Optoelectronics

Emmanuel Rosencher and
Borge Vinter

CAMBRIDGE

more information - www.cambridge.org/0521771293

This page intentionally left blank

Optoelectronics

Optoelectronics is a practical and self-contained graduate-level textbook and reference, which will be of great value to both students and practising engineers in the field. Sophisticated concepts are introduced by the authors in a clear and coherent way, including such topics as quantum mechanics of electron–photon interaction, quantization of the electromagnetic field, semiconductor properties, quantum theory of heterostructures, and non-linear optics. The book builds on these concepts to describe the physics, properties, and performances of the main optoelectronic devices: light emitting diodes, quantum well lasers, photodetectors, optical parametric oscillators, and waveguides. Emphasis is placed on the unifying theoretical analogies of optoelectronics, such as equivalence of quantization in heterostructure wells and waveguide modes, entanglement of blackbody radiation and semiconductor statistics. The book concludes by presenting the latest devices, including vertical surface emitting lasers, quantum well infrared photodetectors, quantum cascade lasers, and optical frequency converters.

Optoelectronics

Emmanuel Rosencher

Research Director
French Aerospace Research Agency (ONERA, France)
Professor at the Ecole Polytechnique (Paris, France)

Borge Vinter

Senior Scientist
THALES Research and Technology

Translated by Dr Paul G. Piva

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published in printed format 2002

ISBN 0-511-03423-7 eBook (Adobe Reader)
ISBN 0-521-77129-3 hardback
ISBN 0-521-77813-1 paperback

Originally published in French as *Optoélectronique* by Emmanuel Rosencher and Borge Vinter, Paris 1998 and © Masson 1998

For Nadia, Anne, Julien, and Clara, for their patience with all my love.

For Nadia who understands so many other things.

Contents

Preface

The field of optoelectronics is currently in full expansion, drawing to its classrooms and laboratories numerous science and engineering students eager to master the discipline. From the lecturer's perspective, optoelectronics is a considerable challenge to teach as it emerges from a complex interplay of separate and often seemingly disjointed subjects such as quantum optics, semiconductor band structure, or the physics of carrier transport in electronic devices. As a result, the student (or lecturer) is left to navigate through a vast literature, often found to be confusing and incoherent.

The aim of this text is to teach optoelectronics as a science in itself. To do so, a tailored presentation of its various sub-disciplines is required, emphasizing within each of these, those concepts which are key to the study of optoelectronics. Also, we were determined to offer a partial description of quantum mechanics oriented towards its application in optoelectronics. We have therefore limited ourselves to a utilitarian treatment without elaborating on many fundamental concepts such as electron spin or spherical harmonic solutions to the hydrogen atom. On the other hand, we have placed emphasis on developing formalisms such as those involved in the quantization of the electromagnetic field (well suited to a discussion of spontaneous emission), or the density matrix formalism (of value in treating problems in non-linear optics).

Similarly, our treatment of semiconductor physics ignores any discussion of the effect of the crystallographic structure in these materials. Rather, a priori use is made of the semiconductor band structures which implicitly incorporate these effects on the electrical and optical properties of these materials. In carrying out our rather utilitarian-minded presentation of these disciplines, we have claimed as ours Erwin Schrödinger's maxim that it mattered little whether his theory be an exact description of reality insofar as it proved itself useful.

We have sought in this work to underline wherever possible the coherence of the concepts touched on in each of these different areas of physics, as it is from this vantage point that optoelectronics may be seen as a science in its own right. There exists, for instance, a profound parallel between the behaviour of an electron in a quantum well and that of an electromagnetic wave in an optical waveguide. As well, one finds between the photon statistics of black bodies, the mechanics of quantum transitions within semiconductor band structures and the statistics of

charge carriers in these materials, an entanglement of concepts comprising the basis for infrared detection. In the same spirit, this work does not pretend to present an exhaustive list of all known optoelectronic devices. Such an effort could only come at the cost of the overall coherence aimed at in this work, and add to the type of confusion we have claimed as our enemy. The goal is rather to present those optoelectronic concepts which will allow an overall understanding of principles necessary in solving problems of a general or device-specific nature. Thus, only the analysis of *generic classes* of optoelectronic components will be undertaken here without entering into the labyrinth offered by more particular applications.

Lastly, regarding the problem of notation (a problem inherent to any multidisciplinary study), we have chosen simply to follow the lead of standard physics notation in any given chapter. Thus, the symbol 'ε' may be used indiscriminately to represent the permittivity, the quantum confinement energy, or the saturation coefficient of a semiconductor laser. We could have attempted the introduction of various notations for each of these different uses based on the Latin, Greek, and Hebrew character sets, but we realized that even these would have soon been exhausted. We have thus chosen merely to redefine in each chapter the correspondence between the symbols and their respective notions.

The authors wish to thank all those having assisted with the preparation of this manuscript, such as Erwan LeCochec, Andrea Fiore, Arnaud Fily, Jean-Yves Duboz, Eric Costard, Florence Binet, Eric Herniou, Jean-Dominique Orwen, Anna Rakovska, and Anne Rosencher among many others. This work could never have seen the light of day without the support of ONERA and THALES (ex THOMSON-CSF) and most particularly the encouragement of Mr Pierre Tournois, formerly scientific director of THOMSON-CSF. Finally, the authors are deeply indebted to Paul Piva, whose translation from French to English reflects his competence, intelligence, and culture.

Properties of common semiconductors

	Si	Ge	GaAs	AlAs	InAs	GaP	InP	GaSb	InSb
Bandgap									
E_g (eV)	indirect	indirect	direct	indirect	direct	indirect	direct	direct	direct
@ $T = 0$ K	1.170	0.744	1.519	2.229	0.418	2.350	1.424		0.236
@ $T = 300$ K	1.124	0.664	1.424	2.17	0.354	2.272	1.344	0.70	0.18
Lattice constant, a_0 Å	5.43095	5.64613	5.6533	5.6600	6.0583	5.4505	5.8688	6.096	6.4794
Relative permittivity, $\varepsilon_{st}/\varepsilon_0$	11.9	16.2	13.1	10.06	15.15	11.1	12.56	15.69	16.8
Effective mass									
Electron longitudinal, m_{el}/m_0	0.9163	1.59	0.067	0.15(Γ)	0.023	0.254	0.073	0.047	0.014
Electron traverse, m_{et}/m_0	0.1905	0.0823				4.8			
Heavy hole, m_{hh}/m_0	0.537	0.284	0.50	0.79	0.40	0.67	0.60	0.8	0.42
Light hole, m_{lh}/m_0	0.153	0.043	0.087	0.15	0.026	0.17	0.12	0.05	0.016
Luttinger parameters									
γ_1	4.25	13.4	7.0	3.45	20.4	4.05	5.04	13.3	40.1
γ_2	0.32	4.3	2.3	0.68	8.3	0.49	1.6	4.4	18.1
γ_3	1.45	5.7	2.9	1.3	9.1	1.25	2.4	6.2	19.2
Intrinsic density, n_i (cm^{-3})	1.5×10^{10}	2.4×10^{13}	1.8×10^{6}		1.3×10^{15}	3.0×10^{6}	1.2×10^{8}	4.3×10^{12}	2.0×10^{16}
Mobility									
Electron, μ_e ($cm^2\,Vs^{-1}$)	1450	3900	8000	400	30 000	200	5000	5000	80 000
Hole, μ_h ($cm^2\,V^{-1}\,s^{-1}$)	370	1800	400	100	480	150	180	1500	1500

Further reading

General references useful in obtaining values for semiconductor properties:

K. H. Hellwege, ed., *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology*, Springer, Berlin.

O. Madelung, ed., *Semiconductors, Group IV Elements and III-V Compounds, in Data in Science and Technology*, Springer, Berlin (1996).

Recent review works:

B. L. Weiss, ed., *EMIS Dataviews Series*, INSPEC, London.

1 Quantum mechanics of the electron

1.1 Introduction

This chapter reviews the fundamental principles and techniques of quantum mechanics that are necessary to understand the subject of optoelectronics. Often, concepts are not presented in depth: the aim, rather, is to provide the tools and notation required to work through this book. Thus, in spite of their immense importance in other areas of physics, and the severe scientific injustice resulting from their being placed aside, we shall pass almost entirely in silence over Heisenberg's uncertainty principle, spherical harmonics, electron spin, etc. The reader wishing to deepen his/her understanding of these concepts is greatly encouraged to read or reread the remarkable work by C. Cohen-Tannoudji et al. (1992).

1.2 The postulates of quantum mechanics

We consider an electron of charge q and mass m_e subjected to a generalized potential of the form $V(\boldsymbol{r}, t)$ varying in three-dimensional space $\boldsymbol{r}$, and time t. Quantum mechanics tells us that the notion of a classical electron trajectory loses its meaning when the distance over which this potential varies is of the order of the *de Broglie wavelength* (λ_{DB}). This length is given by:

$$\lambda_{\mathrm{DB}} = \frac{2\pi\hbar}{\sqrt{2m_e E}} \approx \frac{1.23\,(\mathrm{nm})}{\sqrt{V\,(\mathrm{V})}} \tag{1.1}$$

where $\hbar$ is *Planck's constant* ($1.04 \times 10^{-34}\,\mathrm{J\,s^{-1}}$), V is the average potential experienced by the particle, and E is the energy of the particle. We will see that in a crystalline solid where electrons are subjected to spatially varying potentials of the order of 5 eV ($1\,\mathrm{eV} = 1.6 \times 10^{-19}\,\mathrm{J}$), their *de Broglie* wavelength turns out to be of the order of 5 Å. As this length corresponds to the interatomic distance between atoms in a crystalline lattice, conduction electrons in this medium will be expected to display interference effects specific to the mechanics of wave-motion. These effects (studied in Chapter 5) are the origin of the semiconductor band gap, and cannot readily be discussed in terms of classical theories based upon the notion of a well-defined trajectory.

Quantum mechanics also teaches us that we must forgo the idea of a trajectory

in favour of a more subtle description in terms of *quantum states* and *wavefunctions*. The electron is then represented by a *state vector* evolving in time $|\psi(t)\rangle$. One of the strongest postulates of quantum mechanics is that all these state vectors span a *Hilbert space*. For instance, the existence of linear combinations of states (which leads to dramatic effects such as molecular stability, energy bandgaps, . . .) is a direct consequence of this postulate. This vector space possesses a *Hermitian scalar product*, whose physical significance will be given later. We will use *Dirac notation* to represent the scalar product between two vector states $|\psi_1\rangle$ and $|\psi_2\rangle$ as $\langle\psi_2|\psi_1\rangle$. Now, we recall the properties of a Hermitian scalar product:

$$\left.\begin{array}{l}\langle\phi|\psi\rangle = \langle\psi|\phi\rangle^* \\ \langle\phi|\alpha\psi_1 + \beta\psi_2\rangle = \alpha\langle\phi|\psi_1\rangle + \beta\langle\phi|\psi_2\rangle \\ \langle\alpha\phi_1 + \beta\phi_2|\psi\rangle = \alpha^*\langle\phi_1|\psi\rangle + \beta^*\langle\phi_2|\psi\rangle \\ \langle\psi|\psi\rangle \text{ real, positive, and zero if and only if } |\psi\rangle = 0\end{array}\right\} \tag{1.2}$$

where the asterisk indicates that the complex conjugate is taken. By definition a physical state possesses a norm of unity, which is to say that $|\psi(t)\rangle$ is a physical state if:

$$\langle\psi(t)|\psi(t)\rangle = 1 \tag{1.3}$$

A certain number of linear operators act within this Hilbert space. A second postulate of quantum mechanics is that classically measurable quantities such as position, energy, etc. are represented by Hermitian operators A (i.e. operators such that $A^\dagger = A$, where $\dagger$ is the *adjoint* or *Hermitian conjugate*) called *observables*, and that the result of the measurement of such an observable can only be one of the eigenvalues associated with the observable. If the ensemble of eigenvalues of the observable A forms a discrete set, then the set of all possible *measurements* of a system are given by the a_n solutions of the eigenvalue equation:

$$A|\psi_n\rangle = a_n|\psi_n\rangle \tag{1.4}$$

As the observable operators are Hermitian, it follows that their eigenvalues are necessarily real (consistent with the familiar fact that the result of a physical measurement is a real number). We also define the commutator of two operators A and B as:

$$[A, B] = AB - BA \tag{1.5}$$

It can be shown that if two operators commute (i.e. if their commutator equals zero), then they share a complete set of simultaneous eigenvectors. A noteworthy consequence of this is that physical states exist in which the results of measurement of both of these observables (A and B) can be obtained simultaneously with certainty: these are their common eigenstates.

If the orthonormal eigenvector basis of observable A is complete, then any physical state $|\psi(t)\rangle$ of the electron can be described in terms of a linear combination of eigenvectors:

$$|\psi(t)\rangle = \sum_n c_n(t)|\psi_n\rangle \tag{1.6}$$

where the coefficients c_n are given by:

$$c_n(t) = \langle\psi_n|\psi(t)\rangle \tag{1.7}$$

The *probabilistic interpretation* of quantum mechanics states that the square of the norm of the coefficient $|c_n(t)|^2$ gives the probability of finding the electron in the $|\psi_n\rangle$ state at time t (implying that measurement of the observable A at that time will yield the value of a_n with equal probability $|c_n(t)|^2$). A further postulate is that, immediately after a measurement of observable A has been performed, the state function resides entirely in one of the eigenstates of the observable A (i.e. $c_n(t) = 1$ or $|\psi\rangle = |\psi_n\rangle$). In the event that a particular eigenvalue is degenerate, the state function after measurement is restricted to the subspace spanned by the degenerate eigenstates. The latter postulate, which is still the subject of intense investigation, is necessary for the coherence of quantum mechanics.

It is therefore implicit in the probabilistic interpretation that we may not, in general, know the outcome of a measurement with certainty. We can, however, extract the average value of an observable A taken over the course of a statistically significant number of independent measurements. This value will then correspond to an average value of all possible measurement outcomes a_n of an observable weighted by the individual probabilities $|c_n(t)|^2$ of finding the system in an eigenstate $|\psi_n\rangle$ associated with this particular eigenvalue a_n:

$$\langle A\rangle(t) = \sum_n a_n|c_n(t)|^2 \tag{1.8}$$

This average value is easily found to be:

$$\langle A\rangle(t) = \langle\psi(t)|A|\psi(t)\rangle \tag{1.9}$$

Some of these A observables may be vectorial, such as the position $\hat{\mathbf{r}} = (x, y, z)$ and momentum $\hat{\mathbf{p}}$ operators. For these operators, the eigenvalues belong to a continuum of values. Therefore, the eigenvector $|\mathbf{r}\rangle$ of the position operator $\hat{\mathbf{r}}$ is interpreted as describing the state of the system once the measurement of the position has yielded a particular value $\mathbf{r}$. We then say that the particle may be found at $\mathbf{r}$ with certainty.

The decomposition of a state vector onto any particular basis set of eigenvectors

is called a *representation*. One important representation is the projection of the state vector onto the eigenstates of the position operator $\hat{\mathbf{r}}$. Each component of this projection is the wavefunction $\psi(\mathbf{r}, t)$ given by:

$$\psi(\mathbf{r}, t) = \langle \mathbf{r} | \psi(t) \rangle \tag{1.10}$$

Referring back to the probabilistic interpretation of quantum mechanics, we see that the norm of the wavefunction $|\psi(\mathbf{r}, t)|^2$ gives the probability of finding an electron at $\mathbf{r}$ at time t. Furthermore, in the $\mathbf{r}$ representation, the inner product of the two states $|\psi_1\rangle$ and $|\psi_2\rangle$ may be shown to be written as:

$$\langle \psi_1 | \psi_2 \rangle = \int \psi_1^*(\mathbf{r}) \psi_2(\mathbf{r}) \, \mathrm{d}^3\mathbf{r} \tag{1.11}$$

where the integral is evaluated over all space. Finally, evolution of the state of the system with time is given by Schrödinger's equation:

$$\mathrm{i}\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \hat{H}(t) |\psi(t)\rangle \tag{1.12}$$

Schrödinger's equation

where $\hat{H}(t)$ is the Hamiltonian of the system, which yields as an observable the *energy*[1] of the system. Its general expression takes the form:

$$\hat{H}(t) = \frac{\hat{\mathbf{p}}^2}{2m_e} + V(\mathbf{r}, t) \tag{1.13}$$

Hamiltonian for a particle with mass m_e subject to a potential V

where $\hat{\mathbf{p}}$ is the momentum operator. In the $\mathbf{r}$ representation (i.e. projected onto the position eigenvectors of $|\mathbf{r}\rangle$), the *correspondence principle* gives the following expression for the $\hat{\mathbf{p}}$ operator:

$$\hat{\mathbf{p}} = \frac{\hbar}{\mathrm{i}} \nabla = \frac{\hbar}{\mathrm{i}} \begin{bmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{bmatrix} \tag{1.14}$$

and in the $\mathbf{r}$ representation, takes the following form when acting upon a wavefunction $\psi(\mathbf{r}, t)$:

[1] The symbol ˆ is generally used when confusion may arise between a classical physical quantity (such as position $\mathbf{r}$) and its corresponding quantum observable ($\hat{\mathbf{r}}$).

$$\langle \mathbf{r}|\hat{\mathbf{p}}|\psi(t)\rangle = \hat{\mathbf{p}}\psi(\mathbf{r},t) = \frac{\hbar}{\mathrm{i}}\begin{bmatrix} \dfrac{\partial}{\partial x}\psi(\mathbf{r},t) \\ \dfrac{\partial}{\partial y}\psi(\mathbf{r},t) \\ \dfrac{\partial}{\partial z}\psi(\mathbf{r},t) \end{bmatrix} \tag{1.15}$$

This correspondence results from the requirement that the $\hat{\mathbf{p}}$ operator acting upon the *de Broglie* matter-waves ($e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$) yields an associated momentum eigenvalue of $\mathbf{p} = \hbar\mathbf{k}$, i.e. satisfying the de Broglie relation. The momentum operator $\hat{\mathbf{p}}$ therefore takes on the form $\hbar/\mathrm{i}\nabla$.

