearly, if H is constant over the optical path, we have:

$$\rho = VHL$$

From the discussion in section A5.3.7, we see that this is a magnetic-field-induced circular birefringence.

The physical reason for the effect is easy to understand in qualitative terms. When a magnetic field is applied to a medium the atomic electrons find it easier to rotate in one direction around the field than in the other: the Lorentz force acts on a moving charge in a magnetic field and this will act radially on the electron as it circles the field. The force will be outward for one direction of rotation and inward for the other. The consequent electron displacement will lead to two different radii of rotation and thus two different rotational frequencies; and electric permittivities. Hence the field will result in two different refractive indices, and thus to circular birefringence. Light which is circularly polarized in the 'easy' (say clockwise) direction will travel faster than that polarized in the 'hard' direction (anti-clockwise), leading to the observed effect

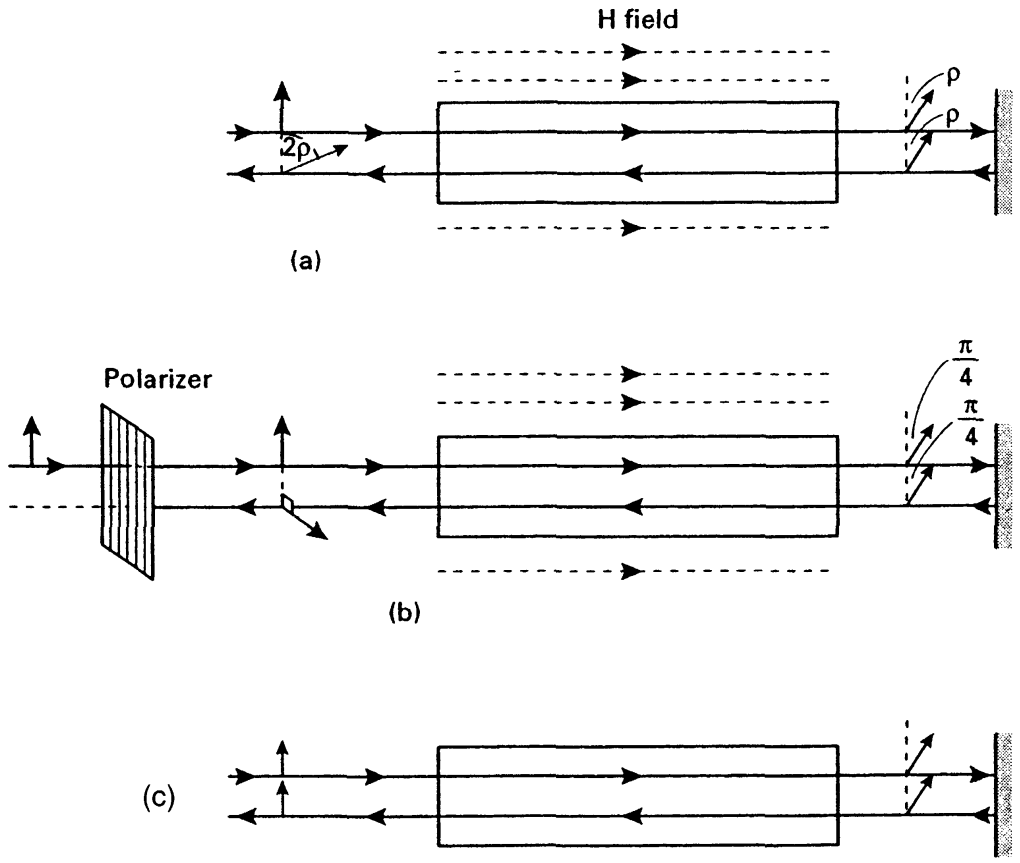


Figure A5.34. Reciprocal and non-reciprocal polarization rotation. (*a*) Non-identical rotation (Faraday effect). Rotation in same direction in relation to the magnetic field. (*b*) Optical isolator action. Total rotation of $\pi/2$ for polarization blocking. (*c*) Reciprocal rotation (optical activity). Rotation in same direction in relation to propagation direction. (From Rogers A 1997 *Essentials of Optoelectronics* (Cheltenham: Nelson Thornes) with permission.)

(figure A5.33(*b*)). Another important aspect of the Faraday magneto-optic effect is that it is 'non-reciprocal'. This means that linearly polarized light (for example) is always rotated in the same absolute direction in space, independently of the direction of propagation of the light (figure A5.34(*a*)). For an optically active crystal this is not the case: if the polarization direction is rotated from right to left (say) on forward passage (as viewed by a fixed observer) it will be rotated from left to right on backward passage (as viewed by the same observer), so that back-reflection of light through an optically active crystal will result in light with zero final rotation, the two rotations having cancelled out (figure A5.34(*c*)). This is called a reciprocal rotation because the rotation looks the same for an observer who always looks in the direction of propagation of the light (figure A5.34(*c*)).

For the Faraday magneto-optic case, however, the rotation always takes place in the same direction with respect to the magnetic field (not the propagation direction) since it is this which determines 'easy' and 'hard' directions. Hence, an observer always looking in the direction of light propagation will see different directions of rotation since he/she is, in one case, looking along the field and, in the other, against it. It is a non-reciprocal effect. The Faraday effect has a number of practical applications. It can be used to modulate light, although it is less convenient for this than the electro-optic effect. This is a result of the greater difficulty of producing and manipulating large and rapidly varying (for high modulation bandwidth) magnetic fields when compared with electric fields (large solenoids have large inductance!).

The Faraday magneto-optic effect can valuably be used in optical isolators, however. In these devices light from a source passes through a linear polarizer and then through a magneto-optic element which rotates

the polarization direction through 45°. Any light which is back-reflected by the ensuing optical system suffers a further 45° rotation during the backward passage, in the same rotational direction, thus arriving back at the polarizer having been rotated through 90°; it is therefore blocked by the polarizer (figure A5.34(*b*)). Hence the source is isolated from back-reflections by the magneto-optic element/polarizer combination which is thus known as a Faraday magneto-optic isolator. This is very valuable for use with devices whose stability is sensitive to backreflection, such as lasers and optical amplifiers, and it effectively protects them from feedback effects. The Faraday magneto-optic effect can also be used to measure magnetic fields and the electric currents which give rise to them [13]. There are other magneto-optic effects (e.g. Kerr, Cotton–Mouton, Voigt) but the Faraday effect is by far the most important for opto-electronics.

A5.4 Conclusions

In the first part of this chapter we began by noting that the light wave comprises field vibrations which take place transversely to the propagation direction. This makes it possible to explain satisfactorily the phenomenon of optical interference. With this understanding we saw also how to design useful interferometers for the analysis and control of light.

We have also looked at the conditions necessary for optical waves to interfere in a consistent and measurable way, with themselves and with other waves. We have seen that the conditions relate to the extent to which the properties such as amplitude, phase, frequency and polarization remain constant in time and space, i.e. the extent to which knowledge of the properties at one point in time or space tells us about these properties at other points.

Any interference pattern will remain detectable only as long as coherence persists and, by studying the rise and fall of interference patterns, much can be learned about the sources themselves and about the processes which act upon the light from them.

Coherence also relates critically to the information-carrying capacity of light and to our ability to control and manipulate it sensibly. The design and performance of any device or system that relies on interference or diffraction phenomena must take into account the coherence properties of the sources to be used; some of these work to the designer's disadvantage, but others do not.

In the second part of this chapter we have looked closely at the directionality possessed by the optical transverse electric field, i.e. we have looked at optical polarization. We have seen how to describe it, to characterize it, to control it, to analyse it and how, in some ways, to use it.

We have also looked at the ways in which the transverse electric and magnetic fields interact with directionalities (anisotropies) in material media through which the light propagates; and we looked briefly at the ways in which these material interactions allow us to control light: to modulate it and, perhaps, to analyse it.

All of these ideas bear upon more advanced phenomena such as those which allow light to switch light and to process light, opening up a new range of possibilities in the world of very fast (femtosecond, $\sim 10^{-15}$ s) phenomena.

Acknowledgments

Much of the material in this chapter was first presented in 'Essentials of Optoelectronics' (Rogers), published by Chapman and Hall, 1997, and is included here with permission.

References

[1] Michelson A A 1882 Interference phenomena in a new form of refractometer *Am. J. Sci.* **23** 395–400
[2] Michelson A A and Morley E W 1887 On the relative motion of the earth and the luminiferous aether *Phil. Mag.* **24** 449–63
[3] Born M and Wolf E 1975 The Fabry–Pérot interferometer *Principles of Optics* 5th edn (Oxford: Pergamon) section 7.6.2, pp 329–33

[4] Twyman F and Green A 1916 British Patent No 103832
[5] Michelson A A 1920 An interferometer for measurement of stellar diameters *Astrophys. J.* **51** 263
[6] Vali V and Shorthill R W 1976 Fiber ring interferometer *Appl. Opt.* **15** 1099–100
[7] Lefevre H 1993 *The Fiber-Optic Gyroscope* (Boston: Artech House)
[8] Nye J F 1976 *Physical Properties of Crystals* (Oxford: Clarendon)
[9] Clark Jones R 1941 A new calculus for the treatment of optical systems *J. Opt. Soc. Am.* **38** 671–85
[10] Jerrard H G 1954 Transmission of light through birefringent and optically-active media *J. Opt. Soc. Am.* **44** 634–40
[11] Barlow A J, Ramskov-Hansen J J and Payne D N 1982 Anisotropy in spun single-mode fibres *Electron. Lett.* **18** 200–2
[12] Varnham P *et al* 1983 Single polarization operation of highly-Birefringent 'bow-tie' optical fibres *Electron. Lett.* **19** 246–7
[13] Rogers A J 1988 Optical-fibre current measurement *Int. J. Optoelectron.* **3** 391–407

A6
Optical waveguide theory

G Stewart

A6.1 Introduction

Over the last 30 years or so, much effort has been devoted to advancing the theory and practice of optical waveguides. This has been mainly driven by the widespread deployment of optical fibre communication systems where system performance has been dramatically improved over the last three decades. In addition to the optical fibres themselves, a number of optical components, such as diode lasers, couplers and external modulators, are formed on optical waveguide structures and their development has demanded a thorough understanding of waveguide theory. On a smaller scale, and particularly since the early 1980s, there has been a growing research programme on optical sensor technology which has also provided a stimulus for the development of new types of waveguide devices and the associated theoretical models.

In this chapter, after presenting a brief review of the various types of optical waveguides, we outline the key principles and parameters which describe and define the operation of optical waveguides and fibres. The ways in which propagation through optical fibres affects the properties of the guided waves are discussed, including dispersion and non linear effects. Power transfer between propagating waves is essential to the operation of a number of components and the fundamentals of coupling theory are reviewed. In summary, the theory given provides the foundation for understanding the detailed operation of a wide variety of optical components and systems based on optical fibre technology.

A6.2 Basic types of optical waveguides

The simplest form of optical waveguide is the three-layer *planar* or *slab guide* shown in figure A6.1(*a*) consisting of a central guiding layer sandwiched between lower index layers. If the layers on either side are of equal index, it is known as a *symmetric* planar guide, otherwise it is an *asymmetric* guide. The lower layer (on which the waveguide may be formed in practice) is called the *substrate* and the top layer the *superstrate*. A further distinction arises depending on the index distribution within the guide. If the layers are of uniform index then the guide is referred to as a *step-index* structure, whereas if the index varies (usually within the central layer) it is known as a *graded-index* guide. As discussed later, the central (waveguiding) layer is typically a few micrometres in thickness for single-mode operation.

Typical *rectangular* or *two-dimensional (2D)* guides are illustrated in figures A6.1(*b*) and (*c*). Here the central (waveguiding) layer is confined to a narrow channel or strip a few micrometres in dimension. Depending on the fabrication technology employed, the 2D guide may have a ridge structure or be embedded within a planar substrate. As with planar guides, the 2D guide is classified as step or graded index in structure.

Optical fibre waveguides, shown in figure A6.1(*d*), (*e*) and (*f*), consist of a circular core surrounded by a *lower-index* cladding. Typically, standard *single-mode* fibres have a core/cladding diameter of $9/125$ μm whereas *multi-mode* fibres have dimensions of $50/125$ μm or $62.5/125$ μm and may be step or graded

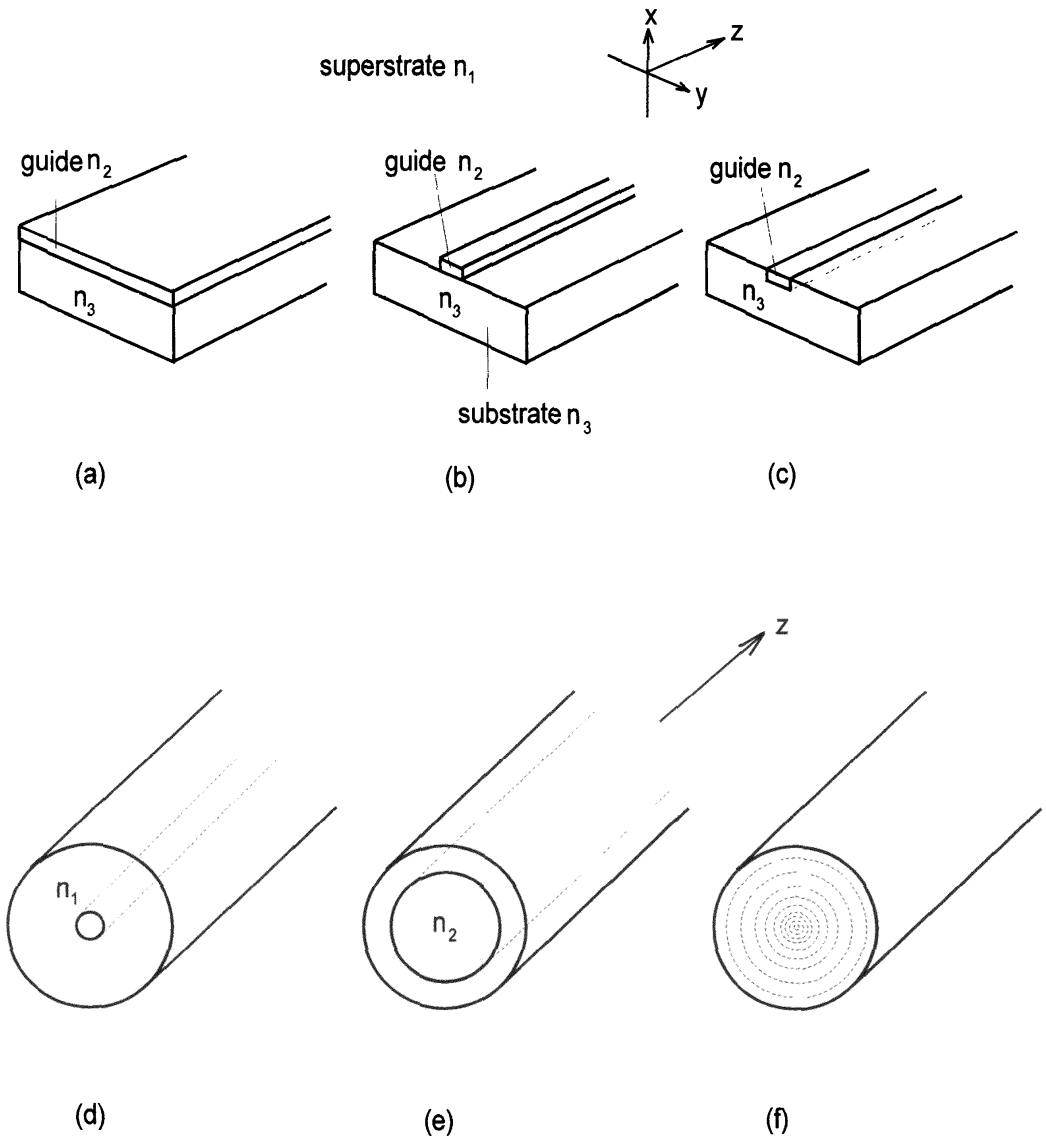


Figure A6.1. Various types of optical waveguide: (*a*) planar, (*b*) ridge, (*c*) embedded, (*d*) single-mode fibre, (*e*) multi-mode step-index fibre and (*f*) graded-index fibre.

index in construction. In terms of transmission data rates, *single-mode* fibres are superior, followed by the graded-index multi-mode type.

A6.3 Planar and rectangular guides

A6.3.1 Planar guides

The simple planar guide, with a ray optics approach, provides a useful starting point for understanding the key properties of optical fibres and waveguides in general. If we first consider a light ray incident on the boundary between two materials of index n_1 and n_2, where $n_1 < n_2$, as shown in figure A6.2(*a*), then if θ exceeds the critical angle, θ_c, given by $\sin\theta_c = n_1/n_2$, total internal reflection (TIR) occurs. In TIR, the light actually penetrates a short distance beyond the boundary, i.e. the field does not abruptly drop to zero at the boundary but decays exponentially on the lower index side. This exponential field is called the *evanescent field* and has an associated penetration (or $1/e$) depth. As a result, light which has undergone TIR is phase-shifted from

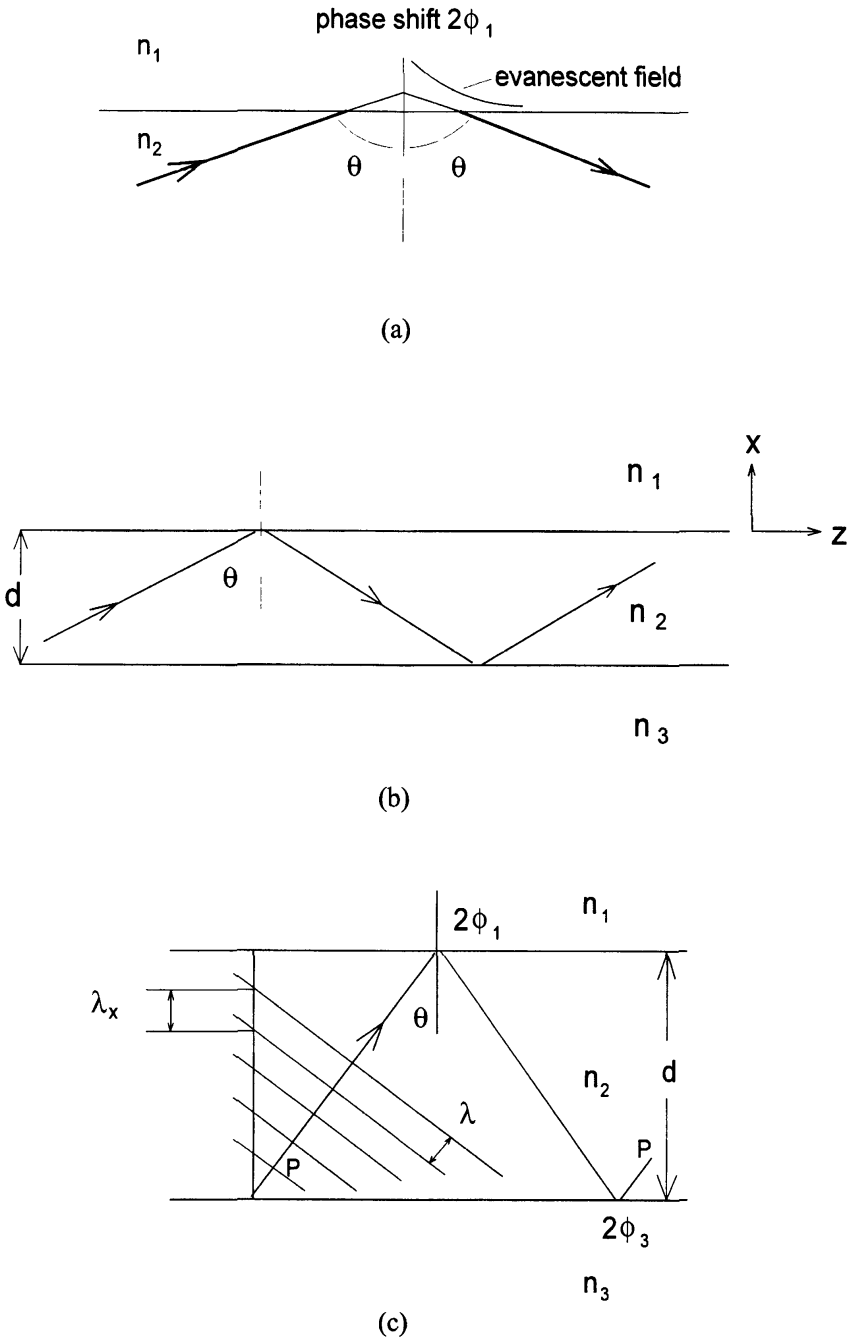


Figure A6.2. (*a*) Expanded view of total internal reflection (TIR), (*b*) light trapped in guide by TIR and (*c*) transverse resonance condition for guided modes.

the incident light by an amount $2\phi_1$ where

$$\tan\phi_1 = \xi \frac{\sqrt{\sin^2\theta - sin^2\theta c}}{\cos\theta} \tag{A6.1}$$

In equation (A6.1), $\xi = 1$ for s-polarization (i.e. an electric field perpendicular to the plane of incidence) and $\xi = n_2^2/n_1^2$ for p-polarized light.

As noted earlier, a planar optical waveguide is formed by a higher index layer sandwiched between regions of lower index and so light rays may be trapped in the core layer by TIR whenever $\theta > \theta_c$ as shown in figure A6.2(*b*). However, for guiding to occur, a standing-wave pattern must be established across the guide

Table A6.1. Cut-off thickness d_C for several modes of a slab guide (parameters in text).

Mode	TE_0	TM_0	TE_1	TM_1	TE_2	TM_2
d_C (μm)	0.51	0.58	1.81	1.89	3.11	3.19

since the wave must be confined between the boundaries. Put in another way, the round-trip phase change in the transverse (x-direction, see figure A6.2(c)) must satisfy

$$-\left[\frac{2\pi}{\lambda_x} \times 2d\right] + 2\phi_1 + 2\phi_3 = \pm 2m\pi \tag{A6.2}$$

where m is an integer and λ_x is the spacing of the wavefronts in the x-direction, $\lambda_x = \lambda/\cos\theta = \lambda_0/(n_2\cos\theta)$. Hence

$$k_0 d n_2 \cos\theta = m\pi + \phi_1 + \phi_3 \tag{A6.3}$$

where $m = 0, 1, 2, \ldots$ is the mode order, with $m = 0$ for the fundamental mode and $k_0 = 2\pi/\lambda_0$.

Since m is an integer, this condition implies that only a *discrete* set of values of θ are allowed, $\theta_0, \theta_1, \theta_2, \ldots, \theta_m$ and each allowed value of θ corresponds to a certain transverse standing-wave pattern or *guided mode*. (Compare standing waves on a string, fixed at both ends, giving rise to different modes of vibration.) The modes which arise from s-polarized rays are transverse-electric (TE) modes because the electric field (but not the magnetic field) is entirely transverse to the propagation direction (z-direction) of the mode. Similarly p-polarized rays give rise to transverse-magnetic (TM) modes. Hence, in general, a multi-mode guide will support two classes of modes, $TE_0, TE_1, TE_2, \ldots$, etc and $TM_0, TM_1, TM_2, \ldots$, etc. Since ϕ is polarization-dependent, corresponding TE and TM modes have slightly differing values of θ and, hence, propagation constant (this is *waveguide birefringence*).

A very useful definition for describing the propagation of a guided mode is its *effective-index* value. With reference to figure A6.1(c), the phase velocity of the ray in the guide is c/n_2 but since the ray zig-zags at an angle θ, the wave-fronts of the guided mode propagate in the z-direction with a phase velocity of $c/(n_2\sin\theta_m) = c/n_e$ where $n_e = n_2\sin\theta_m$ is the effective index of the mode. Because $\theta > \theta_c$ for all guided modes, the allowed range of n_e is $n_2 > n_e >$ (n_1 and n_3) and the propagation constant of the mode, $\beta_m = k_0 n_e$. With this definition, the eigenvalue equation for the modes (A6.3) can also be written in the form:

$$k_0 d\sqrt{n_2^2 - n_e^2} = m\pi + \phi_1 + \phi_3. \tag{A6.4}$$

The condition $n_e = n_1$ or $n_e = n_3$ (whichever is the greater) corresponds to the *cut-off* condition for a mode, since at that point the ray is travelling at the critical angle at one of the boundaries. This condition may be used in equation (A6.4) to determine the cut-off thickness (or cut-off wavelength, if the thickness is known) given the other waveguide parameters, for a particular mode of order m. To illustrate, table A6.1 gives typical cut-off values calculated from equation (A6.4) for a film of index $n_2 = 1.5$ on a quartz substrate ($n_3 = 1.45$) with air superstrate ($n_1 = 1$) at a wavelength of 1 μm.

The cut-off thickness for a mode is the minimum thickness which will support that mode. Using the data in table A6.1 as an example, if the thickness is in the range $0.51 < d < 0.58$ μm then the guide will support the TE_0 mode only. Data of this type are used to design waveguides for single-mode operation or for operation in a selected number of modes, by choosing the appropriate guide dimensions.

At this point, let us consider in some detail the symmetric slab (which provides a very simple planar model for an optical fibre) and introduce a parameter known as the *V-number* or normalized frequency. The

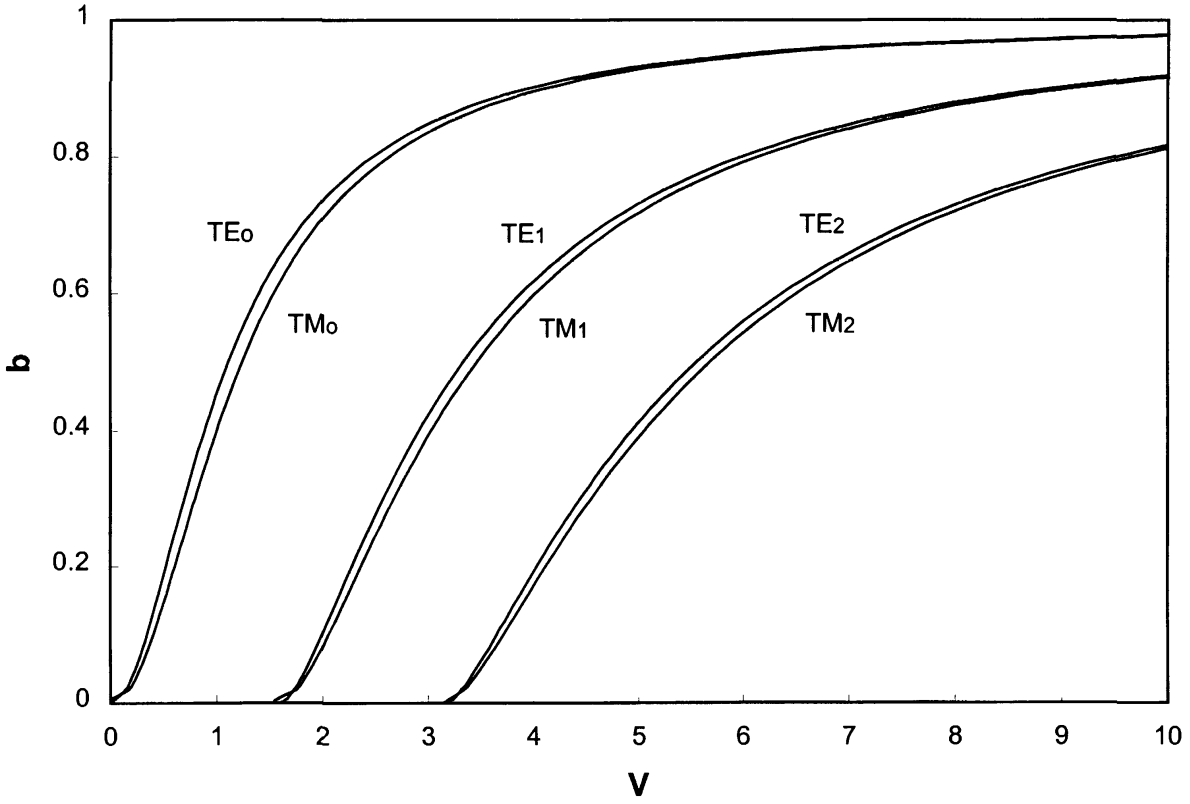


Figure A6.3. Universal dispersion curves for a symmetric planar guide.

V-number is a very important parameter for determining the number of modes supported by a guide. By analogy with an optical fibre, the V-number for the planar (slab) guide may be defined as

$$V_{\text{slab}} = k_0 \left(\frac{d}{2}\right) \sqrt{n_2^2 - n_1^2}. \tag{A6.5}$$

(Note that $d/2$ is the half-width of the planar guide which is analogous to the fibre radius a as used in the definition of the fibre V-number in equation (A6.12).)

In addition, it is customary to define a normalized propagation constant, or b-parameter, which lies in the range $0 < b < 1$, as follows:

$$b = \frac{n_e^2 - n_1^2}{n_2^2 - n_1^2} \tag{A6.6}$$

With these definitions, equation (A6.4) can be recast in a form which applies to any symmetric slab guide:

$$2V\sqrt{1-b} = m\pi + 2\tan^{-1}\xi\sqrt{\frac{b}{1-b}}. \tag{A6.7}$$

Figure A6.3 shows universal b–V curves (or dispersion curves) plotted from equation (A6.7) with $\xi = n_2^2/n_1^2$ chosen as 1.2 to show waveguide birefringence. Note that mode cut-off occurs at $b = 0$, giving the cut-off value for a mode of order m as $V_C = m\pi/2$ and that TE and TM modes have the same cut-off points in a symmetric guide. Also for single-mode operation, $0 < V < \pi/2$, so that in theory the fundamental mode is always supported in a symmetric guide (compare the asymmetric slab as illustrated in table A6.1 where TE and TM have different cut off points and the fundamental mode is cut off below a certain value).

Waveguides which are made by diffusion or ion-exchange processes generally have a graded-index core. In terms of ray optics, we can visualize the ray as being continuously refracted and travelling in a curved path

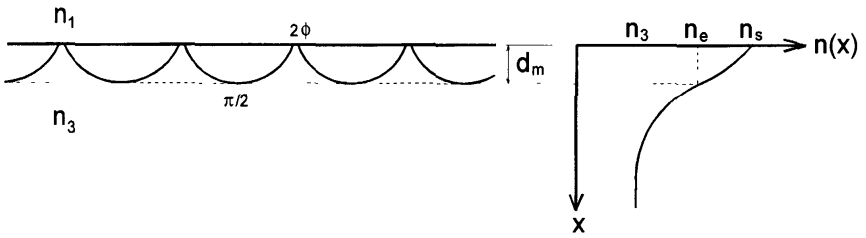


Figure A6.4. Graded-index slab guide.

through the graded index medium as illustrated in figure A6.4. By applying the same procedure of requiring the round-trip phase change in the transverse direction to be an integer number of 2π, the following equation may be derived [1] for the modes of a graded-index guide:

$$k_0 \int_0^{d_m} \sqrt{n^2(x) - n_e^2}\, dx = m\pi + \tfrac{1}{4}\pi + \phi \tag{A6.8}$$

where $n(x)$ is the index profile of the guide, d_m is the mode depth given by $n(d_m) = n_e$ and 2ϕ is the usual phase shift at the superstrate interface. Note that if the guide is buried into the substrate (e.g. a symmetric parabolic profile with maximum index below the surface), then $2\phi = \pi/2$. For certain profile forms, such as a linear or parabolic variation in index, the integral in equation (A6.8) may be evaluated in closed form [1,2].

So far, all the results discussed are derived on the basis of a simple ray-optics model of a planar waveguide. A more rigorous method is to obtain the field solutions by application of Maxwell's equations. The guided modes are assumed to propagate in the z-direction as described by a travelling-wave term of the form: $\exp i\{\omega t - \beta_m z\}$. For a planar guide, the transverse-field distribution is assumed to vary in only the x-dimension so that the Maxwell wave equation is simplified by the assumption that the $\partial^2/\partial y^2$ term is zero. From the one-dimensional wave equation, one can then obtain the transverse field component (for example, E_y for TE modes) in the form of either a decaying exponential for the cladding regions or as a co-sinusoidal distribution in the guiding core. The other field components may then be obtained from E_y by application of Maxwell's curl equations. When continuity conditions are applied on tangential field components at the boundaries of the guide the same eigenvalue equation (A6.4) is obtained for the guided modes. For TE modes in a step-index planar guide, the TE field component, E_y, can thus be obtained in the form:

$$\begin{aligned}
&\text{Substrate } (x < -d): && E_y = \pm C \frac{\cos\phi_3}{\cos\phi_1} \exp\{\gamma_3(x+d)\} \\
&\text{Core } (-d < x < 0): && E_y = \frac{C}{\cos\phi_1} \cos\{k_x x + \phi_1\} \\
&\text{Superstrate } (x > 0): && E_y = C \exp\{-\gamma_1 x\}
\end{aligned} \tag{A6.9}$$

where $+$ is for even-mode orders and $-$ is for odd-mode orders, C is a arbitrary constant, γ is the decay constant of the evanescent field, $\gamma_i = k_0\sqrt{n_e^2 - n_i^2}$ and $k_x = k_0\sqrt{n_2^2 - n_e^2}$. Note that in these equations the field is matched at the boundaries $x = 0$ and $x = -d$ and all the fields should be multiplied by the travelling-wave term: $\exp i\{\omega t - \beta_m z\}$. The constant C is the field amplitude at the $x = 0$ boundary and may be related to the total power in the guide.

Figure A6.5 shows the field distribution, E_y, for the first few TE modes of an asymmetric step-index slab guide. Note the cosine distribution of the field in the core, forming a standing-wave pattern, with an increasing number of cycles for increasing mode order. The cosine distribution is connected smoothly (through the continuity conditions as noted earlier) to the evanescent field tails which penetrate the substrate and superstrate. The evanescent field penetration depth, defined by $d_p = 1/\gamma$, increases with mode order

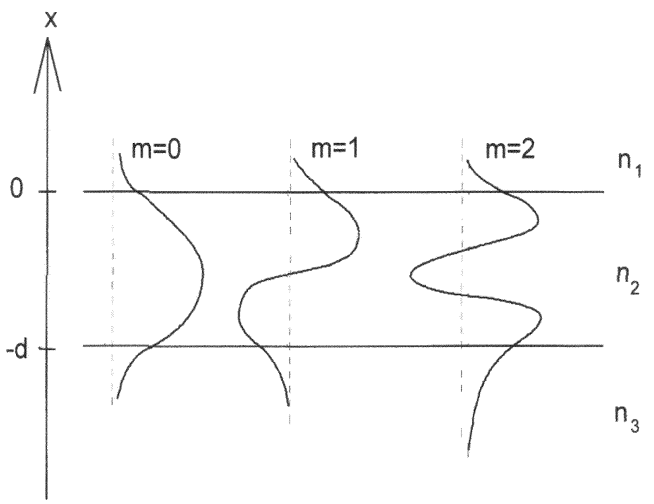


Figure A6.5. Field distribution of modes in a planar step-index slab guide.

and is larger in the substrate for $n_3 > n_1$. Each of these field patterns represent a particular way in which the light is guided by the structure and, hence, is described as a *mode* of the guide.