The operators $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ are linked by the important commutator relation:

$$\hat{\mathbf{r}}\cdot\hat{\mathbf{p}}\psi(\mathbf{r},t) = \frac{\hbar}{\mathrm{i}}\left[x\frac{\partial}{\partial x}\psi(\mathbf{r},t) + y\frac{\partial}{\partial y}\psi(\mathbf{r},t) + z\frac{\partial}{\partial z}\psi(\mathbf{r},t)\right] \tag{1.16}$$

and

$$\hat{\mathbf{p}}\cdot\hat{\mathbf{r}}\psi(\mathbf{r},t) = \frac{\hbar}{\mathrm{i}}\left\{\frac{\partial}{\partial x}[x\psi(\mathbf{r},t)] + \frac{\partial}{\partial y}[y\psi(\mathbf{r},t)] + \frac{\partial}{\partial z}[z\psi(\mathbf{r},t)]\right\} \tag{1.17}$$

from which we deduce the commutation relation:

$$[\hat{x}_i, \hat{p}_j] = \mathrm{i}\hbar\delta_{ij} \tag{1.18}$$

Anticommutation of position and momentum observables leading to the first of the Heisenberg uncertainty relations

A corollary of the properties stated earlier for commuting observables, is that non-commuting observables cannot share a common basis set of eigenvectors. Therefore, neither of these position or momentum observables may be known simultaneously with arbitrary precision. This is the first of Heisenberg's uncertainty principles, which can be shown to lead to the following relationships between the momentum and position uncertainties:

$$\left.\begin{aligned} \Delta x\Delta p_x &\geq \hbar/2 \\ \Delta y\Delta p_y &\geq \hbar/2 \\ \Delta z\Delta p_z &\geq \hbar/2 \end{aligned}\right\} \tag{1.19}$$

Returning to Schrödinger's equation in the position representation, we may now write:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\psi(\mathbf{r},t) = -\frac{\hbar^2}{2m_e}\Delta\psi(\mathbf{r},t) + V(\mathbf{r},t)\psi(\mathbf{r},t) \tag{1.20}$$

where Δ is the Laplacian operator $((\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2))$. Once given the

space and time evolution of the potential, this last equation allows one, in principle, to calculate the evolution of electron probability in the structure. We note that this equation preserves the norm of a function, which is consistent with the fact that every physical state evolves in time and space to some other physical state.

1.3 The time-independent Schrödinger equation

1.3.1 Stationary states

We will interest ourselves, in this section, with the description of the physical state of an electron subjected to a time-independent potential (i.e. a *conservative* system). This system could be a hydrogen atom, in which case the potential $V(\mathbf{r})$ is a Coulomb field localized in space, or a crystal, where the potential $V(\mathbf{r})$ is periodic (corresponding to the regular spacing of the constituent atoms). Schrödinger's equation may then be written as:

$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle = \left[\frac{\hat{\mathbf{p}}^2}{2m_e} + V(\mathbf{r})\right]|\psi(t)\rangle \tag{1.21}$$

Let us first begin by considering the eigenstates of the Hamiltonian:

$$\hat{H}|\psi_n(t)\rangle = E_n|\psi_n(t)\rangle \tag{1.22}$$

Time-independent Schrödinger equation

For the time being we will suppose that these states are:

- discrete, i.e. they can be denoted by integers;
- non-degenerate, i.e. no two or more distinct quantum states may have the same energy;
- complete, i.e. any physical state may be projected in a unique fashion onto the basis set formed by the eigenfunctions of $\hat{H}$ of type (1.6).

Substituting Eq. (1.22) into (1.21), we find the time evolution of an eigenstate $|\psi_n\rangle$ to be:

$$|\psi_n(t)\rangle = |\psi_n(0)\rangle e^{-i\omega_n t} \tag{1.23}$$

where

$$E_n = \hbar\omega_n \tag{1.24}$$

and ω_n is the *Bohr oscillation frequency* associated with the state $|\psi_n\rangle$. Equation (1.23) is noteworthy as it allows an important prediction to be made. Let us suppose the system is in an eigenstate $|\psi_n\rangle$ and that we seek the average value of some observable A:

$$\langle A\rangle(t) = \langle \psi_n e^{-i\omega_n t}|A|\psi_n e^{-i\omega_n t}\rangle = \langle \psi_n|A|\psi_n\rangle \tag{1.25}$$

This average value therefore does not vary over time, i.e. the eigenstates are *stationary* states for all observables. These stationary states are particularly important as they form states which yield unchanging values for observables. Additionally, they allow a description of the time evolution of a non-stationary state. Let us suppose an arbitrary state $|\psi(t)\rangle$, for which we know its projection at $t = 0$ onto the basis set of stationary states $|\psi_n\rangle$:

$$|\psi(0)\rangle = \sum_n c_n|\psi_n\rangle \tag{1.26}$$

We then determine the time evolution of the coefficients $c_n(t)$. To do this we substitute the $|\psi(t)\rangle$ stationary state decomposition into the time-independent Schrödinger equation (1.21), which gives:

$$\sum_n i\hbar \frac{d}{dt} c_n(t)|\psi_n\rangle = \hat{H}\left[\sum_n c_n(t)|\psi_n\rangle\right] = \sum_n c_n(t)E_n|\psi_n\rangle \tag{1.27}$$

Projecting this equation onto each eigenvector $|\psi_n\rangle$ we find that:

$$c_n(t) = c_n e^{-i\omega_n t} \tag{1.28}$$

Therefore, once we know the effect of decomposition of the state function at $t = 0$ on the stationary states of the system, we will know the state function at any ulterior time t.

$$|\psi(t)\rangle = \sum_n c_n e^{-i\omega_n t}|\psi_n\rangle \tag{1.29}$$

This decomposition may be generalized for a basis set consisting of degenerate eigenstates and/or forming a continuum. This generalization comes, however, at the cost of a more cumbersome notation, and so we shall limit ourselves to its use only in those situations in which such a treatment cannot be avoided.

1.3.2 Calculation of stationary states in a one-dimensional potential

Let us consider a one-dimensional space mapped by the co-ordinate x and let us suppose a confinement potential $V(x)$, such that $V(x) < 0$ over all space, and $V(x) \to 0$ as $x \to \pm\infty$. In x representation, the time-independent one-dimensional Schrödinger equation for stationary states may be written as:

$$-\frac{\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi_n(x) + [V(x) - E_n]\psi_n(x) = 0 \tag{1.30}$$

Time-independent Schrödinger equation in *x* representation

We recall that the unknowns are the eigenvalues E_n, and the stationary state wavefunctions $\psi_n(x)$. For each value of E_n, Eq. (1.30) becomes a second-order differential equation. We can show that the $\psi_n(x)$ solutions of this equation are continuous, as are their first derivatives $\mathrm{d}\psi_n(x)/\mathrm{d}x$ over all space.[2] These two conditions, added to the normalization requirement of all physical states, lead to the quantization of energy, i.e. the existence of discrete energy levels. *It is therefore the wave nature of the wavefunctions and their integrability and continuity requirements which lead to the quantized nature of the energy levels.* We will illustrate this point with a precise example which plays a primordial role in the remainder of this text – the quantum well.

1.4 The quantum well

1.4.1 The general case

We now consider an electron subject to a potential well as described in Fig. 1.1, i.e. defined by:

$$\begin{aligned} V(x) &= 0, \qquad \text{if } |x| > \frac{a}{2} \\ V(x) &= -V_0, \qquad \text{if } |x| < \frac{a}{2} \end{aligned} \tag{1.31}$$

The first region ($|x| > a/2$) defines the potential *barrier*, whereas the second region ($|x| < a/2$) defines the *well*. The Schrödinger equation which governs the electron in this structure is:

$$\begin{aligned} &\frac{\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) - E\psi(x) = 0, \text{ for } |x| > \frac{a}{2} \\ &-\frac{\hbar^2}{2m_e}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) - (V_0 + E)\psi(x) = 0, \text{ for } |x| < \frac{a}{2} \end{aligned} \tag{1.32}$$

We first seek solutions to this equation having energies less than the potential barrier, i.e. $E < 0$. For this, we introduce three quantities, k, κ, and k_0, having as

[2] When the mass of a particle varies as a function of position x, in a semiconductor heterostructure for example, it is the quantity $1/m(x)\mathrm{d}/\mathrm{d}x$ which is conserved.

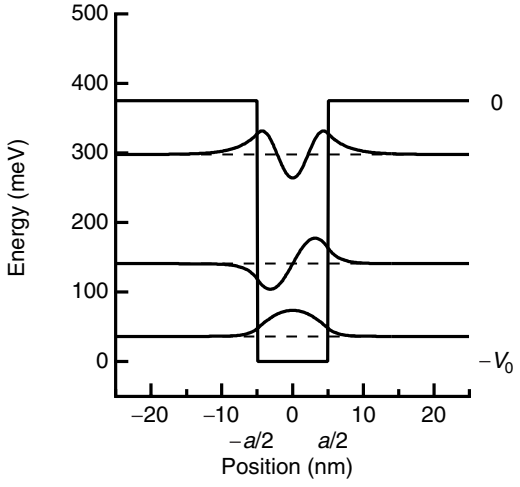


Fig. 1.1. A one-dimensional quantum well. Represented are the eigenenergies and wavefunctions associated with the three bound states of the system. This particular quantum well may be implemented in the $GaAs/Al_{0.45}Ga_{0.55}As$ system. The difference between the first two energy levels is 104 meV and leads to photon absorption at 11.9 μm.

dimension the inverse of a length, i.e. having the dimensions of a wavevector (the number of spatial periods in 2π), defined by:

$$E = -\frac{\hbar^2\kappa^2}{2m_e}$$

$$V_0 + E = \frac{\hbar^2 k^2}{2m_e} \tag{1.33}$$

$$V_0 = \frac{\hbar^2 k_0^2}{2m_e}$$

We note that $2\pi/k_0$ is the de Broglie wavelength associated with the energy V_0 of the confining potential.

Using this notation, the most general solutions to (1.32) are:

$$\psi_c(x) = A_c e^{ikx} + B_c e^{-ikx}, \text{ for } |x| < \frac{a}{2}$$

$$\psi_l(x) = A_l e^{\kappa x} + B_l e^{-\kappa x}, \text{ for } x < -\frac{a}{2} \tag{1.34}$$

$$\psi_r(x) = A_r e^{\kappa x} + B_r e^{-\kappa x}, \text{ for } x > \frac{a}{2}$$

where c, l, and r designate the centre, left, and right regions, respectively. We will

now illustrate the process of quantization by propagating the continuity conditions of the wavefunction and its first derivative (also referred to as *boundary conditions*) from $-\infty$ to $+\infty$ and, furthermore, by requiring that the results be normalized.

As the wavefunction must be normalized, its value cannot diverge as $x \to -\infty$. Therefore, $B_l = 0$. Additionally, the boundary conditions at $x = -a/2$ lead to:

$$\begin{aligned} A_c &= \frac{\kappa + \mathrm{i}k}{2\mathrm{i}k} \mathrm{e}^{(-\kappa + \mathrm{i}k)a/2} A_l \\ B_c &= -\frac{\kappa - \mathrm{i}k}{2\mathrm{i}k} \mathrm{e}^{-(\kappa + \mathrm{i}k)a/2} A_l \end{aligned} \tag{1.35}$$

A_c and B_c are related by the following useful equation:

$$\frac{A_c}{B_c} = -\frac{\kappa + \mathrm{i}k}{\kappa - \mathrm{i}k} \mathrm{e}^{\mathrm{i}ka} \tag{1.36}$$

The boundary conditions at $x = a/2$ give:

$$\begin{aligned} A_r \mathrm{e}^{\kappa a/2} + B_r \mathrm{e}^{-\kappa a/2} &= A_c \mathrm{e}^{\mathrm{i}ka/2} + B_c \mathrm{e}^{-\mathrm{i}ka/2} \\ A_r \mathrm{e}^{\kappa a/2} - B_r \mathrm{e}^{-(\kappa a/2)} &= \frac{\mathrm{i}k}{\kappa} (A_c \mathrm{e}^{\mathrm{i}ka/2} - B_c \mathrm{e}^{-(\mathrm{i}ka/2)}) \end{aligned} \tag{1.37}$$

We propagate the boundary conditions by bringing (1.37) into (1.35) where:

$$\begin{aligned} A_r &= \frac{[(\kappa + \mathrm{i}k)^2 \mathrm{e}^{\mathrm{i}ka} - (\kappa - \mathrm{i}k)^2 \mathrm{e}^{-\mathrm{i}ka}]}{4\mathrm{i}k\kappa} \mathrm{e}^{-\kappa a} A_l \\ B_r &= \frac{\kappa^2 + k^2}{2k\kappa} \sin ka A_l \end{aligned} \tag{1.38}$$

As the wavefunction must remain finite as $x \to +\infty$, this requires that $A_r = 0$ or that:

$$\left(\frac{\kappa - \mathrm{i}k}{\kappa + \mathrm{i}k} \right)^2 = \mathrm{e}^{2\mathrm{i}ka} \tag{1.39}$$

which may also be expressed as:

$$\frac{\kappa}{k} = \tan\left(\frac{ka}{2} \right) \tag{1.40}$$

or

$$\cos^2\left(\frac{ka}{2} \right) = \left(\frac{k}{k_0} \right)^2 \tag{1.41}$$

These solutions, of which there are two types, are expressed in the transcendental equations that follow.

(1) Even solutions

$$\left(\frac{\kappa - \mathrm{i}k}{\kappa + \mathrm{i}k}\right) = -\mathrm{e}^{\mathrm{i}ka} \tag{1.42}$$

or

$$\left|\cos\left(\frac{ka}{2}\right)\right| = \frac{k}{k_0}$$
$$\tan\left(\frac{ka}{2}\right) > 0 \tag{1.43}$$

Equation (1.36) informs us that $A_c = B_c$, or that the solutions are even. The energy levels, solutions of Schrödinger's equation, can then be determined from Fig. 1.2, and are represented by the intersection points where the line of slope $1/k_0$ meets the sinusoidal arches (dotted lines). Therefore, the energies accessible to an electron with total energy less than that of the potential barrier constitute a discrete spectrum (implying the energy levels are quantized).

The wavefunction then takes the form:

$$\psi_{n,c}(x) = A_n \cos k_n x, \text{ for } |x| < \frac{a}{2}$$
$$\psi_{n,l}(x) = B_n \mathrm{e}^{\kappa x}, \text{ for } x < -\frac{a}{2} \tag{1.44a}$$
$$\psi_{n,r}(x) = B_n \mathrm{e}^{-\kappa x}, \text{ for } x > \frac{a}{2}$$

where n designates the nth even solution of the equation. The values for A_n and B_n are obtained by noting that the integral of the square of $\psi_n(x)$ from $-\infty$ to $+\infty$ equals 1. For the ground state ($n = 1$), we obtain:

$$A_1 = \left(\frac{2}{a + 2/\kappa}\right)^{1/2}$$
$$B_1 = \left(\frac{2}{a + 2/\kappa}\right)^{1/2} \frac{k_1}{k_0} \mathrm{e}^{-\kappa a/2} \tag{1.44b}$$

where k_1 is the wavevector for the ground state from (1.43). Equation (1.44) shows that the electron wavefunction penetrates into the barrier over a distance given by $1/\kappa$, which signifies that the probability of finding the electron in the barrier region is non-zero (see Fig. 1.1). This phenomenon, known as *tunnelling*, possesses no

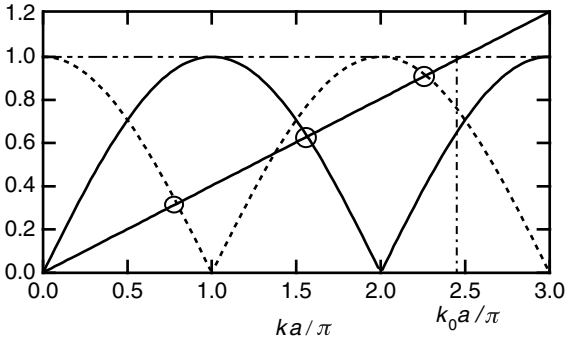


Fig. 1.2. Graphical determination of the quantized states for a symmetric quantum well using Eqs. (1.43) and (1.46), with $k_0 = 0.78\,\mathrm{nm}^{-1}$, and a well width $a = 10\,\mathrm{nm}$ (see the example).

classical equivalent and results from the fundamental wave nature of the electron and recalls analogous behaviour in light. We now recall the equation relating the energy of the eigenstate to its penetration depth into the barrier region:

$$\lambda_{\text{tunnel}} = \frac{1}{\kappa} = \frac{\hbar}{\sqrt{2m_e(-E)}} \tag{1.45}$$

(2) Odd solutions

These correspond to an alternative solution to (1.39):

$$\left(\frac{\kappa - \mathrm{i}k}{\kappa + \mathrm{i}k}\right) = \mathrm{e}^{\mathrm{i}ka}$$

Namely:

$$\left|\sin\left(\frac{ka}{2}\right)\right| = \frac{k}{k_0}$$
$$\tan\left(\frac{ka}{2}\right) < 0 \tag{1.46}$$

This time Eq. (1.36) tells us that $A_c = -B_c$, i.e. that the solutions are odd. The energy levels are now given by the intersection of the same line, with slope $1/k_0$, with the other series of sinusoidal arches appearing as solid lines in Fig. 1.2.

It is also interesting to calculate the number of quantum levels within the well. Inspection of Fig. 1.2 gives

$$N = 1 + \mathrm{Int}\left[\frac{\sqrt{2m_eV_0}}{\hbar}\frac{a}{\pi}\right] \tag{1.47}$$

where 'Int' designates the 'integer function'.

Also, no matter how shallow the well is, there is always at least one quantized state which lies within it. While this is a general observation which applies to all one-dimensional wells, this may not hold in three dimensions. The quantized levels are also referred to as: *localized*, as the wavefunctions have a non-negligible amplitude only in the vicinity of the well; and *bound*, as the probability of finding the electron is only significant near the well (the electrons are not mobile and cannot participate in current flow). The energy levels residing above the barrier ($E > 0$) are called *delocalized* or *free* (consult Complement 1.A for further details).

It is important to note that this line of reasoning may be generalized to any sort of potential: i.e. that quantification of the energy levels results from propagation of the boundary conditions from $-\infty$ to $+\infty$, and from the requirement that the amplitudes of the wavefunctions vanish at infinity.

Example

We will later see in Chapter 8, that an electron in a semiconductor heterostructure fabricated with $GaAs/Al_{0.45}Ga_{0.55}As$ is subjected to a potential well of 360 meV depth. Furthermore, the interaction of the electron with the periodic potential of the GaAs host crystal is taken into account by multiplication of the electron mass by a coefficient equal to 0.067. The result of this product corresponds to the *effective mass* of the electron $m^* = 0.067m_e$. Application of Eq. (1.33) allows us to solve for the wavevector k_0:

$$k_0 = \sqrt{(2 \times 0.067 \times 0.9 \times 10^{-30}\,(\text{kg}) \times 0.36\,(\text{eV}) \times 1.6 \times 10^{19}\text{C})/ \\ 1.05 \times 10^{-34}\,\text{J s}}$$

or

$$k_0 = 0.78\,\text{nm}^{-1}$$

which corresponds to a wavelength of $\lambda_0 = 8.05\,\text{nm}$.

Let us now consider a quantum well with a width of 10 nm. As the well width is of the order of the de Broglie wavelength λ_0 associated with $V_0 = 360\,\text{meV}$, we may expect the system to exhibit quantization. Using Eq. (1.47), we see that we can expect three bound states in this particular system (i.e. 1 + Int(0.78 × 10/3.14)). The wavefunctions corresponding to each of these states are shown in Fig. 1.1.

The MATHEMATICA program below is very useful for solving quantum confinement problems:

```
m0=0.91 10^-30 (*kg*);hbar=1.05 10^-34 (*J.s*);
q=1.6 10^-19 (*C*);
meff=0.067 (* effective electron mass in GaAs*);
V0=.36 (*well depth in eV*);
a=10. (*well width in nm*);
k0=Sqrt[2*meff*m0*q*V0]*10^-9/hbar (*in nm-1*)
```

```
eq1=Cos[k*a/2];
eq2=Sin[k*a/2];
eq3=k/k0;
plot1=Plot[Abs[eq1],{k,0,k0}]
plot2=Plot[Abs[eq2],{k,0,k0}]
plot3=Plot[Abs[eq3],{k,0,k0}]
Show[plot1,plot2,plot3]
FindRoot[eq1==eq3,{k,0.2}]
E1 =hbar^2*(k*10^9)^2/(2*meff*m0*q)/.%
FindRoot[eq2==eq3,{k,0.5}]
E2 =hbar^2*(k*10^9)^2/(2*meff*m0*q)/.%
hnu=E2-E1 (*optical transition energy in eV*)
```

1.4.2 The infinite square well

A particularly important case worth investigating is that of the infinite square well (see Fig. 1.3). In this case, the solution to Schrödinger's equation is found immediately:

$$\begin{aligned} ka &= n \text{ odd}, \qquad \psi_n(x) = \sqrt{\frac{2}{a}} \cos n\pi \frac{x}{a} \\ ka &= n \text{ even}, \qquad \psi_n(x) = \sqrt{\frac{2}{a}} \sin n\pi \frac{x}{a} \end{aligned} \tag{1.48}$$

and in both cases:

$$E_n = n^2 \frac{\hbar^2 \pi^2}{2m_e a^2} = n^2 E_0 \tag{1.49}$$

Energy levels for the infinite square well

E_0 is the *confinement energy*. We thereby uncover an alternate interpretation of the *de Broglie* wavelength given in (1.1), i.e. it is the width required of an infinite square well to yield a confinement energy E_0 equal to the energy of the particle. An important definition is the *thermal de Broglie wavelength* $\lambda_{\mathrm{DB,th}}$: this is the width of an infinite square well necessary for a confinement energy equal to the thermal energy kT:

$$\lambda_{\mathrm{DB,th}} = \frac{2\pi\hbar}{\sqrt{2m_e kT}} \tag{1.50}$$

Therefore, potential wells having widths less than $\lambda_{\mathrm{DB,th}}$ at $T = 300\,\mathrm{K}$ will show quantum effects unhindered by thermal vibrations in the system. Only in these cases may we speak of quantum wells at room temperature.

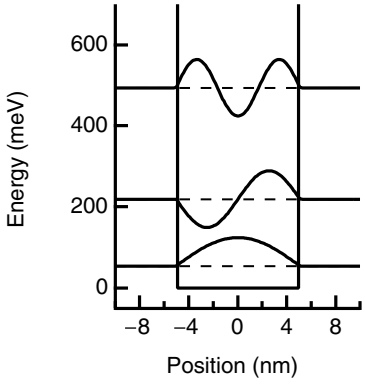


Fig. 1.3. Infinite quantum well with a width of 10 nm. In contrast to the finite square well solutions depicted in Fig. 1.1, the wavefunctions do not penetrate into the barriers and the energy levels are lifted up relative to the base of the well as a result.

Example

Assuming an infinite (although unachievable in practice) square well consisting of GaAs, at 300 K we calculate a $\lambda_{\mathrm{DB,th}} = 6.28 \times 1.05 \times 10^{-34}\,(\mathrm{J\,s^{-1}})/\sqrt{(2 \times 0.067 \times 0.9 \times 10^{-30}\,(\mathrm{kg}) \times 0.0259\,(\mathrm{eV}) \times 1.6 \times 10^{-19}\,\mathrm{C})} \approx 12\,\mathrm{nm}$.

Therefore, quantum effects will only be discernible in GaAs layers thinner than this value and only in these cases will we be able to speak of GaAs quantum wells existing at room temperature.

1.5 Time-independent perturbation theory

Very few physical systems present solutions as simple as those afforded by quantum wells. We find among such analytically tractable systems, the hydrogen atom (not treated in the present text) and the harmonic oscillator (treated in Complement 1.D). More general systems seldom have analytical solutions. However, by elaborating on simpler systems possessing better known solutions, we will attempt to approximate solutions to those that are more complex. The most popular (and arguably the most fruitful in terms of its success in expanding our conceptual understanding of many physical systems) is time-independent perturbation theory.

Consider an electron in a system described by a time-independent Hamiltonian H_0 for which the complete basis set of stationary states $\{|\psi_n\rangle\}$ consists of solutions to Schrödinger's equation:

$$H_0|\psi_n\rangle = E_n|\psi_n\rangle \tag{1.51}$$

(Note that from this point onwards, to simplify the notation, we will drop the "^" used earlier to identify operators, as we assume the reader is now able to

distinguish between an operator and a variable.) For the present, we will suppose that the states are discrete and non-degenerate. An important case involving the extension of perturbation theory to degenerate systems is given in Complement 1.B. We will now submit this system to a small additional perturbation $W = \alpha U$, such as may be achieved by the application of an electric field to a quantum well. By small, we mean that $\alpha \ll 1$ and that the eigenvalues of U are of the order E_n (i.e. that $U \approx H_0$ or that the eigenenergies of U are roughly the same size as those of the unperturbed Hamiltonian H_0). The eigenvalues of the new Hamiltonian $H = H_0 + W$ are:

$$(H_0 + \alpha U)|\psi_n(\alpha)\rangle = E_n(\alpha)|\psi_n(\alpha)\rangle \tag{1.52}$$

Then, make the important hypothesis that a sufficiently weak perturbation will allow us to consider the solutions of the modified system in terms of the original levels of the unperturbed system (i.e. that such a small perturbation has not distorted the original energy spectrum of the system beyond recognition). The new eigenvalues and eigenvectors of the perturbed system are then written in terms of the original eigenenergies and eigenvectors and the perturbation coefficient α:

$$\begin{aligned} E_n(\alpha) &= \varepsilon_0 + \alpha\varepsilon_1 + \alpha^2\varepsilon_2 + \cdots \\ |\psi_n(\alpha)\rangle &= |0\rangle + \alpha|1\rangle + \alpha^2|2\rangle + \cdots \end{aligned} \tag{1.53}$$

Substitute (1.53) into (1.52) and obtain by identifying like terms in powers of α:

$$\text{Order 0} \qquad H_0|0\rangle = \varepsilon_0|0\rangle \tag{1.54a}$$

$$\text{Order 1} \qquad (H_0 - \varepsilon_0)|1\rangle + (U - \varepsilon_1)|0\rangle = 0 \tag{1.54b}$$

$$\text{Order 2} \qquad (H_0 - \varepsilon_0)|2\rangle + (U - \varepsilon_1)|1\rangle - \varepsilon_2|0\rangle = 0 \tag{1.54c}$$

0th order

As we have assumed that the levels are non-degenerate, Eq. (1.54a) shows that $|0\rangle$ is an eigenstate of H_0. By continuity, as $\alpha \to 0$, we find that $|0\rangle = |\psi_n\rangle$. This is not true when the levels are degenerate, as Eq. (1.54a) no longer corresponds to a single quantum level.

1st order

Project (1.54b) on $|0\rangle = |\psi_n\rangle$ and use the identity:

$$\langle 0|H_0 - \varepsilon_0|1\rangle = 0 \tag{1.55}$$

to find the first-order energy correction:

$$\varepsilon_1 = \langle\psi_n|U|\psi_n\rangle \tag{1.56}$$

or, in terms of earlier definitions:

$$E'_n = E_n + \langle \psi_n | W | \psi_n \rangle \tag{1.57}$$

First-order energy perturbation

where the perturbed energy E'_n is expressed without reference to α.

To find a limited expansion for the eigenvector, we need only project (1.54b) onto the other states $|\psi_p\rangle$ with $p \neq n$:

$$(E_p - E_n)\langle \psi_p | 1 \rangle + \langle \psi_p | U | \psi_n \rangle = 0 \tag{1.58}$$

We then obtain for the perturbed eigenvectors the following first-order expansion:

$$|\psi'_n\rangle = |\psi_n\rangle + \sum_{p \neq n} \frac{\langle \psi_p | W | \psi_n \rangle}{E_n - E_p} |\psi_p\rangle \tag{1.59}$$

First-order perturbation of the eigenstates

We notice that the unperturbed stationary state $|\psi_n\rangle$ is *contaminated* by other eigenstates $|\psi_p\rangle$, and all the more so for those states $|\psi_p\rangle$ closest to $|\psi_n\rangle$ in energy. Therefore, in describing the effect of a perturbation, we will be content to limit ourselves to a description in terms of those levels closest in energy (see, for example, the treatment of the Stark effect given in Complement 1.C).

2nd order

In a certain number of cases, the first-order perturbation will be null when:

$$\langle \psi_n | W | \psi_n \rangle = 0 \tag{1.60}$$

This occurs as a result of symmetry considerations (as, for instance, in the case of the perturbation of a quantum well confinement potential by an electric field). As a result, it is often necessary to continue the perturbation expansion to higher orders. Projecting (1.54c) onto $|\psi_n\rangle$, we find:

$$\varepsilon_2 = \langle \psi_n | U | 1 \rangle \tag{1.61}$$

after which using (1.59) we may write for the second-order perturbation:

$$E'_n = E_n + \langle \psi_n | W | \psi_n \rangle + \sum_{p \neq n} \frac{|\langle \psi_p | W | \psi_n \rangle|^2}{E_n - E_p} \tag{1.62}$$

Second-order energy perturbation

where again we note that the magnitude of the contribution of any given state increases for those closest to $|\psi_n\rangle$ in energy.

1.6 Time-dependent perturbations and transition probabilities

1.6.1 The general case

Situations where exact solutions may be found to Schrödinger's time-dependent equation (1.12) are unfortunately few and far between. The time-dependent behaviour of an electron in a quantum well is worth citing; it may be worked out as an exercise. Generally, we employ a perturbative approach, which will enable definition of the transition rate. Let us consider a system described by the Hamiltonian H_0 which is in an initial state $|\psi_i\rangle$ at time 0. At time $t = 0$ we *turn on* a perturbation $W(t) = \alpha U(t)$, where the conditions placed on α and $U(t)$ are same as in the preceding section (namely that $\alpha \ll 1$ and $U \approx H_0$). In order to solve Schrödinger's time-dependent equation:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle = [H_0 + W(t)]|\psi(t)\rangle \tag{1.63}$$

to describe the evolution of the system, we can expand $|\psi(t)\rangle$ in terms of the basis of stationary states, as described in (1.6):

$$|\psi(t)\rangle = \sum_n c_n(t)|\psi_n\rangle \tag{1.64}$$

Substituting (1.64) into (1.63) and identifying like terms, we obtain a system of coupled differential equations, relating the coefficients $c_n(t)$ to one another:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}c_n(t) = E_n c_n(t) + \sum_p \alpha U_{np}(t)c_p(t) \tag{1.65}$$

where U_{np} are the elements in the matrix:

$$U_{np}(t) = \langle\psi_n|U(t)|\psi_p\rangle \tag{1.66}$$

We will suppose that, for reasons of symmetry, $U_{nn} = 0$ for any given level n. We then make the following change of variables:

$$b_n(t) = c_n(t)\mathrm{e}^{+\mathrm{i}E_n t/\hbar} \tag{1.67}$$

which leads us to:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}b_n(t) = \alpha\sum_p \mathrm{e}^{\mathrm{i}\omega_{np}t}U_{np}(t)b_p(t) \tag{1.68}$$

where $\omega_{np} = (E_n - E_p)/\hbar$ is the *Bohr oscillation frequency* for the transition $n \to p$. As in Section 1.5, we perform a limited expansion:

$$b_n(t) = b_n^{(0)}(t) + \alpha b_n^{(1)}(t) + \alpha^2 b_n^{(2)}(t) + \cdots \tag{1.69}$$

allowing us to identify like terms in α after substitution of (1.69) into (1.68).

0th-order term
We find that $b_n(0)$ is a constant which corresponds to the stationary state solutions given by (1.29).

***q*th-order term**
We obtain:

$$\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t} b_n^{(q)}(t) = \sum_p \mathrm{e}^{\mathrm{i}\omega_{np}t} U_{np}(t) b_p^{(q-1)}(t) \tag{1.70}$$

Therefore, once the zeroth-order solution is known, we may calculate the first-order solution and then any other order solution by recurrence. We will interest ourselves in the remainder of this chapter with first-order perturbations. Second-order perturbation will be developed in Chapter 12, in the context of non-linear optics.

At $t = 0$, the system is in the state $|\psi_i\rangle$, with initial conditions:

$$\begin{aligned} &b_i^{(0)}(t=0) = 1 \\ &b_n^{(0)}(t=0) = 0, \quad \text{for } i \neq n \end{aligned} \tag{1.71}$$

To zeroth order, these values remain constant with respect to time. Inserting these values into (1.70), we obtain the first-order time evolution equation:

$$\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t} b_n^{(1)}(t) = \mathrm{e}^{\mathrm{i}\omega_{ni}t} U_{ni}(t) \tag{1.72}$$

which takes the integral form:

$$b_n^{(1)}(t) = \frac{1}{\mathrm{i}\hbar} \int_0^t \mathrm{e}^{\mathrm{i}\omega_{ni}t'} U_{ni}(t') \mathrm{d}t' \tag{1.73}$$

We are now in a position to calculate the probability $P_{if}(t)$ of finding the system in a final stationary state $|\psi_f\rangle$ at time t. Following the probabilistic interpretation of quantum mechanics, this is obtained by evaluating $|b_f(t)|^2$ or:

$$P_{if}(t) = \frac{1}{\hbar^2}\left|\int_0^t \mathrm{e}^{\mathrm{i}\omega_{fi}t'} W_{fi}(t')\mathrm{d}t'\right|^2 \tag{1.74a}$$

Transition probability between levels *i* and *f* under the effect of a time-varying perturbation

where

$$W_{fi}(t) = \langle\psi_f|W(t)|\psi_i\rangle \tag{1.74b}$$

This formula is one of the most important in quantum mechanics and will be referred to throughout this book. We will presently apply it to the particularly interesting and useful problem of a time-varying sinusoidal perturbation.

1.6.2 Sinusoidal perturbation

This perturbation potential may be written as:

$$W(\mathbf{r}, t) = W(\mathbf{r})\sin\omega t \tag{1.75}$$

Equation (1.74) leads immediately to a time-dependent transition probability $P_{if}(t)$ between initial and final states:

$$P_{if}(t) = \frac{|W_{fi}|^2}{4\hbar^2}\left|\frac{1-\mathrm{e}^{\mathrm{i}(\omega_{fi}+\omega)t}}{\omega_{fi}+\omega} - \frac{1-\mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t}}{\omega_{fi}-\omega}\right|^2 \tag{1.76}$$

We therefore make what is classically referred to as the *rotating phase* or the *quasi-resonance* approximation, which ignores the contribution of the term possessing the larger denominator $\omega_{fi} + \omega$ in favour of that with $\omega_{fi} - \omega$. Thus, keeping only the second term in (1.76) we obtain:

$$P_{if}(t) = \frac{|W_{fi}|^2}{4\hbar^2}\left|\frac{1-\mathrm{e}^{\mathrm{i}(\omega_{fi}-\omega)t}}{\omega_{fi}-\omega}\right|^2 = \frac{|W_{fi}|^2}{4\hbar^2}\left|\left[\sin\frac{(\omega_{fi}-\omega)t}{2}\right]\Big/\frac{(\omega_{fi}-\omega)}{2}\right|^2 \tag{1.77}$$

Figure 1.4 shows the evolution of this probability as a function of time for different frequencies (or *detuning*) between the perturbing field and resonant transition frequency $\omega - \omega_{fi}$. We note that as the frequency of the perturbation field approaches that of the resonant Bohr oscillation frequency (i.e. $\omega \to \omega_{fi}$), the time dependence of the transition amplitude changes from a sinusoidally varying function to a parabola in t. In a complementary fashion, we show in Fig. 1.5 the spectral distribution of the transition probability as a function of detuning for various times t. This function is a sinus cardinal multiplied by t^2, which tends towards a Dirac delta function as $t \to \infty$. We therefore rewrite (1.77) as:

$$P_{if}(t) = \frac{|W_{fi}|^2}{4\hbar^2}t^2\left[\mathrm{sinc}\frac{(\omega_{fi}-\omega)t}{2}\right]^2 \tag{1.78}$$

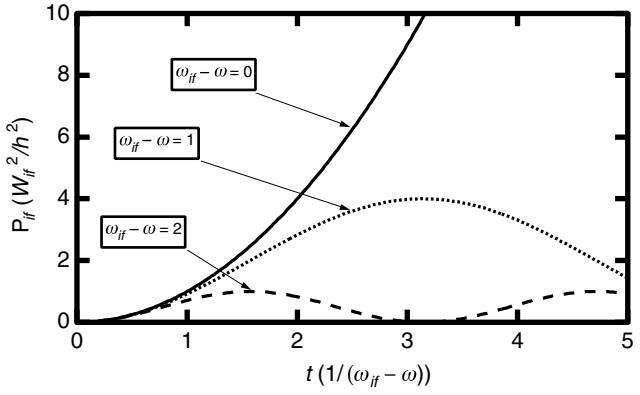


Fig. 1.4. Time evolution of the transition probability between levels i and f for different detuning values of $\omega_{if} - \omega$. In off-resonance conditions, the electrons oscillate between both levels.

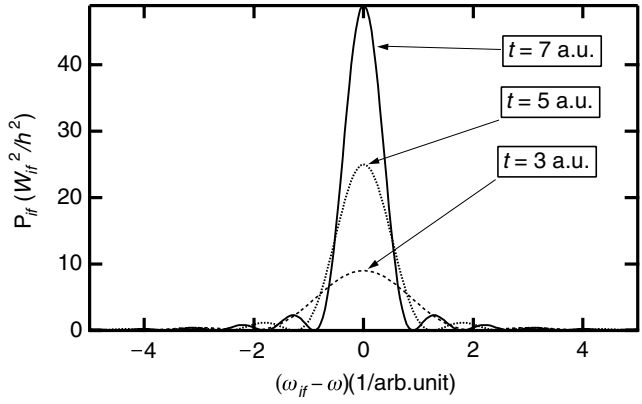


Fig. 1.5. Transition probability between two levels i and f as a function of detuning frequency for different observation times t (arb. units). At longer times, only transitions between states satisfying the requirement of energy conservation are accessible. This behaviour is in accordance with Heisenberg's second uncertainty principle.

where sinc(x) is the sinus cardinal $\sin x/x$. Equation (1.78), while appearing simple, is in fact rather difficult to grasp in its entirety, as it is a function of two intimately related quantities, namely frequency and time. To investigate its behaviour better, we will distinguish between three different cases.