A6.3.2 Two-dimensional guides

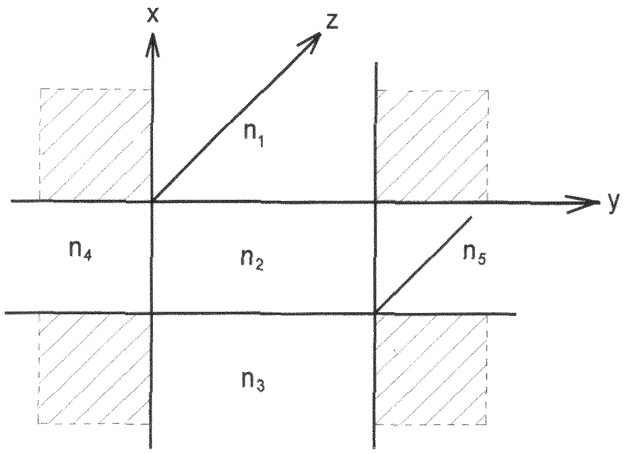
Planar structures are generally impractical for guiding light because of diffraction spreading of the beam in the plane, so 2D or channel guides are required for the manufacture of integrated optic guided wave devices such as Y-junctions, branching elements, couplers, interferometers and modulators [3]. However, the theoretical description and analysis of 2D guides is complex and, in general, a rigorous electromagnetic solution is required. There are some approximate methods that can be applied, mainly based on the extension of planar waveguide concepts, which we briefly consider here.

First of all, Marcatili [4] provided an approximate electromagnetic solution in closed form for the general structure shown in figure A6.6(a) under the assumption of small index differences and rigorous field matching along the sides only (not in the shaded corners where the field is relatively negligible). Under these conditions one can regard the light as travelling nearly axially down the guide and consequently the modes are essentially TEM in character with electric field polarization along either the x- or y-axis. The field distribution in the x–y plane (transverse to the propagation direction) may then be constructed in the form of a product: $E(x, y) = E_1(x)E_2(y)$ where $E_1(x)$ and $E_2(y)$ have the form of planar-type solutions as in equation (A6.9). For example, for the y-polarized mode we would write:

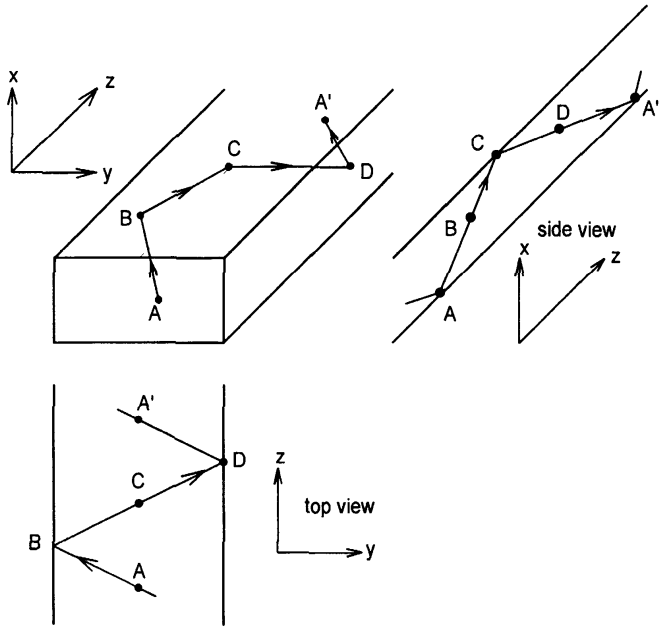
$$\begin{aligned} &\text{in the core:} && E_y = C\cos(k_x x + \phi_x)\cos(k_y y + \phi_y) \\ &\text{in the region:} x > 0: && E_y = C\cos(\phi_x)\cos(k_y y + \phi_y)\exp\{-\gamma_1 x\} \end{aligned} \tag{A6.10}$$

and so on.

Following on from the Marcatili approach, the mode parameters may be obtained by the *effective index method* [1,5] which makes use of planar mode conditions. To understand this, figure A6.6(b) shows a simple ray-optics picture of a mode in the 2D guide where the ray undergoes total internal reflection in the sequence: A (bottom) → B (left side) → C (top) → D (right side) → A′ (bottom), and so on. Viewed from above or from the side, the ray describes a zig-zag path as in a planar guide and the 2D guide appears as a combination of two planar guides. In the effective-index method, the two planar guides are linked by using the effective index calculated from the first guide to replace the core index of the second. The procedure is as follows (see figure A6.6(c)).



(a)

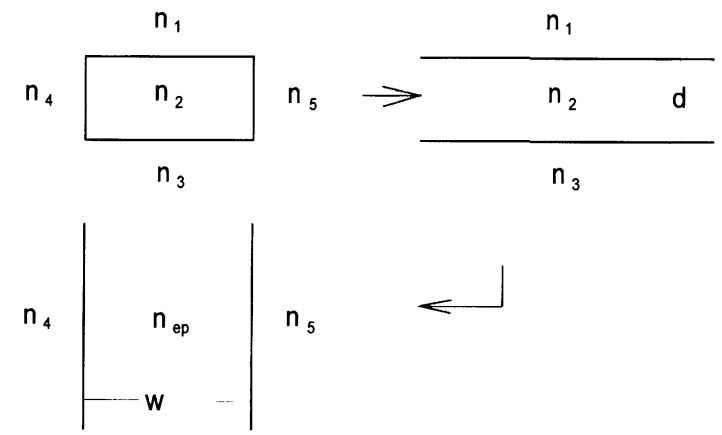


(b)

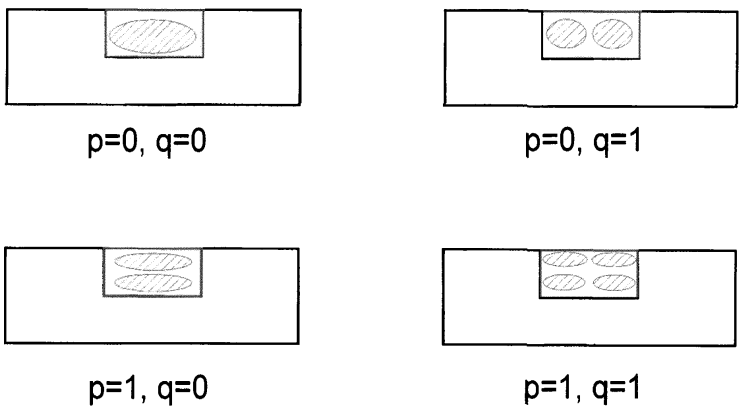
Figure A6.6. (*a*) Model for the general analysis of 2D guides, (*b*) ray paths in a 2D guide, (*c*) effective-index method for 2D guides and (*d*) field and intensity distribution of modes in a 2D guide.

(i) Let the long dimension of the guide $\rightarrow \infty$ to obtain the first planar guide. Use the characteristic equation (A6.4) for a planar guide (or the b–V characteristic from equation (A6.7) if the guide is symmetrical) to obtain the effective index, n_{ep}, for each of the allowed modes with mode order $p(= 0, 1, 2, \ldots)$.

(ii) Construct the second planar guide in the other direction as shown, but replacing n_2 with each value of n_{ep} in turn. Again calculate the effective index n_{epq} for each allowed mode with mode order q $(= 0, 1, 2, \ldots)$ from the planar guide equation (A6.4).



(c)



(d)

Figure A6.6. (Continued.)

The set of calculated effective indices, n_{epq}, provides an approximate description of the set (p, q) of 2D guided modes. Field parameters, γ_1, γ_3, γ_4, γ_5 and k_x, k_y may also be calculated from the effective-index values. Figure A6.6(d) shows the typical field distribution for several modes of the 2D guide.

Note that in applying the method, the appropriate mode polarization in the planar guides must be used. For example, for 2D modes polarized along the y-axis, TE modes of the first planar guide would be used with TM modes of the second (*vice versa* for x-polarized 2D modes). The effective-index method can also be applied to graded-index structures [5], in which case the eigenvalue equation (A6.8) would be used.

A6.4 Optical fibres

Optical fibres are the most widely used form of waveguide because of their flexibility, low cost, small dimensions and ability to transport optical data over long lengths with low loss ($<$0.5 dB km^{-1}). As noted earlier, depending on the core size, optical fibres are either multi-mode (supporting perhaps many thousands of modes) or single mode in operation. Single-mode fibres are the most important in communication systems because of their high information-carrying capacity but multi-mode fibres have found several important

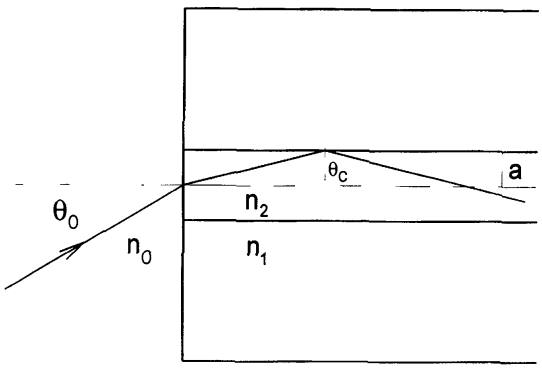


Figure A6.7. Optical fibre parameters and the numerical aperture.

applications in sensor technology, such as in evanescent field sensors and in sensors exploiting mode-coupling through micro- or macro-bending. Graded-index multi-mode fibres may also be used for lower capacity or shorter length communication links.

Three important parameters often quoted for fibres are the numerical aperture (NA), V-number and birefringence (B) defined as follows (see figure A6.7).

$$NA = n_0 \sin\theta_0 = \sqrt{n_2^2 - n_1^2} \tag{A6.11}$$

$$V = k_0 a \sqrt{n_2^2 - n_1^2} \tag{A6.12}$$

$$B = (n_{ex} - n_{ey}). \tag{A6.13}$$

The numerical aperture defines the cone of light accepted into guided modes by the fibre and is particularly useful for multi-mode fibres. The V-number, similar to that defined for planar waveguides, may be used to determine the number of modes supported by the fibre and to determine the range for single-mode operation. The birefringence defines the difference in effective index between the two polarization states and will be discussed later.

A6.4.1 Description of the modes and fields in optical fibres

Under the assumption of a core surrounded by an infinite cladding, the electric and magnetic fields of guided waves in optical fibres, as well as the eigenvalue equations for the modes, may be obtained from application of Maxwell's wave equations to the cylindrical structure [6–8]. In general, the fibre supports several types of modes classified as: transverse electric (TE) modes where $E_z = 0$, transverse magnetic (TM) modes where $H_z = 0$ and hybrid modes (HE and EH) where both E_z and H_z are non-zero. The fundamental mode is designated the HE_{11} mode. The problem with this approach is that the mode description, and the associated eigenvalue equations, are rather complicated mathematically and not easy to apply in practical applications.

A much more useful description can be obtained by realizing that most fibres used in practice have a relatively small index difference between core and cladding[1], i.e. $\Delta = (n_2 - n_1)/n_1 \ll 1$. Under this condition of weak guidance, modes propagate nearly axially down the fibre and field components along the propagation direction (z-direction) are, therefore, small. Simplified solutions for these modes, where the fields are essentially transverse in nature, were first derived by Gloge [9] who called them *linearly polarized* or *LP* modes. For polarization along either the x- or y-axis, the transverse electric field has the form (see

[1] Sometimes Δ is defined as $\Delta = (n_2^2 - n_1^2)/2n_1^2$ but the two definitions are the same in weakly guiding fibre where the refractive indices are similar.

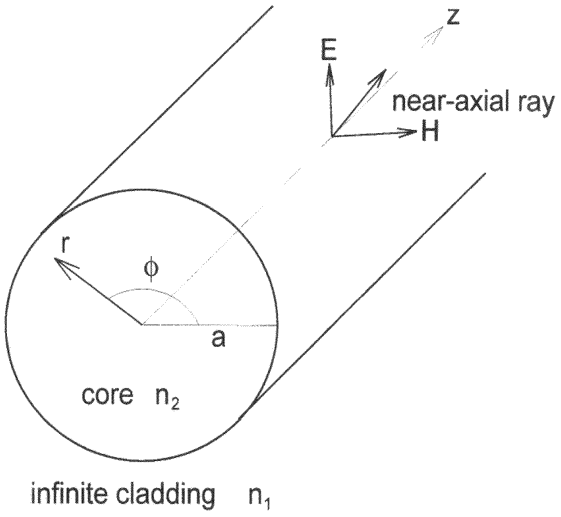


Figure A6.8. Fibre coordinates for describing LP modes.

figure A6.8)

$$E_{\text{core}} = E_l \frac{J_l(ur/a)}{J_l(u)} \cos(l\phi)$$
$$E_{\text{core}} = E_l \frac{K_l(wr/a)}{K_l(w)} \cos(l\phi) \tag{A6.14}$$

where l is an integer ($l = 0, 1, 2, \ldots$), J_l is a Bessel function of the first kind of order l and K_l is the modified Bessel function of order l, E_l is the field at the core/cladding interface and the parameters u and w are defined by:

$$u = k_0 a\sqrt{(n_2^2 - n_e^2)} \qquad w = k_0 a\sqrt{(n_e^2 - n_1^2)} \qquad V^2 = u^2 + w^2$$

with n_e the effective index of a mode as defined for planar guides. Note that these equations are multiplied by $\exp i(\omega t - \beta z)$ for mode propagation along the z-axis.

Bessel and modified Bessel functions are shown in figure A6.9. Note the comparison with planar guides where the field in the guiding layer is described by a sinusoidal function (compare the Bessel function for the fibre-core field) and in the substrate by an exponentially decaying field (compare the modified Bessel function for the fibre-cladding field).

The various modes are given the designation LP_{lm} where m is also an integer, $m = 1, 2, \ldots$, which indicates the zero of the Bessel function at which the mode is cut off as explained later. The fundamental mode is LP_{01} which corresponds to the HE_{11} mode from the exact theory. Apart from the fundamental, all the LP modes are, in fact, combinations of several exact modes which become degenerate (i.e. have nearly the same propagation constant) under the weakly guiding approximation; for example $\text{HE}_{21} + \text{TE}_{01}$ (or TM_{01}) $\rightarrow \text{LP}_{11}$. Figure A6.10 shows the intensity distribution and polarization for several LP modes. Note that l determines the number of field zeros in the azimuth direction through the $\cos(l\phi)$ term while m determines the number of zeros in the radial direction through the number of zero crossings of the Bessel function.

As in the case of planar waveguides, the electric and magnetic fields must satisfy boundary conditions at the core/cladding interface which leads to a characteristic or eigenvalue equation for the allowed modes. As figure A6.10 shows, the various modes correspond to standing-wave patterns in the plane transverse to propagation. For LP modes, the characteristic equation for the allowed modes is [9]:

$$u\frac{J_{l-1}(u)}{J_l(u)} = -w\frac{K_{l-1}(w)}{K_l(w)}. \tag{A6.15}$$

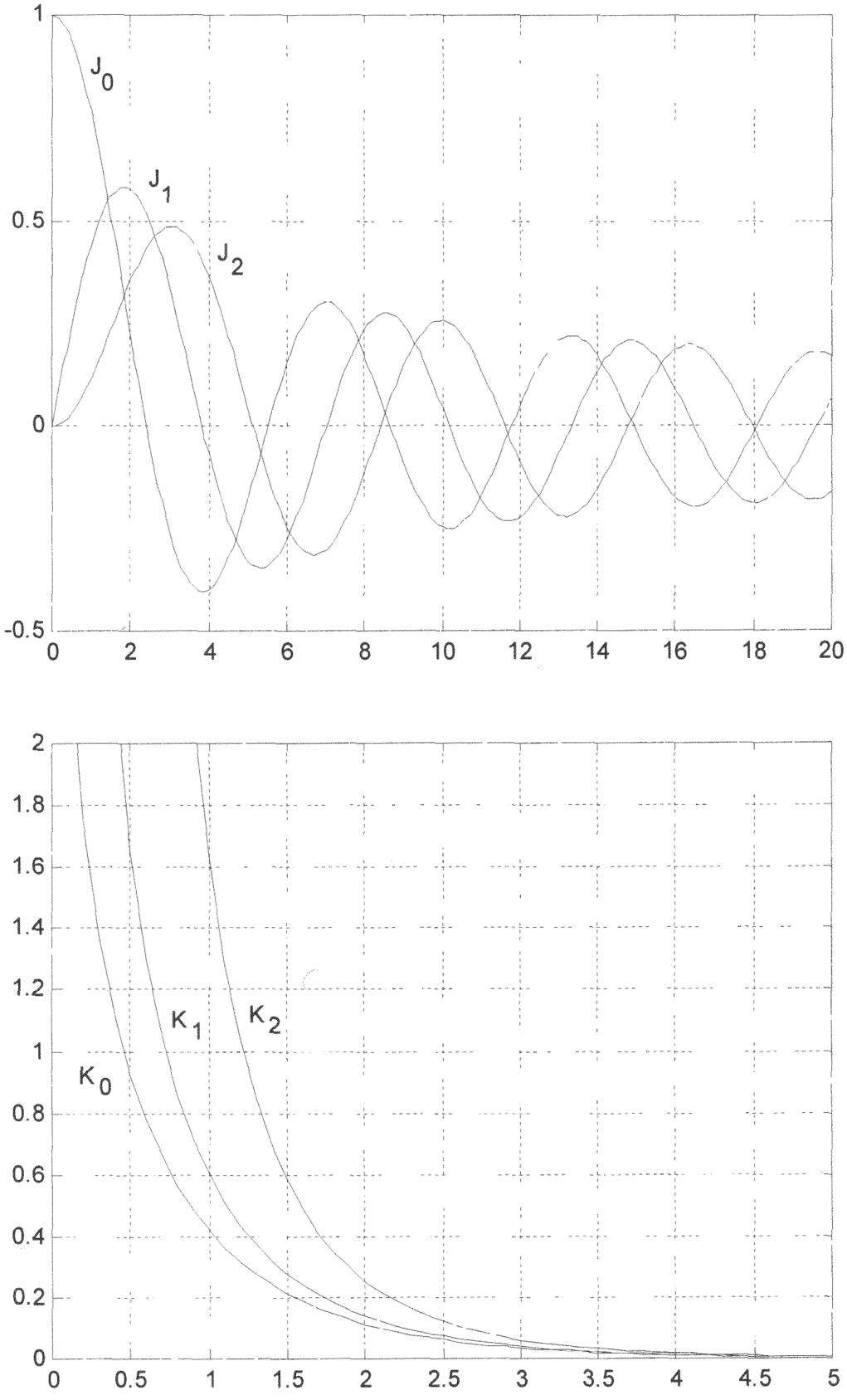


Figure A6.9. Bessel and modified Bessel functions.

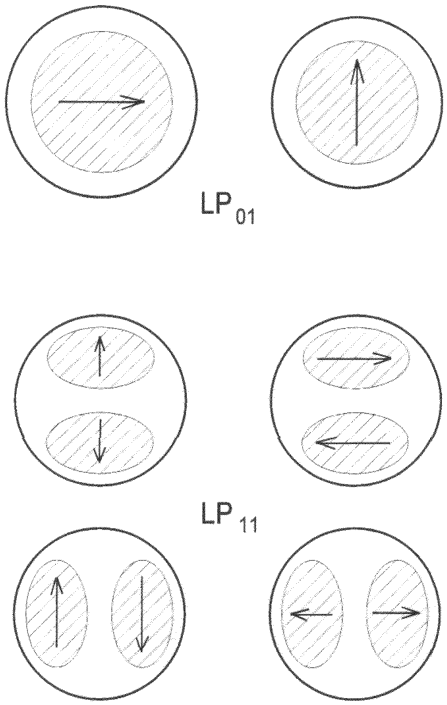


Figure A6.10. Intensity patterns for several LP modes.

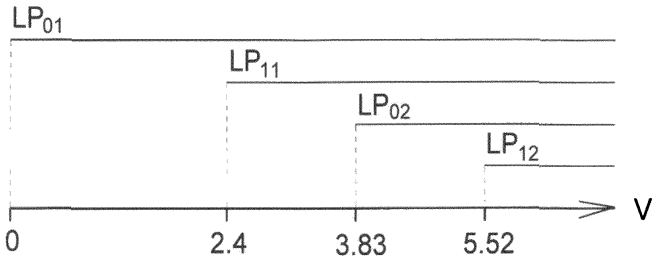


Figure A6.11. Mode cut-off points.

This equation is very useful for determining mode cut-off conditions. The cut-off is the point when the mode just ceases to be guided and occurs when the mode effective index equals the cladding index, $n_e = n_1$ or, equivalently, $w = 0$ which means that $V = u$. From equation (A6.15), this gives the general cut-off condition as $J_{l-1}(V) = 0$. The cut-off conditions for the various modes are, therefore (see figure A6.11):

LP_{0m} modes: since $l = 0$, the cut-off condition becomes: $J_{-1}(V) = -J_1(V) = 0$

$$LP_{01} \quad m = 1 \rightarrow \text{first zero crossing of } J_1 \Rightarrow V = 0$$
$$LP_{02} \quad m = 2 \rightarrow \text{second zero crossing of } J_1 \Rightarrow V = 3.832$$

etc.

LP_{1m} modes: since $l = 1$, the cut-off condition becomes: $J_0(V) = 0$

$$LP_{11} \quad m = 1 \rightarrow \text{first zero crossing of } J_0 \Rightarrow V = 2.405$$
$$LP_{12} \quad m = 2 \rightarrow \text{second zero crossing of } J_0 \Rightarrow V = 5.52$$

etc.

From this, the condition for only the LP_{01} mode to be guided is: $0 < V < 2.405$. This condition is very

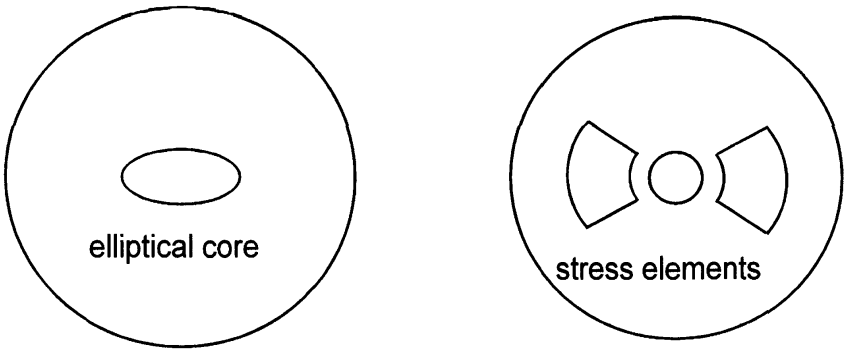


Figure A6.12. Polarization-maintaining fibres.

important in fibre design to determine the appropriate range of wavelengths or core diameters for single-mode operation. Two-mode operation (LP_{01} and LP_{11} only) occurs over the range $2.405 < V < 3.832$.

As in the case of planar guides, the characteristic equation (A6.15) may be used to obtain b–V dispersion curves for the various modes with mode cut-off at $b = 0$. Gloge [9] provides an approximate solution for the characteristic equation and presents the b–V curves for a large range of LP modes.

A6.4.2 Modal birefringence and polarization-maintaining fibres

As noted earlier, the modal birefringence, B, is the difference in effective indices between the orthogonally polarized modes, $B = (n_{ex} - n_{ey})$. In a perfectly symmetric circular core fibre, the x- and y-polarized modes would be identical so the propagation constants or effective indices would be equal and $B = 0$. In practice, the lack of perfect symmetry in conventional single-mode fibres means that two nearly-degenerate modes are propagated, with principal x- and y-axes defined by the asymmetry of the cross section (through geometrical, composition or strain-induced factors) and B ranges from about 10^{-5} to 10^{-6} with a lowest value of $\sim 10^{-9}$. In addition, perturbations arising during fibre manufacture or from twists and bends in operation will cause the principal axes to vary randomly along the fibre length. Coupling between the polarization states will, therefore, occur and such fibres cannot maintain a fixed polarization state for more than a few metres [10].

In certain applications, it is sometimes desirable to maintain a fixed polarization state on propagation through the optical fibre path, for example, in interferometric optical fibre sensors or to avoid noise arising from random fluctuations in the output polarization state. Polarization-maintaining (PM) fibres with $B \leq 10^{-4}$ can be manufactured by making use of either geometrical birefringence, as in the elliptical core fibre, or strain-induced birefringence by incorporating stress elements in the fibre as illustrated in figure A6.12.

With PM fibre, if linearly polarized light is launched into one of the polarization eigenmodes (i.e. polarized along the x- or y-axis) then the PM fibre will maintain the polarization state over a considerable distance (depending on the value of B) (see section A5.3). If, however, the input polarization direction is not parallel to either the x- or y-axis, both eigenmodes will be launched into the fibre. Since the x- and y-eigenmodes have different effective indices (or phase velocities), the phase difference between the modes will change along the length of the fibre and the polarization state will evolve from linear to elliptical to linear as illustrated in figure A6.13. After a certain distance, called the beat length, L_B, the phase difference will be 2π and the original polarization state will be recovered. The beat length is thus given by

$$\Delta\phi = (\beta_x - \beta_y)L_B = 2\pi$$

giving the beat length $L_B = \lambda_0/B$. The beat length can be observed visually in a PM fibre from the periodic variation of the scattered light along the fibre length. The Rayleigh scattered light can be thought of as originating from dipole radiators. The radiation pattern of a dipole has a minimum along the dipole axis and

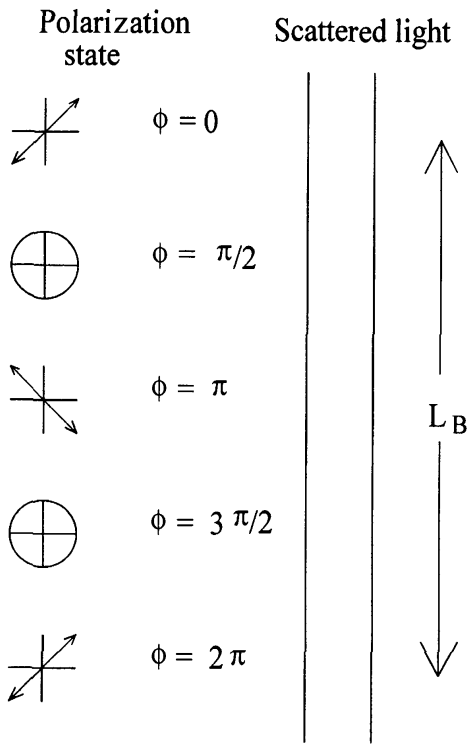


Figure A6.13. Evolution of polarization state along a PM fibre.

a maximum transverse to the axis so, if the fibre is viewed from any particular direction, the scattered light will vary periodically as the dipole orientation changes due to the changing polarization state. Observation of the beat length provides a useful way of measuring the birefringence of a fibre.

A6.5 Propagation effects in optical fibres

The most widespread use of optical fibres is for data and communication systems involving transmission of light pulses at rates of up to hundreds of gigabits per second. Consequently there is a need to consider the effects of propagation on the data pulses. We consider here three physical factors that can modify or distort the data pulses as a result of propagation through the fibre, namely

- attenuation which causes a reduction in the pulse intensity leading to reduced signal-to-noise ratios.
- dispersion which broadens the pulses, leading to possible overlap of pulses and errors in detection.
- nonlinear effects which can modify the pulse shape and also cause cross-talk between neighbouring channels in wavelength-division-multiplexed (WDM) systems.

A6.5.1 Attenuation in optical fibres

Figure A6.14 illustrates the attenuation characteristic of silica fibres, showing the operational wavelengths around 850, 1300 and 1550 nm of first-, second- and third-generation optical fibre communication systems respectively. The shape of the curve arises from several factors [6, 11].

On the short wavelength side, the attenuation is dominated by Rayleigh scattering and, to a lesser extent, by the tail of the UV absorption band. Rayleigh scattering has a strong wavelength dependence of the form $1/\lambda^4$ and arises from index variations in the glass over distances that are small in relation to the wavelength. Index variations occur because of fluctuations in the density and composition of the glass (especially since dopants are added) and from inhomogeneities during manufacture. UV absorption occurs when photons

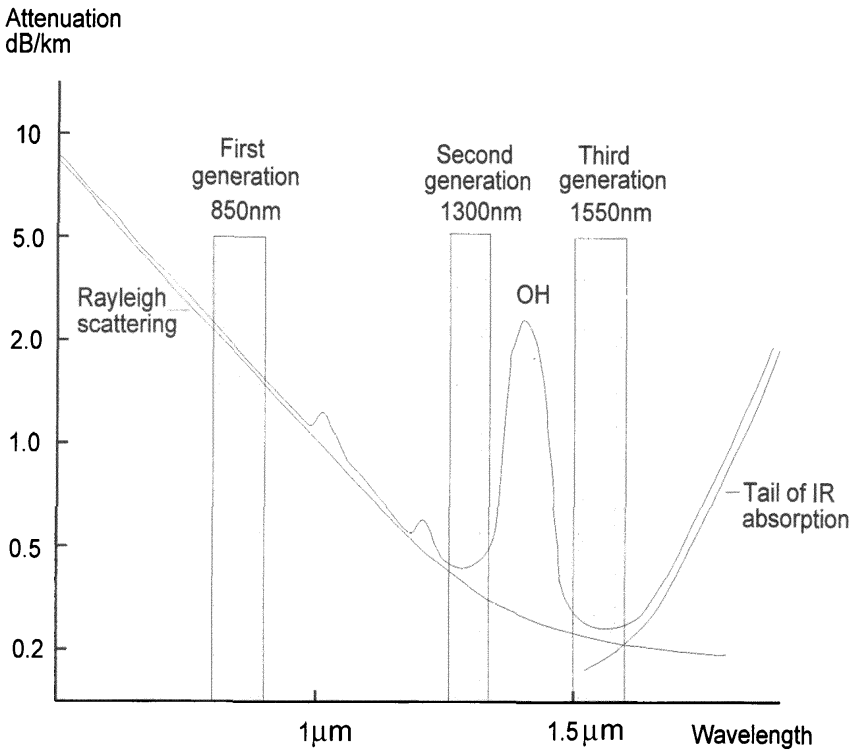


Figure A6.14. Loss characteristic for silica fibres.

excite electrons from the valence to the conduction band of the silica material and the tail of this band extends into the 1 μm region adding a small contribution to the loss. On the long wavelength side, the loss characteristic is dominated by the tail of IR absorption bands from the fundamental molecular vibrations in the 7–12 μm region of bonds such as Si–O, Ge–O, B–O, P–O. Various impurities and dopants also contribute to the loss. Water in the form of the hydroxyl group (OH) bonded into the silica structure gives overtone absorption lines at 0.72, 0.95 and 1.38 μm and combination lines (with silica vibrations) at 0.88, 1.13 and 1.24 μm. Other sources of loss are ionic impurities of transition metals including Cr, Fe, Cu, Ni, Mn and V.

More recently (in 1998), Lucent Technologies developed a new purifying process for virtually eliminating all the OH from the fibre, removing the attenuation peak around 1.38 μm in figure A6.14. This fibre was specifically designed for metropolitan networks and has the trade name *AllWave*^{TM} fibre [6].

Finally it should be noted that in the practical use of fibres, other factors may contribute, to a greater or lesser extent, to the attenuation experienced by guided light in the fibre. If the fibre is bent beyond a certain critical radius, *macrobending* losses become significant as a result of radiation from the evanescent wave tail on the outer curved side of the fibre. *Microbending* losses as a result of mode coupling occur when a fibre is subjected to periodic or repetitive variations in curvature along the fibre axis (microbends) which may happen during manufacture, cable installation or in service.

A6.5.2 Dispersion in optical fibres

In general, dispersion in optical fibres arises from three factors, namely, inter-modal dispersion, chromatic (or intra-modal) dispersion and polarization-mode dispersion. The effect on pulse broadening from the different dispersion factors can be combined according to the relation [11]

$$\Delta t_{\text{tot}} = \sqrt{\Delta t_{\text{i}}^2 + \Delta t_{\text{c}}^2 + \Delta t_{\text{p}}^2}$$

but usually only one (or possibly two) factors are significant, depending on the fibre type and the application.

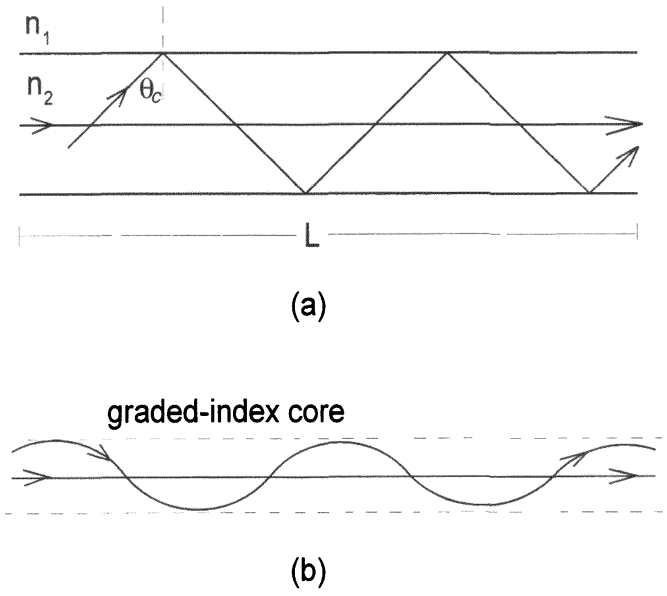


Figure A6.15. (*a*) Fastest and slowest modes in a step-index fibre and (*b*) ray paths in a graded-index fibre.

Intermodal dispersion

The greatest dispersion occurs with the step-index multi-mode fibre where inter-modal dispersion is dominant. Here the pulse power at the fibre input is distributed over the (very) large number of modes supported by the fibre. From a simple ray-optics description, the ray experiences multiple reflections as it travels along the fibre and the path length increases with mode order. Hence, each mode has its own transit time through the fibre link so there is a spread in the arrival times of the pulse power from each mode at the output. The two extreme cases are illustrated in figure A6.15(*a*) which shows the lowest-order mode making a direct transit in the shortest time and the highest-order mode taking the longest time. The pulse broadening can be estimated from the difference in transit time of these two modes (ignoring skew rays). For the fastest (axial) mode the transit time is n_2L/c, whereas for the extreme meridional ray (at the critical angle θ_c) it is $n_2L/(c\sin\theta_c)$. Hence, the inter-modal pulse broadening is:

$$\Delta t_i = \frac{L}{c}\frac{n_2}{n_1}(n_2 - n_1) \approx \frac{nL\Delta}{c} \tag{A6.16}$$

where $\Delta = (n_2 - n_1)/n_1$ is small. For $\Delta \approx 0.01$, then $\Delta t_i \simeq 50$ ns km^{-1}. However, a useful definition is the rms pulse broadening for inter-modal dispersion [11], given by $nL\Delta/(2\sqrt{3}c)$ giving a typical value of $\sim$14 ns km^{-1}.

Inter-modal dispersion can be dramatically reduced by using a graded-index core fibre with a parabolic or near-parabolic index profile as illustrated in figure A6.15(*b*). The transit times for the various rays are equalized by using the fact that off-axis rays experience a lower index in the outer regions of the core where the group velocity is higher, thus compensating for the longer path length. For the best case with optimum index profile, the rms pulse broadening can be reduced by a factor of $\sim\Delta/10$ compared to the step-index case [12], giving a value of $\sim$14 ps km^{-1}. In practice, slight deviations of the profile from its ideal shape due to manufacturing difficulties can greatly increase this value. Note that with graded-index fibre, both inter-modal and chromatic (intra-modal) dispersion may be significant. Although graded-index fibres have improved dispersion properties over the step-index type, they are not commonly used now for telecommunications purposes because of the superior performance of single-mode fibres, as discussed in the following section.

Chromatic dispersion

Single mode fibres are used in all modern high-data-rate communication systems due to their superior performance as a result of no inter-modal dispersion but the effects of chromatic dispersion on pulse broadening must still be considered. Because of the wavelength spread (linewidth) of the source, the pulse is composed of a range of wavelength components which disperse on propagation through the fibre due to their different group velocities. Because of the great importance of single-mode fibres in communication systems, chromatic dispersion is considered here in some detail.

Before examining dispersion in single-mode fibres, consider first the propagation of a pulse through a (dispersive) medium where the refractive index, $n(\lambda_0)$, depends on wavelength. A pulse travels a distance L in a time, $t = L/v_g$ where v_g is the group velocity given by $v_g = \mathrm{d}\omega/\mathrm{d}\beta$ and β is the propagation constant, $\beta = (2\pi/\lambda_0)n = (\omega/c)n$, so

$$v_g = \left(\frac{\mathrm{d}\beta}{\mathrm{d}\omega}\right)^{-1} = \left(\frac{n}{c} + \frac{\omega}{c}\frac{\mathrm{d}n}{\mathrm{d}\omega}\right)^{-1} = \frac{c}{N_g} \tag{A6.17}$$

where the group index N_g is given by

$$N_g = \left\{n + \omega\frac{\mathrm{d}n}{\mathrm{d}\omega}\right\} = \left\{n - \lambda_0\frac{\mathrm{d}n}{\mathrm{d}\lambda_0}\right\}. \tag{A6.18}$$

The transit time (or *group delay*) for the pulse is therefore $t = (L/c)N_g$. If the pulse is generated by a source with a wavelength spread of $\Delta\lambda = (\lambda_2 - \lambda_1)$, then the transit time for the pulse will also have a spread in values because the group index is wavelength dependent. The pulse broadening as a result can thus be approximated by

$$\Delta t = (t_2 - t_1) = \frac{L}{c}\{N_g(\lambda_2) - N_g(\lambda_1)\} \cong \frac{L}{c}\frac{\mathrm{d}N_g}{\mathrm{d}\lambda_0}\Delta\lambda. \tag{A6.19}$$

Using equation (A6.18) for the group index gives the result:

$$\Delta t = -\frac{L}{c}\frac{\Delta\lambda}{\lambda_0}\left[\lambda_0^2\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda_0^2}\right]. \tag{A6.20}$$

This expression shows that the pulse broadening depends on the relative spectral linewidth of the source and the chromatic dispersion coefficient $\lambda_0^2(\mathrm{d}^2 n/\mathrm{d}\lambda_0^2)$. Note that pulse broadening occurs irrespective of the sign of Δt (positive $\Rightarrow$ longer wavelengths have longer transit times, negative $\Rightarrow$ longer wavelengths have shorter times). The dispersion is often expressed in terms of a parameter, D, with units of ps nm^{-1} km^{-1} through the definition:

$$D = -\frac{1}{c\lambda_0}\left[\lambda_0^2\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda_0^2}\right] = -\frac{\lambda_0}{c}\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda_0^2}. \tag{A6.21}$$

This definition means that pulse broadening is simply given by: $\Delta t = DL\Delta\lambda$. Figure A6.16 shows typical values for the material dispersion parameter D_m for conventional silica optical fibres. Note that the coefficient is negative for $\lambda_0 <\sim 1270$ nm and positive for $\lambda_0 >\sim 1270$ nm.

Consider now the case of pulse propagation in a waveguide. The propagation constant of a guided mode is $\beta_g = (2\pi/\lambda_0)n_e$ where the effective index, n_e, is a function of λ_0 and so equation (A6.20) describes pulse broadening where the dispersion coefficient is now $\lambda_0^2(\mathrm{d}^2 n_e/\mathrm{d}\lambda_0^2)$. This dispersion coefficient may, in principle, be derived from the characteristic equation for the modes of a guide (for example equation (A6.4) for planar guides), taking into account the wavelength dependence of the refractive indices.

For this purpose, it is convenient to consider the dispersion in terms of two contributions, namely material dispersion and waveguide dispersion. As already noted, material dispersion arises from the wavelength

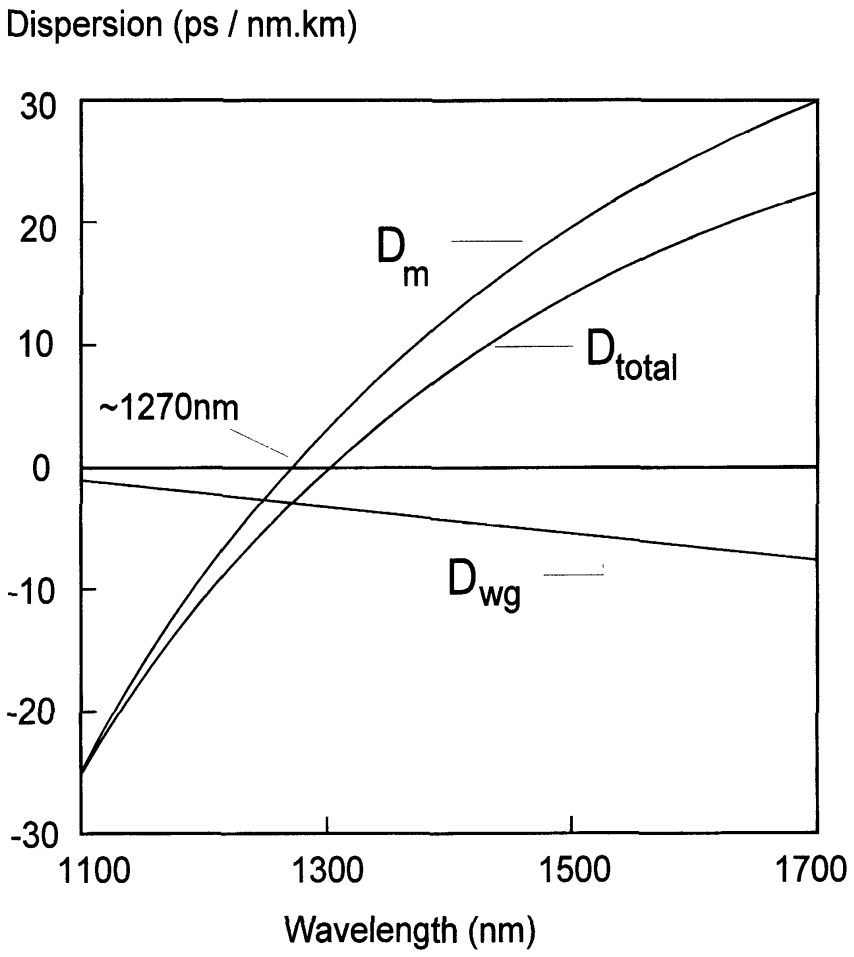


Figure A6.16. Material, waveguide and total dispersion for a conventional single-mode fibre.

dependence of the refractive indices of the glasses making up the guide, whereas waveguide dispersion arises from the nature of guided modes (the b–V curves shown earlier are essentially waveguide dispersion curves).

Consider the most important case, that of dispersion in a single-mode fibre [6, 9, 11]. Assuming a small index difference between core and cladding (as for LP modes), $\Delta = (n_2 - n_1)/n_1 \ll 1$, then from equation (A6.6) the effective index can be written as

$$n_e \cong n(b\Delta + 1)$$

and, hence

$$\frac{dn_e}{d\lambda_0} \cong \frac{dn}{d\lambda_0} + n\Delta\frac{db}{d\lambda_0} \tag{A6.22}$$

where $n = n_1 \approx n_2$ and it has been assumed that the dispersion of n_1 and n_2 are similar so that Δ is not a function of λ_0.

The dispersion parameter can now be expressed as a sum of material and waveguide dispersion effects in the form:

$$D = -\frac{\lambda_0}{c}\frac{d^2n_e}{d\lambda_0^2} \approx -\frac{\lambda_0}{c}\frac{d^2n}{d\lambda_0^2} - n\Delta\frac{\lambda_0}{c}\frac{d^2b}{d\lambda_0^2} = D_m + D_{wg} \tag{A6.23}$$

where D_m is the material dispersion as given earlier and D_{wg} is the waveguide dispersion given by:

$$D_{wg} = -n\Delta\frac{\lambda_0}{c}\frac{d^2b}{d\lambda_0^2} = -\frac{n\Delta}{c\lambda_0}V\frac{d^2(bV)}{dV^2} \tag{A6.24}$$

where the approximation $V \cong k_0 a n\sqrt{2\Delta}$ is used. (Note that for the approximation of equation (A6.23) a term involving $dn/d\lambda_0$ is ignored [11].)

Figure A6.16 shows typical material, waveguide and total dispersion values for conventional single-mode fibre. Note that waveguide dispersion is negative over the range shown. Hence, material and waveguide dispersion are of opposite sign beyond ~1270 nm and the total dispersion is zero around 1300 nm. The point of zero total dispersion can, however, be shifted to longer wavelengths by modifying the waveguide dispersion through the use of special refractive index profiles for the fibre core and cladding [6] to give *dispersion-shifted fibres*. In this way, the point of minimum dispersion can be made to coincide with the minimum loss region around 1550 nm. Alternatively, special profiles can also be used to flatten the dispersion minimum over a wider range to give *dispersion-flattened fibres*.

A problem can arise in wavelength-division-multiplexed (WDM) communication systems using near-zero dispersion fibre. In WDM systems, a number of closely-spaced wavelength channels (grid interval 100 GHz or ~0.8 nm) are sent along the same fibre to multiply its information-carrying capacity. With zero dispersion fibre, nearby wavelength channels have similar group velocities and the consequent phase matching between them enhances cross-phase modulation and four-wave mixing effects [13], giving rise to severe cross-talk between channels. Rather than using zero-dispersion fibre, an alternative solution is to employ *dispersion compensation* whereby the (low) positive dispersion of the transmission fibre is compensated at appropriate intervals by shorter lengths of *dispersion-compensating fibre* which has a large negative dispersion factor. In this way the total net dispersion is zero but there is always dispersion present at any point to inhibit phase matching between channels.

Polarization-mode dispersion

For high bit-rate long-haul systems using fibres operating around the zero-dispersion point or where compensation techniques are employed to minimize chromatic dispersion, *polarization mode dispersion* (PMD) becomes the limiting factor on the maximum bit rate from dispersion. The origin of this dispersion, which is typically of the order of 0.1–1.0 ps $km^{-1/2}$, can be explained as follows [6, 14].

As we noted in section A6.4.2, although a perfectly circular fibre has zero birefringence, in practice factors such as geometric irregularities, stress variations and other perturbations arising during fibre manufacture result in a small birefringence. This is further exacerbated by twists, bends or pinching of the fibre during operation. As a result of the birefringence, the fibre has two principal polarization states and an input pulse is split into the two states which propagate with slightly differing group velocities, giving rise to dispersion. In fact, only a short, straight and undisturbed length of fibre can be described in this way with uniform birefringence and fixed principal axes of polarization. In practice the fibre is more accurately modelled as a chain of birefringent segments 'spliced' at random angles with respect to their principal axes. At the 'splices', mode coupling occurs when the output polarization modes of one segment is decomposed into the polarization states of the next. Environmental factors (such as temperature or wind effects for aerial cables) cause fluctuation both in the birefringence and in the degree of mode coupling. Hence, unlike chromatic dispersion which is relatively stable, the polarization mode delay varies randomly in time and a time average value is used to characterize the PMD (which accounts for the square root dependence on the fibre length). The PMD also varies with wavelength, which means that different channels in a WDM system experience different amounts of pulse spreading.

In describing the PMD at a particular wavelength and time, a pair of *principal states of polarization* (PSP) can be defined at the input and output which correspond to the fast and slow modes of propagation of the fibre link. The polarization transformation between input and output states is independent of wavelength over a small bandwidth (average size referred to as the *PSP bandwidth*). The dispersion performance of a channel is then dependent on the relative intensities launched into these fast and slow modes and is worst when both are equally excited. Various methods are currently under development for mitigation of PMD. One

method uses a polarization controller at the fibre link output and a length of polarization-maintaining fibre to compensate for the differential delay from the PMD. Since the PMD has a time-varying nature, the output impairment must be continually monitored and a feedback loop used to adjust the polarization controller to track the varying PMD.

A6.5.3 Nonlinear effects in optical fibres and solitons

Light pulses propagating through optical fibres also experience several nonlinear effects which can impair performance, especially in WDM communication systems (see section A4.1). The most important are *self-phase modulation* (SPM), *cross-phase modulation* (XPM), *four-wave mixing* (FWM) and *stimulated Raman scattering* (SRS). Briefly, SPM converts power changes in a propagating wave to phase changes in the same wave; XPM converts power fluctuation in one channel of a WDM system to phase fluctuations in the neighbouring channels; FWM generates new frequency components in WDM systems which may coincide with existing channels and cause severe cross-talk; and SRS results in transfer of power from shorter to longer wavelengths through interaction with vibrational modes of the silica molecules [6, 13].

Consider SPM, which arises from the Kerr nonlinearity where the refractive index has a (weak) dependence on the optical intensity, I, according to the relation

$$n = n_0 + n_{\mathrm{nl}} I \tag{A6.25}$$

where n_{nl} is the nonlinear index coefficient, $n_{\mathrm{nl}} \simeq 3 \times 10^{-16}$ cm^2 W^{-1}.

Following from equation (A6.25), the increasing intensity on the rising edge of a pulse will induce a positive $\mathrm{d}n/\mathrm{d}t$, whereas the falling edge will cause a negative $\mathrm{d}n/\mathrm{d}t$. This time variation in index produces a frequency change across the pulse (frequency chirp) with the leading edge reduced in frequency and the trailing edge increased.

If the pulse is travelling in a fibre where the dispersion parameter, as defined earlier, is negative (i.e. the normal dispersion regime where longer wavelengths have a greater group velocity), then the group velocity will be increased for the leading edge and reduced for the trailing edge. The effect is to redistribute the pulse power from the centre to the sides, broadening the pulse and impairing the system performance. However if the pulse is travelling in a fibre where the dispersion parameter is positive (anomalous dispersion regime), the leading edge will be reduced in group velocity and the trailing edge increased. The result is that the pulse will be compressed, which may compensate for the normal dispersion broadening.

The latter case is the situation which leads to the existence of *temporal* soliton pulses, where compression from SPM balances dispersion broadening and the pulse shape is retained without broadening over an indefinitely long transmission path. (For further details see chapter A4: nonlinear optics, section A4.8). In fact, propagation of temporal solitons along an optical fibre occurs in two forms—the fundamental soliton maintains its original shape indefinitely whereas periodic solitons change shape but return periodically to their original shape, with a typical period of 100 km [6,7]. Soliton pulses are typically of 10–50 ps in duration with a few milliwatts of peak power, but only became viable with the advent of the optical fibre amplifier to maintain the required power levels. The unique properties of solitons are of interest for communication systems as return-to-zero (RZ) pulses with the prospect of attaining data rates per channel in excess of 10 Gbit s^{-1} [15]. One of the key problems in the practical application of solitons for very high data rate systems has been controlling timing jitter. Timing jitter arises from random fluctuation in the soliton's central frequency (from the Gordon–Haus [16] effect or from soliton collisions in WDM systems) which is transformed into timing jitter through dispersion. Recently a solution to this problem has been demonstrated using the same methods of dispersion management described previously, namely, alternating sections of fibre with positive and negative dispersion factors [17].

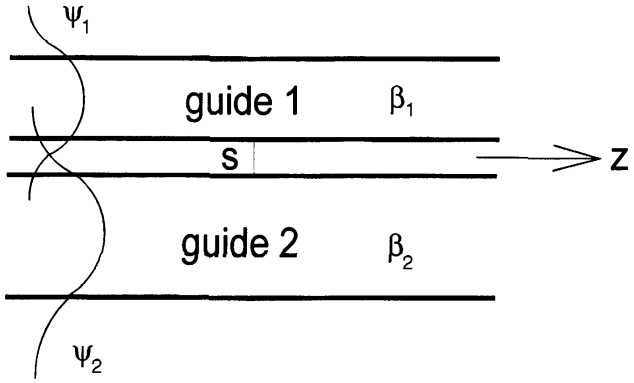


Figure A6.17. Waveguide coupling through the evanescent field of the guided modes.

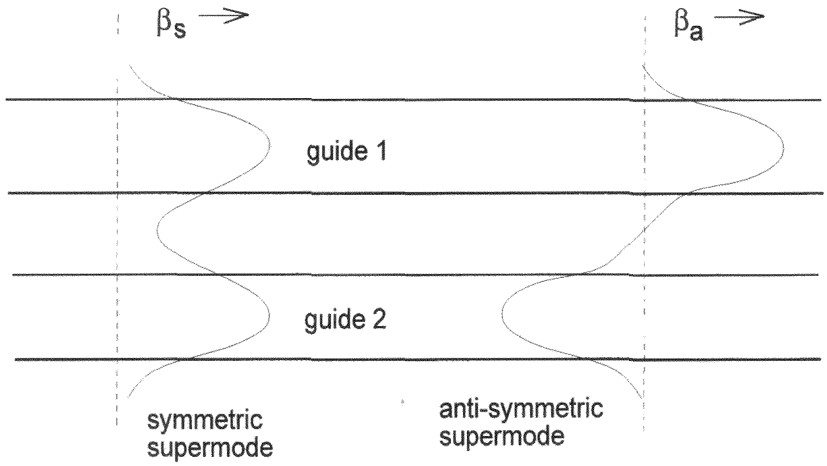


Figure A6.18. Symmetric and anti-symmetric supermodes of a five-layer guiding structure.

A6.6 Mode coupling

Coupling of power between the modes of two optical waveguides is a very important phenomenon and is widely used to make integrated- and fibre-optic components such as power dividers, wavelength selective couplers and modulator devices.

Consider the situation shown in figure A6.17 where two single-mode waveguides (planar, 2D or fibre guides), with mode propagation constants of β_1 and β_2 (in isolation), are placed side by side. If β_1 and β_2 are equal or closely matched and the separation, s, between the waveguides is reduced so that the evanescent field of one guide is able to penetrate the other, power is transferred periodically between the guides over their interaction length. The minimum length required for the transfer of maximum power from one guide to the other is called the *coupling length, L.*

The reason for this behaviour can be understood by examining the properties of the structure as a single, composite (five-layer) waveguide. If Maxwell's wave equations are applied to the whole structure and appropriate field-matching conditions are applied at each of the boundaries, then the allowed modes are described by a characteristic (or eigenvalue) equation for a five-layer system. For the case considered, we would find that the structure supports two modes, referred to as symmetric and anti-symmetric *supermodes*, illustrated in figure A6.18. The propagation constants of these supermodes are: $\beta_s = \beta_{av} + \kappa'$ (slow mode) and $\beta_a = \beta_{av} - \kappa'$ (fast mode), respectively, where $\beta_{av} = (\beta_1 + \beta_2)/2$ and κ' is a constant to be defined later.

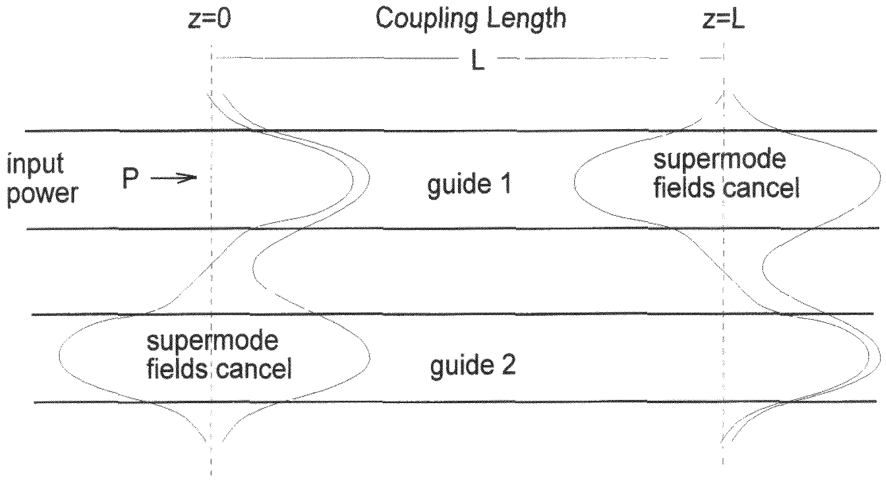


Figure A6.19. *Periodic power transfer between two guides from the addition of the supermode fields.*

Suppose now that at some point $z = 0$, power is launched into guide 1 with no power in guide 2. Since the only way that light can be guided by the structure is in supermode(s), the input field requires both supermodes to be excited with phases such that their field summation corresponds to power in guide 1 and zero power in guide 2, as illustrated in figure A6.19. As the supermodes propagate along the z-axis, their phase difference, $(\beta_s z - \beta_a z)$, increases with z due to their different propagation constants. After a distance L where $(\beta_s - \beta_a)L = \pi$ and the phase difference has reached a value of π, the summation of the fields gives maximum power in guide 2 and minimum in guide 1. Using the values given for the propagation constants, β_s and β_a, the coupling length is, therefore

$$L = \frac{\pi}{2\kappa'}. \tag{A6.26}$$

Note that, after another coupling length, the phase difference becomes 2π and so maximum power is back in guide 1. The power is thus cyclically transferred between the guides over their interaction length.

A very useful approximate technique for the theoretical analysis of mode coupling is based on the concept of *weak coupling* which leads to the *coupled mode equations* [3, 18–20]. Returning to figure A6.17, if the coupling between the guides is weak, the (transverse) field distribution of each guide in isolation will only be slightly perturbed by the presence of the other guide. Hence, the field distribution, ψ, of the whole structure can be approximated by

$$\psi(x, y, z) \approx A_1(z)\psi_1(x, y)\exp(-i\beta_1 z) + A_2(z)\psi_2(x, y)\exp(-i\beta_2 z) \tag{A6.27}$$

where ψ_1 and ψ_2 are the normalized transverse field distributions of guides 1 and 2 in isolation and $A_1(z)$ and $A_2(z)$ indicate that the field amplitudes vary with distance z along the guides due to the coupling.

Under these assumptions, ψ, ψ_1 and ψ_2 must all satisfy the wave equation since they all represent solutions of a guiding structure. When these functions are substituted into the wave equation (with some approximations which are valid for weak coupling), the following two relationships are obtained [18]:

$$\begin{aligned} \frac{dA_1}{dz} &= -i\kappa_{12} \cdot A_2 \exp(-i\Delta\beta z) \\ \frac{dA_2}{dz} &= -i\kappa_{21} \cdot A_1 \exp(+i\Delta\beta z) \end{aligned} \tag{A6.28}$$

where $\Delta\beta = (\beta_2 - \beta_1)$ and $\kappa = \sqrt{\kappa_{12}\kappa_{21}}$ is the *coupling coefficient* which depends on the overlap between the fields ψ_1, ψ_2 of the two guides.

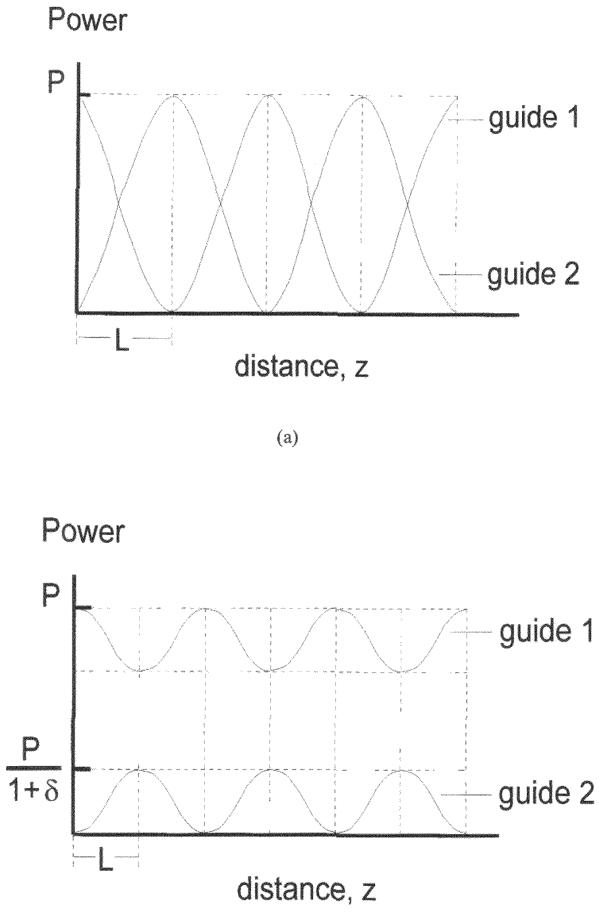


Figure A6.20. Power in each guide for a directional coupler: (*a*) phase-matched case and (*b*) mismatched case.

Equations (A6.28) are known as the *coupled mode equations* and show that variations in the amplitude in one guide are linked to the amplitude in the other guide through the coupling coefficient. Note that if $\kappa = 0$ (i.e. no interaction between the guides), then the amplitudes in each guide, A_1 and A_2, remain constant along the z-direction, as expected.

For the particular case illustrated in figure A6.19 where at $z = 0$ power, P, is launched into guide 1 with no power in guide 2, the solution of the coupled mode equations gives the power in each guide as a function of z:

$$P_1(z) = P\left[1 - \frac{1}{1+\delta}\sin^2 \kappa' z\right]$$
$$P_2(z) = P\left[\frac{1}{1+\delta}\sin^2 \kappa' z\right] \tag{A6.29}$$

where $\kappa' = \kappa\sqrt{(1+\delta)}$ and $\delta = \left(\frac{\Delta\beta}{2\kappa}\right)^2$. Equations (A6.29) reveal the dependence of the coupling length and the maximum power transferred to guide 2 on the degree of mismatch, δ, between the guides. Note that the power fraction transferred is $1/(1+\delta)$ at a coupling length of $\pi/2\kappa'$. For efficient power transfer in a directional coupler, δ must be small, i.e. $\Delta\beta \ll 2\kappa$. Figure A6.20 shows examples of two cases: (*a*) phase-matched case where $\beta_1 = \beta_2$ and $\delta = 0$; and (*b*) mismatched case where $\beta_1 \neq \beta_2$.

The coupling coefficient, κ, is determined from the overlap of the fields of the individual guides [19] and may be calculated either analytically or numerically if the field distributions of the individual guides are

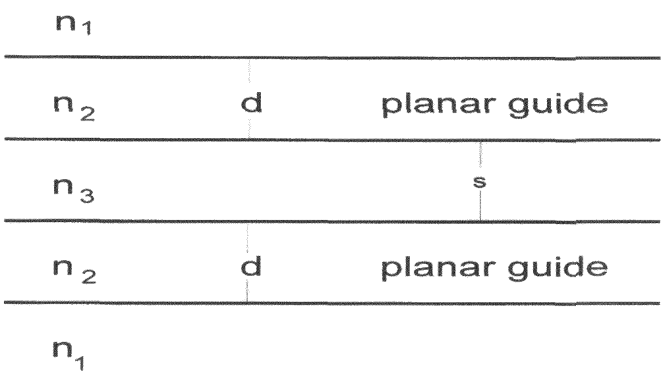


Figure A6.21. Coupling between two single-mode planar waveguides.

known [21]. For example, using the field distributions given in equation (6.9) the coupling coefficient for the TE_0 mode of the two planar guides illustrated in figure A6.21 is:

$$\kappa = \frac{2\gamma_3 \exp(-\gamma_3 s)}{\beta_0 d_e \left[1 + \left(\frac{\gamma_3}{k_x}\right)^2\right]} \tag{A6.30}$$

where:

$$\gamma_3 = k_0\sqrt{n_e^2 - n_3^2} \qquad k_x = k_0\sqrt{n_2^2 - n_e^2} \qquad \text{and} \qquad d_e = d + \frac{1}{\gamma_1} + \frac{1}{\gamma_3}.$$

These principles underpin the operation of a variety of integrated and fibre optic devices based on directional coupling. For example, by designing the guides with an appropriate interaction length in relation to the coupling length, fibre or integrated optic couplers with different power-splitting ratios may be manufactured. Also, since the coupling is stronger at longer wavelengths, as seen by the $\exp(-\gamma_3 s)$ term in equation (A6.30), wavelength splitters can be designed by choosing an interaction length which corresponds to the coupling length (or a multiple of it) for λ_1 but not for λ_2, so that maximum cross-coupling occurs at λ_1 and minimum at λ_2. Active devices are made by using directional couplers in electro-optic materials such as lithium niobate. The index change from the applied voltage is used to match or mismatch the propagation constants of the two guides, thus forming a modulator or switch.

This discussion has dealt with coupling between modes which are matched or nearly matched in propagation constants. It is also possible to couple modes which have substantially different propagation constants if periodic coupling is introduced [22] by, for example, a diffraction grating. This can be seen from equation (A6.28) by introducing a periodic coupling coefficient of the form: $\kappa_{21}\kappa_{12}^* = \kappa \exp(-ik_c z)$. Combining this with the exponential terms in equation (A6.28) gives the terms: $\exp \pm i(\Delta\beta - k_c)$ so that 'matching' now occurs when $(\Delta\beta - k_c) = 0$ where $k_c = (2\pi/\Lambda)$ and Λ is the periodicity of the perturbation. This phenomenon may cause unwanted coupling between the modes within a waveguide or may be exploited in certain types of fiber optic sensors and narrow-linewidth fibre Bragg gratings for use in optical transmission networks [23, 24].

A6.7 Conclusion

This chapter has outlined the fundamental concepts and principles involved in guiding light in various types of structures. Knowledge of these principles and the various parameters that have been defined is essential in the analysis and design of optical waveguides and waveguide-based components. Further detailed information on fibre optic systems, waveguide components and their applications in fibre- and integrated-optics may be found in the list of further reading.

References

[1] Lee D L 1986 *Electromagnetic Principles of Integrated Optics* (New York: Wiley) pp 116–35

[2] Stewart G, Millar C A, Laybourn P J R, Wilkinson C D W and De La Rue R M 1977 Planar optical waveguides formed by silver-ion migration in glass *IEEE J. Quantum Electron.* **QE-13** 192–200

[3] Syms R and Cozens J 1992 *Optical Guided Waves and Devices* (London: McGraw-Hill) pp 217–49 and 253–6

[4] Marcatili E A J 1969 Dielectric rectangular waveguide and directional coupler for integrated optics *Bell Syst. Tech. J.* **48** 2071–102

[5] Hocker G B and Burns W K 1977 Mode dispersion in diffused channel waveguides by the effective index method *Appl. Opt.* **16** 113–18

[6] Keiser G 2000 *Optical Fiber Communications* 3rd edn (New York: McGraw-Hill) pp 10, 43–56, 92–103, 109–15, 123–5, 496–500, 505–13

[7] Belanger P 1993 *Optical Fiber Theory* (Singapore: World Scientific) pp 100–21, 52–7

[8] Cherin A H 1987 *An Introduction to Optical Fibers* (Singapore: McGraw-Hill) pp 85–100

[9] Gloge D 1971 Weakly guiding fibers *Appl. Opt.* **10** 2252–8

[10] Kaminow I P 1981 Polarization in optical fibers *IEEE J. Quantum Electron.* **QE-17** 15–22

[11] Senior J M 1992 *Optical Fiber Communications Principles and Practice* (New York: Prentice-Hall) pp 88–94, 113, 121–30, 905–6

[12] Olshansky R and Keck D B 1976 Pulse broadening in graded index optical fibres *Appl. Opt.* **15** 483–91

[13] Willner A E 1997 Mining the optical bandwidth for a terabit per second *IEEE Spectrum* **34** 32–41

[14] Hernday P 2001 PMD posts a speed limit for high speed fiber networks *Laser Focus World* **37** 171–8

[15] Georges T and Faul J 2000 Soliton transport gives backbones more speed *Fibre Systems* **4** 51–4

[16] Gordon J P and Haus H A 1986 Random walk of coherently amplified solitons in optical fiber transmission *Opt. Lett.* **11** 665–7

[17] Forysiak W, Nijhof J H B and Doran N J 2000 Dispersion managed solitons: the key to terabit per second optical fiber communication systems *Opt. Photon. News* **11** 35–9

[18] Yariv A 1985 *Optical Electronics* 3rd edn (New York: Holt-Saunders) pp 413–49

[19] Ghatak A K and Thyagarajan K 1989 *Optical Electronics* (Cambridge: Cambridge University Press) pp 447–54, 609–12

[20] Marcuse D 1971 The coupling of degenerate modes in two parallel dielectric waveguides *Bell Syst. Tech. J.* **50** 1791–816

[21] Marcuse D 1987 Directional couplers made of non-identical asymmetric slabs *IEEE J. Lightwave Technol.* **5** 113–18

[22] Miller S E 1969 Some theory and applications of periodically coupled waves *Bell Syst. Tech. J.* **48** 2189–219

[23] Othonos A and Kalli K 1999 *Fibre Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing* (London: Artech House)

[24] Kashyap R 1999 *Fibre Bragg Gratings* (San Diego, CA: Academic)

Further reading

Mynbaer D K and Scheiner L L 2001 *Fibre-optic Communications Technology* (New Jersey: Prentice-Hall)

Provides a useful starter text for learning about fibre optic communications technology. Describes the operation and characteristics of key fibre components with specific, commercial examples. The book is also suitable for technician training in fibre optic systems.

Keiser G 2000 *Optical Fiber Communications* 3rd edn (New York: McGraw-Hill)

Gives a comprehensive account of all aspects of the design and practice of modern fibre communications systems and networks, including a readable account of fibre theory, signal degradation, non-linear effects and measurements in optical fibres.

Senior J M 1992 *Optical Fiber Communications Principles and Practice* (New York: Prentice-Hall)

Similar in level to Keiser but due for an update.