Case 1: transitions induced between discrete levels by single frequency excitation
In this case, the resonant transition largely dominates the behaviour whereby:

$$P_{if}(t) = \frac{|W_{fi}|^2}{4\hbar^2} t^2 \tag{1.79}$$

As the transition probability thus stated increases quadratically with time, this

description is clearly an approximation as its value cannot exceed unity. We will see in Complement 1.E that this approximation holds only for very short times over which the zeroth-order expansion employed in (1.69) may be seen as valid. We note that the resonance condition $\omega = \omega_{fi}$ may be written, alternately, as

$$\hbar\omega = E_f - E_i = \hbar\omega_{fi} \tag{1.80}$$

Bohr frequency

This equation describes the conservation of energy between the energy quantum transferred to the system and the energy difference of the system between the final and initial states $\Delta E = E_f - E_i$. From a technical point of view concerning the calculation of these quantities, we note that the ordering of the indices in these equations ($|W_{if}| = |W_{fi}|$) plays no explicit role owing to the properties of Hermitian products. This, however, is not the case for the Bohr oscillation frequency $\omega_{if} = -\omega_{fi}$!

Additionally, Fig. 1.5 shows that the transition probability becomes negligible once:

$$\Delta E t > \hbar \tag{1.81}$$

This last condition is also known as Heisenberg's *second uncertainty relation* and it allows the classical restriction of energy conservation to be violated by excitations acting over short time periods.

Case 2: transitions induced between a discrete level and a continuum state by single frequency excitation

In this case, the final states form a continuum described by the continuous variable ω_{fi}, and the transition probability between the discrete level and the continuum $P_{ic}(t)$ is calculated by summing the probabilities over the density of final states $\rho(\omega_{fi})$:

$$P_{ic}(t) = \frac{1}{4\hbar^2} t^2 \int_{-\infty}^{+\infty} |W_{if}(\omega_{fi})|^2 \left[\operatorname{sinc}\frac{(\omega_{fi} - \omega)t}{2}\right]^2 \rho(\omega_{fi}) \mathrm{d}\omega_{fi} \tag{1.82}$$

Distribution theory tells us that if a function $|W_{if}(\omega_{fi})|^2 \rho(\omega_{fi})$ is well behaved (i.e. square normalizable and slowly varying), then:

$$\lim_{t\to\infty} \left\{\operatorname{sinc}\left[\frac{1}{2}(\omega_{fi} - \omega)t\right]\right\}^2 = \frac{2\pi}{t}\delta(\omega_{fi} - \omega) \tag{1.83}$$

where δ is the Dirac delta function. Therefore, for long times, Eq. (1.82) takes the form:

$$P_{ic}(t) = \frac{\pi}{2\hbar^2} |W_{if}(\omega_{fi})|^2 \rho(\omega = \omega_{fi}) t \tag{1.84}$$

The above equation tells us that when a transition occurs from a discrete state to a continuum, the *transition rate* $G_{ic} = \mathrm{d}P_{ic}/\mathrm{d}t$ is constant as a function of time and has a value given by:

$$G_{ic}(\omega) = \frac{\pi}{2\hbar^2}|W_{if}(\omega_{fi})|^2\rho(\omega = \omega_{fi}) \tag{1.85a}$$

Fermi's golden rule in terms of frequency

or

$$G_{ic}(\hbar\omega) = \frac{\pi}{2\hbar}|W_{if}|^2\rho(\hbar\omega = E_f - E_i) \tag{1.85b}$$

Fermi's golden rule in terms of energy

This important equation is referred to as *Fermi's golden rule*. It stipulates that under the influence of monochromatic excitation $\hbar\omega$, only continuum levels having energy $E_f = E_i + \hbar\omega$ will be populated by the optical excitation with a transition rate given by the above equation.

Case 3: transitions induced between two discrete levels by multi-frequency excitation

In this case, the perturbation consists of a continuum of excitation frequencies:

$$W_{fi}(t) = \int_0^\infty g(\omega)W_{fi}(\omega)\sin(\omega t)\mathrm{d}\omega \tag{1.86}$$

where $g(\omega)$ is the excitation spectrum and W_{fi} is the matrix element of the interaction Hamiltonian at each particular wavelength. A development strictly equivalent to the one given above leads to a transition rate:

$$G_{if}(\hbar\omega) = \frac{\pi}{2\hbar}|W_{if}|^2 g(\hbar\omega = E_f - E_i) \tag{1.87}$$

Transition rate for broad frequency excitation

1.7 The density matrix

Two kinds of uncertainty coexist within the description of a physical system. There is a purely quantum uncertainty related to the probabilistic interpretation of the results of the operator algebra applied to the system. There is also uncertainty resulting from the thermal agitation of the system's constituent parts, which is described by statistical mechanics. *Density matrix* formalism presents itself as a

very powerful and elegant framework which integrates these two notions into a single mathematical description.

1.7.1 Pure quantum ensembles

Let us consider a quantum system in a state $|\psi(t)\rangle$ described by (1.64). We wish to know the average value of an operator A. Following Eq. (1.8), this average value $\langle A\rangle$ may be written in a particular basis $|\psi_k\rangle$ as:

$$\langle A\rangle = \sum_{n,m} A_{mn} c_m^*(t) c_n(t) \tag{1.88}$$

where A_{mn} is an element of the matrix:

$$A_{mn} = \langle \psi_m | A | \psi_n \rangle \tag{1.89}$$

Equation (1.88) is somewhat deceiving as it seems to suggest that a privileged basis set exists in which to carry out the decomposition. If we change this basis, however, the A_{mn}, $c_m(t)$, and $c_n(t)$ entries will change in such a manner as to leave (1.88) invariant. This inconvenience is eliminated by introducing the *density matrix*, whose elements are:

$$\rho_{nm}(t) = c_m^*(t) c_n(t) \tag{1.90}$$

In fact, the matrix $\rho(t)$ may be written as:

$$\rho(t) = |\psi(t)\rangle\langle\psi(t)| \tag{1.91}$$

With this definition, Eq. (1.88) then becomes:

$$\langle A\rangle = \mathrm{Tr}(\rho A) \tag{1.92}$$

which is independent of the decomposition basis $\rho = |\psi(t)\rangle\langle\psi(t)|$ as its trace is a linear operator whose value is independent of the basis in which it is evaluated. Furthermore, using (1.91), we immediately see that the evolution of $\rho(t)$ as a function of time is given by:

$$\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = [H(t), \rho(t)] = H(t)\rho(t) - \rho(t)H(t) \tag{1.93}$$

Schrödinger's equation in density matrix formalism

1.7.2 Mixed quantum ensembles

We now consider a system consisting of a statistically distributed mixture of states $\{|\phi_i\rangle\}$. This system has a *thermodynamic probability* p_i of being in a state $|\phi_i\rangle$.

Neglecting quantum interferences between thermodynamically blurred states, it seems natural to define the average value of observable A as:

$$\overline{\langle A\rangle} = \sum_i p_i \langle A\rangle_{\phi_i} \tag{1.94}$$

where $\langle A\rangle_{\phi i}$ is the average value of observable A when the system is in the state $|\phi_i\rangle$. Following Eq. (1.92), this may be written as:

$$\overline{\langle A\rangle} = \sum_i p_i \mathrm{Tr}(\rho_i A) = \mathrm{Tr}(\bar{\rho} A) \tag{1.95}$$

where $\bar{\rho}$ is the density matrix of the *mixed ensemble*:

$$\bar{\rho} = \sum_i p_i \rho_i = \sum_i p_i |\phi_i\rangle\langle\phi_i| \tag{1.96}$$

In Eq. (1.96) we see the advantage obtained by introducing the density matrix. It is the linear dependence of $\langle A\rangle$ on the density matrix ρ which allows the introduction of the density or averaging operator $\bar{\rho}$.

As each matrix $\bar{\rho}$ allows the same time-evolution equation (1.93), the density matrix for the mixed ensemble may be written:

$$\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t}\bar{\rho}(t) = [H, \bar{\rho}(t)] \tag{1.97}$$

Schrödinger's equation for a mixed ensemble

The fundamental equations of the density matrix are (1.95)–(1.97). Within the density matrix itself, we may differentiate between two conceptually distinct constituents.

(a) Diagonal elements

From equations (1.90) to (1.96), the diagonal terms may be expressed in the stationary state basis as:

$$\bar{\rho}_{kk} = \sum_i p_i |c_k^i|^2$$

where c_k^i is the $|\phi^i\rangle$ component in the $|\psi_k\rangle$ basis. An immediate physical interpretation of the diagonal terms in $\bar{\rho}_{kk}$ is that they represent the probability of finding the system, upon measurement, in a stationary state $|\psi_k\rangle$ *given both the quantum and statistical uncertainties*. Therefore, ρ_{kk} represents the population of the state $|\psi_k\rangle$. As these elements result from the summation of positive terms, they may not be zero unless the value of each of these terms is zero (i.e. that the occupation of each state is null).

(b) Off-diagonal elements

The significance of these terms is a little more difficult to understand. They are sometimes referred to as the *coherence elements*: they describe the quantum behaviour of the system. When thermal fluctuations completely smear out the quantum interference effects, these are the terms that become zero.

In Section 1.7.3, we will give an example of a two-level system which will allow us to grasp the usefulness of this powerful and elegant formalism better. We note in closing that the distinction between population and coherence depends on the decomposition basis $\{|\psi_k\rangle\}$.

1.7.3 Density matrix and relaxation time for a two-level system

We consider a two-level system with a Hamiltonian H_0, possessing eigenenergies E_1 and E_2, and stationary states $|1\rangle$ and $|2\rangle$ (i.e. $H_0|i\rangle = E_i|i\rangle$). In the stationary state basis, the Hamiltonian H_0 may be written:

$$H_0 = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \tag{1.98}$$

We subject this system at time $t = t_0$ to a sinusoidal perturbation $W(t)$ which may be expressed in the basis of $|1\rangle$ and $|2\rangle$ as:

$$W = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \cos \omega t \tag{1.99}$$

where $m_{ij} = \langle i|W|j\rangle$. We may assume by symmetry, that the elements m_{11} and m_{22} are null, and that the terms m_{12} and m_{21} are real and thus equal. The general case may be determined as an exercise. Equation (1.97) then may be written as:

$$\begin{aligned} \frac{\mathrm{d}\rho_{11}}{\mathrm{d}t} &= -\mathrm{i}\frac{m_{12}}{\hbar}(\rho_{21} - \rho_{12})\cos\omega t \\ \frac{\mathrm{d}}{\mathrm{d}t}(\rho_{11} + \rho_{22}) &= 0 \\ \frac{\mathrm{d}\rho_{21}}{\mathrm{d}t} &= -\mathrm{i}\omega_{21}\rho_{21} + \mathrm{i}\frac{m_{12}}{\hbar}(\rho_{22} - \rho_{11})\cos\omega t \end{aligned} \tag{1.100}$$

The second equation of (1.100) states that the total population is conserved (i.e. $\rho_{11} + \rho_{22} = 1$). Solutions to this very important set of coupled differential equations are given in Complement 1.E. Nonetheless, we may investigate the transitory behaviour of the system at this point. For instance, it is clear that the terms in $\cos\omega t$ will act to drive the system into oscillation. If the excitation ceases (i.e.

supposing we set m_{12} to zero), the diagonal terms will remain constant and the off-diagonal terms will continue to oscillate with frequency ω_{21}.

Intuitively, we may expect that once the excitation stops, the populations ρ_{ii} will tend toward their thermodynamic equilibrium levels ρ_{ii}^{eq} with a certain time constant resulting from stochastic interactions. This time constant is often referred to as the *diagonal relaxation time* or the *population lifetime*, to name but a few. It is written as T_1 when its value is independent of i in ρ_{ii}, i.e. level independent. In converse situations, the custom is to make use of the relaxation rate Γ_{ii}. In the same way, we expect the off-diagonal elements to lose coherence with a time constant of Γ_{ij}^{-1} or T_2 if this time is independent of ij.[3] Introducing these different relaxation times, the equations in the density matrix become:

$$\frac{\mathrm{d}\rho_{11}}{\mathrm{d}t} = -\mathrm{i}\frac{m_{12}}{\hbar}(\rho_{21}-\rho_{12})\cos\omega t - \frac{\rho_{11}-\rho_{11}^{\mathrm{eq}}}{T_1}$$

$$\frac{\mathrm{d}\rho_{22}}{\mathrm{d}t} = \mathrm{i}\frac{m_{12}}{\hbar}(\rho_{21}-\rho_{12})\cos\omega t - \frac{\rho_{22}-\rho_{22}^{\mathrm{eq}}}{T_1} \tag{1.101a}$$

$$\frac{\mathrm{d}\rho_{21}}{\mathrm{d}t} = -\mathrm{i}\omega_{21}\rho_{21} + \mathrm{i}\frac{m_{12}}{\hbar}(\rho_{22}-\rho_{11})\cos\omega t - \frac{\rho_{21}}{T_2}$$

Time-evolution of elements in a density matrix for a two-level system

or

$$\frac{\mathrm{d}\rho_{11}}{\mathrm{d}t} = -\mathrm{i}\frac{m_{12}}{\hbar}(\rho_{21}-\rho_{12})\cos\omega t - \Gamma_1(\rho_{11}-\rho_{11}^{\mathrm{eq}})$$

$$\frac{\mathrm{d}\rho_{22}}{\mathrm{d}t} = \mathrm{i}\frac{m_{12}}{\hbar}(\rho_{21}-\rho_{12})\cos\omega t - \Gamma_1(\rho_{22}-\rho_{22}^{\mathrm{eq}}) \tag{1.101b}$$

$$\frac{\mathrm{d}\rho_{21}}{\mathrm{d}t} = -\mathrm{i}(\omega_{21}-\mathrm{i}\Gamma_{12})\rho_{21} + \mathrm{i}\frac{m_{12}}{\hbar}(\rho_{22}-\rho_{11})\cos\omega t$$

Time-evolution of elements in a density matrix for a two-level system

These last two expressions are one of the major conclusions from this first chapter and they will be used intensively throughout this book. Complement 1.E gives as

[3] Interestingly, the introduction of a relaxation time reintroduces the set of stationary states as a privileged basis. This observation is of theoretical interest, however, and will not receive further consideration by us.

an example an application whose treatment using this theory leads to the optical Bloch equations.

FURTHER READING

C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Wiley, New York (1992).

C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Atom–Photon Interactions*: *Basic Processes and Applications*, Wiley, New York (1998).

R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, *Vol III*: *Quantum Mechanics*, Addison-Wesley, Reading, Mass. (1966).

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).

E. Merzbacher, *Quantum Mechanics*, Wiley, New York (1970).

A. Messiah, *Quantum Mechanics*, vols 1 and 2, Wiley, New York (1966).

L. I. Schiff, *Quantum Mechanics*, 2nd edn, McGraw-Hill, New York (1955).

Complement to Chapter 1

1.A Problems posed by continuums: the fictitious quantum box and the density of states

The quantum description of delocalized states, which therefore belong to a continuum, makes reference to the theory of distributions. It attempts to deal with difficult problems such as the normalization of wavefunctions in a null potential from $-\infty$ to $+\infty$. In this book, systematic use is made of the theoretical trick afforded by introduction of a *fictitious infinite square well* of width L in which the motion of the continuum electrons are later shown to be pseudo-quantified. What we mean by pseudo-quantification is that when we take L tending towards infinity in the expressions obtained, the dependence of L will conveniently disappear from physical predictions. There is no moral in this; only the tutelar protection of distribution theory! We now proceed to an illustrative example: photoemission from a one-dimensional well.

We consider a quantum well of width d as represented in Fig. 1.A.1. This quantum well admits a quantized level $|i\rangle$ described by a square integrable wavefunction $\psi_i(z)$ and a quantized energy level $-E_I$ (where the index I stands for *ionization*, for reasons which shall soon be clear). We further presume that the well is sufficiently deep for $\psi_i(z)$ to be considered the wavefunction for the ground state of the infinite well such that:

$$\psi_i(z) = \sqrt{\frac{2}{d}} \cos \frac{\pi}{d} z \tag{1.A.1}$$

This well also admits delocalized states, where the electrons may take on any value of positive energy. We will neglect the influence of the well on the free electrons, i.e. we will suppose that the free electrons are subject to a null potential once they are in the continuum. To avoid problems involving the normalization of these wavefunctions, we introduce a fictitious square well of width L within which the continuum electrons are trapped. The corresponding eigenenergies and eigenfunctions of the unbound states are:

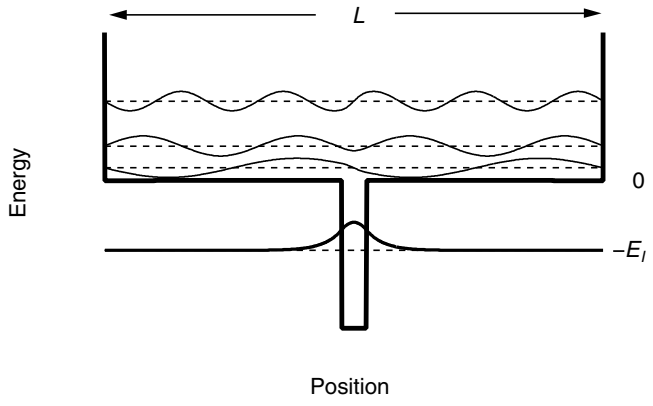


Fig. 1.A.1. Procedure for pseudo-quantification of the potential barrier states. The width L of the infinite quantum well is arbitrary.

$$e_n = n^2 e_0$$

$$\psi_n(z) = \sqrt{\frac{2}{L}} \sin nk_L z, \text{ if } n \text{ is even} \tag{1.A.2}$$

$$\psi_n(z) = \sqrt{\frac{2}{L}} \cos nk_L z, \text{ if } n \text{ is odd}$$

where e_0 is the confinement energy of the fictitious well:

$$e_0 = \frac{\hbar^2}{2m^*} k_L^2 \tag{1.A.3}$$

with wavevector k_L

$$k_L = \frac{\pi}{L} \tag{1.A.4}$$

If L takes on dimensions of centimetres, then e_0 is of the order of 10^{-11} eV. In this sense, such a box would be fictitious as the energy level spacings would be infinitesimal in comparison with typical interaction or thermal energies (of the order of meV). The energy levels given in (1.A.2) are so close that rather than attempting to take each one into account individually, we group them together by means of infinitesimal batches of the density of states.

Let us consider a certain wavevector interval dk. In this interval the individual states of the fictitious box are separated in the wavevector by π/L. *Without taking into account electron spin*, the number of states in this interval is clearly:

$$\mathrm{d}n = \frac{L}{\pi} \mathrm{d}k \tag{1.A.5}$$

The density of states dn/dk is then given by:

$$\rho(k) = \frac{dn}{dk} = \frac{L}{\pi} \tag{1.A.6}$$

In this same interval, as obtained by differentiation of (1.A.3), the corresponding change in energy relates to dk as:

$$dk = \frac{1}{2}\frac{\sqrt{2m^*}}{\hbar}\frac{dE}{E} \tag{1.A.7}$$

The energy density of states dn/dE is then finally:

$$\rho(E) = \frac{L}{2\pi}\frac{\sqrt{2m^*}}{\hbar}\frac{1}{\sqrt{E}} \tag{1.A.8}$$

One-dimensional density of states without spin

We note that as L tends towards infinity, $\rho(E)$ increases without bound. This is to be expected as more and more states become available over the same energy range as the energy separation between levels decreases.