Belanger P 1993 *Optical Fiber Theory* (Singapore: World Scientific)

Provides a more advanced and detailed description of the theory of optical fibres from the electromagnetic approach, including the effects of dispersion and nonlinearity on pulse propagation. Theory is well illustrated by graphical plots and a number of exercises are included.

Cherin A H 1987 *An Introduction to Optical Fibers* (Singapore: McGraw-Hill)

Somewhat dated but provides a good account of optical fiber theory using the electromagnetic approach along with some practical aspects of optical fibre technology.

Zappe H P 1995 *Introduction to Semiconductor Integrated Optics* (Norwood, MA: Artech House)

Gives a comprehensive account of semiconductor properties and fabrication technology as applied to integrated optics and describes the construction and operation of a number of important optical components in semiconductor materials including waveguides, modulators, lasers and detectors.

Lee D L 1986 *Electromagnetic Principles of Integrated Optics* (New York: Wiley)

As the title suggests, this book presents the electromagnetic theory of planar and rectangular guides and mode coupling, but also includes optical fibers. Examples of integrated optics devices are given as well as the basic fabrication techniques.

Snyder A W and Love J D 1983 *Optical Waveguide Theory* (London: Chapman and Hall)

A well-known and classic textbook in the area. Part I of the book gives a full treatment of ray optics in multimode fibres which is not so relevant for modern optical communications systems, but Parts II and III provide a comprehensive account of the electromagnetic analysis of optical waveguides, including analysis of bends and perturbations in waveguides and mode coupling between guides.

Syms R and Cozens J 1992 *Optical Guided Waves and Devices* (London: McGraw-Hill)

Provides a fairly descriptive account of optoelectronic devices including both fibre- and integrated-optic components as well as semiconductor devices with illustrations and applications. The background theory for understanding waveguide and component operation is also presented.

Hunsperger R G 1995 *Integrated Optics: Theory and Technology* 4th edn (Berlin: Springer)

Presents the basic theory of waveguides and couplers but concentrates on components and technology and describes in detail the construction and operation of a number of devices including couplers, modulators and lasers.

A7
Optical detection and noise

Gerald Buller and Jason Smith

A7.1 Introduction

The reliable detection of optical radiation is an essential element of laser technology. Indeed, almost everywhere a laser is used, a photodetector will also be employed either to sample the laser output itself or to detect radiation produced by some other system as a result of laser excitation.

By far the most common method by which the intensity of laser light is measured is by conversion to a real-time electrical signal. There are three processes commonly used in photodetection, by which this conversion takes place:

(i) photoelectric emission,

(ii) internal photoelectric (photoconductive, photocurrent) and

(iii) photothermal/thermoelectric conversion.

Each of these processes will be discussed in detail in the following sections. For ease of recognition, the section on internal photoelectric effect detectors is entitled 'semiconductor detectors'; the devices described therein rely explicitly on excitation of electrons either across the semiconductor bandgap or between levels in a quantum well heterostructure. It should be noted that semiconductors are also used both in photoemission and photothermal detection.

Each process has a range of optical wavelengths in which it is suitable for photodetection, as depicted in figure A7.1.1. Broadly speaking, photoemissive detectors are best suited to wavelengths from the UV to the near infrared. In the UV and visible, photoemissive and internal photoelectric detectors are comparable in performance and selection will depend on the particular application. Beyond about 0.9 μm, internal photoelectric detectors gain the upper hand. Photothermal detectors, whilst sensitive throughout the visible, currently only offer superior detectivity at wavelengths longer than about 20 μm. With semiconductor quantum well infrared photodetector (QWIP) technology advancing at a rapid pace, it is likely that this crossover wavelength will grow ever longer in the coming years.

Photodetection in all its manifestations and detail is a vast subject and here we intend only to provide a brief introduction to the science, an overview of the existing technology and some basic theory aimed at the user (rather than the designer) of photodetection systems. For this reason quite a lot of space is dedicated to the subject of noise—an area in which a clear understanding by the user can lead to real improvements in performance.

Most standard optoelectronics textbooks cover the subject of photodetection to some degree. The further reading list at the end of this chapter provides a selection of such references, along with some more specialist texts.

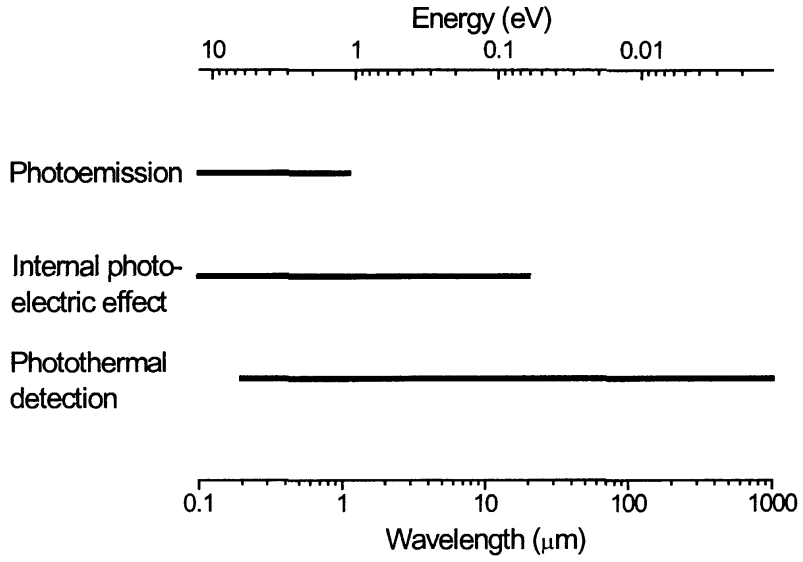


Figure A7.1. Sensitive spectral ranges of the three principal photodetection processes.

A7.1.1 Nomenclature and figures of merit

In order to discuss the relative performance of different types of photodetector in sections A7.2 through A7.4, it will be helpful first to provide definitions of some widely used figures of merit. A full discussion of the sources of noise and their relevance to the different detector types is left until section A7.5.

A7.1.1.1 Signal-to-noise ratio (SNR)

The output from a real-time photodetector subject to an optical signal of constant mean intensity can be represented as a time-dependent current i with a mean value $\langle i \rangle$ and a variance $\langle \delta i^2 \rangle = \langle i^2 \rangle - \langle i \rangle^2$. It is the variance that is most commonly used as the quantitative assessment of the absolute noise level. The often-quoted root-mean-square (rms) noise current, i_{rms}, is equal to the square root of the current variance as defined above.

The total output current, i_{tot}, is the sum of the current generated by the optical signal i_{sig}, and any background current, i_{bkgnd}, caused by non-signal light or leakage currents in the detector. These currents both contribute to the noise in the total current:

$$\langle \delta i_{\mathrm{tot}}^2 \rangle = \langle \delta (i_{\mathrm{sig}} + i_{\mathrm{bkgnd}})^2 \rangle. \tag{A7.1}$$

The simplest figure of merit for the quality of the photodetection signal is the *signal-to-noise ratio* (SNR) defined as the ratio of the square of the mean signal current to the variance in the total current,

$$\mathrm{SNR} = \frac{\langle i_{\mathrm{sig}} \rangle^2}{\langle \delta i_{\mathrm{tot}}^2 \rangle}. \tag{A7.2}$$

A7.1.1.2 Noise-equivalent power (NEP)

The sensitivity of a photodetector can be defined in terms of the incident optical power required to achieve a given SNR. A figure of merit commonly used is the *noise-equivalent* power (NEP), defined as the mean optical power required to be incident on the detector to generate an SNR of unity. NEP in its most simple form, therefore, has units of watts and is specific to a particular detector, taking into account such parameters as the sensitive detection area A and the detection bandwidth Δf.

A7.1.1.3 Detectivity (D) and specific detectivity (D)*

The *detectivity D* is the reciprocal of the NEP as previously defined. It is, therefore, specific to a particular detector operating with a particular bandwidth, and has units of W^{-1}.

In order to generate a figure of merit for detector sensitivity that is independent of detection bandwidth, it is often assumed that the detector noise is independent of frequency or is *white noise* (cf white light). To this end, the NEP is often quoted as having units of W $Hz^{-1/2}$.

The *specific* or *normalized detectivity (D*)* is defined by assuming also that the background signal is proportional to the device area, so that $D = D^*/\sqrt{A\Delta f}$. Specific detectivity is, therefore, usually quoted in units of cm $Hz^{1/2}$ W^{-1}.

We shall see in the following sections that the assumption that detector noise is white is valid in many cases but that there are also common situations in which non-white noise sources must be taken into account. In all situations, however, the bandwidth is an important consideration when determining noise characteristics.

A7.1.1.4 Responsivity

The responsivity of a detector indicates the change in the output of the detector per unit change in the incident optical power and is, therefore, unrelated to noise considerations. In a photodiode, for instance, the output is a photocurrent and so the responsivity will be quoted in A W^{-1}. In a pyroelectric detector, the output is a voltage and units of V W^{-1} are appropriate.

A7.2 Photoemissive detectors

A7.2.1 The photoemissive effect

When photons of sufficient energy are incident on a solid, electrons are emitted from the surface. This phenomenon is known as the photoemissive or photoelectric effect. Figure A7.2 represents this effect schematically in terms of energy levels for (*a*) a metal, (*b*) a semiconductor with positive electron affinity and (*c*) a semiconductor with negative electron affinity. We shall discuss each of these briefly in turn.

In metals, the Fermi energy E_F is situated within an energy band of allowed electron states, and so electrons occupy a quasi-continuum of states up to that energy[1] The minimum energy that a photon must provide in order to eject an electron into the vacuum is

$$h\nu_{min} = e\phi \tag{A7.3}$$

where $e\phi = E_{vacuum} - E_F$ is known as the *work function* of the material.

Since the majority of electrons exist well below the Fermi energy, and since energy can be lost by inelastic scattering of the excited electron prior to emission, a photon energy higher than $h\nu_{min}$ will usually be required.

The probability that an incident photon will cause an electron to be emitted (known as the quantum efficiency, or quantum yield, and represented by η) is, therefore, a strong function of the photon energy, falling off rapidly for $h\nu < e\phi$. The function $\eta(h\nu, T)$ is the most important consideration when selecting a photoemissive material for a detector. In an intrinsic semiconductor, depicted in figure A7.2(*b*), the highest energy electrons exist not at the Fermi energy, which lies within the bandgap but at the valence band edge. Instead of being given by the work function, the minimum energy that a photon must provide to emit an electron from a semiconductor is given by

$$h\nu_{min} = E_g + \chi \tag{A7.4}$$

[1] This is only strictly true at low temperatures. The energy distribution of electron state occupancy at finite temperatures is given by the Fermi–Dirac function $p(E, T) = 1/[\exp[(E - E_F)/kT] + 1]$.

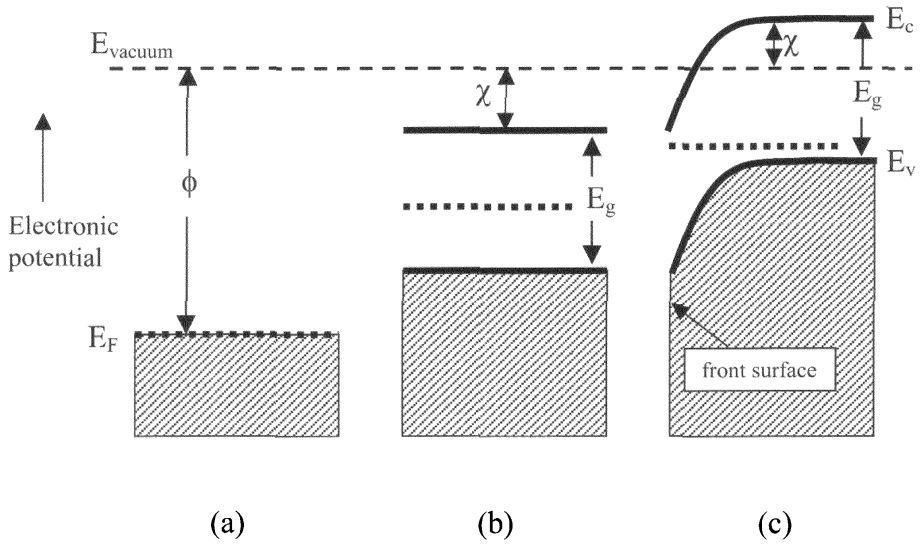


Figure A7.2. Work functions, ϕ, and electron affinities, χ, for (*a*) a metal and for a semiconductor with (*b*) positive and (*c*) negative electron affinity. The semiconductor bandgap is labelled E_g.

Table A7.1. Common photodiode materials with international S codes and sensitive spectral ranges (see also figure A7.3).

Photocathode material	Code	Wavelength range(μm)
Cs_3Sb	S-11	0.3–0.6
[Cs]Na_2K Sb	S-20	0.3–0.8
Ag–O–Cs	S-1	0.25–1.2
GaAs(Cs_2O)		0.2–0.9
$In_{0.18}Ga_{0.82}As(Cs_2O)$		0.2–1.1
$In_{0.52}Ga_{0.48}As$/InP		0.3–1.7

where E_g is the bandgap energy and $\chi = E_{vacuum} - E_c$ is the *electron affinity*. χ is positive for intrinsic semiconductors. Direct-gap semiconductors such as GaAs and $In_xGa_{1-x}As$ are particularly attractive as photoemitters since their higher optical absorption coefficients allow greater absorption near to the surface and so higher photoemission efficiencies.

Negative electron affinity (NEA) photocathodes are created by applying a layer of n-type impurities (commonly Cs or Cs_2O) to the surface of a heavily doped p-type semiconductor. Charge redistribution leads to bending of the electronic bands near to the surface, so that the conduction band minimum in the bulk material lies above the vacuum energy, as shown in figure A7.2(*c*). If the region of band bending is thinner than the electron mean free path through the crystal, efficient photoemission can, therefore, occur for any photon energy above $h\nu = E_g$. Cs_2O provides the largest shift in bulk band energies but also forms a potential barrier at the surface through which the electrons must tunnel in order to be emitted. This barrier limits the quantum efficiency in narrow-gap NEA semiconductors. Note that NEA semiconductors work only as reflective and not as transmissive photocathodes.

Some common photocathode materials are listed in table A7.2.1 and their responsivity spectra are shown in figure A7.3. The internationally agreed S code identifying a particular photocathode design is also included in the table.

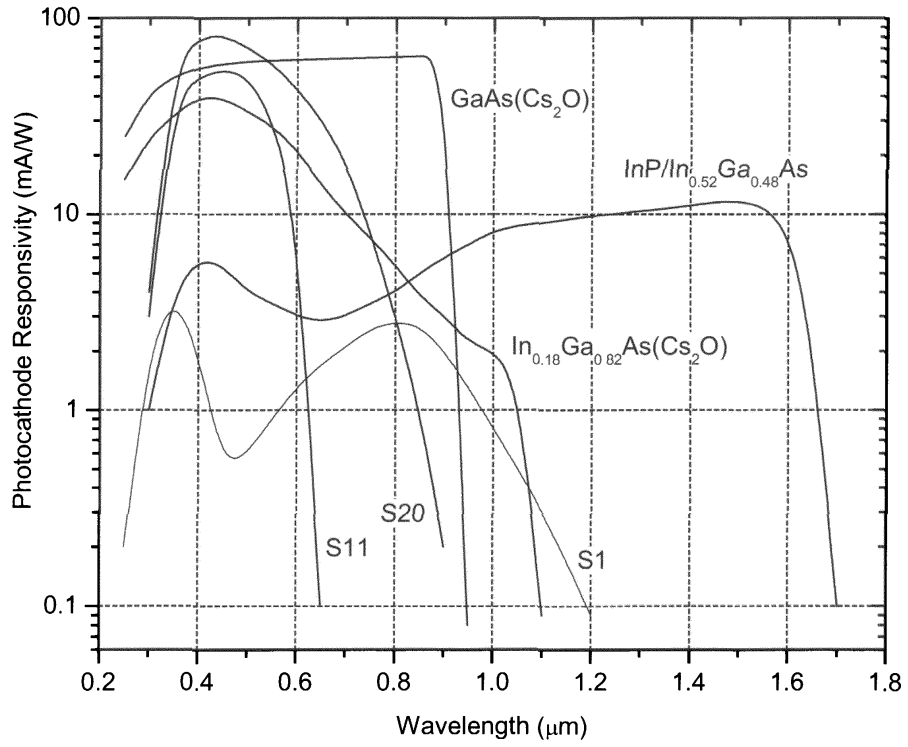


Figure A7.3. Responsivity spectra for some common photocathodes (see also table A7.1).

A7.2.2 Photomultipliers

To create a simple photodetection circuit, the photoemissive material is negatively biased (and called a photocathode) and emitted electrons are accelerated towards and collected by a positively biased electrode (anode) whereupon a measurable current is generated. In order to improve detectivity from this basic principle, a multiplication process is used, the result being known as a photomultiplier. A generic diagram of a photomultiplier tube (PMT) is shown in figure A7.4.

In a PMT, a series of electrodes, known as dynodes, are held at progressively higher potentials under high vacuum conditions. Electrons emitted by the photocathode are accelerated towards the first dynode by a large potential difference (typically >100 V). Each 'primary' electron gains sufficient kinetic energy to eject a number of 'secondary' electrons from the first dynode. These secondary electrons are then accelerated towards the second dynode where the gain process is repeated. A typical PMT contains eight to twelve dynodes, and offers current gains of 50–70 dB.

Dynode materials are chosen for their low work functions to enable efficient secondary emission. To this end, many of the materials that are used for photocathodes are also suitable for dynodes and, in particular, CsSb is commonly used.

The design of the PMT and, in particular, the dynode configuration is crucial in ensuring the efficient transfer of electrons between dynodes and in optimizing the time response of the device by minimizing the range of electron pathlengths—the transit time spread—during the multiplication process. Figure A7.5 shows three popular designs. In each case a semi-transparent photocathode is employed and radiation is incident from the left of the diagram. Design (*a*) has dynodes in a 'venetian blind' configuration, with each dynode consisting of a number of small plates facing at 45° to the tube axis. Alternate dynodes face in opposite directions to maximize collection efficiency. This design tends to have high stability in gain but poor time response. Design (*b*), the 'box and grid' design, has very high efficiency of transfer between dynodes, whilst in design (*c*) each dynode is curved so as to focus the secondary electrons onto the centre of the subsequent

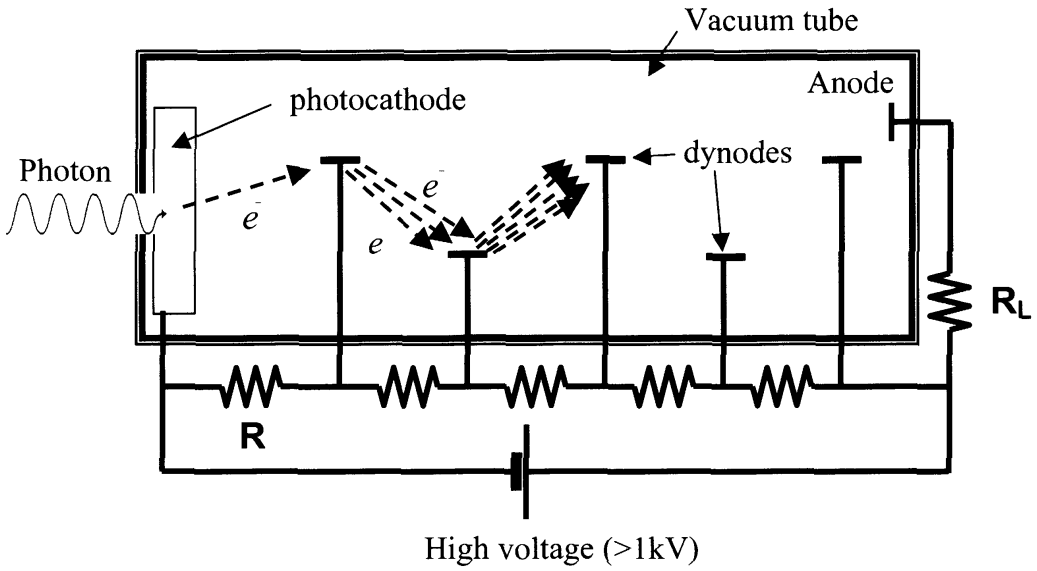


Figure A7.4. Schematic diagram of a photomultipier. Electrons emitted from the photocathode are accelerated in turn to a series of dynodes at progressively higher potential. At each dynode a single incident electron causes emission of several secondary electrons and so the electron current increases exponentially through the device. The signal voltage is measured across load resistor R_L.

dynode, resulting in lessened transit time spreading and, therefore, achieving a fast time response. 'Focused dynode chain' PMTs with risetimes of a few nanoseconds are readily available.

The high gains achieved using photomultipliers make them suitable for single photon detection. In this configuration, the anode current is monitored and a 'count' is recorded whenever it exceeds a specified threshold value.

An adaptation of the standard photomultiplier design that achieves faster response times and allows imaging is the multi-channel plate (MCP). A MCP is a thin disc consisting of many thin glass channels ($\sim$10 μm in diameter) fused in parallel in a 2D array. The interior of each capillary is coated with a photoemissive material and biased at each end, thus acting as a continuous dynode. Each channel operates as an independent electron multiplier, hence reducing the effects of transit time spreading and improving the rise time of the detector to <200 ps. A typical MCP contains 10^6–10^8 channels. Imaging is facilitated in some designs by the inclusion of independent anodes for the array of capillaries.

With no light incident, thermionic emission at the photocathode will give rise to a 'dark current'. For a photocathode of area A at temperature T in a photomultiplier of gain G, the dark current is given by the expression

$$i_T = G\alpha AT^2 \exp\left(-\frac{h\nu_{\min}}{kT}\right) \tag{A7.5}$$

where α is a constant (for pure metals, $\alpha = 1.2 \times 10^{-6}$ A m^{-2} K^{-1}) and k is Boltzmann's constant. It is clear that, for minimized dark current, low temperature operation is necessary—especially for low work function materials employed for detection at longer optical wavelengths. For example, the PMT that incorporates the InGaAs/InP photocathode in figure A7.2 is operated at −80 °C. 'Shot' and 'gain' noise in the dark current (see section A7.5.2) and Johnson noise in the anode (section A7.5.3) are the key factors which limit the detectivity of photomultipliers.

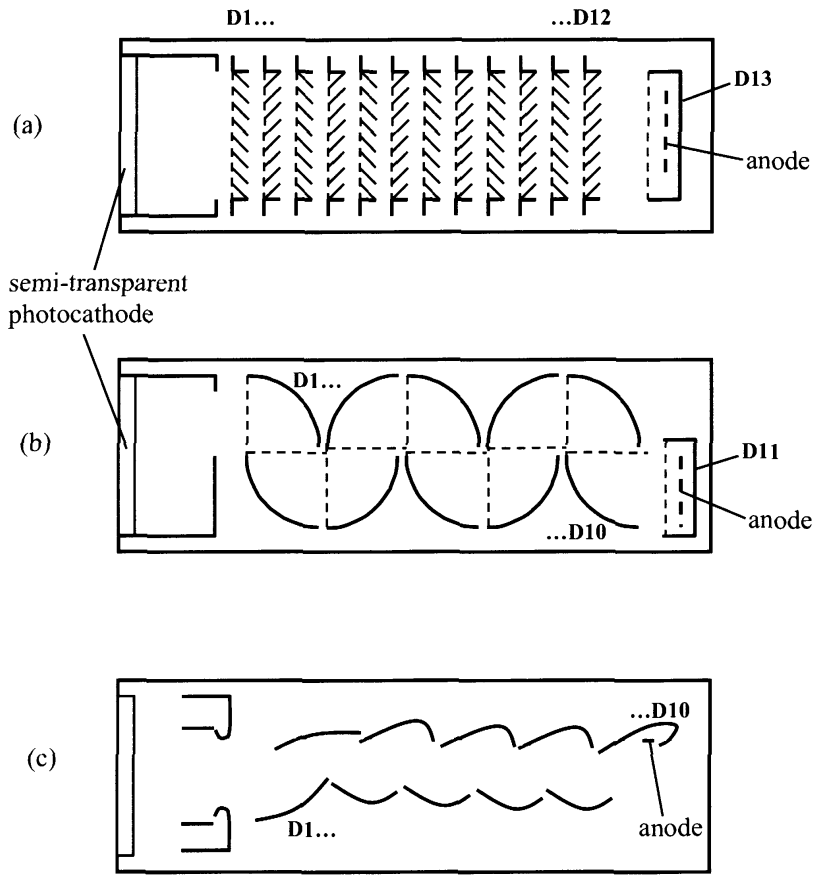


Figure A7.5. Three common photomultiplier dynode configurations: (*a*) venetian blind, (*b*) box and grid and (*c*) focused dynode chain.

A7.3 Semiconductor detectors

A7.3.1 Photoelectric absorption

Absorption of a photon by an electron in a semiconductor can radically alter its conductive properties, and enable photodetection. Assessing the relative suitability of different semiconductors for photodetection (or any other optoelectronic application) relies on a thorough understanding of the electronic band structure of the materials. The reader is referred to the bibliography for introductory texts.

The most commonly exploited process is that in which an electron is excited from the valence band, where it is bound to an atom, to the conduction band, where it is free to move around the crystal and act as a conducting 'carrier'. The 'hole' left behind in the valence band also acts as a carrier, and we speak of an 'electron–hole pair' being created upon absorption of a photon. This process is depicted in figure A7.6(*a*). In a photoconductor, the resulting change in conductivity is measured, whilst in a photodiode, the excited carriers are accelerated by an electric field to generate a photocurrent. These methods are effective for optical wavelengths from the UV to the mid-infrared, the long wavelength limit of sensitivity being determined by the band gap E_g of the absorptive material. Several semiconductors commonly used for interband absorption in photodetectors, and their electronic band gaps, are listed in table A7.2. Optimum detectivity will often be achieved by selecting a material with a bandgap only slightly smaller than the photon energy to be detected, since the narrower-gap materials suffer much higher background signals due to thermal carrier generation.

Intraband absorption—in which a carrier electron (hole) is excited to a higher energy state within the conduction (valence) band—can also be exploited for photodetection. In quantum well infrared photodetectors (QWIPs), carriers confined in quantum wells must be excited either into a higher quantum well sub-band or into

Table A7.2. Some semiconductor photodiode materials and their electronic bandgaps at 300 K. Both the bandgap energy and the corresponding optical wavelength are shown.

	Bandgap at 300 K	
	(eV)	(μm)
Si	1.14[a]	1.09
Ge	0.67[a]	1.86
GaAs	1.43	0.87
InP	1.35	0.92
$In_{0.53}Ga_{0.47}As$[b]	0.75	1.66
InAs	0.35	3.56
InSb	0.18	6.93
$Hg_{1-x}Cd_xTe$	$0 < E_g < 1.44$	$0.86 < \lambda < \infty$

[a] Indirect bandgap.
[b] Lattice matched to InP.

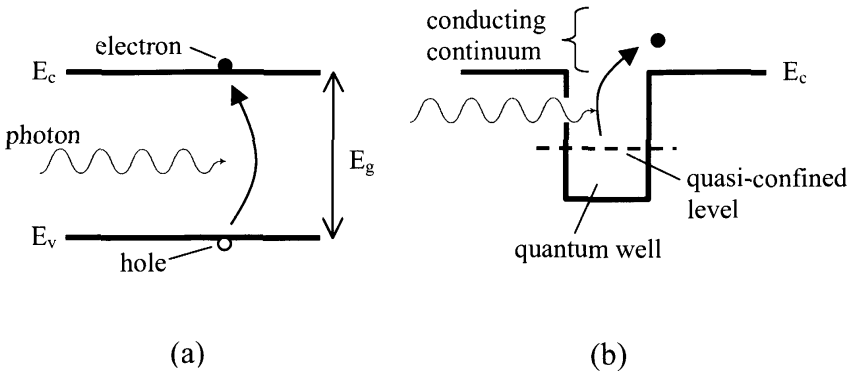


Figure A7.6. (*a*) Inter-band absorption of a photon to create an electron–hole pair and (*b*) intra-band absorption from a quasi-confined state into a conducting continuum. The conduction and valence band edges are labelled E_c and E_v respectively.

a continuum state above the potential barrier, in order to contribute to the electrical current (figure A7.6(*b*)). They can be designed to detect light at any wavelength longer than about 1 μm, whilst avoiding the need to use narrow-gap materials for many of which the processing technology is relatively immature. QWIPs will be discussed further in section A7.3.6.

A7.3.2 pn and pin photodiodes

The simplest design for a photodiode is a pn junction, an example of which is shown in figure A7.7(*a*). Here the n-type surface has been coated with an opaque metal contact, while the p-type surface has an annular ohmic contact with a transparent dielectric window—usually coated to reduce surface reflection—through which the light may enter. The device is shown under a reverse bias V_A.

The potential profile of the conduction and valence band edges and of the electric field ξ through the device are shown in figures A7.7(*b*) and (*c*) respectively. The region of the device that is under electric field (i.e. $dV/dz \neq 0$) is known as the depletion region and within it, electrons (holes) in the conduction (valence) band drift towards the n-type (p-type) contact. The total electrical current flow is then the sum of these

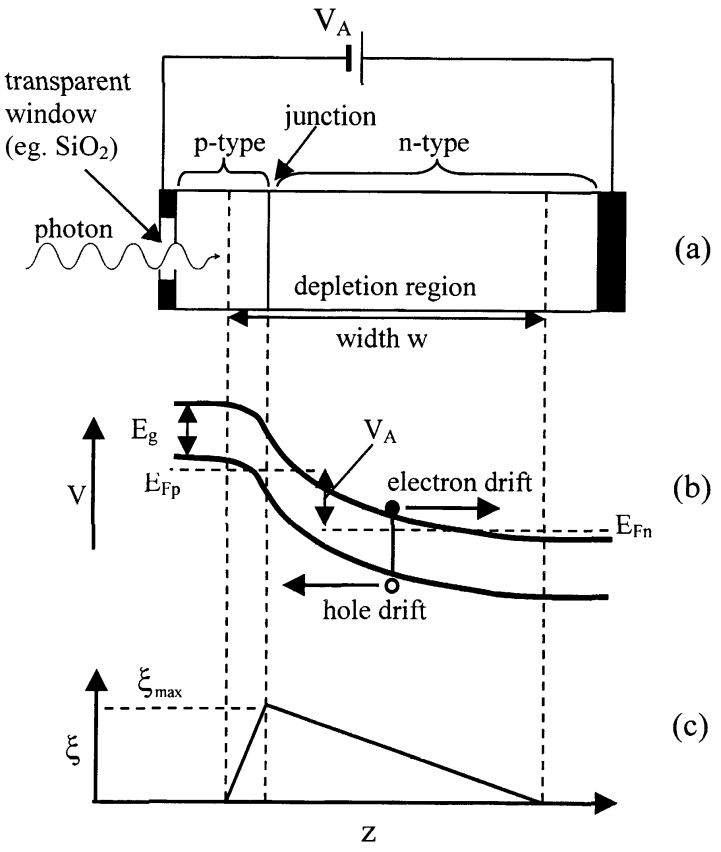


Figure A7.7. The pn junction semiconductor photodiode, (*a*) schematic diagram of a typical device structure. The n-type material is less heavily doped and so contains most of the depletion region (*b*) Profile of the electronic potential through the device. E_g is the bandgap energy, while the applied reverse bias V_A equals the difference between Fermi levels in the p and n-type contact regions, E_{Fp} and E_{Fn}. Photogenerated electrons (holes) drift towards the cathode (anode). (*c*) Electric field profile through the device. At each position on the z-axis, the doping type and density determines the electric field gradient.

electron and hole currents. In the device shown, the p-side of the junction is much more heavily doped than the n-side, so the majority of the depletion region is in the n-type layer.

A typical current–voltage characteristic for a pn photodiode under dark and illuminated conditions is shown in figure A7.8. Four regions are identified on the 'bias' axis, corresponding to different modes of photodetection using pn junction-based devices. Bias region I provides the most straightforward photodetection, offering low dark currents and linear photocurrent response. We shall discuss operation in this bias region first.

The photocurrent—the difference in current between the 'dark' and 'illuminated' curves in figure A7.8—is approximately proportional to the rate of interband absorption in the depletion region. Carriers generated in the regions of the device not under field can contribute to the photocurrent by diffusing into the depletion region, but this process is by far both less efficient and slower and should be kept to a minimum. Since the optical absorption length of any given material reduces with increased photon energy, the detectivity falls away at short wavelengths as more of the signal is absorbed in the top contact region.

Thermal carrier generation in the depletion region is a primary source of dark current in these devices. It depends upon the material band-gap and the operating temperature according to the expression

$$i_{\text{therm}} \propto e^{-E_g/2kT}. \tag{A7.6}$$

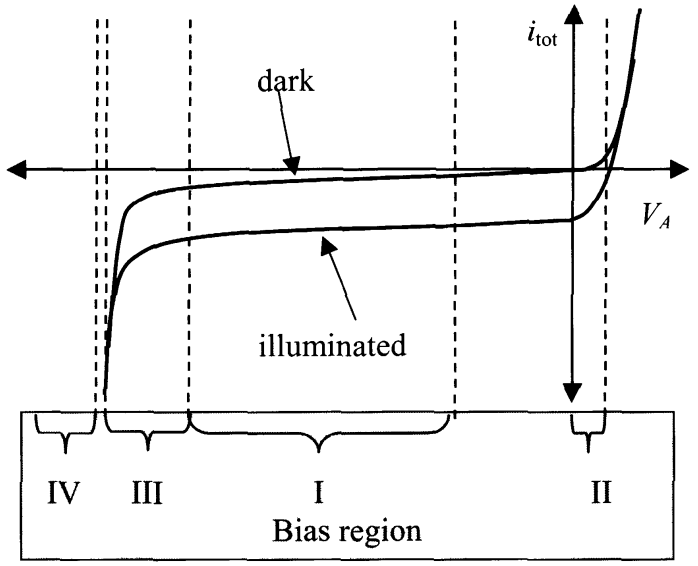


Figure A7.8. Typical current–voltage characteristic for a pn junction photodiode, both dark and illuminated. Four bias regions are identified for different modes of photodetection (see text for details).

Noise in the dark current limits the detectivity and so cooling of photodiodes—especially those utilizing narrow-gap materials—is commonplace. For example, in a germanium pn photodiode near to ambient temperature, a reduction in dark current by a factor of two (and a corresponding increase in detectivity by a factor of $\sqrt{2}$) is obtained for each 7 °C reduction in detector temperature.

The responsivity of the photodiode can be adjusted somewhat by changing the bias V_A across the device, which changes the depletion width W. For an abrupt pn junction the relationship is given by

$$W = \sqrt{\frac{2\varepsilon(V_0 + V_A)}{e}\left(\frac{N_A - N_D}{N_A N_D}\right)} \tag{A7.7}$$

where ε is the permittivity of the detector material, V_0 is the built-in potential difference that remains within each band when $V_A = 0$ (substituting E_g in place of V_0 is often a reasonable approximation) and N_A and N_D are the acceptor and donor doping concentration in the p- and n-type regions respectively. Note that the detectivity is—to a first approximation—less affected by small changes in bias, since the dark current due to thermal carrier generation scales with W as does the photocurrent.