We will now calculate the transition probability between an initial quantized state $|i\rangle$ and the continuum under the effect of a sinusoidally varying dipole perturbation:

$$W(z,t) = -qFz\cos(\omega t) \tag{1.A.9}$$

From Fermi's golden rule (1.85b), the transition probability may be written as:

$$G_{ic}(\hbar\omega) = \frac{\pi q^2 F^2}{2\hbar}|z_{if}|^2\rho(\hbar\omega = E_f - E_i) \tag{1.A.10}$$

The transition element z_{if} is non-zero only for odd parity final states and is given by:

$$|z_{if}|^2 = |\langle\psi_i|z|\psi_f\rangle|^2 = \left|\frac{2}{\sqrt{Ld}}\int_{-d/2}^{d/2}\cos\left(\frac{\pi}{d}z\right)\sin(k_f z)z\,dz\right|^2 \tag{1.A.11}$$

or

$$|z_{if}|^2 = 4\frac{d^3}{L}f^2(E_f) \tag{1.A.12}$$

where $f(E)$ is the dimensionless integral in equation (1.A.11) and is found to be:

$$f(E_f) = \frac{\pi}{\pi^2 - d^2k_f^2}\left(\sin\frac{k_f d}{2} - \frac{4k_f d}{\pi^2 - d^2k_f^2}\cos\frac{k_f d}{2}\right) \tag{1.A.13}$$

and k_f is the wavevector of the associated state with energy E_f:

$$E_f = \hbar\omega - E_I = \frac{\hbar^2 k_f^2}{2m^*} \tag{1.A.14}$$

We then substitute the expression for $|z_{if}|$ into (1.A.10). Notice that the width of the fictitious box L, *which appears in the denominator of the transition element and in the numerator of the density of states*, cancels out as advertised. In physical terms, the wider the width of the pseudo-well, the greater the final density of states; however, this effect on the transition probability is cancelled out as the increased width also serves to dilute the probability density of the electron states above the quantum well (of width d) by a similar amount.

Taking into account the fact that the density of final states is only one-half of the expression we derived in (1.A.8) (because of the spin of the electrons, which is conserved in the transition) and since only the odd parity wavefunctions participate in the transitions, we therefore find for the transition probability from an initial quantized state to the continuum:

$$G_{if}(\hbar\omega) = q^2 F^2 d^3 \frac{\sqrt{m^*}}{\hbar^2} \frac{f^{\,2}(\hbar\omega - E_I)}{\sqrt{2(\hbar\omega - E_I)}} \tag{1.A.15}$$

The behaviour of the system is therefore found to be independent of the size of the fictitious box we introduced at the onset. This technique is referred to as *pseudo-quantification*, and is in fact a very powerful tool in spite of its simplistic appearance. Figure 1.A.2 shows the variation of the transition rate as a function of the excitation frequency ω.

We notice the presence of an ionization threshold for the transition probability. The cut-off energy for detected photons corresponds to the ionization energy E_I. Furthermore, the absorption near the detection threshold, i.e. for photons with $\hbar\omega \approx E_I$ is given by:

$$G_{if} \propto \sqrt{\hbar\omega - E_I}, \text{ for } \hbar\omega \approx E_I \tag{1.A.16}$$

A second characteristic is the quasi-resonant nature of the transition probability near the energy threshold for photoionization. This quasi-resonance results from decreases in both the density of states (in k_f^{-1}) and in the dipole moment (in k_f^{-2}) which leads to:

$$G_{if} \propto \frac{1}{(\hbar\omega - E_I)^{5/2}}, \text{ for } \hbar\omega \gg E_I \tag{1.A.17}$$

These expressions give a reasonable description of the spectral response of quantum well based detectors.

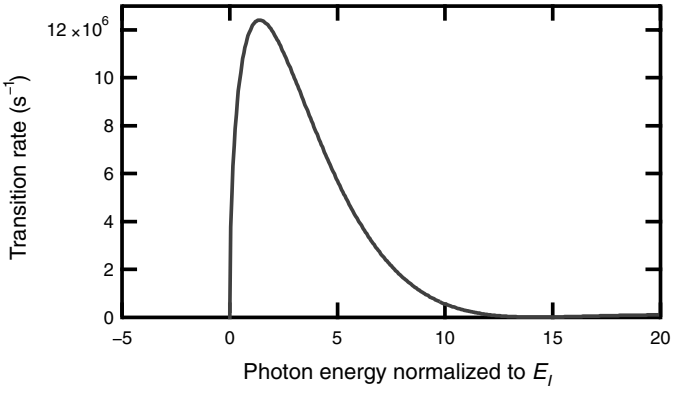


Fig. 1.A.2. Ionization transition rate (s^{-1}) of a quantum well as a function of incident photon energy in multiples of the ionization energy, E_I.

1.B Perturbation on a degenerate state

Consider a system described by a Hamiltonian H_0 which possesses a degenerate state E_n with degeneracy g_n (i.e. that g_n independent eigenvectors $\{|\psi_n^i\rangle, i = 1,\ldots,g_n\}$, forming an eigenvector subspace, share the same eigenenergy E_n). We seek the perturbation induced in this eigenvector subspace by a perturbing field $W = \alpha U$. Equation (1.54a) stipulates that the perturbed state always belongs to the eigenvector subspace but does not allow one to find the perturbed state $|0\rangle$ since all linear combinations of the eigenvectors $|\psi_n^i\rangle$ form possible solutions of this equation. It is therefore necessary to use (1.54b) to obtain the perturbed state $|0\rangle$. Projecting (1.54b) onto the vectors $|\psi_n^i\rangle$, we obtain:

$$\langle\psi_n^i|U|0\rangle = \varepsilon_1\langle\psi_n^i|0\rangle \tag{1.B.1}$$

We recall that the unknowns are the new perturbed state $|0\rangle$ and the perturbation $\Delta E = \alpha\varepsilon_1$. Equation (1.B.1) is nothing other than the eigenvalue and eigenvector equation of the perturbation operator W in the eigenvector subspace $\{|\psi_n^i\rangle, i = 1,\ldots,g_n\}$. More convincingly, for the time being let us designate c_i as the component of $|0\rangle$ on the basis $|\psi_n^i\rangle$ (i.e. $c_i = \langle\psi_n^i|0\rangle$) and w_{ij} as the element of the perturbation matrix $w_{ij} = \langle\psi_n^i|W|\psi_n^j\rangle$. Written in matrix form (1.B.1) then becomes:

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & w_{ij} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}\begin{bmatrix} \cdot \\ c_i \\ \cdot \end{bmatrix} \Delta E \begin{bmatrix} \cdot \\ c_i \\ \cdot \end{bmatrix} \tag{1.B.2}$$

We recognize the above secular equation as corresponding to the diagonalization of the perturbation operator W in the subspace spanned by the eigenvectors $|\psi_n^i\rangle$.

In order to refine our ideas, we now turn to an example illustrating the use of this formalism. We consider two identical physical systems far removed from one another (see Fig. 1.B.1). The system on the *left* is centred at r_l, while the system on the *right* is centred at r_r. Their respective Hamiltonians are:

$$\begin{aligned} H_l &= \frac{p^2}{2m} + V(r_l) \\ H_r &= \frac{p^2}{2m} + V(r_r) \end{aligned} \tag{1.B.3}$$

These two systems share similar Hamiltonians and therefore have identical eigenenergies. For example, both $|l\rangle$ and $|r\rangle$ share the same ground state energy:

$$\begin{aligned} H_l|l\rangle &= E|l\rangle \\ H_r|r\rangle &= E|r\rangle \end{aligned} \tag{1.B.4}$$

Let us now bring these two systems into closer proximity with one another in such a manner as to cause each system to engender a perturbing influence on the other. We may therefore represent this effect by a perturbation Hamilton W. Equation (1.B.2) then takes the form:

$$\begin{bmatrix} w_{rr} & w_{rl} \\ w_{lr} & w_{ll} \end{bmatrix} \begin{bmatrix} c_r \\ c_l \end{bmatrix} = \Delta E \begin{bmatrix} c_r \\ c_l \end{bmatrix} \tag{1.B.5}$$

Without loss of generality, we may suppose $w_{rr} = w_{ll} = 0$; and w_{rl} to be real and equal to A. Equation (1.B.5) may then be rewritten as:

$$\begin{bmatrix} 0 & A \\ A & 0 \end{bmatrix} \begin{bmatrix} c_r \\ c_l \end{bmatrix} = \Delta E \begin{bmatrix} c_r \\ c_l \end{bmatrix} \tag{1.B.6}$$

From which the system admits the following eigenenergies and eigenvectors:

$$\begin{aligned} E^+ &= E + A \qquad |+\rangle = \frac{1}{\sqrt{2}}(|r\rangle + |l\rangle) \\ E^- &= E - A \qquad |-\rangle = \frac{1}{\sqrt{2}}(|r\rangle - |l\rangle) \end{aligned} \tag{1.B.7}$$

The ground state energy E is then no longer an eigenenergy of the states $|l\rangle$ and $|r\rangle$. We say that the interaction between the systems has *lifted the degeneracy* and that the states $|l\rangle$ and $|r\rangle$ have *hybridized*. This mechanism is represented schematically in Fig. 1.B.2. This mechanism forms the basis for chemical bonding. Let us

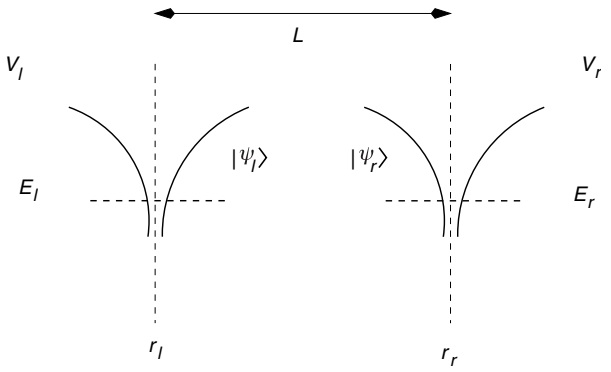


Fig. 1.B.1. A degenerate system consisting of two identical atoms separated spatially by L.

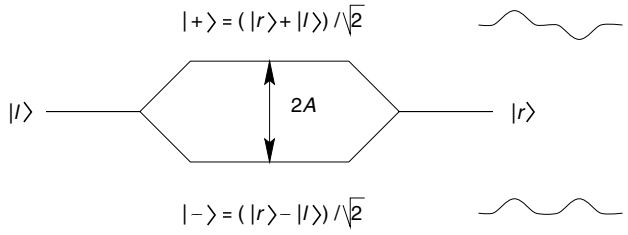


Fig. 1.B.2. When the two states $|l\rangle$ and $|r\rangle$ are brought into close proximity, they hybridize, thereby lifting the degeneracy of the system.

consider the simple example of two hydrogen atoms. Far away from one another, the energy of their two electrons is $2E$ (where E represents each electron's ground state energy). As they are brought together, the strength of their interaction W increases, lifting the degeneracy and the two electrons on the $|-\rangle$ level have now only a total energy of $2(E - A)$ (thanks to spin degeneracy). We continue below with a particularly illuminating example.

Example: coupled quantum wells

Advances in the growth of high quality, crystalline semiconductor heterostructures have allowed the growth of quantum wells separated by barriers measuring on the order of lattice constants (a few nanometres or more). We will consider two such quantum wells of width a, centred at $+b$ and $-b$, respectively (see Fig. 1.B.3). The energy depth of these wells is taken to be V_0. The total Hamiltonian for an electron in this system is then given by:

$$H_T = \frac{\hat{p}^2}{2m} + V(x - b) + V(x + b) \tag{1.B.8}$$

where V is a step function which equals $-V_0$ between $-a/2$ and $a/2$, and zero everywhere else. We have seen that we may write the total Hamiltonian as a sum of

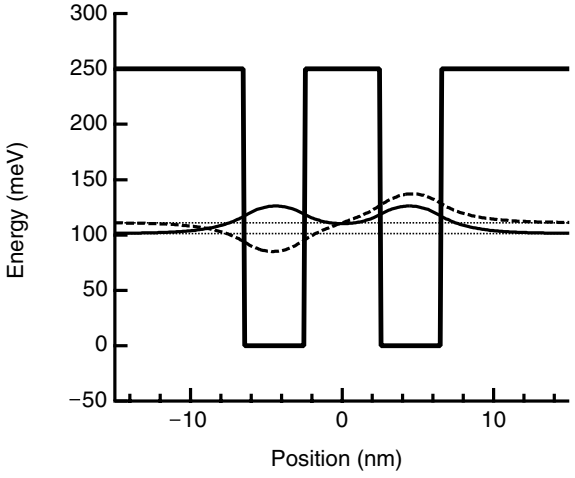


Fig. 1.B.3. Lifting of the degeneracy in two coupled quantum wells by electron tunnelling.

the unperturbed Hamiltonians in addition to that of the interaction perturbation:

$$H_T = H_r + V(x+b) = H_l + V(x-b) \tag{1.B.9}$$

Each well possesses the same ground state energy E_1 solution to Schrödinger's equation. If the depth V_0 of the quantum wells is sufficient, the wavefunctions for the stationary states in each of the wells may be written from the limited development of (1.44b) as:

$$\begin{aligned} \phi_l(x) &\approx \sqrt{\frac{2}{a}}\cos k(x+b), \text{ for } -(L/2+a) < x < -L/2 \\ \phi_l(x) &\approx \sqrt{\frac{2}{a}}\frac{\pi}{\kappa a}\mathrm{e}^{-\kappa(x+L/2)}, \text{ for } x > -L/2 \end{aligned} \tag{1.B.10}$$

and

$$\begin{aligned} \phi_r(x) &\approx \sqrt{\frac{2}{a}}\cos k(x-b), \text{ for } L/2 < x < L/2+a \\ \phi_r(x) &\approx \sqrt{\frac{2}{a}}\frac{\pi}{\kappa a}\mathrm{e}^{-\kappa(x-L/2)}, \text{ for } x < L/2 \end{aligned} \tag{1.B.11}$$

where k and κ are, respectively, the electron wavelength and the tunnelling penetration coefficient given in (1.33) with $L = 2b - a$. The matrix elements (1.B.5) are therefore easy to calculate. The off-diagonal terms are:

$$w_{lr} = -V_0 \int_{L/2}^{L/2+a} \phi_r(x)\phi_l(x)\mathrm{d}x$$
$$= -\frac{2\pi}{\kappa a^2} V_0 \int_0^a \mathrm{e}^{-\kappa u} \cos k(u - a/2)\mathrm{d}u \tag{1.B.12}$$

and yield with little difficulty:

$$w_{lr} \approx -\frac{4}{\kappa a} E_1 \mathrm{e}^{-\kappa L} \tag{1.B.13}$$

A similar calculation allows one to obtain the values for the diagonal elements. These are proportional to $\mathrm{e}^{-2\kappa L}$ and are negligible in comparison to the off-diagonal terms. The tunnelling effect coupling between the two quantum wells therefore lifts the degeneracy of the system by $2w_{lr}$ or:

$$\Delta E \approx \frac{8\mathrm{e}^{-\kappa L}}{\kappa a} E_1 \tag{1.B.14}$$

Lifting of degeneracy for two coupled quantum wells

For two 40 Å wide quantum wells, with potential depths of 250 meV, (1.43) gives a confinement energy E_1 of 106 meV and a tunnelling attenuation coefficient of $0.502\,\mathrm{nm}^{-1}$. If these two wells are separated by 5 nm, the degeneracy will be lifted by $(8 \times \mathrm{e}^{-0.502\times5})/(0.502 \times 4) \times 106$ meV or 34 meV. Figure 1.B.3 shows the result from an exact calculation, indicating an actual value for the splitting of 10 meV. The discrepancy between these two values originates from the fact that the infinite square well approximation is not valid for these finite square wells.

FURTHER READING

C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Wiley, New York (1992).
C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures*, Academic Press, Boston (1991).

1.C The quantum confined Stark effect

We consider a GaAs/AlGaAs quantum well where the growth axis is defined along the Oz direction (see Chapter 8). The electrons trapped in the quantum well conduction bands are described by the localized wavefunctions $|i\rangle$ and quantization energy E_i (Fig. 1.1). We are interested in uncovering the effect of a static

electric field F, applied across this structure, on the energy levels of the electrons situated within the well. We employ a dipole perturbation Hamiltonian of the form:

$$W = -qFz \tag{1.C.1}$$

If the quantum well is symmetric, the perturbation is clearly null to first order. Neglecting the contribution from states with high quantum numbers ($i > 2$), (1.62) then gives the second-order perturbation correction to E_i from the electric field:

$$\Delta E_1 = E_1' - E_1 = -\Delta E_2 = q^2F^2\frac{|z_{12}|^2}{E_1 - E_2} \tag{1.C.2}$$

where $z_{12} = \langle 1|z|2\rangle$ is the element of the dipole matrix. This calculation may be performed simply if one uses the infinite square well approximation. Equation (1.49) gives the expressions for the energy levels E_1 and E_2, and (1.48) gives those for the corresponding wavefunctions. Integration along the width of the well leads us to the dipole element:

$$|z_{12}| = \frac{2^4}{3^2}\frac{a}{\pi^2} \tag{1.C.3}$$

The energy levels E_1 and E_2 become increasingly separated as the field strength F is augmented (see Figure 1.C.1).

In spectroscopic terms, we say that the resonant optical transition connecting these two levels is *blue-shifted* in energy by:

$$\Delta E = 2|\Delta E_1| = \frac{2^{10}}{3^5\pi^6}\frac{q^2F^2a^4m^*}{\hbar^2} \tag{1.C.4}$$

or after some degree of simplification:

$$\Delta E = \frac{2^9}{3^5\pi^4}\left(\frac{qFa}{E_0}\right)qFa \tag{1.C.5}$$

where E_0 is the confinement energy from (1.49). We recall that m^* is the effective mass of the electron and is equal to $0.067m_e$ in GaAs. We see, therefore, that the effect begins to become important when the potential drop across the quantum well qFa becomes non-negligible in comparison to the quantum confinement energy E_0. Perturbation theory clearly breaks down when the electron wavefunction does not vanish outside the quantum well (see Fig. 1.C.1). In that case, the electron in the ground state can tunnel through the triangular barrier, leading to electric field ionization of the quantum well.

Example

In a GaAs/AlGaAs quantum well with a thickness of 12 nm and possessing an

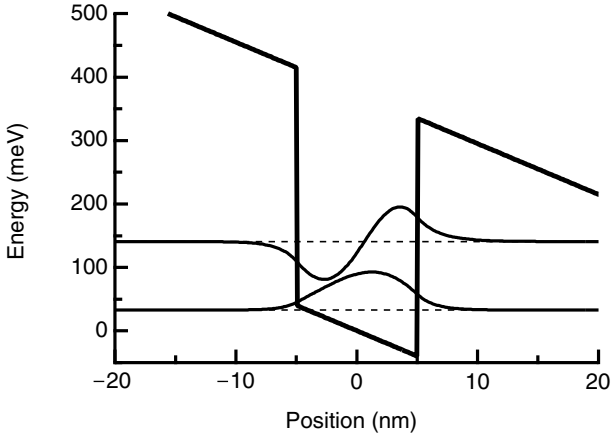


Fig. 1.C.1. A quantum well identical to that in Fig. 1.1, under the influence of a static electric field. The energy separation between the first two energy levels increases as a result of the applied field. This effect is the *blue shift* due to the Stark effect.

effective electron mass of $m^* = 0.067m_e$, a Stark shift of $\Delta E = 2^9/(3^5 \times \pi^4) \times (10^5\,\mathrm{V\,cm^{-1}} \times 10^{-6}\,\mathrm{cm})^2/0.115\,\mathrm{eV}$ or 2 meV, is expected for an applied electric field of $10\,\mathrm{V\,\mu m^{-1}}$. This shift is enormous in comparison with the magnitude of shifts which may be induced in atoms. This is a direct result of the low effective mass of electrons in GaAs and of the much greater electric field strengths which may be applied to condensed matter structures with the aid of conventional microfabrication techniques (A. Harwit and J. J. S. Harris, *Appl. Phys. Lett.* **50**, 685 (1987)).

This shift is very important as it can be larger than the linewidth of the $E_1 \rightarrow E_2$ transition (typically 5 meV), and was observed for the first time by Harwit and Harris (1987). Figure 1.C.2 compares their experimental results with theoretical values. Their results for the symmetric well are seen to be compatible with a parabolic dependence of the Stark shift on the electric field strength.

If the quantum well is asymmetric (see Fig. 1.C.3), then the first-order perturbation is non-zero and is equal to:

$$\Delta E = qF\delta_{12} \tag{1.C.6}$$

where $\delta_{12} = \langle 1|z|1\rangle - \langle 2|z|2\rangle$ is the displacement of the average position of the electron from the $|1\rangle$ state to the $|2\rangle$ state.

Example

For the quantum well in Fig. 1.C.3, the solution to Schrödinger's equation yields a displacement dipole $\delta_{12} = 4.2\,\mathrm{nm}$. We derive a Stark shift for the transition $E_1 \rightarrow E_2$: $\Delta E = 42\,\mathrm{meV}$ for an applied field of $10\,\mathrm{V\,\mu m^{-1}}$. The effect, to first order

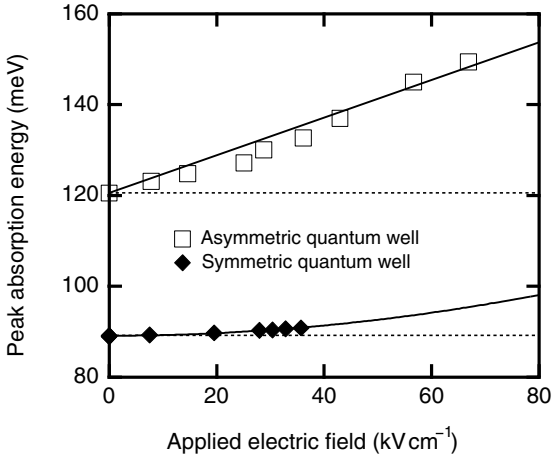
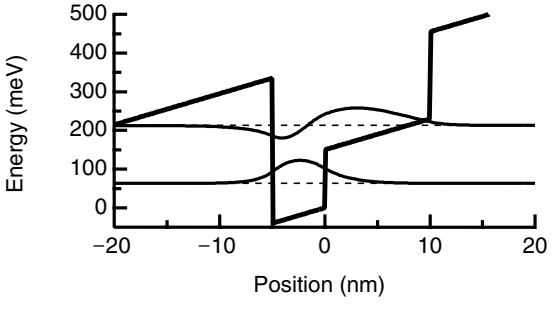
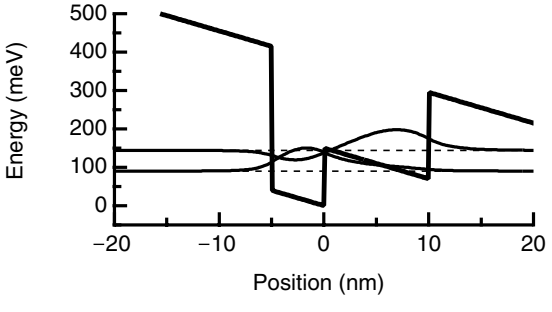


Fig. 1.C.2. Induced Stark shift to the optical transition energy in symmetric and asymmetric quantum wells. The differential behaviour of these wells is discussed in the two examples earlier.



(*a*)



(*b*)

Fig. 1.C.3. Asymmetric quantum well under the influence of an electric field.

then, is much more important than in the preceding example involving the symmetric well (E. Martinet, F. Luc, E. Rosencher, P. Bois, E. Costard, S. Delaître, and E. Böckenhoff, in *Intersubband Transitions in Quantum Wells*, edited by E. Rosencher, B. Vinter, and B. Levine, Plenum, London (1992) p. 299).

This experiment has been performed in asymmetric GaAs/AlGaAs structures. Figure 1.C.2 shows the experimental results for the Stark shift as a function of the applied field. In these data, the linear dependence of the transition resonance shift on field strength is apparent and in agreement with the predictions of (1.C.6) for the parameters given in the example above. This effect has been harnessed in the new generation of infrared detectors and will be explored further in Chapter 11.

FURTHER READING

G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Wiley, New York (1991).

E. Rosencher, B. Vinter, and B. Levine, Eds, *Intersubband Transitions in Quantum Wells*, Plenum Press, New York (1992).

C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures*, Academic Press, Boston (1991).

1.D The harmonic oscillator

The study of the harmonic oscillator in quantum mechanics provides one with an exceptional wealth of physical insights. We shall therefore investigate the quantum mechanical description of an electron bound elastically to an attractive centre. This system is of equal use whether describing the behaviour of an electron bound to a nucleus, the vibration of atoms in a crystal, or even, as we shall soon see in Chapter 2, the mode of an electromagnetic wave, otherwise referred to as a *photon*.

We will now consider a particle of mass m subjected to an attractive one-dimensional parabolic potential:

$$V(x) = \frac{1}{2}kx^2 \tag{1.D.1}$$

where k is the force constant of the quantum mechanical spring (see Fig. 1.D.1). (While generalization of this problem to three dimensions does not pose any great theoretical difficulties, the plethora of indices required to do so would add little more than clutter to the present discussion.) The particle's Hamiltonian:

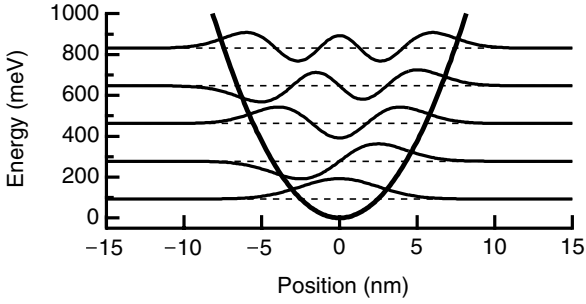


Fig. 1.D.1. Harmonic oscillator potential showing confined energy levels and wavefunctions for its corresponding eigenstates.

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 x^2 \tag{1.D.2}$$

Classical Hamiltonian for the harmonic oscillator

where ω_0 is the resonant frequency of the system:

$$\omega_0 = \sqrt{\frac{k}{m}} \tag{1.D.3}$$

The classical behaviour in this case is well known: the system oscillates and the position of the particle is given by:

$$x_{\text{class}}(t) = x_0 \sin(\omega_0 t + \phi) \tag{1.D.4}$$

where x_0 is the amplitude of the oscillation and ϕ is the phase. To obtain the quantum mechanical behaviour of the particle, the stationary states are found by solving:

$$\left(\frac{\hat{p}^2}{2m} + \frac{1}{2}m\omega_0^2\hat{x}^2\right)|\psi\rangle = E|\psi\rangle \tag{1.D.5}$$

The correspondence principle gives us the form of Schrödinger's equation in the $\hat{\mathbf{r}}$ ($\hat{\mathbf{p}} = -\mathrm{i}\hbar\,(\mathrm{d}/\mathrm{d}x)$) representation:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + \frac{1}{2}m\omega_0^2 x^2\psi(x) = E\psi(x) \tag{1.D.6}$$

Schrödinger's equation for a stationary one-dimensional harmonic oscillator

We need to find physically admissible solutions to this equation (i.e. square integrable solutions). This differential equation can be solved by many diverse means (the polynomial method, for instance), but it is much more powerful to tame

this problem using operator algebra as first developed by P. A. M. Dirac. We begin by introducing the dimensionless operators:

$$\tilde{x} = \sqrt{\frac{m\omega_0}{\hbar}}\hat{x}$$

$$\tilde{p} = \frac{1}{\sqrt{m\hbar\omega_0}}\hat{p} \tag{1.D.7}$$

The Hamiltonian may then be written:

$$H = \hbar\omega_0\tilde{H} \tag{1.D.8}$$

where the dimensionless operator $\tilde{H}$ is given by:

$$\tilde{H} = \frac{1}{2}[\tilde{x}^2 + \tilde{p}^2] \tag{1.D.9}$$

We recall that the anticommutation relation between the *position* and *momentum* operators is given by:

$$[\hat{x}, \hat{p}] = \hat{x}\hat{p} - \hat{p}\hat{x} = \mathrm{i}\hbar \tag{1.D.10}$$

where for dimensionless operators:

$$[\tilde{x}, \tilde{p}] = \mathrm{i} \tag{1.D.11}$$

We will now show that this anticommutation relation – which is nothing more than a reformulation of the first of Heisenberg's uncertainty relations – also leads to energy quantization. We seek therefore to solve the system:

$$\frac{1}{2}(\tilde{x}^2 + \tilde{p}^2)|\psi\rangle = \varepsilon|\psi\rangle$$
$$[\tilde{x}, \tilde{p}] = \mathrm{i} \tag{1.D.12}$$

To do this, we need two new operators:

$$a = \frac{1}{\sqrt{2}}(\tilde{x} + \mathrm{i}\tilde{p}) = \frac{1}{\sqrt{2}}\left(\sqrt{\frac{m\omega_0}{\hbar}}\hat{x} + \mathrm{i}\frac{1}{\sqrt{m\hbar\omega_0}}\hat{p}\right)$$
$$a^+ = \frac{1}{\sqrt{2}}(\tilde{x} - \mathrm{i}\tilde{p}) = \frac{1}{\sqrt{2}}\left(\sqrt{\frac{m\omega_0}{\hbar}}\hat{x} - \mathrm{i}\frac{1}{\sqrt{m\hbar\omega_0}}\hat{p}\right) \tag{1.D.13}$$

Definition of the creation and annihilation operators

The position and momentum operators are given by:

$$\hat{x} = \sqrt{\frac{\hbar}{2m\omega_0}}(a + a^+)$$

$$\hat{p} = \frac{1}{\mathrm{i}}\sqrt{\frac{\hbar m\omega_0}{2}}(a - a^+) \tag{1.D.14}$$

For reasons which shall become apparent shortly, a will be referred to as the *annihilation operator* and a^+ the *creation operator*. Let us calculate the product a^+a:

$$a^+a = \frac{1}{2}(\tilde{x} - \mathrm{i}\tilde{p})(\tilde{x} + \mathrm{i}\tilde{p}) = \frac{1}{2}[\tilde{x}^2 + \tilde{p}^2 + \mathrm{i}(\tilde{x}\tilde{p} - \tilde{p}\tilde{x})]$$

or

$$a^+a = \frac{1}{2}(\tilde{x}^2 + \tilde{p}^2 + \mathrm{i}[\tilde{x}, \tilde{p}])$$

so that:

$$\tilde{H} = a^+a + \frac{1}{2} \tag{1.D.15}$$

We now calculate the commutator of the annihilation and creation operators:

$$[a, a^+] = \frac{1}{2}[\tilde{x} + \mathrm{i}\tilde{p}, \tilde{x} - \mathrm{i}\tilde{p}] = \frac{1}{2}([\tilde{x}, -\mathrm{i}\tilde{p}] - [\tilde{x}, +\mathrm{i}\tilde{p}]) = \frac{1}{2}[(-\mathrm{i})\mathrm{i} - \mathrm{i}(+\mathrm{i})]$$

so that:

$$[a, a^+] = 1 \tag{1.D.16}$$

Equation (1.D.16) is therefore an expression for the uncertainty principle in terms of the language of annihilation and creation operators. We then introduce the operator N:

$$N = a^+a \tag{1.D.17}$$

N is referred to as the *number operator*. The Hamiltonian for the harmonic oscillator and the number operator are related by:

$$H = \hbar\omega_0\left(N + \frac{1}{2}\right) \tag{1.D.18}$$

We will now show that the eigenvalues of N are either positive integer numbers or equal to zero. To do so, however, we will require the proof of three separate theorems. Let $|v\rangle$ be an eigenvector of N with eigenvalue v or:

$$N|v\rangle = v|v\rangle \tag{1.D.19}$$

Theorem 1: The eigenvalues of the N operator are either positive or null.

Rewriting (1.D.19) as:

$a^+ a|v\rangle = v|v\rangle$

we may project this equation on $|v\rangle$, to yield:

$\langle v|a^+ a|v\rangle = v\langle v|v\rangle = v$

This may also be written:

$\langle v|a^+ a|v\rangle = (a|v\rangle)^+(a|v\rangle) = |(a|v\rangle)|^2$

so that:

$v = |(a|v\rangle)|^2$

From which we deduce that v is non-negative and that:

$$v = 0, \quad \text{if and only if } a|v\rangle = 0 \tag{1.D.20}$$

Theorem 2: If $|v\rangle$ is an eigenvector of N with eigenvalue v, then $a|v\rangle$ is also an eigenvector of N with eigenvalue $v - 1$.
Indeed,

$Na|v\rangle = (a^+ a)a|v\rangle$

Since the anticommutation relation stipulates that $aa^+ - a^+ a = 1$, this last relationship may be written:

$Na|v\rangle = (aa^+ - 1)a|v\rangle = a(a^+ a - 1)|v\rangle = a(N - 1)|v\rangle = (v - 1)a|v\rangle$

which is what we wished to show.

Theorem 3: If $|v\rangle$ is an eigenvector of N with eigenvalue v, then $a^+|v\rangle$ is also an eigenvector of N with eigenvalue $v + 1$.
In the same fashion then:

$$\begin{aligned} Na^+|v\rangle &= (a^+ a)a^+|v\rangle = a^+(aa^+)|v\rangle \\ &= a^+(1 + a^+ a)|v\rangle = a^+(1 + N)|v\rangle \\ &= (v + 1)a^+|v\rangle \end{aligned}$$

We are now in a position to find the eigenvalues of N. Let us suppose that v, the eigenvalue of N, is non-integer and not equal to zero, with $n = \text{Int}(v)$, where 'Int' means 'integer part of'. Then $a|v\rangle$ is an eigenvector of N with eigenvalue $v - 1$; $a^2|v\rangle$ is an eigenvector with eigenvalue $v - 2, \ldots$; $a^n|v\rangle$ is an eigenvector with eigenvalue $v - n = v - \text{Int}(v)$, which must lie between 0 and 1 (see Fig. 1.D.2). As v is non-integer, $v - n$ must be different from 0, and by extension $a^{n+1}|v\rangle$ would have an eigenvalue of $v - n - 1$, which must be negative. This, however, is not possible given theorem 1. The value of v must be such that the vector $a^{n+1}|v\rangle$ cannot exist,

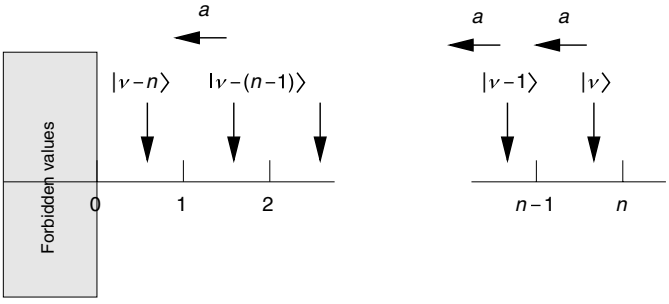


Fig. 1.D.2. Each application of the annihilation operator a destroys a single excitation quantum.

i.e. $a^n|v\rangle = 0$. Therefore, the eigenvalues v of N are either positive integers or equal to zero. As we may generate from any eigenvector $|m\rangle$ any other eigenvector $|n\rangle$ by the application of $m - n$ lowering or raising operators a or a^+, N allows as eigenvalues the total ensemble of positive integers, i.e:

$$N|n\rangle = n|n\rangle, n = 0, 1, 2, \ldots \tag{1.D.21}$$

From (1.D.18), we see that the energies for the stationary states of the harmonic oscillator are quantized and given by:

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega_0, n = 0, 1, 2, \ldots \tag{1.D.22}$$

Energy levels for the one-dimensional harmonic oscillator

and its stationary states are the eigenstates $|n\rangle$ of the *number* operator. The energy spectrum of the harmonic oscillator is then composed of a series of equidistant levels separated in energy by $\hbar\omega_0$ (see Fig. 1.D.1).