Slight increases in the detectivity of photodiodes can often be realized by tilting the detector by a small angle relative to the incident signal, whereupon the optical pathlength through the depletion layer becomes somewhat greater than W. An increase in photocurrent of several percent can often be obtained by this method, with no associated noise penalty.

The response time of a pn photodiode is determined by two important factors: (1) the transit time of carriers across the junction; and (2) the RC time constant of the complete detection circuit.

A rough estimate of the transit time can be made using the expression

$$\tau_{drift} = \frac{W^2}{\mu(V_0 + V_A)} \tag{A7.8}$$

where μ is the carrier mobility. The quadratic dependence of equation (A7.8) on the depletion thickness shows that a compromise has to be reached between high responsivity and high speed and it is common for pn photodiodes to be designed either for one or the other. Comparing equations (A7.7) and (A7.8), we can see

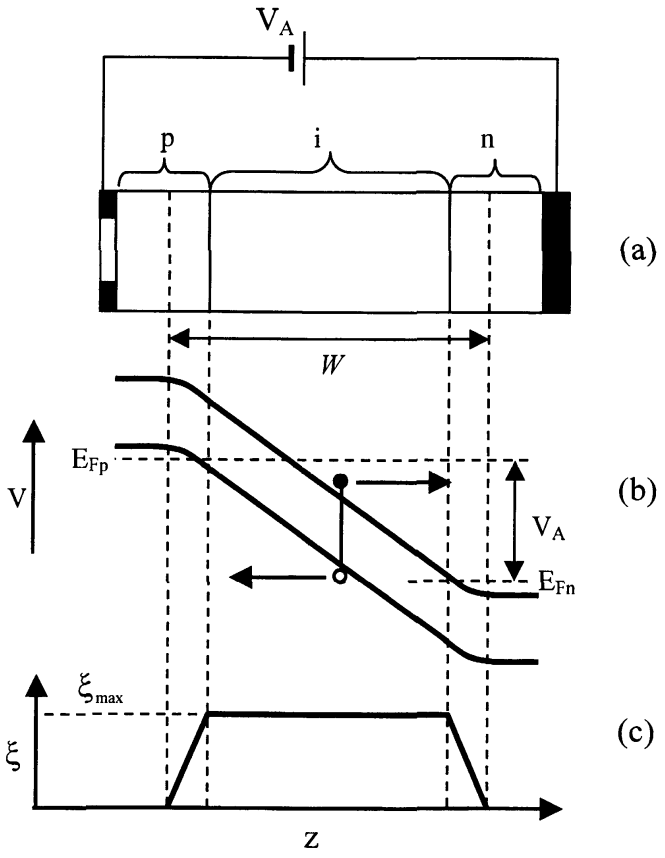


Figure A7.9. The pin photodiode, shown for comparison with figure A7.7: (*a*) schematic diagram of a typical device structure; (*b*) profile of the electronic potential; and (*c*) electric field profile through the device. Since most of the bias is across the intrinsic region, modification of V_A primarily changes ξ_{max} and has little effect on W.

that, to a first approximation, the transit time is independent of the bias voltage, so the user should consider the detection speed required when selecting a device.

The appropriate RC time constant is the product of the device capacitance C_{pn} and the load resistance R_L:

$$\tau_{RC} = R_L \cdot C_{pn} = R_L \frac{\varepsilon A}{W} \tag{A7.9}$$

where A is the device area. Fast devices, therefore, tend to have narrow depletion widths, or 'shallow junctions', to minimize τ_{drift} but consequently rather small active areas to keep τ_{RC} from limiting the bandwidth.

To achieve greater control over the depletion depth and to maximize the mobility, an undoped layer is often grown between the p-type and n-type layers (which are then usually both highly doped). This is known as a pin photodiode and is shown in figure A7.9. It operates in exactly the same way as a pn photodiode but the region of the device under field is determined by the thickness of the intrinsic layer rather than the depletion region. The electric field is uniform in the intrinsic region and so equation (A7.8) is a better approximation than for a pn structure. Variation of the reverse bias, to a first approximation, modifies only the electric field in these devices. The capacitance of a pin structure is given approximately by $C_{pin} \approx \varepsilon A / w_i$, where w_i is the width of the intrinsic layer. The lack of dopant centres in the intrinsic region results in significantly increased carrier mobilities and, therefore, smaller transit times. Pin photodiodes operating at tens of GHz are now widely available, their development having been driven in large part by the telecommunications industry.

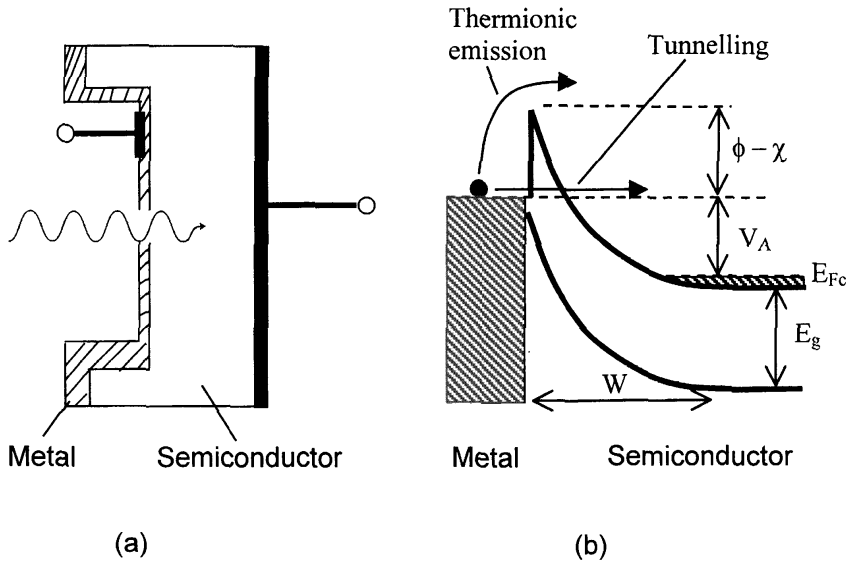


Figure A7.10. The Schottky photodiode: (*a*) schematic diagram of a typical device structure; and (*b*) the profile of the electronic potential through the device. Photocurrent is generated by absorption in the semiconductor depletion region W. Electron transport is inhibited by a triangular potential barrier of height $\phi - \chi$. The two principal mechanisms leading to the dark current—thermionic emission over and tunnelling through the potential barrier—are also indicated.

A7.3.2.1 Photovoltaic mode

A special case occurs in a circuit containing just a pn (or pin) photodiode and a load resistor. The built-in electric field in the diode means that a photocurrent is still generated in the reverse direction, causing a potential drop across the load. This manifests itself as a forward bias across the photodiode (bias region II in figure A7.8), creating negative feedback in the photocurrent, with equilibrium occurring when $V_{\text{forward}} = i_{\text{tot}} R_{\text{L}}$. Electrical power is generated in the load resistor equal to $P_{\text{PV}} = V_{\text{forward}} \cdot i_{\text{tot}}$. This is the principle of operation of a solar, or photovoltaic, cell.

A7.3.3 Schottky diode detectors

One of the problems in pn photodiode operation is that, with high absorption coefficients (often $\alpha > 10^4$ cm^{-1}), photons are absorbed close to the semiconductor surface, potentially lowering the photodiode quantum efficiency due to high surface recombination rates. One solution to this problem is the Schottky photodiode, shown in figure A7.10, which consists of a thin metal layer deposited on to a lightly doped semiconductor. The materials are chosen such that the work function of the metal is greater than the electron affinity in the semiconductor, with the result that band bending occurs in the semiconductor similar to that on the lightly doped side of a pn junction. At the semiconductor–metal interface a triangular potential barrier—the Schottky barrier—inhibits the conduction of majority carriers through the device and gives the device its diode-like electrical characteristics. The height and width of this barrier are fundamental to the device performance, since thermionic emission over and tunnelling through the barrier are the principal origins of dark current in the device. Optical access is through the deposited metal layer, which is typically only a few tens of nanometres thick. For modelling purposes, a Schottky photodiode can be treated like an asymmetric pn junction (equations (A7.6)–(A7.9)), the semiconductor representing the 'lesser doped' side of the junction. Like pin photodiodes, Schottky barrier detectors can be made to operate at very high speeds.

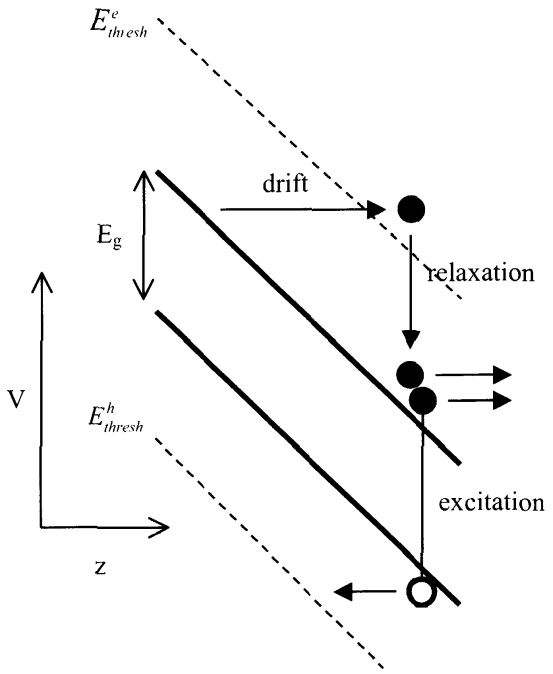


Figure A7.11. The (electron-initiated) impact ionization process. An electron obtains sufficient kinetic energy from the electric field to exceed the threshold energy, E^{e}_{thresh}. It can then relax to the conduction band edge, whilst exciting a further electron across the bandgap. The resulting two electrons and one hole then repeat the process, the hole needing to exceed kinetic energy E^{h}_{thresh} to instigate ionization.

A7.3.4 Avalanche photodiodes (APDs)

Avalanche gain is a convenient means by which a photocurrent can be amplified within the detector itself and can offer better noise performance than amplification by external electronics (see section A7.5)

The avalanche process in semiconductors relies on the multiplication of carriers by impact ionization. If a carrier can obtain sufficient kinetic energy when drifting in an electric field, it can undergo intra-band relaxation whilst using the released kinetic energy to excite a further electron–hole pair. Figure A7.11 shows this process schematically as it occurs when instigated by an electron. The three resulting carriers are then accelerated in the field and have the opportunity to instigate similar processes, and so on. In this way an 'avalanche' of carriers may be generated.

Electric fields in excess of 10^5 V cm^{-1} are generally required to instigate impact ionization in semiconductors, since the saturation velocity $\mu\xi$ must be high enough that carriers reach the threshold kinetic energy required to satisfy the energy and momentum conservation requirements of the microscopic process. The threshold kinetic energy depends on the effective masses of the participant electrons and holes, and other bandstructure details, and is generally in the range $E_g < E_{thresh} < 3E_g/2$. The key parameters commonly used in describing the macroscopic multiplication are the empirically determined electron and hole impact ionization coefficients, α_e and α_h, which specify the ionization probability per cm of carrier drift, and are expressed as functions of the local electric field.

The net effect of the avalanche process in a device is the multiplication of the photocurrent by the factor

$$M = \frac{1}{1 - f_M} \tag{A7.10}$$

where f_M is the *McIntyre function* which is determined by an integral of the electric-field-dependent ionization coefficients throughout the device [1–3]. Most usage of avalanche photodiodes is in 'analogue' mode whereby $0 < f_M < 1$.

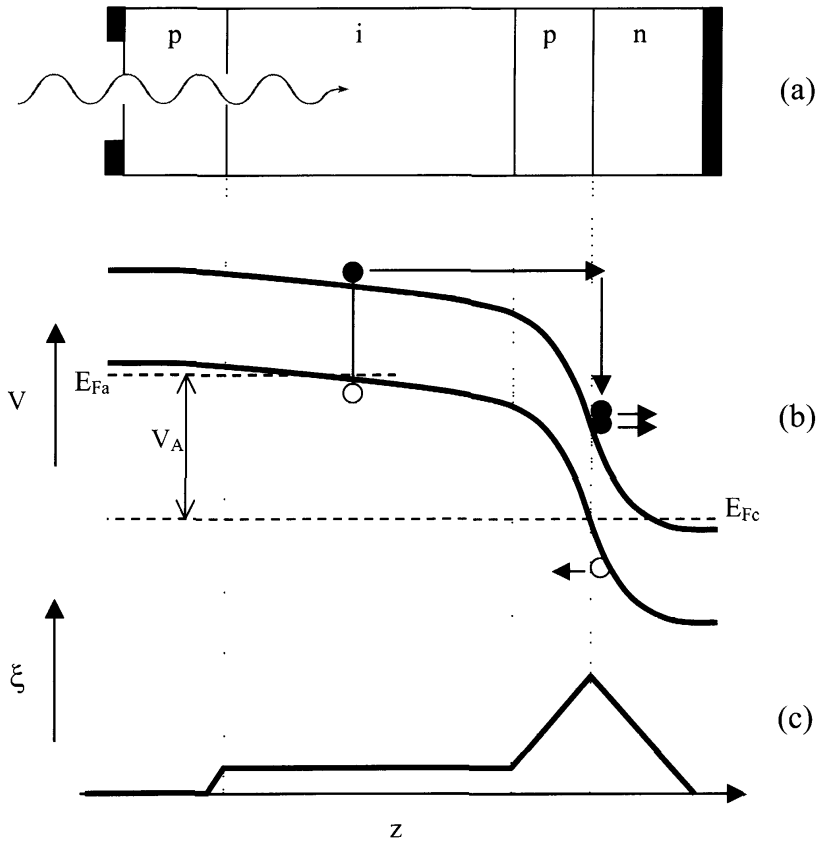


Figure A7.12. The reach-through avalanche photodiode (with electron injection into the gain region): (*a*) typical device structure; and (*b*) electrical potential profile through the device. Optical absorption occurs throughout the device but avalanche multiplication only occurs in the high field region near the pn junction. (*c*) Electric field profile.

The singularity in equation (A7.10) at $f_M = 1$ corresponds to avalanche breakdown, where the avalanche process becomes self-sustaining. In analogue-mode operation, APDs are usually biased at around a volt below this breakdown condition (bias region III in figure A7.8), where M in excess of 100 can be achieved in high quality devices.

Avalanche gain adversely affects both the bandwidth and the noise of a photodiode. The former is a result of the time taken for the avalanche current to build and to decay, approximately equal to the gain multiplied by the single carrier transit time (as such, the gain–bandwidth product is an important figure of merit for avalanche photodiodes). The latter is due to statistical fluctuations in the avalanche gain, and will be discussed further in section A7.5.3.

The desire for high absorption efficiency combined with wide bandwidths has led to APDs of the 'reach through' design, in which gain is limited to a narrow region of the device by careful control of the doping profile (figure A7.12).

A common modification of this design is employed for construction of avalanche photodiodes with sensitivity at optical wavelengths well into the near infrared. In semiconductors with bandgaps narrower than about 1 eV, the electric fields needed for avalanche multiplication also result in significant increases in dark current due to inter-band tunnelling. For this reason, devices have been developed with separate absorption and multiplication regions (known as SAM devices). The structure is the same as a reach-through device but with a narrow-gap semiconductor occupying the intrinsic region. Absorption at photon energies between those of the two bandgaps occurs only in the narrow-band material, which is kept under low enough electric field to prevent inter-band tunnelling of carriers. The principal exponent of the SAM design is the InGaAs/InP avalanche photodiode, commonly used for optical fibre-based telecommunications applications.

Here, photons of wavelengths between 1.3 and 1.6 μm are absorbed in a low-field InGaAs layer and the photogenerated holes are swept into a high-field InP layer where they then multiply by impact ionization. Note the role of holes as the primary ionization species (holes cause ionization more readily than electrons in InP), so that the high-field region of the device is near to the p-type contact, opposite to that shown in figure A7.12.

Avalanche breakdown offers a particularly effective means of single photon detection. If the high-electric-field region of a device is free of conducting carriers when biased above the reverse bias breakdown voltage, no avalanche can occur and no current flows. By injecting a single carrier optically, however, an exponentially growing avalanche current is generated, which can easily be recorded. This current must then be switched off, or 'quenched', before the device overheats and the carriers allowed to drain from the active region before the bias is restored to the above-breakdown value ready for the next photon to arrive. Operated in this mode, avalanche photodiodes are commonly referred to as single-photon avalanche diodes or SPADs.

For photon-counting applications in which the signal can be focused on to a small detector area (spot diameter $\simeq$ 100 μm or less) and the photon flux is less than about $10^6\,s^{-1}$, commercially available SPADs offer detectivity about a 100 times higher than the best photon-counting photomultipliers. A thermoelectrically cooled silicon SPAD can offer a dark count rate as low as a few counts per second coupled with a quantum efficiency of 70%, equating to a NEP of the order of 10^{-18} W $Hz^{-1/2}$. Photomultipliers make better large-area devices, however, and are still standard equipment for sensitive fluorescence measurements in which the signal cannot be focused. Photomultipliers' lack of need for quenching also means that the maximum measurable photon flux is limited only by the discrimination of distinct current pulses and 10^7 photons s^{-1} or more is readily achievable.

A7.3.5 Photoconductive detectors

Photoconductive detectors rely on the fact that the conductivity σ of a semiconductor is proportional to the densities n of carriers therein:

$$\sigma = e(\mu_e n_e + \mu_h n_h) \tag{A7.11}$$

where subscripts e and h refer to electrons and holes. For an intrinsic semiconductor under illumination, $n_e = n_h$, and the change in conductivity due to the optical signal can be shown to be

$$\Delta\sigma = \frac{e\eta\tau_{rec}F(\mu_e + \mu_h)}{xA} \tag{A7.12}$$

where η is the quantum (absorption) efficiency, τ_{rec} is the recombination lifetime, F is the photon flux and x is the electrode spacing.

In the common situation where $\mu_e \gg \mu_h$, the current passed by the photoconductor may be expressed conveniently as

$$i_P = e\eta F\frac{\tau_{rec}}{\tau_e} \tag{A7.13}$$

where τ_e is the transit time for electron drift between the electrodes. For each photon absorbed, the number of electrons that can drift between the electrodes before recombination occurs with the hole and the gain of the photoconductor is, therefore, given by $G = \tau_{rec}/\tau_e$.

Careful design of the detector can achieve $G > 1$. For example, the structure should minimize the distance between electrodes both to minimize the transit distance and to maximize the electric field (and thus the drift velocity) for a given bias. Device geometry, therefore, usually consists of interdigitated electrodes deposited onto the semiconductor surface, as shown in figure A7.13, allowing high gain whilst maintaining a reasonable active area. Measurement of the conductivity is achieved by monitoring the voltage across a small load resistor in series with the detector. In this way the biasing of the photoconductor remains constant and the output voltage is (approximately) proportional to the incident light intensity.

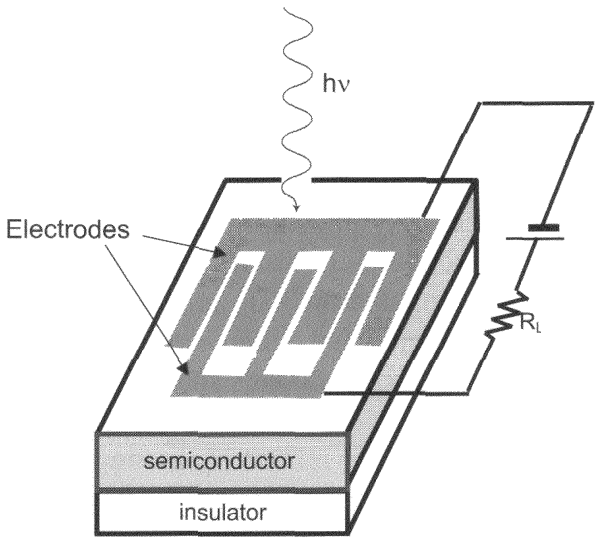


Figure A7.13. Schematic diagram of a photoconductive detector with an inter-digitated electrode structure.

Basic photoconductors can be produced extremely cheaply compared with photodiodes, since less care needs to be taken over material quality and contacting. Whilst offering a simple method to achieve signal gain, their detectivity is limited by generation–recombination noise (see section A7.5.2). They, therefore, tend to be employed in imaging applications where large-scale arrays are needed but low sensitivity is acceptable, such as in photocopiers.

A7.3.6 Intra-band detectors or QWIPs

Intra-band electronic transitions—those that occur within the conduction band or within the valence band—can be utilized to permit photodetection using a doped semiconductor regardless of its band-gap. The threshold photon energy above which strong absorption occurs is determined by the energetic separation of the sub-bands or the energy required to eject a carrier from the quantum well altogether. This is determined partly by the properties of the materials used but also significantly by the width of the quantum wells. QWIPs usually consist of multiple quantum wells, which are doped n- or p-type to provide a source of carriers in the lowest confined level. Common well/barrier material combinations are GaAs/AlGaAs, SiGe/Si and AlGaInAs/InP. The same doping species is used in the contact regions and the device is held under a bias of typically less than 1 V. Since the same intra-band transitions that are caused by photoabsorption can equally be caused thermally, QWIPs are invariably cooled. Detectivities approaching 10^{11} cm $Hz^{1/2}$ W^{-1} have been reported for a QWIP detecting light of $\sim$10 μm wavelength [4]. This figure is some two orders of magnitude higher than the best pyroelectric detectors.

The most common application for QWIPs is in infrared imaging at wavelengths between 2 and 30 μm, for which they are fabricated into focal plane arrays (FPAs). Most commercial QWIP FPAs operate in the temperature range 20–80 K, cooled by Stirling closed-cycle coolers which are compact and light enough to be contained in a hand-held camera.

QWIPs that are designed to utilize a bound-to-continuum excitation are sensitive over a broad wavelength range with a sharp cut-off at long wavelength. Those that utilize a bound-to-bound transition have a narrow range of sensitive wavelengths ($\Delta\lambda \simeq 0.2\lambda$ is common). Two or more differently tuned bound-to-bound QWIPs can, therefore, be used in parallel to generate 'colour' images.

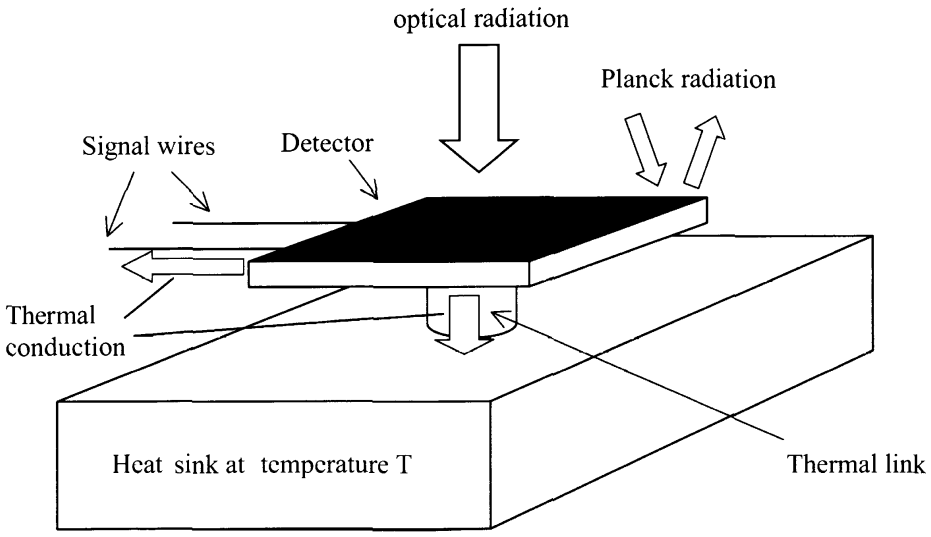


Figure A7.14. Heat balance in a thermal detector. The detector itself is coated with a 'black' surface for maximum absorption of radiation and is small in mass for minimum heat capacitance. Principal heat loss is to a heat sink through a thermal link and through any signal wires attached.

A7.4 Thermal detectors

Thermal detectors contain two principal elements: an absorption region, the temperature of which is a function of the incident optical power, and a transducer to convert the temperature variation in the detector to an electrical signal. The latter takes several different forms in different types of detector. In this section we will discuss the three most common of these: thermocouples and thermopiles; bolometers and thermistors; and pyroelectric detectors.

Thermal detectors are most commonly used for detection of optical radiation at wavelengths in the mid infrared and longer, as the thermal detection process is slow and inefficient compared with the photoelectric processes described previously, and the latter are usually preferred within their range of sensitivity. However, the ability of thermal detectors to provide very uniform detectivity over a wide range of optical wavelengths has proved useful also for radiometric calibration purposes. In this chapter, we restrict ourselves to a brief introduction to the workings of thermal detectors. More detailed descriptions can be found in the references at the end of the chapter [5, 6].

The change in temperature θ caused by incident radiation can be calculated by considering the thermal processes in a typical device. Figure A7.14 shows a schematic diagram that illustrates these processes. Heat is gained by the detector with the absorption of incident radiation and is lost through contact with a heat sink and through heat conduction by any electrical wires connected to the detector (radiation of heat at the surface is generally negligible by comparison).

The heat balance equation is, therefore,

$$H\frac{\mathrm{d}\theta}{\mathrm{d}t} = \eta P - G\theta \tag{A7.14}$$

where H is the heat capacity of the detector, G is the thermal conductivity of the electrical wires and the thermal link, P is the incident radiation power, and η is the absorptivity of the surface.

The temperature responsivity $S_\theta = \mathrm{d}\theta/\mathrm{d}P$ for a modulated incident signal is, therefore,

$$S_\theta = \frac{\eta}{G\sqrt{1+\omega^2\tau_\mathrm{h}^2}} \tag{A7.15}$$

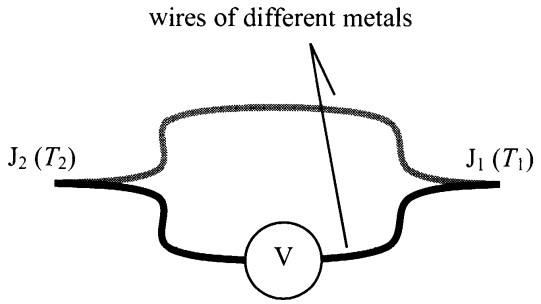


Figure A7.15. A two-junction thermocouple. The voltage is proportional to the temperature difference between the junctions.

where τ_h is the thermal time constant ($= H/G$), which limits the detector response time (pyroelectric detectors offer something of an exception to this—see section A7.4.3). Under constant illumination, $\omega = 0$, and $\theta = \eta P/G$. Principal design considerations are to maximize θ_0 by maximizing η and to minimize τ_h by minimizing H. The former is achieved by coating the absorption region with a 'black' material for maximum absorption across a wide spectral range, whilst the latter is achieved by using a thin wafer of material with a low specific heat capacity C (since $H = Cm$ for a body of mass m). The ultimate detectivity of thermal detectors is limited by fluctuations in the detector temperature. 'Temperature noise' will be discussed in section A7.5.2.6.

A7.4.1 Thermocouples and thermopiles

A thermocouple is a thermoelectric transducer based on the Seebeck effect, which describes the current generated in an electrical circuit containing two or more different conductors when the junctions are held at different temperatures. A simple two-junction thermocouple is shown in figure A7.15. The voltage measured by the voltmeter in the diagram is $V = S(T_1 - T_2)$, where S is the Seebeck coefficient. For temperature measurement, one of the junctions is held at known temperature while the other is held at the temperature to be measured. For improved sensitivity, junctions are often connected in series to make a thermopile. The very best thermopiles can provide detectivities as high as 10^9 W^{-1} $Hz^{1/2}$ and offer response times of a few milliseconds.

A7.4.2 Bolometers and thermistors

In bolometers and thermistors the temperature change resulting from optical absorption is measured as a change in conductivity. In a metal, the conductivity decreases with temperature and the device is called a bolometer, whereas in a semiconductor, conductivity increases with temperature and the device is called a thermistor. Semiconductors offer temperature coefficients (α) of the order of -4%/K, whilst metals offer coefficients of the order of $+0.5\%$/K but metals have the advantage of being easier to fabricate into micrometre thin flakes which offer smaller thermal capacitance.

The voltage responsivity $S_V = \mathrm{d}V/\mathrm{d}P$ of a bolometer or thermistor is related to the temperature responsivity by

$$S_V = I\alpha R S_\theta \tag{A7.16}$$

where I is the device current and R is the device resitance.

The material to be used usually forms a thin flake that is coated directly with a highly absorbent 'black' material. Since it is a change in resistance that must be measured, the optical signal is generally chopped mechanically in front of the detector. Often a bridge circuit is used in which a second 'control' flake is placed

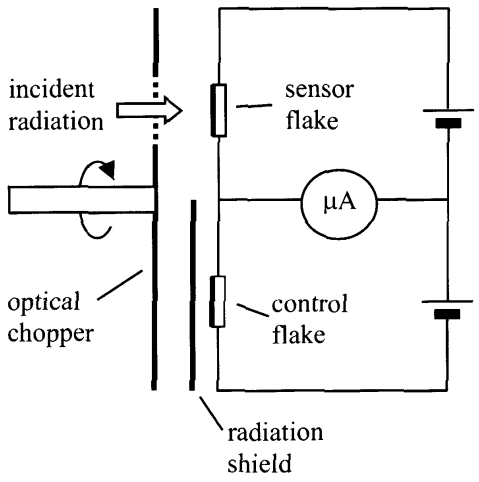


Figure A7.16. *Measurement set-up using bolometers or thermistors. The microammeter measures the difference between the currents passing through the two flakes. In addition, the optical signal is modulated using a chopper to minimize the background signal.*

close to the detecting flake but shielded from the radiation. This method serves to remove most of the dark current and reduces the effects of fluctuations in the ambient temperature on the measured signal. A typical measurement bridge circuit is shown in figure A7.16.

Bolometers and thermistors offer time responses similar to those of thermopiles. Detectivities within an order of magnitude of the thermodynamic limit (see section A7.5.2.6) are possible, and sensor flake sizes range from 1 cm^2 to as small as 0.01 mm^2.

A7.4.3 Pyroelectric detectors

The pyroelectric effect consists of a change in the surface charge of a material in response to a change in temperature. Since the detectors electrical characteristic is that of a capacitor, only fluctuations in temperature can be measured and the optical signal must always be modulated. Materials are chosen for their high pyroelectric coefficient, p, and for their high Curie temperature, above which the electrical polarization is lost. An example of a popular high-p material is triglycine sulphate (TGS), which has $p \simeq 30$ nC cm^{-2} K^{-1} and a Curie temperature of 49 °C. Also popular are ceramics such as lithium tantalate, $LiTaO_3$, which offers a higher Curie temperature of $T_C \approx 620$ °C combined with a pyroelectric coefficient of $p \approx 6$ nC cm^{-2} K^{-1}.

For a modulated temperature excess θ_ω, the pyroelectric detector produces an alternating current equal to

$$I_P = \omega p A \theta_\omega \tag{A7.17}$$

where ω is the angular frequency, p is the pyroelectric coefficient (in C cm^{-2} K^{-1}) and A is the device area.

The responsivity of the illuminated pyroelectric detector is found by considering its equivalent circuit, shown in figure A7.17. The detector is represented as a capacitor C_D, a resistor R_D, and an alternating current source I_P in parallel. The output, V_s, is usually amplified, whereby the amplifier input impedance must also be considered. The voltage responsivity of the complete detection circuit is found to be

$$S_V = \frac{\omega p A}{\sqrt{(1 + \omega^2 \tau_{RC}^2)}} S_\theta \tag{A7.18}$$

where τ_{RC} is the RC time constant of the equivalent circuit.

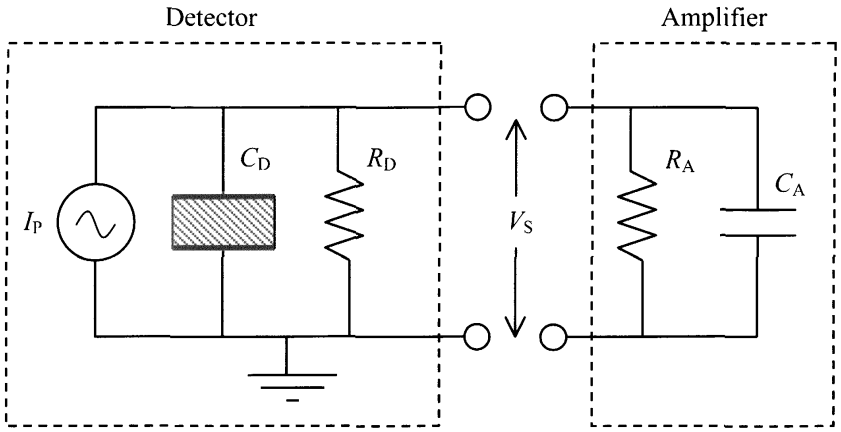


Figure A7.17. Equivalent circuit for an illuminated pyroelectric detector and amplifier input stage. Consideration of the complete circuit yields the frequency response of the system.

Equation (A7.18) presents a very different frequency response to that of the other photothermal detectors. Assuming that $\tau_{RC} \ll \tau_h$, we find that (a) for $\omega \ll 1/\tau_h$, $S_V \propto \omega$ reminding us that no signal occurs under constant illumination; (b) in the frequency range $1/\tau_h \ll \omega \ll 1/\tau_{RC}$, the response is approximately flat; and (c) for $\omega \gg 1/\tau_{RC}$ it falls off as $1/\omega$. The ability of pyroelectric detectors to respond to optical transients much faster than τ_h means that they are the only detectors capable of detecting nanosecond pulses into the far and extreme infrared, albeit with a much reduced responsivity.

A7.5 Noise in photodetection

Whenever photodetection is performed, as with all continuous physical measurements, the signal trace will contain a certain level of noise or random fluctuations in the detector output, which limits the accuracy that can be attained. The level of noise experienced depends upon a number of factors and, in many cases, steps can be taken to 'clean up' a noisy signal. Whilst some improvement can often be achieved by post-measurement data processing, better results will invariably be achieved through careful design and use of the measurement apparatus. In particular, we shall see that restriction of the measurement bandwidth can lead to a dramatic reduction in noise levels.

This section is intended as an introduction to the various sources of noise that may be encountered when performing photodetection. In each case the physical source of the noise will be explained and a simple expression given with the aim of enabling the reader to estimate the magnitude of the effect as it pertains to the measurement in question.

Section A7.1.1 provides an introduction to the nomenclature commonly used in quantifying noise and in defining figures of merit with which to quantify photodetector performance. Sections A7.5.1–A7.5.4 describe the various sources of noise that can arise in the photodetection process. These three sections reflect the sequential nature of the photodetection process, starting with the optical signal itself, through noise generated within common detector types and finishing with noise generated in the electrical detection circuit. Section A7.5.4 describes how to combine these noise contributions to generate an aggregate noise spectrum for a photodetection system. Finally section A7.5.5 provides a brief introduction to bandwidth-related methods of noise reduction.

A7.5.1 Noise in the optical signal

Before examining noise generated by the detection process, let us consider the nature of the optical radiation incident on the detector.

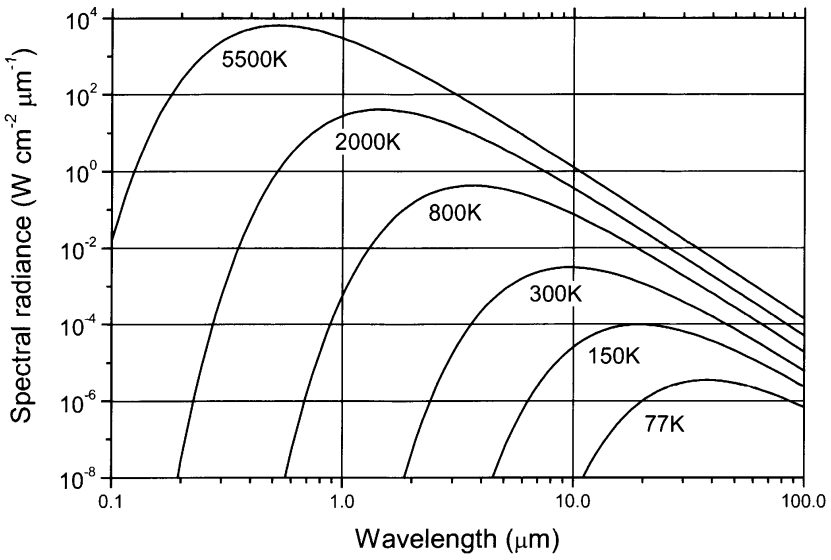


Figure A7.18. Blackbody spectral radiance at temperatures between 77 K and 5500 K (approximate solar temperature). The radiance describes the background optical power incident on a detector situated in a black box at temperature T.

A7.5.1.1 Background noise (blackbody radiation)

In any photodetection situation effort must clearly be taken to minimize the detection of unwanted photons. However, this is not always possible to achieve perfectly since, according to Planck's theory, any surface at a finite temperature will radiate photons that can potentially be detected in addition to the desired signal. The optical power emitted per unit surface area and wavelength is

$$P_{\text{bkgrnd}}(\lambda) = \frac{2\pi hc^2}{\lambda^5(\exp[hc/\lambda kT] - 1)}. \tag{A7.19}$$

This relationship is plotted for several temperatures in figure A7.18. For thermal equilibrium in a black box, the power radiated by any given area must be equal to the power incident on it. Using this approximation we can estimate the optical power incident on a detector as being equal to the radiance by a surface of the same area at the temperature of the surroundings. In most practical situations, temperatures are a few hundred Kelvin or less and the effect is only significant for measurements at wavelengths in the mid-infrared or longer, for which emission from detector housing can be significant.

The background radiation incident on a detector leads to an additional 'dark' signal in the photoresponse—its effect still present when the signal source is switched off—the average of which can be subtracted from the total detector output signal. The signal is not constant, however, and is subject to random fluctuations which can not be subtracted (because they are random!). These fluctuations ultimately limit the sensitivity of the measurements. A detector exhibiting noise dominated by these fluctuations is said to be at the background-limited intrinsic performance (BLIP) limit. Further improvement can only be achieved by reducing the temperature (T in equation (A7.19)) of as much of the detector field of view as possible, for example by utilizing a cooled radiation shield. A shield should be chosen with an aperture corresponding closely to the spatial extent of the signal so as to limit the surface area over which 'uncooled' background radiation is detected.

A7.5.1.2 Photon noise

The behaviour of light as discrete 'particles' means that the arrival of light at a detector is not uniform but made up of pulses corresponding to the arrival of individual photons. Moreover, in most situations the photons

do not arrive at regular intervals and repeated observations of time duration T will reveal that the number of photon arrivals fluctuates about the mean value of $\langle n \rangle = PT/h\nu$, where P is the optical power and $h\nu$ is the photon energy. These fluctuations are known as *photon noise* and can be observed using a detector of sufficient detectivity.

In coherent light from a laser, the interval between photon arrivals is purely random (i.e. if we choose $T \ll h\nu/P$ the probability of a photon arriving in any time window is equal). The number of photon arrivals in a time window T is, therefore, governed by Poissonian statistics and the variance in the photon number is equal to $\langle \delta n^2 \rangle = \langle n \rangle$.

Even with a perfect linear photodetector, therefore, in which each incident photon is converted into an exactly determined photocurrent, we can expect this noise to be present. In such a situation, the SNR is given by

$$\text{SNR} = \frac{\langle n \rangle^2}{\langle \delta n^2 \rangle} = \langle n \rangle \tag{A7.20}$$

and so increases linearly with the photon flux or with the detection time T. Such a relationship between SNR and detection time indicates that photon noise is white noise (the maximum measurement bandwidth is $\Delta f = 1/T$).

For the vast majority of photodetection situations, photon noise constitutes an absolute lower limit on the achievable noise amplitude. Exceptions to this are only achieved by manipulation of the light itself, by techniques known as 'squeezing', which allow photon noise to be 'traded' between complementary measurements in two-dimensional amplitude/phase space [6–8].

A7.5.2 Noise in the photodetector

A7.5.2.1 Photoelectron noise

In most photodetectors the quantum efficiency η is less than unity. Whether or not an individual photon is absorbed is arbitrary and so again Poissonian statistics apply. In the same way that the photon noise is proportional to the average photon flux, the photoelectron noise is proportional to the average photocurrent, I_{ph}.

$$\langle \delta i^2 \rangle = 2eI_{\text{ph}}\Delta f \tag{A7.21}$$

where Δf is the bandwidth of the detector.

The linearity of the expression with Δf indicates that photoelectron noise, like photon noise, is also white noise.

A7.5.2.2 Shot noise

Shot noise is the general term for the noise that results from the varying number of photocarriers that contribute to the photocurrent. In a photodiode, it accounts for both photon and photoelectron noise, along with any other mechanisms that cause random loss of photogenerated electrons in the detector. One common mechanism is thermionic emission across a potential barrier, such as that encountered by photogenerated holes in a heterojunction design InGaAs/InP avalanche photodiode (see section A7.3.4). The expression for the total shot noise is

$$\langle \delta i^2 \rangle = 2eI_0\Delta f \tag{A7.22}$$

where I_0 is the average photocurrent, similar to I_{ph} but including these additional losses. Figure A7.19 shows how the random processes described result in shot noise in the detector photocurrent, on the assumption that each primary photoelectron generates a similar current pulse at the detector output.

The maximum SNR due to shot noise is found to be $I_0/2e\Delta f$ and is independent of any subsequent gain, as both signal and noise are amplified by the same factor.

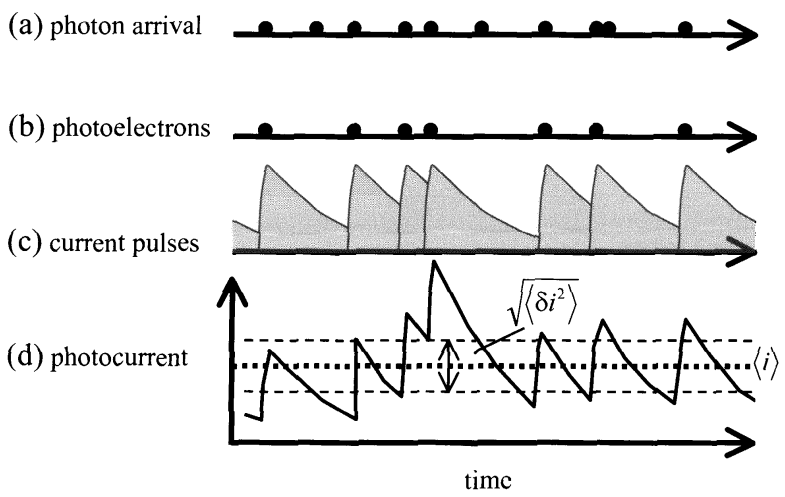


Figure A7.19. The origins of shot noise: (*a*) photons arriving at irregular intervals, some of which are absorbed to generate (*b*) photoelectrons; (*c*) each photoelectron results in a similar detector current pulse, which sum (*d*) to give the total photocurrent.

A7.5.2.3 Generation–recombination (G–R) noise

In a photoconductor, an additional source of shot noise is present, due to random fluctuation of the carrier density caused by the competing processes of thermal carrier generation and recombination. In contrast to the white noise sources encountered thus far, G–R noise reduces rapidly above a cut-off frequency corresponding to the recombination lifetime τ_{rec}

$$\langle \delta i^2 \rangle = \frac{4eGI_0}{1 + 4\pi^2 f^2 \tau_{\text{rec}}^2} \tag{A7.23}$$

where G is the gain of the device. G–R noise can be reduced by cooling the detector, as the thermal generation rate is proportional to $\exp(-E_g/kT)$.

A7.5.2.4 Gain noise

In an amplifying photodetector with deterministic gain G, both signal and noise are amplified by this same factor and the SNR remains unchanged. In most common detectors, such as avalanche photodiodes and photomultipliers, the gain mechanisms are random in nature and so the current pulse generated by each absorbed photon will not be the same, as inferred in figure A7.19, but will be subject to Poissonian fluctuations. This further variance, represented schematically in figure A7.20, forms an additional source of noise, known as *gain noise*. It is characterized by the *excess noise factor*

$$F = \frac{\langle G^2 \rangle}{\langle G \rangle^2} = 1 + \frac{\langle \delta G^2 \rangle}{\langle G \rangle^2} \tag{A7.24}$$

where $\langle \delta G^2 \rangle$ is the variance in the gain. The total variance in the detector photocurrent due to the multiplied shot noise and gain noise is then simply $F\langle G \rangle$ times the shot noise given in equation (A7.22).

In an avalanche photodiode, the excess noise factor is determined primarily by the relative contributions of electrons and holes to the multiplication process and

$$F = kG + \left(2 - \frac{1}{G}\right)(1 - k) \tag{A7.25}$$

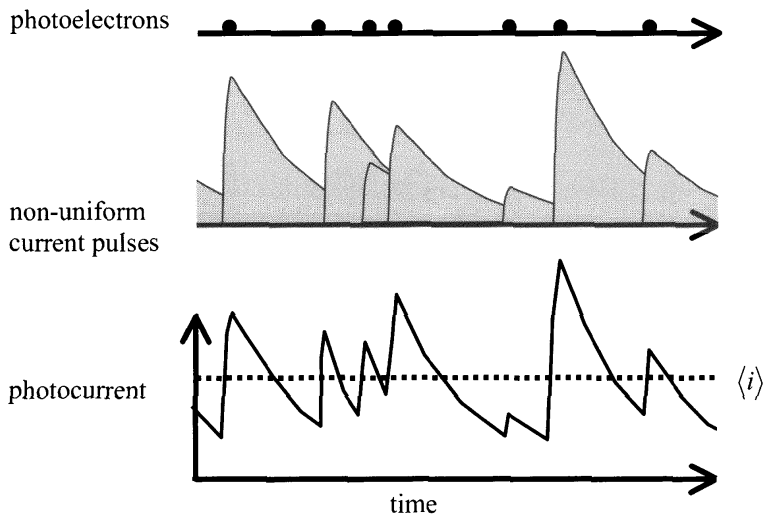


Figure A7.20. Gain noise. Random fluctuations in the gain mean that each photoelectron does not produce the same contribution to the current pulse. The SNR of the resulting photocurrent is smaller than that in figure A7.19.

where $k = \alpha_s/\alpha_p$ is the ratio of the impact ionization coefficients for the secondary and primary carrier types ($0 < k < 1$). The primary carrier type in this context depends upon the semiconductor material from which the avalanche gain region is made. For example, in Si and GaAs, electrons undergo impact ionization more readily than holes ($\alpha_e > \alpha_h$), and so most avalanche photodiodes made from these materials will use electrons as the primary carrier type and $k = \alpha_e/\alpha_h$. In InP, however, $\alpha_h > \alpha_e$, so holes are usually the primary carrier and $k = \alpha_h/\alpha_e$. Equation (A7.25) reveals that the excess noise factor is lowest for $k \approx 0$, i.e. for a large difference between α_e and α_e.

Since the gain process in photomultipliers involves electrons alone, their excess noise factor is given by equation (A7.25) with the substitution $k = 0$, whereby $F = 2 - 1/G$. The one situation in which gain noise can be neglected is that in which G is sufficiently large and the signal sufficiently weak, to perform *single-photon counting*. These digital measurements are subject only to shot noise and are, therefore, well suited to low-frequency photodetection where G–R noise or flicker (see next section) might otherwise dominate. (Note that, in the case of a photomultiplier, the gain distribution may manifest itself as an additional source of shot noise if discrimination of the pulse height from the detector causes some pulses to pass unrecorded.) An additional advantage of single-photon counting its the ability to measure fast optical pulses. Time-correlated single-photon counting (TCSPC) is a powerful technique in which the arrival time of a photon can be measured with picosecond accuracy relative to a synchronization signal, and the time dependence of the optical intensity can thus be acquired by repeating the measurement over many cycles. (For a detailed account of TCSPC, see [9]).

A7.5.2.5 $1/f$ or flicker

At the low-frequency limit of analogue measurements, a further source of noise is universally dominant. It is often known simply as '$1/f$ noise', as its intensity reduces uniformly with increasing frequency. $1/f$ noise exhibits equal power per increment in $\log(f)$ and is sometimes referred to as 'pink noise'. Its presence results in the unwelcome discovery that lower noise cannot necessarily be achieved by increasing the measurement integration time.

$1/f$ noise can be found in a surprising range of physical systems, from cathode ray tubes to tidal movements, to inner city traffic flow! In semiconductor photodetectors, it is generally suspected to arise from

carrier trapping by surface and bulk impurities and is particularly troublesome in devices with poor quality ohmic contacts (for a review, [10]).

A7.5.2.6 Temperature noise

Under certain circumstances, noise can be measured that is due to random fluctuations in the temperature of the detector. This is rarely an issue in detectors based on photoelectric processes but is of crucial importance to photothermal detectors.

A theoretical lower limit in temperature fluctuations is imposed by the absorption and emission of Planck radiation, described in section A7.5.2. The resulting temperature variance is given by statistical thermodynamics considerations:

$$\langle \Delta\theta^2 \rangle = \frac{4kT^2 G \Delta f}{G^2 + 4\pi^2 f^2 H^2} \tag{A7.26}$$

where G and H are the thermal conductance and heat capacitance as defined in section A7.4. At measurement frequencies well below the maximum response frequency of the detector, $f_{\max} = G/H$, temperature noise is white and the noise power delivered to the detector is

$$\langle \Delta\phi^2 \rangle = G^2 \langle \Delta\theta^2 \rangle = 4kT^2 G \Delta f. \tag{A7.27}$$

A thermal detector is said to be 'ideal' when the thermal conductance is due primarily to Planck radiation, whereupon $G = 4\sigma\varepsilon A T^3$, where σ is the Stefan–Boltzmann constant, ε, is the emissivity and A is the detector area. This imposes a theoretical upper limit on the specific detectivity of photothermal detectors,

$$D^*_{\max} = (8\pi\sigma\varepsilon k T^5)^{-1/2}. \tag{A7.28}$$

Assuming $\varepsilon = 1$, at 300 K, $D^*_{\max} = 1.8 \times 10^{10}$ cm $\mathrm{Hz}^{1/2}$ W^{-1}, while at 77 K, $D^*_{\max} = 5.2 \times 10^{11}$ cm $\mathrm{Hz}^{1/2}$ W^{-1}. In real thermal detectors, these lower limits of thermal fluctuation are some way from being achieved, however, and the highest detectivities of uncooled detectors are those of the best pyroelectric detectors and bolometers, at around 10^9 cm $\mathrm{Hz}^{1/2}$ W^{-1}.

A7.5.3 Noise in the measurement circuit

A7.5.3.1 Thermal noise (aka Johnson or Nyquist noise)

The random thermal motion of electrons in a resistor leads to fluctuations in the current passing through it and thus also in the voltage measured across it. This affects all photodetection measurements, the most common sources being the load resistance in the measurement circuit and the output impedence of the photodetector itself.

The expression for the resulting current variance is derived from the Bose–Einstein statistics applied to electromagnetic radiation of frequency f:

$$\langle \delta i^2 \rangle = \frac{4hf}{R(\exp[hf/kT] - 1)} \tag{A7.29}$$

where k is Boltzmann's constant, T is the temperature in kelvin and R is the resistance in ohms. For detection frequencies low compared with kT/h (i.e. below about 10^{12} Hz at room temperature), thermal noise is white and $\langle \delta i^2 \rangle \approx 4kT\Delta f/R$. An important comparison can be made between the level of white noise caused by Johnson noise and that caused by shot noise. Equations (A7.22) and (A7.29) reveal that the Johnson noise level falls below the shot noise level if

$$kT < \frac{eI_0 R}{2}. \tag{A7.30}$$

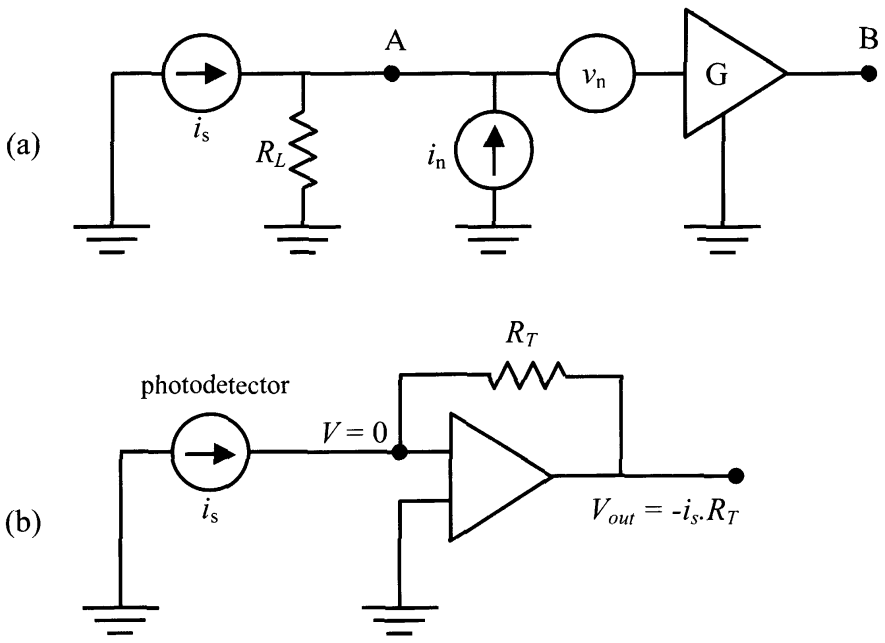


Figure A7.21. Equivalent circuits demonstrating the treatment of amplifier noise. (*a*) shows how noise can be treated where a voltage amplifier is employed. The amplifier circuit (between points A and B) is represented by a noise-free gain G with an additional input noise current i_n and input noise voltage v_n (see text). (*b*) shows an idealized trans-impedance amplifier commonly used with photodiodes and photomultipliers. Here, the only noise contribution by the amplifying circuit is Johnson noise in the feedback resistor.

The Johnson noise in the load resistance seen by the photodetector is thus important in determining the dominant noise source. For example, in a photodiode with a dark current of 100 pA, equation (A7.30) reveals that a load resistance greater than 500 MΩ would be required to achieve shot noise limited detectivity at a temperature of 300 K. Alternatively if we consider a fast photodiode terminated with a 50 Ω load, a photocurrent of at least 1 mA is required for the signal to be shot noise limited.

A7.5.3.2 Amplifier noise and impedance matching

Amplification of the photodetector signal is often essential. Here we look at two common circuits in which amplification is used and consider their noise characteristics. The first, shown in figure A7.21(*a*), concerns a detector, represented as a current source generating current i_s and terminated to earth with a load resistor R_L, the voltage across which is amplified using a voltage amplifier. The total voltage noise at point A in the circuit is, in the absence of amplification, calculated by combining the Johnson noise in R_L with the detector current noise across R_L.

$$\langle \delta v^2 \rangle_A = \langle \delta i_s^2 \rangle R_L^2 + 4kTR_L \Delta f. \tag{A7.31}$$

If the signal at point A is fed into a voltage amplifier, we can represent the further noise contribution by two quantities—an input noise current i_n and an input noise voltage v_n (representing rms values)—in combination with a hypothetical noise-free amplifier G. The *further* variance in the voltage at the input to G, due both to and to passing through R_L is then

$$\langle \delta v_G^2 \rangle = v_n^2 + (i_n R_L)^2 \tag{A7.32}$$

If necessary, v_n and i_n can be measured by recording the noise level in the amplifier output under appropriate limiting input conditions (see [11, 12]).

The factor by which the amplifier stage increases the effective input noise is given by the *noise figure* (NF):

$$\mathrm{NF} = 1 + \frac{\langle \delta v_{\mathrm{G}}^2 \rangle}{\langle \delta v^2 \rangle_{\mathrm{A}}}. \tag{A7.33}$$

It is straightforward to show that if $\langle \delta v^2 \rangle_{\mathrm{A}}$ is dominated by thermal noise in resistor R_{L}, then NF is minimized when $v_{\mathrm{n}}^2 = i_{\mathrm{n}}^2 R_{\mathrm{L}}^2$. Under these conditions the circuit is said to be *impedance matched* for minimum amplifier noise.

For photodiodes and photomultipliers, it is common to use a trans-impedance amplifier to convert the current output of the detector to a voltage. The idealized circuit is shown in figure A7.21(*b*). The ideal operational amplifier has an infinite input impedance and infinite gain, ensuring that all of i_{s} flows through the feedback resistor R_{T} and that the input is a 'virtual earth' held at 0 V. The output voltage is then simply $V_{\mathrm{out}} = -i_{\mathrm{s}} \cdot R_{\mathrm{T}}$. Johnson noise in the feedback resistor is the main source of noise, combining with the noise in the detector current to give a total voltage noise at the output equal to

$$\langle \delta V_{\mathrm{out}}^2 \rangle = \langle \delta i_{\mathrm{s}}^2 \rangle R_{\mathrm{T}}^2 + 4kT R_{\mathrm{T}} \Delta f. \tag{A7.34}$$

The circuit in figure A7.21(*b*) offers a significant advantage of presenting a low load resistance to the detector due to the virtual earth at the input but without a corresponding Johnson noise penalty. Trans-impedance amplifiers are, therefore, widely used as pre-amplifiers, providing at the output a voltage source with low output impedance that can be connected to further amplification or processing components.

A7.5.4 Combining noise sources

More than one source of noise may well be present in a measurement and it is important to know how to combine them to arrive at a figure for the total noise. In equations (A7.31) and (A7.34), variances from different sources have been added to give a 'total variance'. This is the appropriate procedure for noise from *uncorrelated* sources. Most noise sources to be encountered, and all of those discussed previously in this chapter, are uncorrelated and can be combined in this way. Note that adding the variances is equivalent to performing a 'sum of squares' on the rms noise amplitude.

Occasionally, noise sources will be encountered that are correlated. An example of this is electromagnetic pickup from a single source by two components in the detection circuit. In such cases, the rms noise amplitudes are added.

By combining the frequency dependence of the relevant noise sources, a noise spectrum is generated. figure A7.22 shows a typical noise spectrum for a photoconductor, which is subject to flicker, G–R and thermal or shot noise.

In this situation it is clear that the lowest noise will be witnessed for high-frequency signals. For a particular measurement bandwidth, the total noise power is the integral of the noise spectrum across the appropriate range of frequencies.

A7.5.5 Bandwidth-related noise reduction methods

Having optimized the signal and detector properties, the most common method of reducing the noise in a measurement involve manipulation of thc measurement bandwidth.

The simplest example of this, for measurement of a constant signal, is the averaging of the detector output over time. The bandwidth of the measurement is thus reduced and much of the noise power is filtered out.

As figure A7.22 illustrates, however, the detection process may contribute less noise to the detector output if the measurement is performed at higher frequency. To take advantage of this, we must modulate the

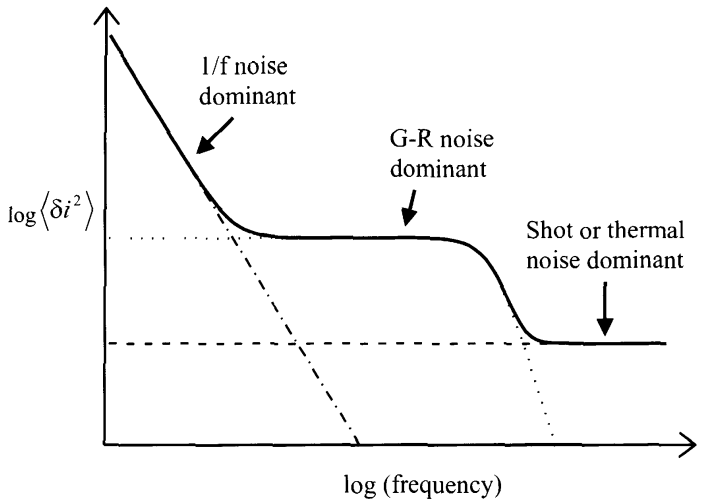


Figure A7.22. A typical frequency spectrum for noise in a photoconductor. At low frequencies (<1 kHz) flicker noise dominates; at intermediate frequencies (1 kHz < f < 1 MHz) generation–recombination noise dominates; and at high frequencies (>1 MHz) shot or thermal noise dominate. Calculation of the total noise power in a measurement is performed by integrating the noise power spectrum over the measurement bandwidth.

optical signal at the desired frequency, then isolate this frequency component in our detection process. This can be achieved readily using an optical chopper coupled with a lock-in amplifier.

Such 'phase-sensitive detection' methods can be extremely useful in identifying small signals amongst large amounts of noise and have the added advantage that unwanted background signals can often be eliminated by careful application of the initial signal modulation (for a more detailed description see [10]).

References

[1] Chuang S L 1995 *Physics of Optoelectronic Devices* (New York: Wiley) (Theory-oriented treatment of basic semiconductor detectors. Clear, friendly notation)

[2] Sze S M 1981 *The Physics of Semiconductor Devices* (New York: Wiley) (General physics of semiconductor photodetectors)

[3] Wood D 1994 *Optoelectronic Semiconductor Devices* (London: Prentice-Hall) (Good general reference, especially on pn and avalanche photodiodes)

[4] Moon J, Li S S and Lee J H 2001 *Electron. Lett.* **37** 1249

[5] Budde W 1983 *Optical Radiation Measurements Vol 4: Physical Detectors of Optical Radiation* (New York: Academic) (Most rigorous and comprehensive of the texts. Excellent coverage of photothermal detectors. A little outdated in some areas)

[6] Donati S 2000 Photodetectors (Englewood Cliffs, NJ: Prentice-Hall) (Good general reference on optical detection, good chapters on photomultipliers and semiconductor detectors. Also discusses non-demolitive detection and optical squeezing)

[7] Bachor H 1998 *A Guide to Experiments in Quantum Optics* (Weinheim: Wiley) (Good practical explanation of squeezing and homodyne detection)

[8] Loudon R 2000 *The Quantum Theory of Light* 3rd edn (Oxford: Oxford University Press) (Quantum optics bible)

[9] O'Connor D V and Phillips D 1984 *Time-Correlated Single Photon Counting* (New York: Academic) (Everything you need to know about TCSPC. Concentrates on photomultipliers rather than semiconductor detectors)

[10] Kogan Sh 1996 *Electronic Noise and Fluctuations in Solids* (Cambridge: Cambridge University Press) (Thorough treatment of noise sources in solids. Good section on $1/f$ noise)

[11] Horowitz P and Hill W 1989 *The Art of Electronics* 2nd edn (Cambridge: Cambridge University Press) (Thorough treatment of noise in amplifier circuits. Excellent overall electronics handbook. Again, not much in the way of physics. Essential reading for circuit designers)

[12] Hartley Jones M 1985 *A Practical Introduction to Electronic Circuits* (Cambridge: Cambridge University Press) (Electronics perspective, with good introduction to low noise amplifier circuits. Equations, but little physical explanation)

[13] Jenkins T E 1987 *Optical Sensing Techniques and Signal Processing* (London: Prentice-Hall) (Excellent introductory text. Especially useful for post-detection signal processing, as few other photodetection texts cover this subject)

Further reading

Battacharya P 1994 Semiconductor Optoelectronic Devices (New York: Prentice-Hall)

Good general introduction to different types of semiconductor photodetector and their performance characteristics.

Dereniak E L and Crowe D G 1984 *Optical Radiation Detectors* (New York: Wiley)

Excellent text on optical detectors.

Saleh B E A and Teich M C 1991 *Fundamentals of Photonics* (New York: Wiley)

Excellent chapter on semiconductor photodetectors and noise therein. Good explanation, equations and diagrams. Nothing on thermal detectors though.

Senior J M 1992 *Optical Fiber Communications* 2nd edn (London: Prentice-Hall).

Good coverage of semiconductor photodiodes for telecommunications.

A8
Introduction to numerical analysis for laser systems

George Lawrence

A8.1 Introduction

The detailed design of laser systems often requires numerical modelling to include all aspects of the optical system: diffraction propagation, nonlinear gain, lenses and mirrors, apertures and other optical elements. As the complexity and variety of laser systems has expanded over the years, the need for powerful analytical methods has become increasingly important. The optical engineer or scientist can determine the end-to-end performance of a complex system based on the characteristics of the lenses and mirrors, propagation distances, apertures, aberrations, laser gain and other effects.

Optical rays were the first type of optical model developed and continue to be of great use in optical design where an optical system is sufficiently short that, from front surface to rear surface, there is little diffraction spreading. In such optics, rays may be used to calculate the optical aberrations with good precision. Optical rays serve well when used in their original role as predictors of aberration error in 'well-behaved' systems, providing many digits of precision in optical path difference calculations. However, rays neglect both near-field diffraction and the interaction between intensity and wavefront gradients. It may be said that geometric ray models are precise but not accurate and that physical optics theory is accurate but not precise (due to sampling limitations).

Among the effects that are difficult to treat with rays are diffraction ripples near edge boundaries, spontaneous emission and speckle, laser modes with discontinuous phase, strong nonlinear gain, nonlinear optics, waveguides and optical fibres and wavefront discontinuities. In particular, speckles cannot be represented well by rays. In many lasers light originates from spontaneous emission and evolves through various stages of speckle size: initially fine-structured speckle and, after passing through a resonator multiple times, the high spatial frequencies are scraped out of the beam and the speckles become larger.

By far the most difficult types of laser systems to model are semiconductor diode lasers. Although the resonators for these lasers can be quite simple, consisting of a rectangular waveguide structure with plane mirrors formed by the surfaces at the ends of the waveguide, the interactions between the charge carriers that produce the laser radiation, the intense laser radiation within the waveguide, and the semiconductor materials that form the waveguide are extremely complex. As a result, even fairly sophisticated software is not capable of accurately modelling many aspects of semiconductor diode laser behaviour. The development of software that can do such modelling is currently an active research topic at a number of universities and companies.

The earliest work in resonator analysis codes was done for optical communications in the 1960s by Fox and Li [1]. The military interest in high-energy lasers and laser fusion stimulated intense development of physical optics modelling codes in the mid 1970s. The work by Siegman in 1973, and Sziklas and Siegman in 1974 studied gas dynamic lasers including diffraction, the active gain medium, apertures and aberration. The first paper by Siegman and Sziklas used an Hermite Gaussian expansion for propagation [2]. The second paper by Sziklas and Siegman used a fast Fourier transform (FFT) method for propagation [3]. A third

method based on finite difference propagation—a direct solution to the differential equation of diffraction—was used by Rench and Chester in 1974 [4]. Over time, the FFT method has become the mainstay of optical propagation codes for system analysis, as much for its modest and well-understood sensitivity to error as for its computational efficiency for many types of problems.

A8.1.1 Representation of the optical beams

For numerical solution of diffraction calculations it is convenient to use the complex amplitude, designated a, which is related to irradiance and to the electric field by $I = |a|^2 = (nc\varepsilon_0/2)|E|^2$, where c is the speed of light in vacuum and $n = \sqrt{\varepsilon/\varepsilon_0}$. In this form, the complex amplitude may take the form of square root watts per unit length.

For a full vector treatment it is necessary to have the complex amplitude fields in all three directions: the two transverse directions as well as the propagation direction. The most common cases that require a full vector treatment are strongly converging or diverging beams, scattering from features comparable to or smaller than the wavelength or dielectric waveguides with high core-to-cladding differences—that is situations in which the diffraction angles are greater than about 10 or 15 degrees for which the cosine of the diffraction angle is no longer close to unit.

Smaller diffraction angles allow a scalar treatment. This may be described as Fresnel diffraction. A small-angle scalar treatment may be generalized to consider the two transverse fields which, for propagation in the z-direction, would be a_x and a_y. Use of the two transverse fields allows representation of any polarization state. This could be considered a small-angle vector treatment but may be better described as Fresnel diffraction with polarization effects considered. By defining the relative amplitudes and phase differences between a_x and a_y, various states of polarization can be defined: linear, circular and general elliptical polarizations. Where different polarization states are not required, calculations may be performed using only one array. Most physical optics code defines two-dimensional computer arrays that represent the transverse distribution of the optical beam at a specific axial point.

The dependence of the optical beam on time may be neglected in many cases: either because the optical beam is so slowly varying that only the steady-state solution is needed; or because the pulse is so short that all physical processes in the system see only the integrated effects of the optical pulse. In either case, the time dependence may be dropped for the purpose of diffraction calculations. Time dependence may be added, if necessary, by breaking a temporal waveform or pulse into discrete time samples and propagating each time sample through the system using diffraction propagation. For example, a Q-switched laser system might be sampled over nanosecond intervals. By including time-varying phase between the temporal samples, finite temporal coherence can be modelled. For broad-band signals, it may be necessary to sample the wavelength spectrum, propagate each wavelength sample through the system independently and add them incoherently. Both temporal and wavelength sampling depend upon proper numerical calculation of strictly coherent propagation.