We now wish to uncover how the operators a and a^+ act. We know, from theorem 2, that $a|n\rangle$ is an eigenvector of N with an eigenvalue of $n - 1$. As we see that the energy levels of the one-dimensional oscillator are non-degenerate (which can easily be proven), this implies that $a|n\rangle$ is proportional to $|n - 1\rangle$. Therefore:

$$a|n\rangle = c|n - 1\rangle, n = 1, 2, \ldots$$

But the norm of $a|n\rangle$ is given by:

$$|(a|n\rangle)|^2 = (a|n\rangle)^+(a|n\rangle) = \langle n|a^+ a|n\rangle = \langle n|N|n\rangle = n$$

so that $c^2 = n$, from which we obtain:

$$a|n\rangle = \sqrt{n}|n - 1\rangle, n = 1, 2, \ldots \tag{1.D.23}$$

In a rigorously identical fashion, we find for a^+:

$$a^+|n\rangle = \sqrt{n+1}|n+1\rangle, n = 0, 1, \ldots \tag{1.D.24}$$

We are now in a better position to understand the treatment of the problem in terms of Dirac's operator algebra. To begin with, each individual quantized state contains greater energy with increasing n. Equation (1.D.23) shows that the effect of the a operator is to initiate a transition between a state with quantum number n to $n-1$. The a operator therefore removes a quantum of excitation corresponding to $\hbar\omega_0$ from the system. In opposing fashion, a^+ adds a quantum of excitation equal to $\hbar\omega_0$. It is important to remember that the operators a and a^+ are not Hermitian and do not possess real number eigenvalues. Equation (1.D.24) can be used as a recurrence relation to generate all the stationary states $|n\rangle$ of H:

$$|n\rangle = \frac{(a^+)^n}{\sqrt{n!}}|0\rangle \tag{1.D.25}$$

We now turn our attention to the $|0\rangle$ state. This is the ground state of the system and it possesses a lowest possible energy of:

$$E_0 = \frac{1}{2}\hbar\omega_0 \tag{1.D.26}$$

In contrast to a classical oscillator, where a minimum energy state of zero (or resting state) is allowed, the minimum energy state of a quantum harmonic oscillator is non-zero. This astonishing observation results from the anticommutation relation (1.D.16) which constitutes a reformulation of Heisenberg's first uncertainty relation. We might therefore expect to be able to derive this result directly from an application of the uncertainty principle, which states that the uncertainties for x and p are related by:

$$\Delta x \Delta p \approx \hbar \tag{1.D.27}$$

In classical terms, the total energy of the ground state of the particle is given by:

$$E = \frac{\Delta p^2}{2m} + \frac{1}{2}m\omega_0^2\Delta x^2 \tag{1.D.28}$$

Substituting (1.D.27) into (1.D.28), we obtain for energy E:

$$E = \frac{\hbar^2}{2m\Delta x^2} + \frac{1}{2}m\omega_0^2\Delta x^2 \tag{1.D.29}$$

The variation of E as a function of Δx is given in Fig. 1.D.3. This curve exhibits a minimum value at:

$$\Delta x = \sqrt{\frac{\hbar}{m\omega_0}}$$

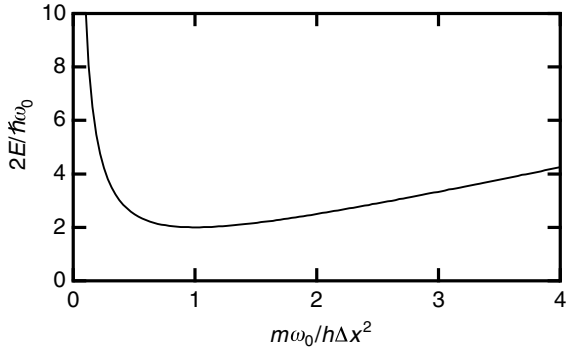


Fig. 1.D.3. Variation of the reduced energy of a harmonic oscillator as a function of the reduced position uncertainty in (1.D.29).

corresponding to minimum energy of $E = \hbar\omega_0$. Not too bad for a heuristic approach!

We have yet to obtain the eigenstates for the oscillator in position representation, i.e. to find the wavefunctions for its stationary states. To do so, we shall start with the ground state and apply the annihilation operator (1.D.20):

$$a|0\rangle = 0$$

or

$$(\tilde{x} + i\tilde{p})|0\rangle = 0$$

Returning to the definition (1.D.7), this gives in r representation:

$$\left(\sqrt{\frac{m\omega_0}{\hbar}}x + \hbar\frac{1}{\sqrt{m\hbar\omega_0}}\frac{\mathrm{d}}{\mathrm{d}x}\right)\phi_0(x) = 0$$

or again

$$\frac{m\omega_0}{\hbar}x\phi_0(x) + \frac{\mathrm{d}}{\mathrm{d}x}\phi_0(x) = 0 \tag{1.D.30}$$

The solution to this classic differential equation is easily found to be a Gaussian given by:

$$\phi_0(x) = \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4} \mathrm{e}^{-(m\omega_0/\hbar)(x^2/2)} \tag{1.D.31}$$

The higher index wavefunctions are then obtained by recurrence. We can now make predictions on the measurement of quantum observables. For instance, the average value of the position of the particle in state $|n\rangle$ is given by:

$$\bar{x} = \langle x \rangle = \langle n|x|n\rangle = \sqrt{\frac{\hbar}{m\omega_0}}\langle n|(a + a^+)|n\rangle$$

Given that $a^+|n\rangle = \sqrt{n}|n-1\rangle$ and since $|n\rangle$ and $|n-1\rangle$ are orthogonal eigenvectors of N possessing different eigenvalues, the average value of x is therefore null (not surprising given the symmetry of the system). To calculate the mean square of the position, we rewrite the operator x^2 in terms of a and a^+:

$$\begin{aligned} x^2 &= \frac{\hbar}{2m\omega_0}(a^2 + a^{+2} + aa^+ + a^+a) \\ &= \frac{\hbar}{2m\omega_0}(a^2 + a^{+2} + 1 + 2a^+a) \end{aligned} \tag{1.D.32}$$

The mean value of the square of the position observable for a given state $|n\rangle$ is then:

$$\overline{x^2} = \langle n|x^2|n\rangle = \frac{\hbar}{2m\omega_0}\langle n|(a^2 + a^{+2} + 1 + 2a^+a)|n\rangle$$

Since a^2 and $(a^+)^2$ send $|n\rangle$ into states which are orthogonal to $\langle n|$, the contribution of these operators is null, leading to:

$$\overline{x_n^2} = \frac{\hbar}{m\omega_0}\left(n + \frac{1}{2}\right) \tag{1.D.33}$$

Therefore, the greater the energy of the quantum oscillator mode, the greater the degree of uncertainty associated with the position of the particle.

We may well find ourselves shocked to discover that the expectation value for the position of the particle in the individual eigenstates of the oscillator *does not oscillate in real space*. We shall see that the creation of states with time-varying expectation values that lie in step with our classical physical intuition, will require the construction of carefully selected linear combinations (Glauber states) of these more fundamental stationary states.

Classical example

We consider a 1 gram sphere coupled to a spring and having a natural oscillation frequency of 1 kHz. The quantum uncertainty on the position in the ground state is given by (1.D.33) with $n = 0$, giving $\Delta x = 10^{-17}$ m! Clearly, the quantum behaviour of such a system is negligible.

Quantum example: Einstein's phonons

We now consider a periodic chain of interacting atoms forming a crystalline solid. The potential wells which bind each of the atoms to the next have an energy minimum at a relative distance r_0 from each atom. In proximity to this equilibrium

position, the Hamiltonian for any pair of neighbouring atoms may be approximated as:

$$H = \frac{p_r^2}{2m} + \frac{1}{2}U''x_r^2 \tag{1.D.34}$$

where x_r and p_r correspond to the respective distances and momenta between each atom, and U'' is the second derivative of the potential at the minimum. Each pair of neighbouring atoms then form a harmonic oscillator possessing a resonant frequency of:

$$\omega_0 = \sqrt{\frac{U''}{m}} \tag{1.D.35}$$

We may approximate the crystalline potential by a periodic field, possessing a typical modulation amplitude of 5 eV, every 5 Å. The potential may then be written:

$$U(x) = 5\,\text{eV}\sin\left(\frac{2\pi x_r}{0.5\,\text{nm}}\right)$$

and the second derivative U'' equals $5\,\text{eV} \times 4\pi^2/(0.5\,\text{nm})^2$ or $126\,\text{kg}\,\text{s}^{-2}$. For atoms consisting of roughly 50 nucleons (i.e. having a mass of $8.35 \times 10^{-26}\,\text{kg}$), the natural oscillation frequency of the atoms is given by (1.D.35) and is equal to $\sqrt{(126\,\text{kg}\,\text{s}^{-2}/8.35 \times 10^{-26}\,\text{kg})}/2\pi$ or $6 \times 10^{12}\,\text{s}^{-1}$. These lattice vibrations (referred to as *longitudinal phonons* or as *Einstein's phonons*) are commensurate with the infrared spectrum, possessing interaction energies of the order of 25 meV.

FURTHER READING

C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Wiley, New York (1992).

1.E Transition probabilities and Rabi oscillations

We saw in Section 1.6 that in a two-level system, a sinusoidal perturbation with an oscillation frequency of ω induces transitions between the two levels with probabilities being:

- proportional to t^2, if the transition is monochromatic and resonant (i.e. if the excitation quantum $\hbar\omega$ is equal to the energy difference between the two levels $E_2 - E_1$);
- proportional to t if the excitation is polychromatic, or if the final state belongs to a continuum.

These results are absurd to the extent that they diverge in the large t limit. This behaviour is clearly not admissible, as a probability value cannot exceed unity. We will show that the density matrix allows for yet another system behaviour that can reconcile these difficulties.

We will suppose, in a first approach, that the relaxation times of the population and the phase (T_1 and T_2, respectively) are infinite, i.e. the systems are not subject to random fluctuations or *damping*. Equation (1.100) is then written:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\rho = -2\mathrm{i}\Omega_{12}(\rho_{21} - \rho_{12})\cos\omega t \tag{1.E.1a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{21} = -\mathrm{i}\omega_{21}\rho_{21} - \mathrm{i}\Omega_{12}\Delta\rho\cos\omega t \tag{1.E.1b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{12} = \mathrm{i}\omega_{21}\rho_{12} + \mathrm{i}\Omega_{12}\Delta\rho\cos\omega t \tag{1.E.1c}$$

where $\Delta\rho$ is the *population difference* $\rho_{11} - \rho_{22}$, and the frequency Ω_{12} is given by:

$$m_{12} = \hbar\Omega_{12} \tag{1.E.2}$$

Expecting a resonant oscillatory behaviour near ω_{21}, we introduce a *detuning* frequency $\delta\omega$:

$$\delta\omega = \omega - \omega_{21} \tag{1.E.3}$$

and the following new variables:

$$\sigma_{21} = \rho_{21}\mathrm{e}^{\mathrm{i}\omega_{21}t} \tag{1.E.4a}$$

$$\sigma_{12} = \rho_{12}\mathrm{e}^{-\mathrm{i}\omega_{21}t} \tag{1.E.4b}$$

Substituting these new variables into (1.E.1) and keeping only those terms in $\delta\omega$ (quasi-resonance approximation):

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\rho = -\mathrm{i}\Omega_{12}(\sigma_{21}\mathrm{e}^{\mathrm{i}\delta\omega t} - \sigma_{12}\mathrm{e}^{-\mathrm{i}\delta\omega t}) \tag{1.E.5a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{21} = -\mathrm{i}\frac{\Omega_{12}}{2}\Delta\rho\,\mathrm{e}^{-\mathrm{i}\delta\omega t} \tag{1.E.5b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{12} = \mathrm{i}\frac{\Omega_{12}}{2}\Delta\rho\,\mathrm{e}^{+i\delta\omega t} \tag{1.E.5c}$$

The optical Bloch equations

Equations (1.E.5a) to (1.E.5c) are called the optical Bloch equations. These may be solved by introducing the trial functions:

$$\Delta\rho = \Delta\rho^0 e^{\lambda t} \tag{1.E.6a}$$

$$\sigma_{21} = \sigma_{21}^0 e^{-i\delta\omega t} e^{\lambda t} \tag{1.E.6b}$$

$$\sigma_{12} = \sigma_{12}^0 e^{+i\delta\omega t} e^{\lambda t} \tag{1.E.6c}$$

Substituting (1.E.6a) to (1.E.6c), into (1.E.5a) to (1.E.5c), the system of equations easily takes on the following matrix form:

$$\begin{bmatrix} \lambda & i\Omega_{12} & -i\Omega_{12} \\ \frac{i\Omega_{12}}{2} & \lambda - i\delta\omega & 0 \\ \frac{-i\Omega_{12}}{2} & 0 & \lambda - i\delta\omega \end{bmatrix} \begin{bmatrix} \Delta\rho^0 \\ \sigma_{21}^0 \\ \sigma_{12}^0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.E.7}$$

In order to obtain non-trivial solutions, the determinant of Eq. (1.E.7) must be zero. Therefore, λ can only take on three possible values:

$$\lambda = 0 \quad \lambda = i\Omega \quad \lambda = -i\Omega$$

where Ω is the *Rabi frequency*, given by:

$$\Omega = \sqrt{(\omega - \omega_{21})^2 + \Omega_{12}^2} \tag{1.E.8}$$

We may therefore find the behaviour of the elements in the density matrix as a function of time. We will return in Chapter 3, to the significance of the off-diagonal elements, σ_{12} and σ_{21}, which provide a basis for the description of absorption phenomena. For the time being, we will interest ourselves in the transition probabilities, i.e. the diagonal elements ρ_{11} and ρ_{22}.

We suppose that at time $t = 0$, the system is in its ground state $|1\rangle$, i.e.:

$$\begin{aligned} &\rho_{11}^0 = 1; \rho_{22}^0 = 0 \\ &\sigma_{12}^0 = \sigma_{12}^0 = 0 \end{aligned} \tag{1.E.9}$$

The probability that the system occupies the first excited state $|2\rangle$ as a function of time is given by:

$$\rho_{22}(t) = \frac{\Omega_{12}^2}{(\omega - \omega_{21})^2 + \Omega_{12}^2} \sin^2\left(\frac{1}{2}\Omega t\right) \tag{1.E.10}$$

Figure 1.E.1 shows this probability for different ratios of $\delta\omega/\Omega_{12}$.

We see that this probability oscillates with a frequency of Ω and that the oscillation amplitude is at a maximum and equal to 1 when $\delta\omega = 0$ (i.e. when the excitation is resonant with the energy separation between the two levels). In this case, (1.E.10) gives:

$$\rho_{22}(t) = \sin^2\left(\frac{1}{2}\Omega_{12}t\right) \tag{1.E.11}$$

For short times ($\Omega_{12}t < 1$), the occupation probability takes the form:

$$\rho_{22}(t) = \frac{1}{4}\Omega_{12}^2 t^2 \tag{1.E.12}$$

which is nothing else but Eq. (1.79) obtained using time-dependent perturbation theory. As may be seen in Fig. 1.E.1, the sine function has a reasonably parabolic shape for short times. Consequently, the density matrix formalisms reflect this parabolic behaviour for small t, fix the limit of applicability for time-dependent perturbation theory, and predict an oscillatory behaviour over extended periods of time. This long-term oscillatory nature of occupation probability has been observed in atoms and is referred to as *Rabi oscillations.*

We now seek to describe the system evolving under the influences of relaxation mechanisms. Equation (1.101) may then be written:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\rho = -2\mathrm{i}\Omega_{12}(\rho_{21} - \rho_{12})\cos\omega t - \frac{\Delta\rho - \Delta\rho^{\mathrm{eq}}}{T_1} \tag{1.E.13a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{21} = -\mathrm{i}\omega_{21}\rho_{21} - \mathrm{i}\Omega_{12}\Delta\rho\cos\omega t - \frac{\rho_{21}}{T_2} \tag{1.E.13b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{12} = \mathrm{i}\omega_{21}\rho_{12} + \mathrm{i}\Omega_{12}\Delta\rho\cos\omega t - \frac{\rho_{12}}{T_2} \tag{1.E.13c}$$

where $\Delta\rho$ is the population difference $\rho_{11} - \rho_{22}$ and $\Delta\rho^{\mathrm{eq}}$ is the population difference at thermodynamic equilibrium. The approach is strictly identical to that employed in equation (1.E.1), with the exception that the equations become significantly more complicated. Limiting ourselves to the particular case of resonance by setting $\delta\omega = 0$ and $T_2 = T_1/2 = \gamma^{-1}$, however, will keep the expressions more manageable. Taking as initial conditions those contained in (1.E.9), we obtain:

$$\rho_{22}(t) = \frac{\frac{1}{2}\Omega_{12}^2}{\Omega_{12}^2 + 2\gamma^2}\left[1 - \left(\cos\lambda t + \frac{3\gamma}{2\lambda}\sin\lambda t\right)\mathrm{e}^{-3\gamma t/2}\right] \tag{1.E.14}$$

where this time λ is given by:

$$\lambda = \sqrt{\Omega_{12}^2 - \frac{1}{4}\gamma^2} \tag{1.E.15}$$

This behaviour is represented in Fig. 1.E.2. We notice that after a transition period

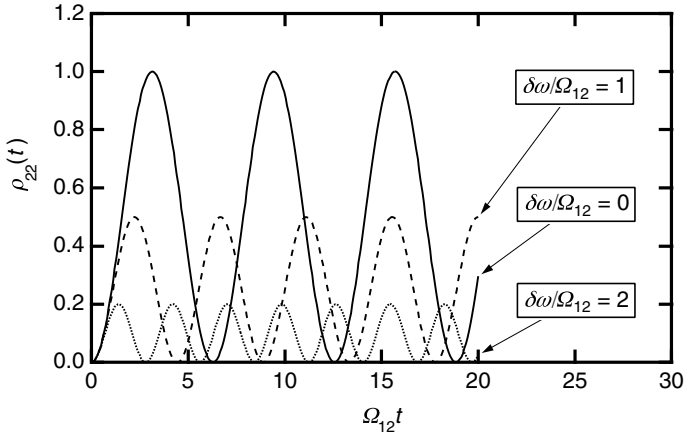


Fig. 1.E.1. Time evolution of the occupation probability ρ_{22} for the $|2\rangle$ level for different ratios of the *detuning* and Rabi frequencies $\delta\omega/\Omega_{12}$.

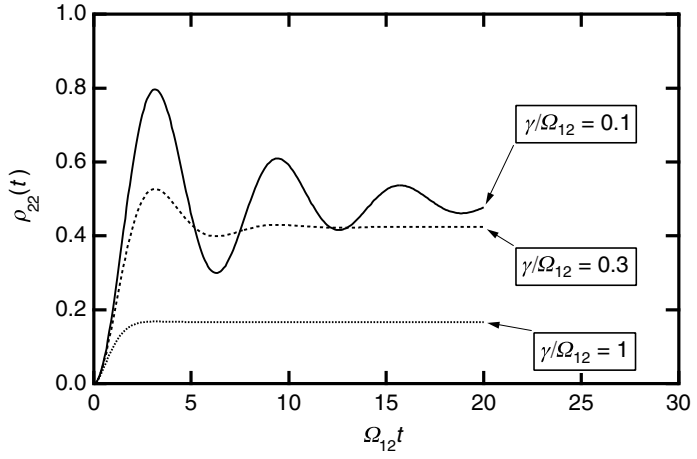


Fig. 1.E.2. Time evolution of occupation probability for the $|2\rangle$ level for zero detuning and different ratios of damping rate and Rabi frequency γ/Ω_{12}.

the system dampens (i.e. coherence is lost between states $|1\rangle$ and $|2\rangle$) and maintains a stationary probability of finding itself in the excited state given by:

$$\rho_{22}^{\text{stat}} = \frac{\frac{1}{2}\Omega_{12}^2}{\Omega_{12}^2 + 2\gamma^2} \tag{1.E.16}$$

Figure 1.E.3 shows the evolution of the populations ρ_{11}^{stat} and ρ_{22}^{stat} as a function of the coupling strength Ω_{12}. We notice that the populations tend to equilibrate to 1/2, when the coupling strength increases. This mechanism, referred to as *optical*

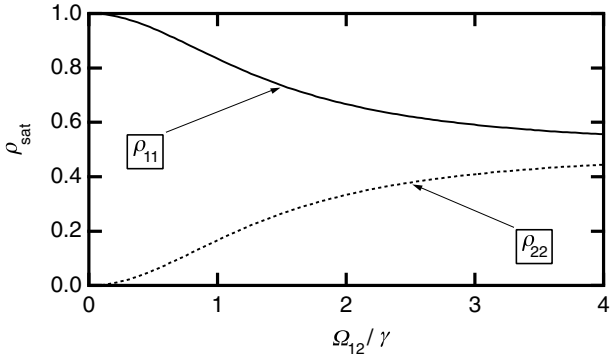


Fig. 1.E.3. Variation of the stationary populations for the two levels as a function of the ratio of the Rabi frequency to the damping rate.

saturation, will be interpreted later in terms of Einstein's corpuscular theory of light when discussing induced and simulated emission.