For small-angle scalar propagation a coherent field may be represented as

$$a(x, y; z) = a_x(x, y; z)\hat{i} + a_y(x, y; z)\hat{j}. \tag{A8.1}$$

The evolution of the optical fields is a function of diffraction and the gain and loss mechanisms in the beam train. For a detailed derivation from Maxwell's equations, the reader is referred to one of the many excellent texts on laser physics such as Sargent *et al* [5]. For many lasers, the differential equation for the optical field may be written as follows

$$\frac{\partial a}{\partial z} = -\mathrm{j}\frac{1}{2k}\nabla_\perp^2 a - \mathrm{j}\frac{\mu\omega^2}{2k}p \tag{A8.2}$$

where p represents the effect of the medium and ignoring the time variation of a and dropping $\mathrm{e}^{-\mathrm{j}\omega t}$. For

nonlinear optical effects, the medium polarization may take a more complex form, according to Bloembergen [6]:

$$p \propto \varepsilon_0 \chi a + \chi^{(2)} aa + \chi^{(3)} aaa + \cdots \tag{A8.3}$$

where the superscripts indicate the linear and various higher-order nonlinear susceptibilities. For example, the polarization term for four-wave mixing takes the form

$$p \propto \chi_{ijk} a_i a_j a_k^* \exp[\mathrm{j}(k_i + k_j - k_k \cdot z)] \tag{A8.4}$$

For linear media, equation (A8.2) takes the form

$$\frac{\partial a}{\partial z} = -\mathrm{j}\frac{1}{2k}\nabla_\perp^2 a - \mathrm{j}\frac{k\chi}{2n^2}a \tag{A8.5}$$

where n is the index of refraction in the medium. Equation (A8.5) describes the propagation of a laser beam in gain media and may also be applied to waveguide and many other forms of material. The first term on the right is the diffraction term; the second is the effect of the medium. In general this equation cannot be solved in closed form. Numerical methods are well understood for solving each of the terms on the right-hand side of equation (A8.5) if taken separately [7].

A8.1.2 Split step method

For equation (A8.5), the field after a small propagation z is given by

$$a(z + \Delta z) = a(z) + \Delta a. \tag{A8.6}$$

For small steps, the term may be separated into a diffraction and a medium term:

$$\Delta a = \Delta a_{\text{diff}} + \Delta a_{\text{medium}} \tag{A8.7}$$

$$\Delta a_{\text{diff}} = -\mathrm{j}\frac{1}{2k}\nabla_\perp^2 a \Delta z \tag{A8.8}$$

and

$$\Delta a_{\text{medium}} = -\mathrm{j}\frac{k\chi}{2n^2}a\Delta z. \tag{A8.9}$$

The solution of equation (A8.8) is performed in separate steps as shown in figure A8.1.

Even when the medium has nonlinear gain or absorption, the effect of the nonlinearities on diffraction effects is often relatively modest. The errors may be reduced to an acceptable level by taking short steps through the medium. This method is often referred to as the split-step method.

A8.1.3 Solving the diffraction part of the split-step method

The two primary means of solving the diffraction part of the split-step method are the finite difference method and angular spectrum decomposition.

A8.1.4 Finite-difference propagation

The finite difference propagator was developed by Rench [4]. The parabolic wave equation may be solved directly:

$$\frac{\partial a}{\partial z} = \frac{1}{2\mathrm{j}k}\left(\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y}\right)a. \tag{A8.10}$$

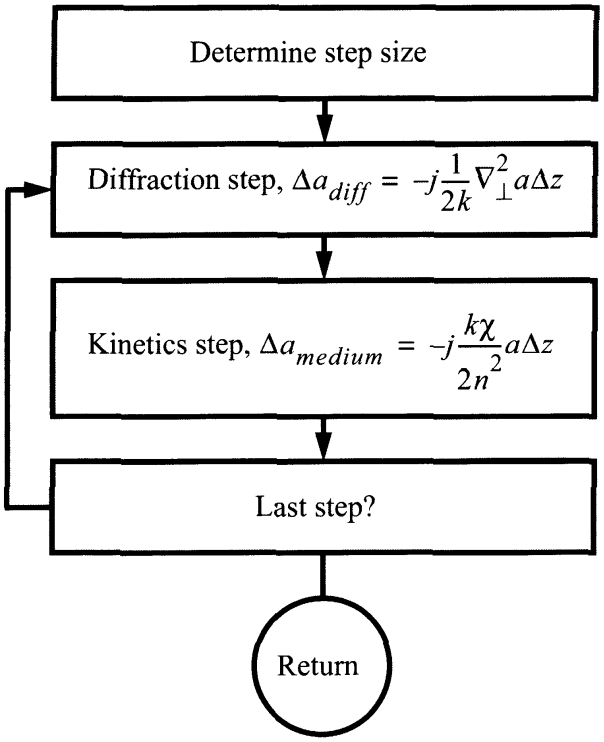


Figure A8.1. Flow chart for diffraction and kinetics routines. Diffraction calculations assume no gain and may be done by FFTs or finite-difference calculations. Kinetics calculations assume no diffraction. The split-step procedure gives a good approximation for a small step and may be repeated to accomplish arbitrarily long distances.

The second derivative is taken by considering only immediately neighbouring points, by

$$\frac{\partial^2 a}{\partial x^2} \approx \frac{a(I+1, J) + a(I-1, J) - 2a(I, J)}{2\Delta x} \tag{A8.11}$$

$$\frac{\partial^2 a}{\partial y^2} \approx \frac{a(I, J+1) + a(I, J-1) - 2a(I, J)}{2\Delta y} \tag{A8.12}$$

where $a(I, J)$ is an element of the complex amplitude array and I and J are the indices. To calculate a single point in two dimensions, four sums are needed: two subtractions and two divisions. The method is extremely fast—its principal virtue.

The longest single allowable propagation step was calculated by Rench to be

$$\Delta z < \frac{k}{2}\Delta x^2. \tag{A8.13}$$

To propagate any significant distance, the algorithm must be repeated many times. For strong nonlinear gain, it may be necessary to calculate diffraction and gain intermittently at many points along the axis, taking steps no longer than the characteristic length. In that case, the requirement for short steps with the finite difference propagator is not a problem.

The finite-difference method, as it is based on the calculation of second derivatives, may prove unstable at apertures and other sources of discontinuities in the optical field. In the case of waveguides, where the optical fields may have no true discontinuities, the method proves to be very valuable.

High Fresnel diffraction and high spatial frequencies require high sampling rates and, necessarily, large arrays. FFT computations are time consuming and the question naturally arises as to whether alternative

methods have advantages. The method most frequently considered is the finite-difference propagator (FDP). The advantages of the FDP are that it is very fast for short steps [4]. The disadvantages are;

1. one must take short steps (repeated application of the algorithm is required for long propagation),
2. the algorithm is numerically unstable at discontinuities and
3. certain diffraction effects are washed out.

This length is called the characteristic diffraction length. Since the iterative solution of the propagation problem in a nonlinear active medium requires recalculation of the diffraction effects for every characteristic length, it may be advantageous to use the finite-difference propagator.

The amplitude–wavefront representation provides an alternative method of representing the optical beam: complex amplitude $a\exp(\mathrm{j}w)$, where a is the amplitude and w is the wavefront error. In the amplitude–wavefront representations, equation (A8.11) takes the form

$$\frac{\partial a}{\partial z} = \frac{a}{2}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right) + \frac{\partial a}{\partial x}\frac{\partial w}{\partial y} + \frac{\partial a}{\partial y}\frac{\partial w}{\partial x} \tag{A8.14}$$

$$\frac{\partial w}{\partial z} = \frac{1}{2}\left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2\right] - \frac{1}{2ak^2}\left(\frac{\partial^2 a}{\partial x^2} + \frac{\partial^2 a}{\partial y^2}\right). \tag{A8.15}$$

Equations (A8.14) and (A8.15) allow direct propagation of the amplitude–wavefront representation of the beam. It is capable of representing large wavefront errors without the problem of higher order intrinsic to the complex amplitude form. It is, however, subject to the same difficulties of finite-difference propagation as equations (A8.11) and (A8.12).

A8.1.5 Angular spectrum propagation

One very effective method of calculating diffraction propagation of an arbitrary complex amplitude distribution is to decompose the distribution into a summation of plane waves, propagate the plane waves individually using the eigenvalues and re-sum the plane waves. This procedure is called the angular spectrum decomposition method [8].

The geometrical representation of the wavefront and propagation may be compared with the complex amplitude and angular spectrum propagation. Geometrical rays are normals to the wavefront. Enough rays are needed to sample the wavefront thoroughly. For example, an optical system may be traced using hundreds of rays. The ray direction is defined by wavenumber unit vector, $\hat{k}$, with a direction perpendicular to the wavefront. For free-space propagation along the ray a distance q, the ray position vector is transformed:

$$r_2 = r_1 + q\hat{k}. \tag{A8.16}$$

Propagation of a plane wave is very similar to geometric propagation. A plane wave of amplitude $A(k)$ is propagated by the equation,

$$a(k; z) = a(k; 0)\mathrm{e}^{\mathrm{j}k\cdot z}. \tag{A8.17}$$

The propagation distance depends on the direction of the plane wave. Evaluating the phase along the z-axis,

$$\mathrm{e}^{\mathrm{j}k\cdot z} = \mathrm{e}^{\mathrm{j}k_z z} \tag{A8.18}$$

where k_x, k_y and k_z are the components of the wavenumber vector:

$$k_x^2 + k_y^2 + k_z^2 = |k|^2 = k^2. \tag{A8.19}$$

We now make the approximation [8]

$$\exp(\mathrm{j}k_z z) = \exp\left[\mathrm{j}z\sqrt{k^2(1-\alpha^2-\beta^2)}\right] \approx \exp(\mathrm{j}kz)\exp[\tfrac{1}{2}\mathrm{j}kz(\alpha^2+\beta^2)]. \tag{A8.20}$$

where k_x/k and k_y/k are the direction cosines in the transverse direction. Equation (A8.20) is the transfer function for a plane wave in homogeneous, isotropic media. The term $\exp(\mathrm{j}kz)$ is generally dropped although it may be important in phased array and coupled resonator studies.

The direction cosines may be associated with spatial frequency variables ξ and η,

$$\xi\lambda = \alpha \qquad \text{and} \qquad \eta\lambda = \beta \tag{A8.21}$$

and the transfer function for a plane wave described in terms of spatial frequency variables is

$$\mathrm{e}^{\mathrm{j}k_z z} \approx \mathrm{e}^{\mathrm{j}kz}\mathrm{e}^{-\mathrm{j}\pi\lambda z\rho^2} \tag{A8.22}$$

where $\rho^2 = \xi^2 + \eta^2$.

Any well-behaved function may be written as a summation of spatial frequency components:

$$a(x, y; 0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\xi, \eta; 0)\mathrm{e}^{\mathrm{j}2\pi(x\xi+y\eta)}\,\mathrm{d}\xi\,\mathrm{d}\eta \tag{A8.23}$$

where $A(\xi, \eta; 0)$ is the spatial frequency component at the zeroth axial position and spatial frequency coordinates (ξ, η). From equation (A8.22),

$$A(\xi, \eta; z) = A(\xi, \eta; 0)\mathrm{e}^{-\mathrm{j}\pi\lambda z\rho^2}. \tag{A8.24}$$

Using equation (A8.24) propagation in homogeneous media can be written in the operator notation:

$$a(x, y; z) = FF^{-1}[T(z)FF[a(x, y; 0)]] \tag{A8.25}$$

where

$$T(z) = \mathrm{e}^{-\mathrm{j}\pi\lambda z\rho^2}. \tag{A8.26}$$

is the transfer function of diffraction propagation. The forward and inverse Fourier transforms, FF and FF^{-1}, are defined by [9]

$$FF[\,] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[\,]\exp[-\mathrm{j}2\pi(x\xi + y\eta)]\,\mathrm{d}x\,\mathrm{d}y \tag{A8.27}$$

$$FF[\,]^{-1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[\,]\exp[\mathrm{j}2\pi(x\xi + y\eta)]\,\mathrm{d}\xi\,\mathrm{d}\eta \tag{A8.28}$$

Propagation may be written as a convolution by taking the Fourier transform of equation (A8.24) [10].

$$a(x_2, y_2; z_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x_1, y_1; z_1)t(x_1 - x_2, y_1 - y_2, z_2 - z_1)\,\mathrm{d}x_1\,\mathrm{d}y_1 \tag{A8.29}$$

$$t(x, y; \Delta z) = \frac{1}{\mathrm{j}\lambda z}\exp[\mathrm{j}(kr^2/2\Delta z)]. \tag{A8.30}$$

The quantity $t(x, y; \Delta z)$ is the point spread function (PSF) or impulse response function at position Δz. Phase factors which are constant over the field have been dropped. The quadratic phase factor of equation (A8.19) can be factored to give equation (A8.31).

$$a(x_2, y_2; z_2) = \frac{1}{\mathrm{j}\lambda z}q(r_2, \Delta z)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x_1, y_1; z_1)q(r_1; \Delta z)\exp\left[-\mathrm{j}\frac{2\pi}{\lambda z}(x_1x_2 + y_1y_2)\right]\mathrm{d}x_1\,\mathrm{d}y_1 \tag{A8.31}$$

Table A8.1. A time slice of the pulselength interacts with a gain region of length L.

	Far-field	Near-field
$\Delta z \to 0$	Rapid	Slow
$\Delta z \to \infty$	Slow	Rapid

where $q(r; z) = \exp[jk(r^2/2z)]$ is a quadratic phase factor and simplifies many of the diffraction equations. In operator notation,

$$a(x_2, y_2; z_2) = \frac{1}{j\lambda\Delta z} q(r_2; \Delta z) FF^s[a(x_1, y_1; z_1) q(r_1; \Delta z)] \tag{A8.32}$$

where $s = \Delta z/|\Delta z|$.

Equations (A8.25) and (A8.32) are the near- and far-field propagation expressions. In the continuous mathematical formulation, there is no difference between the two expressions. In discrete formulation for numerical calculations, errors are reduced if the correct selection of a near- or far-field propagator is made [11]. This arises from the quadratic phase factors that must be evaluated. In the near field, the phase factor is found from equation (A8.26) and in the far field, the phase factor is found from equation (A8.30):

$$T(\Delta z) = e^{-j\pi\lambda\Delta z\rho^2} \qquad \text{and} \qquad t(x, y; \Delta z) = \frac{1}{j\lambda\Delta z} e^{j(kr^2/2\Delta z)}. \tag{A8.33}$$

The phase factor, $T(\Delta z)$, for the near field varies rapidly as $\Delta z \to \infty$ but slowly as $\Delta z \to 0$. However, $t(\Delta z)$ varies slowly as $\Delta z \to \infty$ and rapidly as $\Delta z \to 0$. Rapidly varying phase factors create numerical errors called aliasing, which are described later. The near-field propagator aliases at large propagation distances but is well behaved at short propagation distances. These relationships are summarized in table A8.1.

By using the far-field expression at long propagation distances and the near-field expression at short distances, aliasing can be reduced to a tolerable level in most cases.

In the case of strictly rotationally symmetric functions, Hankel transforms may, in principle, be used to solve the diffraction integrals. The Hankel transform pair may be written:

$$A(\rho) = 2\pi \int_0^\infty a(r') J_0(2\pi\rho r') r' \, dr' \tag{A8.34}$$

$$a(\rho) = 2\pi \int_0^\infty A(r') J_0(2\pi r\rho') \rho' \, d\rho'. \tag{A8.35}$$

Direct solution in terms of the Bessel function representation is relatively slow. The fast-Hankel transform was devised by Siegman to provide a faster method [15–18]. This method has proved successful in many cases but suffers in numerical implementation from a singularity at zero radius in both spatial and frequency domains and nonlinear sample spacing. The zero radius singularity may be minimized by good programming but may still have a tendency to exhibit 'edge droop' or 'sloping shoulders'. The nonlinear sampling results in higher sampling densities at the edge of the array which is an advantage in some cases. Given the high speed and inexpensive memory of modern computers, one may find that the higher accuracy of the methods based on square arrays makes these methods more attractive than rotational propagation methods in spite of slower speed.

An efficient circular propagator with uniform sample spacing and no zero radius singularity results in excellent energy conservation. The method is based on a degenerate form of two-dimensional Fourier

transform. In the general case, a two-dimensional Fourier transform is used to calculate a two-dimensional frequency spectrum of form $A(\xi, \eta)$. If, however, only the frequency spectrum is needed along a single row where $\eta = 0$, the two-dimensional Fourier transform may be simplified into a sum along the y-direction and a one-dimensional transform

$$A(\xi, 0) = \int\int a(x, y)\mathrm{e}^{-\mathrm{j}2\pi(x\xi+y\eta)}\,\mathrm{d}x\,\mathrm{d}y\Big|_{\eta=0} = \int a(x)\mathrm{e}^{-\mathrm{j}2\pi x\xi}\,\mathrm{d}x. \tag{A8.36}$$

where $a(x) = \int a(x)\,\mathrm{d}y \cdot a(x)$ is the sum along the y-direction. The y-sum $a(x)$ can be quickly computed from the centre row of the square array $a(x, 0)$ by interpolation to find the values at the various (x, y) points. The projection method based on y-sums may be used to calculate a Hankel transform pair between $a(x, 0)$ and $A(\xi, 0)$. For an array size of 1024×1024, an improvement in speed of between 20 and 40 times may be realized doing the one-dimensional transformation of $a(x)$ rather than the two-dimensional transforms of $a(x, y)$, but with diminished accuracy. Modern computers are so fast that for most problems the higher accuracy of the two-dimensional method makes it the most convenient choice.

A8.2 Propagation in homogeneous media

The propagation through any well-behaved system can be separated into geometrical aberration calculations and propagation in homogeneous media. The mathematically equivalent expressions of equations (A8.25) and (A8.32) provide a complete description of diffraction propagation in homogeneous media in the Fresnel approximation [11].

In numerical calculations, only discrete points may be represented. Also, only a limited region of space may be considered because of computer memory limitations. Consider a two-dimensional function represented in a rectangular computer array of $M \times N$ points. The sampling intervals for the x- and y-directions are Δx and Δy. In the general case, $M \neq N$ and $x \neq y$. The width of the computer array representation is $M\Delta x$ by $N\Delta y$. Information exists in the computer only at the discrete points defined by the rectangular grid. Any functions to be represented must be truncated by the finite width of the computer array.

The computer points in the spatial domain may be counted with the indices k and l. The indices have the ranges

$$\frac{-M}{2} \leq k \leq \frac{M}{2} - 1 \qquad \frac{-N}{2} \leq l \leq \frac{N}{2} - 1. \tag{A8.37}$$

Note that the centre of the distribution has been chosen to be at $(M/2 + 1, N/2 + 1)$. Many fast Fourier transform (FFT) routines based on arrays with dimensions which are powers of two are implemented with natural centres either at $(1, 1)$ or $(M/2 + 1, N/2 + 1)$ by shifting the array one-half cycle in each direction. The natural centre of the array is defined to be the point at which a delta function will give a perfectly constant real Fourier transform [12]. The physical limits are obtained by multiplying equation (A8.37) by Δx and Δy:

$$-\frac{M\Delta x}{2} \leq k\Delta x \leq \frac{M\Delta x}{2} - \Delta x \qquad -\frac{N\Delta y}{2} \leq l\Delta y \leq \frac{N\Delta y}{2} - \Delta y. \tag{A8.38}$$

Sampling can be represented as multiplication by a special function called the comb function. The comb function is an infinite array of delta functions spaced apart by Δx and Δy [9]:

$$\mathrm{comb}\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right) = |\Delta x||\Delta y| \sum_k \sum_l \delta(x - k\Delta x, y - l\Delta y). \tag{A8.39}$$

The comb function is useful in transforming a continuous function into a discrete representation:

$$a(x, y) \rightarrow a(x, y)\mathrm{comb}\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right). \tag{A8.40}$$

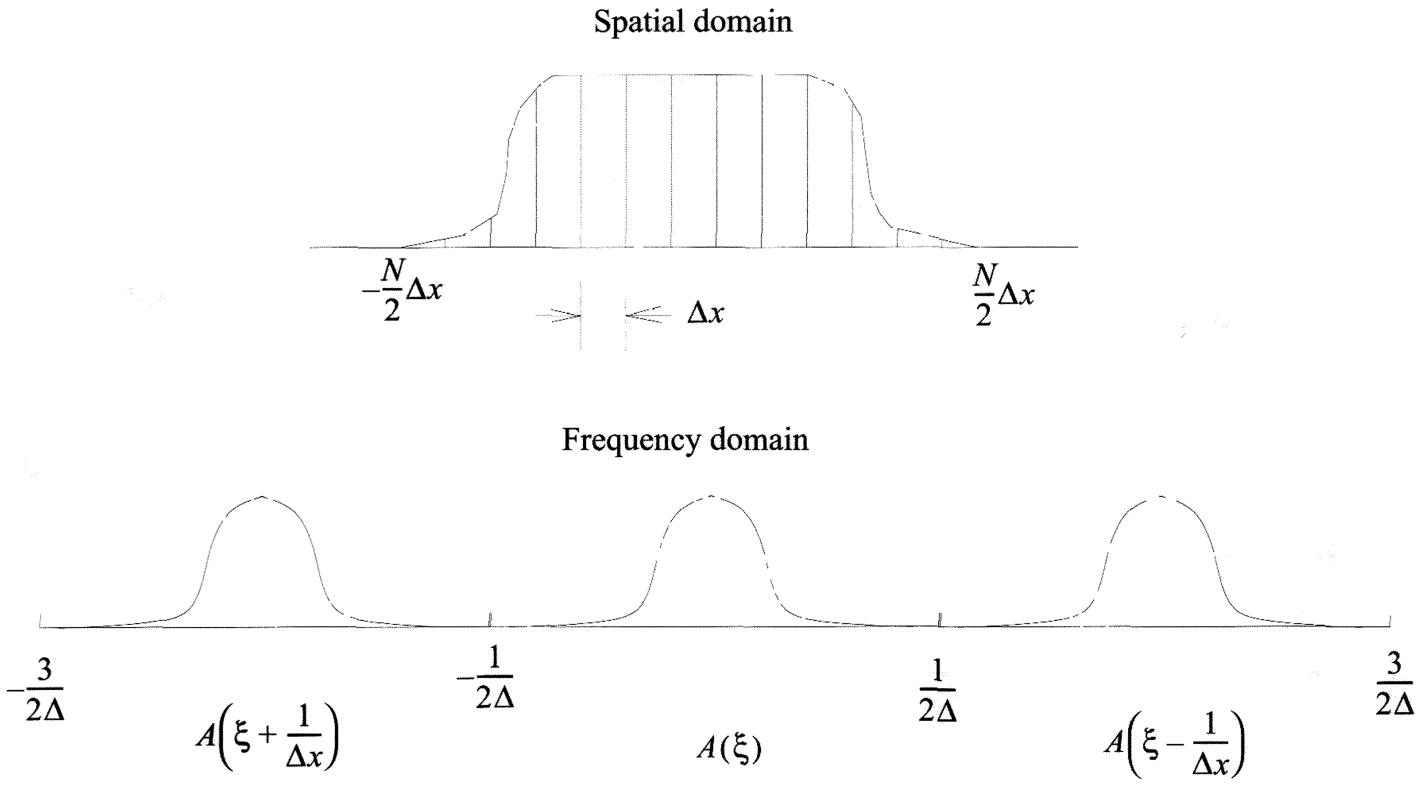


Figure A8.2. Sampling of the spatial domain causes the frequency domain to be periodic with period $(1/\Delta x, 1/\Delta y)$.

where $a(x, y)$ is the continuous function to be sampled. The arrow indicates transformation from continuous to discrete form.

The discrete nature of the spatial domain causes the frequency domain to be periodic (and necessarily of infinite extent). The continuous function $A(\xi, \eta)$, the Fourier transform of $a(x, y)$, is replicated with a period of $(1/\Delta x, 1/\Delta y)$. This is illustrated in figure A8.2.

The Fourier transform domain functions must also be discrete. The most common (and most efficient) form of the FFT has the same dimensions for the spatial and frequency domains. The frequency domain indices m and n have the ranges

$$-\frac{M}{2} \leq m \leq \frac{M}{2} - 1 \qquad -\frac{N}{2} \leq n \leq \frac{N}{2} - 1. \tag{A8.41}$$

Multiplication of equation (A8.41) by $\Delta\xi = 1/(M\Delta x)$ and $\Delta\eta = 1/(N\Delta y)$ gives the frequency range

$$-\frac{1}{2\Delta x} \leq \xi \leq \frac{1}{2\Delta x}\left(1 - \frac{2}{M}\right) \qquad -\frac{1}{2\Delta y} \leq \eta \leq \frac{1}{2\Delta y}\left(1 - \frac{2}{N}\right). \tag{A8.42}$$

The frequency domain bounds are the Nyquist sampling frequencies. The FFT algorithm is occasionally blamed for this restriction but it is more accurately attributed to the discrete sampling process and will exist for any form of propagation of sampled data.

The continuous frequency domain function is also transformed to discrete representation by means of the comb function,

$$A(\xi, \eta) \rightarrow A(\xi, \eta)\text{comb}\left(\frac{\xi}{\Delta\xi}, \frac{\eta}{\Delta\eta}\right) = A(m\Delta\xi, n\Delta\eta). \tag{A8.43}$$

The discrete nature of the frequency domain forces the spatial domain also to be periodic with period $(1/(\Delta\xi), 1/(\Delta\eta))$. These relationships are shown schematically in figure A8.3.

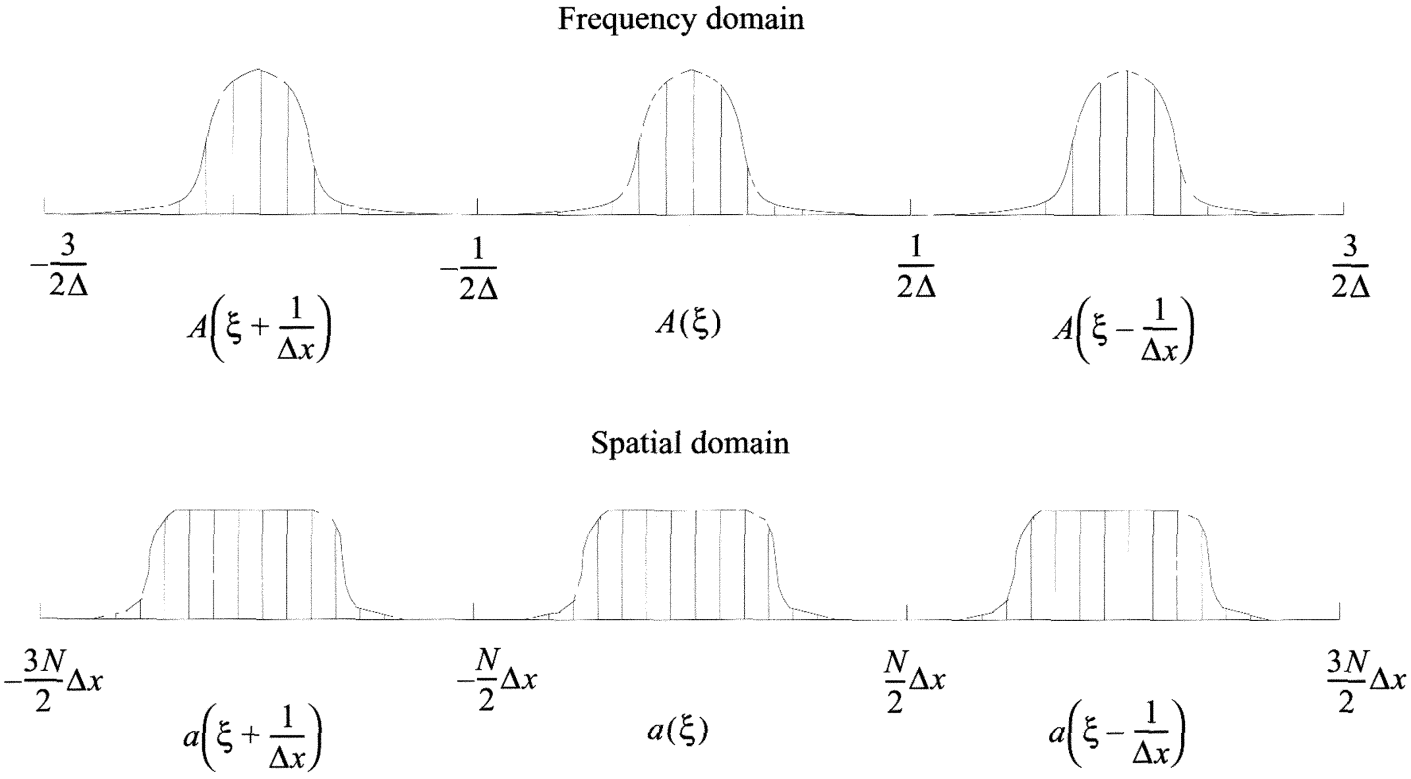


Figure A8.3. Sampling periods in the spatial and frequency domains.

A Fourier transform pair can be defined for modified spatial and frequency functions $a(k\Delta x, l\Delta y)$ and $A(m\Delta\xi, n\Delta\eta)$ such that an exact Fourier relationship is obtained:

$$a(k\Delta x, l\Delta y) \Leftrightarrow A(m\Delta\xi, n\Delta\eta) \tag{A8.44a}$$

$$a(k\Delta x, l\Delta y) \Leftrightarrow a(x, y)\text{comb}\left(\frac{x}{\Delta x}, \frac{y}{\Delta y}\right) ** |\Delta\xi\Delta n|\text{comb}(x\Delta\xi, y\Delta\eta) \tag{A8.44b}$$

$$A(m\Delta\xi, n\Delta\eta) \Leftrightarrow A(\xi, \eta)\text{comb}\left(\frac{\xi}{\Delta\xi}, \frac{\eta}{\Delta\eta}\right) ** |\Delta x\Delta y|\text{comb}(\xi\Delta x, \eta\Delta y) \tag{A8.44c}$$

where $**$ indicates two-dimensional convolution and $\Leftrightarrow$ indicates two-dimensional Fourier transformation pairs.

The function, $\text{comb}(x\Delta\xi, y\Delta\eta)$, causes the spatial domain to be periodic with minimum periods of $M\Delta x$ and $N\Delta y$ in the x- and y-directions. There is, in effect, an infinite rectangular array of functions separated by $M\Delta x$ and $N\Delta y$. Therefore, the frequency domain sampling periods are

$$\Delta\xi = \frac{1}{M\Delta x} \qquad \Delta\eta = \frac{1}{N\Delta y}. \tag{A8.45}$$

The Fourier transform operator can be written in discrete form:

$$FF[\ ] = \sum_k \sum_l [\ ] \exp\left(-sj2\pi\left(\frac{km}{M} + \frac{ln}{N}\right)\right) \tag{A8.46}$$

where s is $+1$ for forward transformation and -1 for inverse. Various forms of algorithms are used in FFTs and some have a normalization step for the forward or inverse transformation.

Evaluation of the far-field expression, equation (A8.32), in discrete terms causes a redefinition of the sampling period,

$$A(m\Delta\xi, n\Delta\eta) = FF[a(k\Delta x, l\Delta y)] \tag{A8.47a}$$

$$\Delta\xi = \frac{1}{M\Delta x_1} \qquad \Delta n = \frac{1}{N\Delta y_1} \tag{A8.47b}$$

The coordinates x_2, y_2 are related to $\Delta\xi$, $\Delta\eta$ by,

$$\Delta\xi = \frac{\Delta x_2}{\lambda|\Delta z|} \qquad \Delta\eta = \frac{\Delta y_2}{\lambda|\Delta z|} \tag{A8.48}$$

based on equation (A8.45). The discrete far-field calculation is, therefore,

$$a(k\Delta z_2, l\Delta z_2) = \frac{1}{\mathrm{j}\lambda\Delta z} q(r_2, \Delta z) FF^s[a(k\Delta x_1, l\Delta y_1) q(r_1, \Delta z)] \tag{A8.49a}$$

$$\Delta x_2 = \frac{\lambda|\Delta z|}{M\Delta x_1} \qquad \Delta y_2 = \frac{\lambda|\Delta z|}{N\Delta y_1} \tag{A8.49b}$$

$$r^2 = (k\Delta x)^2 + (l\Delta y)^2 \tag{A8.49c}$$

$$s = \frac{z}{|\Delta z|}. \tag{A8.49d}$$

Note the scale change of the new sampling periods, Δx_2 and Δy_2. The discrete near-field propagation equation is

$$a(k\Delta x, l\Delta y; z_2) = FF^{-1}[T(\Delta z) FF[a(k\Delta x, l\Delta y; z_1)]]. \tag{A8.50}$$

A8.2.1 Sampling

There are two important and related issues in determining the numerical sampling. The highest spatial frequency which can be represented in the computer is determined by the sample spacing Δx and Δy. The region of space which can be represented is determined by the width of the computer array $M\Delta x$ and $N\Delta y$. First the diffraction phenomenology, which can be observed with a given sample spacing, will be considered.

The Nyquist sampling frequency—the highest frequency which can be represented—is

$$f_{\text{Nyquist}} = \frac{1}{2\Delta x}. \tag{A8.51}$$

Failure to resolve the highest spatial frequencies may not result in an unacceptable representation of the function. In particular, using the near-field propagator for very short distances will show the distribution to be largely unchanged (the correct answer) even though the high spatial frequencies may not be correctly sampled.

Aliasing has a much more serious effect on the accuracy of the information. If the distribution grows outside the bounds of the array, severe aliasing will result which may render the calculation unusable. These errors arise from the finite size of the computer array. With propagation, a collimated beam expands and the complex amplitude grows beyond the bounds of the array and is folded back on itself, as shown in figure A8.4. This folded amplitude is the source of aliasing errors. The folded amplitude causes high spatial frequency errors in the intensity pattern, as shown in figure A8.4.

Often the most severe errors tend to be where the distribution has the highest amplitude, not near the edge of the distribution. This is because the amplitude of the signal and error add rather than the intensities. Consider a nominally top-hat function of unit amplitude and an aliasing contribution of ε. Assume that ε is

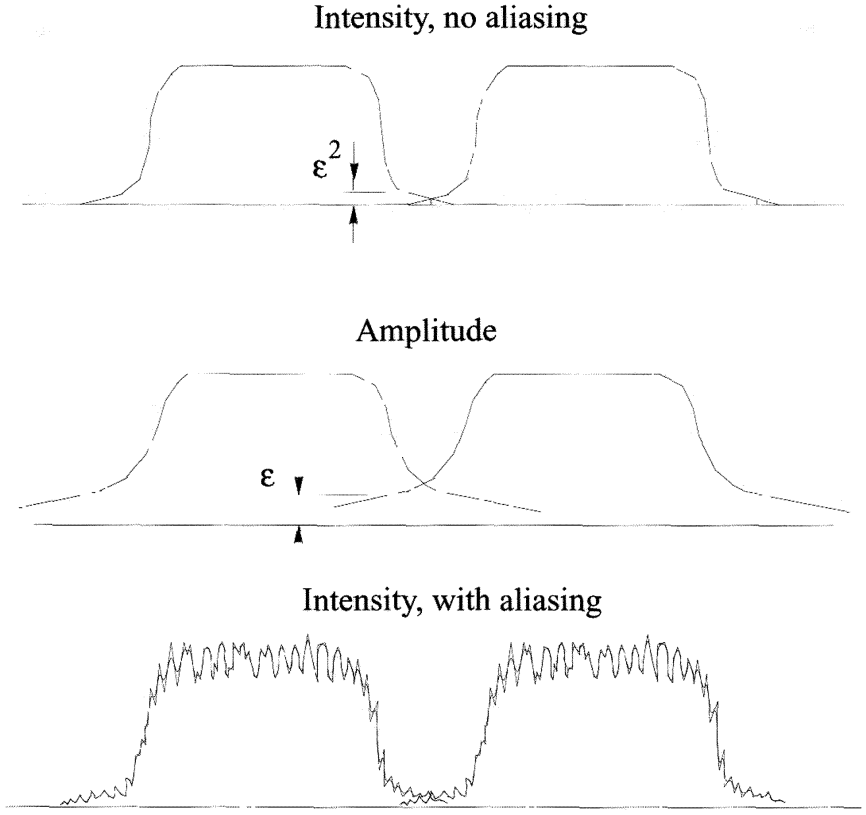


Figure A8.4. Because of the periodic nature of the discrete calculations, the amplitude in the computer array folds over into neighbouring 'ghost' arrays and *vice versa*. The intensity distribution may be at a relatively small value, as indicated in the top drawing by ε^2. However, it is the complex amplitudes that interact at the boundary. The amplitudes decay much more slowly than the intensity, as indicated by ε in the middle drawing. An intensity pattern with aliasing is shown schematically in the bottom drawing. The aliasing errors are largest where the intensity is highest because the signal boosts the amplitude error.

slowly varying across the array. Near the edge of the array the intensity is of the order of magnitude of ε^2. In the centre, the intensity is roughly $1 + 2\varepsilon + \varepsilon^2$. The error in the centre will be of order 2ε—much larger than the error at the edge of order ε^2. It is important not to be deceived into believing that the aliasing errors are negligible by seeing an intensity distribution roll-off at the edge of the array. Approximate guidelines for the magnitude of aliasing errors may be determined for top-hat functions, i.e. uniformly filled circular apertures. The results will be generally characteristic of distributions with strong discontinuities. For top-hat functions there is an exact solution based on Lommel functions [13]. The Lommel functions may be approximated by an asymptotic solution [14]. This approximation enables the calculation of aliasing errors to be made for the bright region inside the geometric aperture area and the dark region in the shadow. Because the amplitude values add—not the irradiance values—the aliasing errors are affected by the signal level. Let ε_b and ε_d be the errors in the bright and dark regions, then approximate expressions for the errors are

$$\varepsilon_b = 8\sqrt{\frac{1}{\pi^2 F_n}\frac{a^3/r^3}{(1-a^2/r^2)}} \qquad \varepsilon_d = \frac{3}{\pi^2 F_n}\frac{a^3/r^3}{(1-a^2/r^2)} \tag{A8.52}$$

These expressions give the order of magnitude of the effects. Aliasing errors are not always immediately distinguishable from diffraction ripples. Generally, high spatial frequency ripples will be manifest in the immediate vicinity of an aperture, but high spatial frequency aliasing errors will be present all over the distribution with the largest errors where the distribution has high intensity.

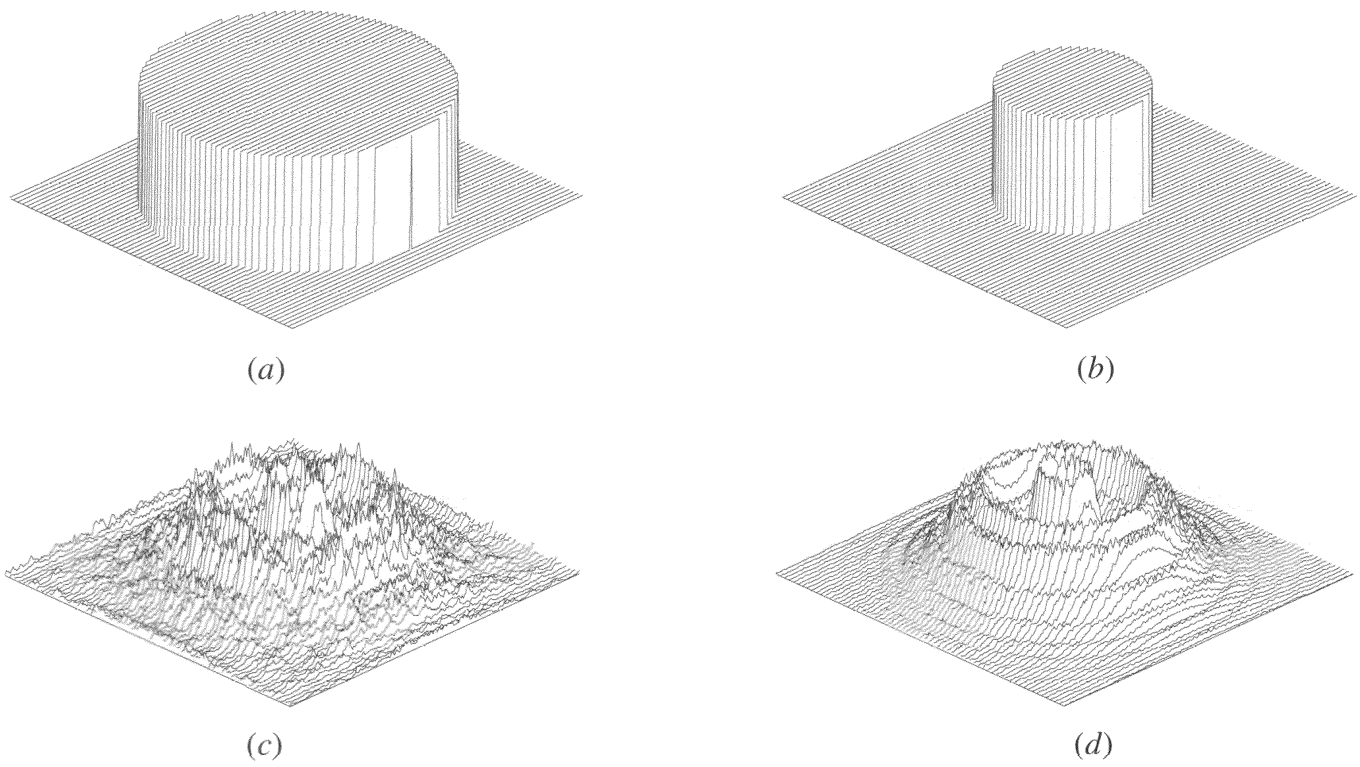


Figure A8.5. (*a*) The aperture is 0.5 cm in diameter and the width is 0.64 cm, leaving only a small guard band. (*b*) The aperture is the same size of 0.5 cm as in (*a*) but the field width is 1.28 cm, so the guard band is much greater. (*c*) After propagating 100 cm (a Fresnel number of about 3.9), the distribution shows large-scale diffraction ripples and many small aliasing ripples. Note that the ripples show up very strongly in the centre and damp down toward the edge of the array. (*d*) The beam with the large guard band, as shown in (c) has been propagated the same distance. The plot shows the distribution re-scaled to cover the same size as (c). for direct comparison. There still remains a large guardband around the distribution which is essentially empty.

Some experimentation with different size arrays and sampling may be required to gain an understanding of the appearance of the two phenomena. Consider a case of a beam of 5 mm diameter, wavelength 1.6 μm and propagation of 100 cm. An array of 128 $\times$ 128 points is selected and in the first case, shown in figure A8.5(*a*), the array is almost completely filled by the aperture (78%). In figure A8.5(*b*) the aperture is less than half the size of the array (39%). Both beams are propagated a distance of 100 cm, which is a Fresnel number of 3.9. Figure A8.5(*c*) shows the diffracted beam with the small guardband. The distribution is filled with high frequency ripples. Figure A8.5(*d*) shows the distribution with the large guardband but windowed to have the same size as figure A8.5(*c*). There is still a relatively large area surrounding the distribution of figure A8.5(*d*) and the high frequency ripples are greatly reduced but not completely absent. Examination of figures A8.5(*c*) and A8.5(*d*) shows that the aliasing errors are most noticeable in the centre of the array and not at the edges as might be expected. This is because the aliasing errors add as complex amplitude numbers to the correct distribution and are effectively boosted by the distribution in the centre of the array. It is clearly not sufficient to judge the degree of aliasing by the level of irradiance at the edge of the aperture.

Close examination of the degree of aliasing either by performing numerical experiments or using equation (A8.52) may, at first, be discouraging since most near-field diffraction calculations have significant amounts of aliasing. In practice, many calculations are not adversely affected by significant levels of aliasing. For specific problems, one can try various guardband values to determine whether results are affected. Ideally, one increases the array size and the guardband width—keeping the same number of sample points across the distribution—until no appreciable change in the results is observed. In practice, one may choose the array

size based on the computational time that is consistent with one's own level of patience.

A8.2.2 Propagation control

The beam spreads due to diffraction and may, therefore, overfill the computer array. Fortunately, the nearfield and far-field propagators may be used to control the size of the array so that the beam aliasing does not change much from the initial state. The sampling period of the near-field is constant. The sampling period of the far-field is

$$\Delta x_2 = \frac{\lambda |\Delta z|}{M \Delta x}. \tag{A8.53}$$

By use of a combination of near-field and far-field propagators, the sampling period may be set to any required value. The far-field propagation has an expanding coordinate system of the form of equation (A8.53) and the near-field propagation has a constant coordinates system of the form, $\Delta x_2 = \Delta x_1$.

A function $f(x, y)$ may be defined as the complex amplitude with respect to the curved reference surface of radius, z, such that

$$f(x, y) = a(x, y) e^{(jkr^2/2z)} \tag{A8.54}$$

where z is still the distance from the waist. Either $a(x, y)$, referenced to a plane surface, or $f(x, y)$, referenced to the curved surface, may be propagated using the equations to be presented. Either $a(x, y)$ or $f(x, y)$ may be selected depending on which has the smaller residual phase. The surrogate Gaussian beam again proves to be useful. At any point in space, the Gaussian beam is

$$a(r) = e^{(-r^2/\omega^2)} e^{(-jkr^2/2R)}. \tag{A8.55}$$

The function $f(r)$, using the curved reference, is

$$f(r) = e^{(-r^2/\omega^2)} e^{(-jkr^2/2R)} e^{(jkr^2/2z)}. \tag{A8.56}$$

The critical question is whether the residual phase of equations (A8.55) or (A8.56) is less. Consider the phase error of the actual wavefront with respect to either a planar or spherical reference surface, evaluated at the $1/e$ point of the Gaussian amplitude:

$$\Delta W_{\text{plane}} = -\frac{h^2}{2}\frac{1}{R}\bigg|_{h=\omega(z)} \qquad \Delta W_{\text{sphere}} = -\frac{h^2}{2}\left(\frac{1}{R} - \frac{1}{z}\right)\bigg|_{h=\omega(z)}. \tag{A8.57}$$

Consider the phase error for a reprcsentative Gaussian beam for the two choices of reference surface:

$$\Delta W_{\text{plane}} = -\frac{\omega_0^2}{2}\frac{z}{z_R^2} \qquad \Delta W_{\text{sphere}} = \frac{\omega_0^2}{2}\frac{1}{z} \tag{A8.58}$$

where the Gaussian beam propagation equations:

$$\omega(z) = \omega_0\sqrt{1 + \frac{z^2}{z_R^2}} \qquad \text{and} \qquad R = z + \frac{z^2}{z_R} \tag{A8.59}$$

have been used.

The phase error ΔW is minimized by choosing a plane reference inside the Rayleigh distance and a spherical reference outside the Rayleigh distance. A system for propagation between any combination of near- or far-field position to any other near- or far-field position, was developed by Lawrence [11].

A8.3 Gain and nonlinear media

Propagation through active media involves both diffraction and gain or absorption. The numerical approach to solution is the split-step method, described in previous sections. In this section, the gain part of the inhomogeneous wave equation is described.

In general, gain is described as a function of the density of the active medium and the intensity of the optical field. Medium density influences the small-signal gain and, in general, has some spatial variation. Because of saturation of the medium, gain is a nonlinear function of the intensity of the optical field.

A8.3.1 Saturated Beer's law gain

A simple model of gain using Beer's law (with a saturation intensity) may be used (see section A1.10). The saturated form of Beer's law may be represented by

$$I(z + \Delta z) = I(z) \exp\left(\frac{g_0 \Delta z}{1 + I(z)/I_{\text{sat}}}\right)^q \tag{A8.60}$$

where g_0 is the small signal gain, I_{sat} is the saturation intensity and $q = 1/2$ for inhomogeneously and $q = 1$ for homogeneously broadened gain.

The gain grows exponentially at low values,

$$\frac{\mathrm{d}I}{\mathrm{d}z} \approx g_0 I. \tag{A8.61}$$

The characteristic gain length is $1/g_0$. When I is comparable to I_{sat}, the homogeneously broadened gain takes the form

$$\frac{\mathrm{d}I}{\mathrm{d}z} \approx g_{\text{sat}} I_{\text{sat}} \tag{A8.62}$$

which is a linear increase in intensity.

A8.3.2 Rate equation model

A rate equation model for the two-level atom provides a more detailed model capable of treating laser startup and transient effects such as Q-switching (see section A1.10). The state of an active medium may be characterized by the density of the medium and the population inversion. Computer arrays may be used to store the populations density of the upper and lower level for a collection of x-, y- and z-points representing samples of the gain region. A series of transverse arrays at different axial positions may represent the gain volume. The constituent transverse arrays may be referred to as gain sheets.

A four-level treatment of gain is one of the most commonly used models, as illustrated in figure A8.6. The rate equations are [20]

$$\Delta N_2 = \left[R_2 - \frac{N_2}{t_2} - (N_2 - N_1) W_i(\nu)\right] \Delta t \tag{A8.63}$$

$$\Delta N_1 = \left[R_1 - \frac{N_2}{t_{10}} + \frac{N_2}{t_{\text{spont}}} + (N_2 - N_1) W_i(\nu)\right] \Delta t \tag{A8.64}$$

where ΔN_1 is the change in population of lower level (atoms cm^{-3}), ΔN_2 the change in population of upper level (atoms cm^{-3}), R_2 the pump rate for upper level (excitations s^{-1} cm^{-3}), R_1 the pump rate for lower level (excitations s^{-1} cm^{-3}), t_{spont} the spontaneous decay lifetime (s), t_{20} the decay time from upper level to ground (s), t_2 the total decay time from upper level to ground (s)($1/t_2 = 1/t_{20} + 1/t_{\text{spont}}$), t_{10} the decay time

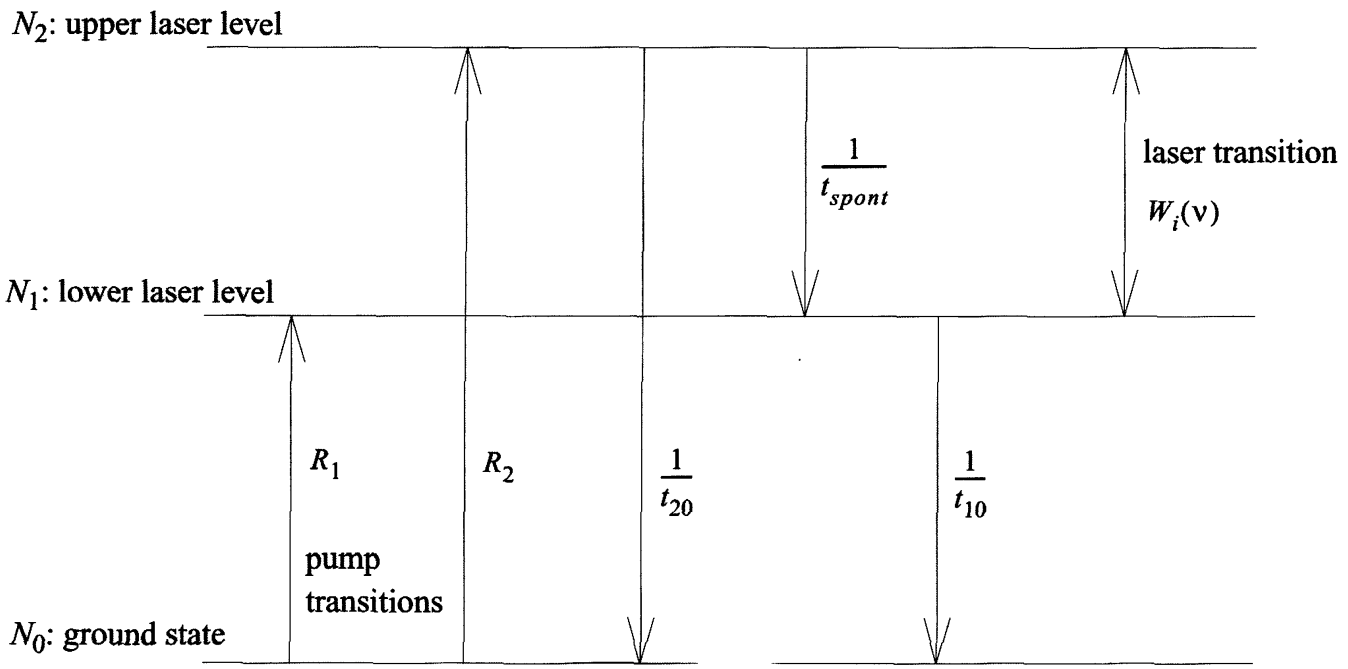


Figure A8.6. Energy transitions for a four-level atom, R_1 and R_2 are the pump rates for levels 1 and 2. t_{spont} is the lifetime for transition between levels 1 and 2. t_{10} is the lifetime for decay from level 1 to the ground state. t_{20} is the transition lifetime from level 2 to the ground state.

from lower level to ground (s), $W_i(\nu)$ the transition probability density (probability s^{-1} cm^{-3}) and Δt the elapsed time.

The transition probability density is

$$W_i(\nu) = \frac{\lambda^2 f(\nu)}{8\pi n^2 h\nu t_{\text{spont}}} I \tag{A8.65}$$

where λ is the wavelength, $f(\nu)$ the normalized lineshape, n the index of refraction, h Planck's constant, ν the frequency of the radiation and I the irradiance of the radiation.

The transition probability of equation (A8.65) may be written in terms of the Einstein B-coefficient (see section A1.6): $W_i(\nu) = B(\nu)\frac{I}{h\nu}$, where

$$B(\nu) = \frac{\lambda^2 f(\nu)}{8\pi n^2 h\nu t_{\text{spont}}} I \tag{A8.66}$$

The small-signal amplification takes the form

$$I(z) = I(0)\text{e}^{B\Delta N z}. \tag{A8.67}$$

Under steady-state conditions, the irradiance of the optical field is constant and equations (A8.63) and (A8.64) lead to the steady-state solution for the population inversion:

$$\Delta N^0 R_2 t_2 - \left(R_1 + \frac{t}{t_{\text{spont}}}\right) R_2 t_{10}. \tag{A8.68}$$

The small-signal gain coefficient is

$$g_0(\nu) = B(\nu)\Delta N^0. \tag{A8.69}$$

The gain coefficient for homogeneous broadening and for arbitrary irradiance magnitude is

$$g(\nu) = \frac{g_0(\nu)}{1 + \frac{I}{I_s}} \tag{A8.70}$$

where,

$$I_s = \frac{8\pi n^2 h\nu}{\left(\frac{t_2}{t_{\text{spont}}}\right)\lambda^2 g(\nu)} = \frac{h\nu}{B(\nu)t_2}. \tag{A8.71}$$

In the case of strong saturation, equation (A8.70) is well approximated by

$$\frac{\mathrm{d}I}{\mathrm{d}z} \approx g_0(\nu)I_s = (B\Delta N^0)\left(\frac{h\nu}{B(\nu)t_2}\right) = \frac{\Delta N^0 h\nu}{t_2}. \tag{A8.72}$$

where the pumping rate into the upper level R_2 dominates the process, equations (A8.68) and (A8.72) give the saturated gain coefficient as

$$\frac{\mathrm{d}I}{\mathrm{d}z} = R_2 h\nu,\ g(\nu) = R_2 h\nu \tag{A8.73}$$

showing, in the case of saturated steady-state gain, a linear growth of irradiance with distance based on the pumping flux density.

A8.3.2.1 Frantz–Nodvik solution

The laser may be treated as consisting of a discrete amplifier and efficiency loss due to outcoupling and other factors. The equation of optical amplification in a laser rod may be represented as [19]

$$\frac{\partial I(z)}{\partial z} = B\Delta N(z)I(z). \tag{A8.74}$$

The population inversion at each point is driven by the transition probability:

$$\frac{\partial \Delta N}{\partial z} = -2\Delta N W_i(\nu) = -2\Delta N \frac{BI}{h\nu} \tag{A8.75}$$

In a coordinate system moving with the optical field, the change of variables $t = zn/c$ may be simplified to

$$\frac{\partial \Delta N(z)}{\partial z} = \frac{-2n}{h\nu c} B\Delta N(z)I(z) \tag{A8.76}$$

where $I(z)$ is the irradiance, $N(z)$ is the population inversion and n is the index of the medium. The constant B is the cross section and has the value

$$B = \frac{\lambda^2}{8\pi t_{\text{spont}}} f(\nu) \tag{A8.77}$$

where λ is the wavelength in the medium (not the vacuum wavelength) and $f(\nu)$ is area-normalized spectral lineshape function,

$$f(\nu) = \frac{\Delta\nu}{2\pi[(V - V_0)^2 + (\frac{\Delta\nu}{2})^2]}. \tag{A8.78}$$

A method that is both fast and robust is possible by re-evaluating the problem. Equations (A8.74) and (A8.76) are appropriate for a small temporal sample of a beam travelling through an optical amplifier. In a resonator, the gain medium interacts with the entire optical field in the device. A single computer array may be used

(or at most two, for the two polarizations) to represent the entire optical field in the resonator. One cannot, therefore, distinguish temporal events which occur on a scale less than the round-trip time of the resonator. One could, in principle, use multiple temporal samples to resolve time events shorter than the round-trip time but this is not necessary for the Q-switch study.

If one considers the optical field to be of intensity I and of duration, Δt, the round-trip time, then the optical field contains a well-defined photon flux. The potential photon flux increase due to the population inversion is

$$\tfrac{1}{2}\Delta N(0)L \tag{A8.79}$$

where L is the length of the gain region, as illustrated in figure A8.7. The net energy of an incident square pulse of irradiance $I(z)$ and temporal length Δt, giving the energy density as $I(0)\Delta t$. The energy density in a gain sheet representing a length of L is $\Delta N(0)h\nu L/2$. The sum of these two energy densities is a constant by conservation of energy:

$$\text{total energy density} = I(0)\Delta t + \tfrac{1}{2}\Delta N(0)h\nu L. \tag{A8.80}$$

By dividing by Δt, one can calculate the maximum possible irradiance if all the population inversion were transformed into light:

$$I_{\max} = I(0) + \frac{\Delta N(0)h\nu L}{2\Delta t} \tag{A8.81}$$

and by dividing by $h\nu L/2$, one has the maximum population inversion if all the light were subsumed by stimulated absorption,

$$\Delta N_{\text{total}} = \Delta N(0) + \frac{2I(0)\Delta t}{h\nu L}. \tag{A8.82}$$

One can use ΔN_{total} to calculate $N(z)$:

$$\Delta N(z) = \Delta N_{\text{total}} - \frac{2I(z)\Delta t}{h\nu L}. \tag{A8.83}$$

Equation (A8.74) now takes the form

$$\frac{\partial I(z)}{\partial z} = B\left[\Delta N_{\text{total}} - \frac{2I(z)\Delta t}{h\nu L}\right] I(z). \tag{A8.84}$$

Equation (A8.84) has the exact solution

$$I(L) = \frac{I_{\max} I(0)}{I(0) + (I_{\max} - I(0))\mathrm{e}^{-B\Delta N_{\text{total}} L}}. \tag{A8.85}$$

At low saturation, equation (A8.85) approaches the expected simple exponential gain. At high saturation, equation (A8.85) approaches $I_{\max}$. Equation (A8.85) works well for both high- and low-energy amplifiers and the rate equation algorithms have been modified to use this expression.

A8.3.2.2 Off-line effects

The gain and off-line index of refraction effects may be represented by a complex index of refraction using χ'_m and χ''_m such that (see section A1.12)

$$n \to n\left(1 + \frac{\chi'_m}{2n^2} + \frac{\chi''_m}{2n^2}\right) \tag{A8.86}$$

$$\chi''_m = (N_1 - N_2)\frac{\lambda^3}{16\pi^3 t_{\text{spont}} n} f(\nu),\ \chi'_m = \frac{2(\nu_{\text{off}} + m\Delta\nu_c)}{\Delta\nu}\chi''_m \tag{A8.87}$$

$$f(\nu_m) = \frac{\Delta\nu}{2\pi[(\nu_{\text{off}} + m\Delta\nu_c)^2 + (\frac{\Delta\nu}{2})]} \tag{A8.88}$$

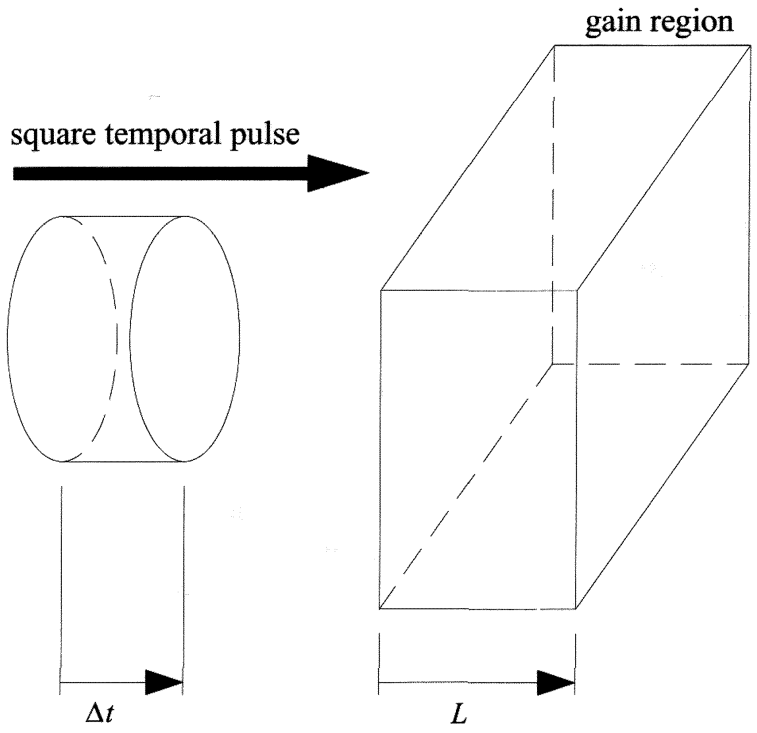


Figure A8.7. A time slice of the pulselength Δt interacts with a gain region of length L.

where $\nu_m - \nu_{cen} = \nu_{off} + m\Delta\nu_c$, and m is the mode number. The optical field, under steady-state conditions, varies as

$$a_m(x, y; \Delta t) = a_m(x, y, 0) \exp[(\mathrm{j}k\chi'_m + k\chi''_m L/2n^2)] \tag{A8.89}$$

where $\chi_m = \chi'_m - \mathrm{j}\chi''_m$ is the electric susceptibility and n is the index of refraction.

A8.3.2.3 Spontaneous emission

Spontaneous emission arises from the decay of the population inversion. This spontaneous emission is a noise source for many laser processes. The noise power injected into each mode in a distance Δz is

$$\Delta I_{noise} = \frac{(N_2 - N_1)h\Delta z}{2t_{spont}} \frac{\lambda^2}{4\pi\Delta x\Delta y} \tag{A8.90}$$

where the solid angle subtended by the computer array is $\Delta\Omega = \lambda^2/4\pi\Delta x\Delta y$ when using sampling intervals of Δx and Δy. This noise is introduced as a delta-correlated, normally distributed random phasor.

A8.4 Laser modelling software

The complexity of diffraction, laser gain, and nonlinear optical calculations are often best handled by computer programs. The consumer will find a great variety of programs available with the choice expanding year-to-year.

A8.4.1 Traditional methods of modelling

By tradition, optical system analysis falls into the general categories of ray-based, hybrid ray-based method using Gaussian beamlets that evolve along optical rays and complex amplitude representations. Ray-based

methods have proved to be the method of choice for conventional optical design where the optical designer is altering radius, thickness, glass type, aspheric coefficients, etc to minimize and balance optical aberrations. Hybrid ray-tracing methods are similar to traditional ray tracing in being capable of calculating optical path differences but have the additional capability of carrying intensity along the optical rays so that surface scattering, multi-faceted optical integrators and other non-imaging applications may be analysed. Both the traditional and hybrid ray methods have the virtue of being able to treat large aberration values without the aliasing difficulties of sampled, complex amplitude descriptions. Rays could be used to represent a flashlight beam. Rays have difficulty in representing the coupling of intensity and phase such as occurs most prominently near a focus region, near-field diffraction, laser gain and nonlinear optics.

Beam propagation methods (BPM) represent the complex amplitude on a point-by-point basis. The research literature shows an overwhelming preference for BPM for analysis of lasers and laser beam trains. BPM may be implemented with propagation by plane-wave decomposition or by finite difference methods. The point-by-point representation of the optical fields facilitates calculation of laser gain and nonlinear effects. More difficult photonic applications, such as fibres and waveguides with high core–cladding index differences, sharp bends, holey fibres and photonic crystals, may require finite difference or finite element methods. The primary limitation of BPM is that the spread of angles associated with strong aberration requires fine sample spacing and angles above about 10 or 15 degrees invalidate the simpler scalar theory.

Many commercial programs are not limited to a single method of analysis, although most programs still have their principal strength in either ray tracing or BPM. The complexity of analysis required for lasers and laser beam trains is vastly greater than for traditional optical design. Most ray-tracing analysis is a repeated application of Snell's law or the law of reflection. One needs little theoretical explanation and just a few examples may suffice.

A8.4.2 Selecting commercial numerical modelling software

Most software vendors will provide a detailed list of features and capabilities. One should consider both the overall emphasis of the software, e.g. ray tracing, BPM, etc, and the detailed list of features. The software vendor should back up claimed capabilities with numerous demonstration examples and associated explanations. A broad range of examples is also a great aid to learning how to use the software. Examples should demonstrate critical phenomena as part of the validation of the program as well as illustrating the use of program features. The thoroughness and clarity of the documentation seems to be a good predictor of overall program quality. Before buying, it is a good idea to evaluate the technical support by posing a question by email or phone.

A8.4.3 Validation of software

It is appropriate for the user to expect the software vendor to provide proof of accuracy and the user should make an effort to be familiar with such validation material. Virtually all models rely on various assumptions to simplify the calculations and such approximate models will have a finite range of parameters for which accuracy is satisfactory. While one will hear it said that certain numerical models may be trusted because they have been 'anchored to experiment', it is unwise to rely on such testimonials (with the exception of measurement of material properties). It is quite possible for unsound computer models to agree with certain data points of certain experiments but to fail for other conditions. Sound computer models will be based on theory and validation exercises, which should be designed to illustrate agreement with critical points of the theory. For example, a gain model may be validated by showing both correct small-signal gain and correct operation under strong saturation. The software builder may rely on the availability of sound and complete theory in virtually all areas of laser technology and diffraction propagation theory. Such a sound theoretical basis is the best way of ensuring accuracy for the broadest range of conditions.

The user is well advised to perform his or her own validation studies to gain familiarity and confidence with the software. For example, one might check near-field diffraction of a circular aperture at even Fresnel number to observe the zero at the centre and the centre structure. It is valuable to perform numerical experiments to observe the detail in the calculated pattern and degree of aliasing for different choices of sampling density and guardband. Laser gain may be checked for a short section of gain media (to minimize effects of diffraction) for weak input intensity to check small-signal gain and high intensity to check saturated gain. The results can be checked against hand calculations from the basic equations. If necessary, any functional feature in physical optics modelling may be checked in isolation against hand calculations.

References

[1] Fox A G and Li T 1961 Resonant modes in a maser interferometer *Bell Syst. Tech. J.* **46** 453
[2] Siegman A E 1973 Hermite Gaussian functions of complex argument as optical beam eigenfunctions *J. Opt. Soc. Am.* **63** 1093
[3] Sziklas E A and Siegman A E 1974 Diffraction calculations using fast Fourier transform methods *Proc. IEEE* **62** 410–12
[4] Rench D B and Chester 1974 Three dimensional unstable resonators with laser medium *Appl. Opt.* **13** 2546–61
[5] Sargent M, Scully M and Lamb W 1974 *Laser Physics* (Reading, MA: Addison-Wesley)
[6] Bloembergen N 1965 *Nonlinear Optics* (Reading, MA: Benjamin)
[7] Hardin R H and Tappert F D 1973 Applications of the split step Fourier method to the numerical solution of nonlinear and variable coefficient wave equations *SIAM Rev.* **15** 423
[8] Goodman J W 1968 *Introduction to Fourier Optics* (New York: McGraw-Hill)
[9] Gaskill J 1976 *Linear Systems, Transforms, and Optics* (New York: Academic) p 139
[10] Kraus H 1989 Huygens Fresnel Kirchoff wave front diffraction formulation: spherical waves *J. Opt. Soc. Am.* A **6** 1196
[11] Lawrence George N 1992 *Optical Modelling (Applied Optics and Optical Engineering XI)* ed R Shannon and J Wyant (New York: Academic) pp 125–200
[12] Hayes J 1992 *Fast Fourier Transforms and their Applications (Applied Optics and Optical Engineering 11)* ed R Shannon and J Wyant (New York: Academic)
[13] Born M and Wolf E 1965 *Principles of Optics* (New York: Pergammon)
[14] Lawrence G 1980 Optical performance analysis of CO_2 laser fusion systems *Doctoral Dissertation* University of Arizona
[15] Siegman A E 1977 Quasi fast Hankel transform *Opt. Lett.* **1** 13–15
[16] Sheng S-C 1980 Studies of laser resonators and beam propagation using fast transfrom methods *PhD Dissertation* ch 3, Department of Applied Physics, Stanford University
[17] Sheng S-C and Siegman A E 1980 Nonlinear optical calculations using fast transform methods: Second harmonic generation with depletion and diffraction *Phys. Rev.* A **21** 599–606
[18] Oppenheim A V, Frisk G V and Martinez G R 1980 Computation of the Hankel transform using projections *J. Acoust. Soc. Am.* **68** 523–9
[19] Frantz L M and Nodvik J S 1963 Theory of pulse propagation in a laser amplifier *J. Appl. Phys.* **34** 2346–9
[20] Yariv A 1976 *Introduction to Optical Electronics* (New York: Holt, Rinehart, and Winston)
[21] Hader J *et al* 2002 Semiconductor quantum-well designer active materials *Opt. Photon. News* special issue 'Photonics in 2002'