We appreciate therefore the versatility and power of the density matrix formalism in describing the non-stationary behaviour of coherent optical transitions, as well as in dealing with transition mechanisms in corpuscular terms.

FURTHER READING

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).
P. Meystre and M. Sargent III, *Elements of Quantum Optics*, Springer-Verlag, Berlin (1989).

2 Quantum mechanics of the photon

2.1 Introduction

The photoelectric cell, one of the first optoelectronic devices, led Albert Einstein in 1905 to quantize optical emission into what later became known as photons. It is therefore natural to encounter the theory of energy quantization on our present path of investigation. The goal of this chapter is to summarize the principles of the quantization of the electromagnetic field. These concepts are necessary to explain the spectral content of blackbody radiation, the phenomena of stimulated and spontaneous emission, and many other elementary topics required in acquiring a fundamental understanding of the operation of optoelectronic components.

2.2 Maxwell's equations in reciprocal space

We know that in a vacuum, the electric field $\mathbf{E}(\mathbf{r}, t)$ and the magnetic field $\mathbf{B}(\mathbf{r}, t)$, are generated by oscillating charge densities $\rho(\mathbf{r}, t)$ and current densities $\mathbf{j}(\mathbf{r}, t)$ ('There is not a single granule of light which is not the fruit of an oscillating charge' (Lorentz)). These fields are coupled by equations derived from various areas of electrostatics and magnetism unified within the framework of electromagnetic theory. These are Maxwell's equations:

$$\nabla \cdot \mathbf{E}(\mathbf{r}, t) = \frac{1}{\varepsilon_0} \rho(\mathbf{r}, t) \tag{2.1a}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0 \tag{2.1b}$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t} \mathbf{B}(\mathbf{r}, t) \tag{2.1c}$$

$$\nabla \times \mathbf{B}(\mathbf{r}, t) = \frac{1}{c^2} \frac{\partial}{\partial t} \mathbf{E}(\mathbf{r}, t) + \frac{1}{\varepsilon_0 c^2} \mathbf{j}(\mathbf{r}, t) \tag{2.1d}$$

Maxwell's equations

The first equation is referred to as the Gauss–Poisson law, the third as Lenz's law, and the fourth as the Faraday–Ampere law, to which Maxwell added a term corresponding to the displacement current $\partial(\varepsilon_0 \mathbf{E})/\partial t$. The second equation repre-

sents the absence of a magnetic monopole. Starting from these equations we may derive the existence of travelling plane waves of the type:

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0(\mathrm{e}^{+\mathrm{i}(\mathbf{k}\cdot\mathbf{r}-\omega t)} + \text{c.c.}) \tag{2.2}$$

with wavevector $\mathbf{k}$, and a radial frequency ω being linked by the vacuum dispersion relation:

$$\omega = c|\mathbf{k}| \tag{2.3}$$

We know that Eq. (2.1a) and (2.1c) lead to the existence of vector and scalar potentials $\mathbf{A}(\mathbf{r},t)$ and $U(\mathbf{r},t)$, defined by:

$$\mathbf{B}(\mathbf{r},t) = \nabla \times \mathbf{A}(\mathbf{r},t) \tag{2.4a}$$

$$\mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t) - \nabla U(\mathbf{r},t) \tag{2.4b}$$

and that these potentials are not uniquely determined, but rather admit an arbitrary choice of potential $F(\mathbf{r},t)$ as in the following *gauge transformation* equations:

$$\mathbf{A}'(\mathbf{r},t) = \mathbf{A}(\mathbf{r},t) + \nabla F(\mathbf{r},t) \tag{2.5a}$$

$$U'(\mathbf{r},t) = U(\mathbf{r},t) - \frac{\partial}{\partial t}F(\mathbf{r},t) \tag{2.5b}$$

all of which are indiscernible by Maxwell's equations.

Thus the electric and magnetic fields are strictly determined once the distributions of charges and electrical currents are known:

$$\rho(\mathbf{r},t) = \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i) \tag{2.6a}$$

$$\mathbf{j}(\mathbf{r},t) = \sum_i q_i \mathbf{v}_i \delta(\mathbf{r} - \mathbf{r}_i) \tag{2.6b}$$

where $\mathbf{r}_i$ is the position of the charge and $\mathbf{v}_i$ its velocity at time t. These fields in turn influence these distributions by means of Lorentz forces:

$$m_i \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{v}_i = q_i[\mathbf{E}(\mathbf{r}_i,t) + \mathbf{v}_i \times \mathbf{B}(\mathbf{r}_i,t)] \tag{2.7}$$

The ensemble of Eqs. (2.1a) to (2.1d) and (2.6a) to (2.7) represent the Maxwell–Lorentz equations and afford a description of the behaviour of any optical medium. The structure of these equations is, however, extremely complex, as they are not spatially localized.

In fact, the linear form of these differential Maxwell–Lorentz equations makes them readily accessible to Fourier transformations. We will show how, when expressed in Fourier space (also referred to as *reciprocal space*), these equations become greatly simplified and local in nature. We briefly review the definition and properties of the Fourier transformation.

2.3 Properties of the Fourier transform

Let $\mathbf{F}(\mathbf{r})$ be a square integrable vector function of $\mathbf{r}$ defined in three-dimensional space. This function may also vary with time – this will be indicated whenever necessary. In order to eliminate problems relating to the convergence of certain integrals, we will introduce a fictitious box with volume $V = L^3$ in which the waves will be confined. As in the method utilized in Complement 1.A, L will be allowed to be arbitrarily large. The Fourier transform of this function is then the function $\mathscr{F}(\mathbf{k})$ defined by:

$$\mathscr{F}(\mathbf{k}) = \frac{1}{(2\pi)^{3/2}} \iiint_V \mathbf{F}(\mathbf{r}) \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \mathrm{d}^3\mathbf{r} \tag{2.8}$$

As the wave amplitudes must be zero at the inner surfaces of the box, we have:

$$k_{nx} = n_x \frac{\pi}{L}; \; k_{ny} = n_y \frac{\pi}{L}; \; k_{nz} = n_z \frac{\pi}{L} \tag{2.9}$$

where n_x, n_y, and n_z are **positive integers**. The variables (k_{nx}, k_{ny}, k_{nz}) form the components of a vector $\mathbf{k}_n$ in reciprocal space, having therefore dimensions of inverse length. The numerical factor appearing in front of (2.8) varies from one author to another. We employ a scaling factor of $1/(2\pi)^{3/2}$ as it allows for more symmetric notation between real and reciprocal space. Expression (2.8) may evidently be applied as easily to a vector function $\mathbf{F}(\mathbf{r}, t)$ as a scalar function $F(\mathbf{r}, t)$. We recall now the important properties of this transformation:

Inverse Fourier transform

$$\mathscr{F}_n = \mathscr{F}(\mathbf{k}_n) = \frac{1}{(2\pi)^{3/2}} \iiint_V \mathbf{F}(\mathbf{r}) \mathrm{e}^{-\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} \, \mathrm{d}^3\mathbf{r} \tag{2.10}$$

$$\mathbf{F}(\mathbf{r}) = \frac{1}{V(2\pi)^{3/2}} \sum_n \mathscr{F}_n \mathrm{e}^{\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} \tag{2.11}$$

Effect on differential operators

If F is scalar, $\nabla F \leftrightarrow \mathrm{i}\mathbf{k}_n \mathscr{F}_n$ (2.12)

If $\mathbf{F}$ is vectorial, $\nabla \cdot \mathbf{F} \leftrightarrow \mathrm{i}\mathbf{k}_n \cdot \mathscr{F}_n$ (2.13)

$$\nabla \times \mathbf{F} \leftrightarrow \mathrm{i}\mathbf{k}_n \times \mathscr{F}_n \tag{2.14}$$

Plancherel–Parseval identity

$$\iiint_V F(\mathbf{r},t)^* G(\mathbf{r},t)\mathrm{d}^3\mathbf{r} = \frac{1}{L^3}\sum_n \mathscr{F}_n^*(t)\mathscr{G}_n(t) \tag{2.15}$$

Properties of the Fourier transform

In Eq. (2.15), the particular case of $F = G$ may be interpreted as a statement of energy conservation:

$$\iiint_V |F(\mathbf{r},t)|^2 \mathrm{d}^3\mathbf{r} = \frac{1}{L^3}\sum_n |\mathscr{F}_n(t)|^2 \tag{2.16}$$

The spatial distribution of energy is therefore redistributed in the spectral distribution.

We consequently note that differential operators have been transformed into algebraic vector operators. This is the fundamental reason for introducing the Fourier transform. Maxwell's equations in reciprocal space now become:

$$\mathrm{i}\mathbf{k}_n \cdot \mathscr{E}_n = \frac{1}{\varepsilon_0}\rho_n \tag{2.17a}$$

$$\mathbf{k}_n \cdot \mathscr{B}_n = 0 \tag{2.17b}$$

$$\mathrm{i}\mathbf{k}_n \times \mathscr{E}_n = -\frac{\partial}{\partial t}\mathscr{B}_n \tag{2.17c}$$

$$\mathrm{i}\mathbf{k}_n \times \mathscr{B}_n = \frac{1}{c^2}\frac{\partial}{\partial t}\mathscr{E}_n + \frac{1}{\varepsilon_0 c^2}\mathbf{j}_n \tag{2.17d}$$

Maxwell's equations in reciprocal space

where evidently ρ_n and $\mathbf{j}_n$ are the transforms of the charge and current densities, respectively. Equation (2.17b) shows how, in reciprocal space, $\mathscr{B}_n$ is orthogonal to the vector $\mathbf{k}_n$ (i.e. the magnetic field has only one component normal to its wavevector $\mathbf{k}_n$). It is therefore natural to decompose the electric field in reciprocal space into two components, a normal component $\mathscr{E}_{\perp n}$:

$$\mathscr{E}_{\perp n} = \mathscr{E}_n - \frac{\mathscr{E}_n \cdot \mathbf{k}_n}{|\mathbf{k}_n|^2}\mathbf{k}_n \tag{2.18a}$$

and a parallel component:

$$\mathscr{E}_{\parallel n} = \mathscr{E}_n - \mathscr{E}_{\perp n} \tag{2.18b}$$

It can be shown that the parallel component $\mathbf{E}_{\parallel}(\mathbf{r}, t)$ of the electric field is the field created *instantaneously* at point $\mathbf{r}$ by the charges. Therefore, the parallel field created at $\mathbf{r}$ by the charge located at $\mathbf{r}_0$ at time t is given by:

$$\mathbf{E}_{\parallel}(\mathbf{r}, t) = \frac{q}{4\pi\varepsilon_0} \frac{\mathbf{r} - \mathbf{r}_0(t)}{|\mathbf{r} - \mathbf{r}_0(t)|^3} \tag{2.19}$$

Expression (2.19) is in fact the usual expression for the electric field (Coulomb's law) without the retardation term $t - |\mathbf{r} - \mathbf{r}_0|/c$. Everything proceeds as though the information on the position of the charge at $\mathbf{r}_0$ arrives instantaneously at every point $\mathbf{r}$ of space. This, however, cannot be in conflict with relativity theory. It can be shown that the field measured at $\mathbf{r}$, which is the sum of the parallel and normal components, is retarded by $|\mathbf{r} - \mathbf{r}_0|/c$ as predicted in theory (i.e. the normal field introduces a term which cancels out the instantaneous contribution in (2.19)).

We are therefore actually interested in the normal component of the electric field which is coupled to the magnetic field by:

$$-\frac{\partial}{\partial t}\mathscr{B}_n = \mathrm{i}\mathbf{k}_n \times \mathscr{E}_{\perp n} \tag{2.20a}$$

$$\frac{\partial}{\partial t}\mathscr{E}_{\perp n} = \mathrm{i}c^2\mathbf{k}_n \times \mathscr{B}_n - \frac{1}{\varepsilon_0}\mathbf{j}_{\perp n} \tag{2.20b}$$

with $\mathbf{j}_{\perp n}$ being the normal component of the electric current density. We note that the introduction of these normal components actually save us from carrying two additional equations. The integration of these equations leads quite easily to an expression for the retarded potential and allows us to find the radiation field for an oscillating dipole. We work through this development in Complement 2.A.

Finally, we introduce the Fourier transform for the vector potential in reciprocal space. Equation (2.4a) gives us its relationship to the magnetic field:

$$\mathscr{B}_n = \mathrm{i}\mathbf{k}_n \times \mathscr{A}_n \tag{2.21}$$

or by introducing the same decomposition used in (2.18b)

$$\mathscr{B}_n = \mathrm{i}\mathbf{k}_n \times \mathscr{A}_{\perp n} \tag{2.22}$$

Equation (2.17c) shows $\mathscr{E}_{\perp n}$ and $\mathscr{A}_{\perp n}$ to be related by:

$$\mathscr{E}_{\perp n} = -\frac{\partial}{\partial t}\mathscr{A}_{\perp n} \tag{2.23}$$

We may also show quite easily that $\mathscr{A}_{\perp n}$ is gauge invariant, making it at least as physically significant a quantity as the electric and magnetic fields.

2.4 Quantization of electromagnetic waves

We now turn our attention to expressing the electromagnetic energy enclosed in a cavity with volume V. The classical Hamiltonian for the electromagnetic field is:

$$H_{\text{em}} = \frac{\varepsilon_0}{2} \iiint_V [|\mathbf{E}(\mathbf{r},t)|^2 + |c\mathbf{B}(\mathbf{r},t)|^2] \mathrm{d}^3\mathbf{r} \tag{2.24}$$

Classical Hamiltonian of an electromagnetic field in real space

or by using the Plancherel–Parseval identity:

$$H_{\text{em}} = \frac{\varepsilon_0}{2L^3} \sum_n [|\mathscr{E}_n|^2 + |c\mathscr{B}_n|^2] \tag{2.25}$$

The classical Hamiltonian may then also be broken down into parallel and normal components:

$$H_{\text{em}} = H_{\perp \text{em}} + H_{\parallel \text{em}} \tag{2.26}$$

notably with:

$$H_{\perp \text{em}} = \frac{\varepsilon_0}{2L^3} \sum_n [|\mathscr{E}_{\perp n}|^2 + |c\mathscr{B}_{\perp n}|^2] \tag{2.27}$$

while the parallel component does not interest us. Taking into account Eq. (2.23) and the dispersion relation $\omega_n = c|\mathbf{k}_n|$, the normal part of the classical Hamiltonian is then:

$$H_{\perp \text{em}} = \frac{\varepsilon_0}{2L^3} \sum_n [|\mathscr{E}_{\perp n}|^2 + \omega_n^2 |\mathscr{A}_{\perp n}|^2] \tag{2.28}$$

$$\mathscr{E}_{\perp n} = -\frac{\partial}{\partial t} \mathscr{A}_{\perp n} \tag{2.29}$$

Classical Hamiltonian for the electromagnetic field in reciprocal space

Equations (2.28) and (2.29) form the basis for quantization of the electromagnetic field. We notice that (2.28) is in fact equivalent to that for the harmonic oscillator (Complement 1.D), with the following correspondences:

$$\begin{aligned}&\text{Frequency,} && \omega_n \\ &\text{Mass,} && m \leftrightarrow \frac{\varepsilon_0}{L^3} \\ &\text{Position,} && \hat{x} \leftrightarrow \hat{\mathscr{A}}_{\perp n} \\ &\text{Momentum,} && \hat{p} \leftrightarrow -\frac{\varepsilon_0}{L^3}\hat{\mathscr{E}}_{\perp n}\end{aligned} \tag{2.30}$$

Nonetheless, we note a difference between the Hamiltonian for the harmonic oscillator, which comprises only real observables, and the Hamiltonian for the electromagnetic field, which may comprise imaginary terms such as $e^{-i\mathbf{k}\cdot\mathbf{r}}$. This may cause certain difficulties which we will not examine here. We introduce the creation a_n^+ and annihilation a_n operators given by:

$$a_n^+ = \sqrt{\frac{\varepsilon_0}{2\hbar\omega_n L^3}}(\omega_n \hat{\mathscr{A}}_{\perp n} + i\hat{\mathscr{E}}_{\perp n}) \tag{2.31a}$$

$$a_n = \sqrt{\frac{\varepsilon_0}{2\hbar\omega_n L^3}}(\omega_n \hat{\mathscr{A}}_{\perp n} - i\hat{\mathscr{E}}_{\perp n}) \tag{2.31b}$$

Relationship between the field operators and the creation and annihilation operators

These operators then act on the n electromagnetic modes. In analogy with the results for the harmonic oscillator, we postulate that these operators fulfil the anticommutation relations associated with (2.29):

$$[a_n, a_m^+] = \delta_{mn} \tag{2.32}$$

As shown in Complement 1.D, the Hamiltonian for the electromagnetic field then takes the form:

$$\hat{H}_{\perp n} = \sum_n \hbar\omega_n \left(a_n^+ a_n + \frac{1}{2}\right) \tag{2.33}$$

Quantum Hamiltonian for the electromagnetic field

Thus, the Hamiltonian for the electromagnetic field in a cavity may be written as a sum of independent harmonic oscillator Hamiltonians, with each oscillator then corresponding to the classical mode of an electromagnetic wave in a cavity with oscillation frequency ω_n. This is the fundamental finding of this chapter.

It is important to remark at this point that time does not appear explicitly in the expression for the Hamiltonian. *The Hamiltonian for the electromagnetic mode is stationary, even if it is destined to be used to describe an oscillating field.* This is a paradox that will be resolved with the introduction of the coherent state in Section 2.5. The inverse form of (2.31a) and (2.31b) will allow us to calculate the various

operators of the electromagnetic field as a function of creation and annihilation operators. Nonetheless, as we have already stated, we have in the expression of a classical Hamiltonian for the electromagnetic field, only the norms of these operators and not their phases. We may show, however, that the operators for the electric and magnetic fields, and the vector potential, are related to the mode n creation and annihilation operators by[1]:

$$\hat{\mathbf{E}}_\perp = \mathrm{i}\sum_n \mathscr{F}_n(a_n \mathrm{e}^{\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} - a_n^\dagger \mathrm{e}^{-\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}})\boldsymbol{\varepsilon}_n \tag{2.34a}$$

$$\hat{\mathbf{A}}_\perp = \mathrm{i}\sum_n \frac{\mathscr{F}_n}{\omega_n}(a_n \mathrm{e}^{\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} - a_n^\dagger \mathrm{e}^{-\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}})\boldsymbol{\varepsilon}_n \tag{2.34b}$$

$$\hat{\mathbf{B}}_\perp = \mathrm{i}\sum_n \mathscr{F}_n \frac{\mathbf{k}\times\boldsymbol{\varepsilon}_n}{\omega_n}(a_n \mathrm{e}^{\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} - a_n^\dagger \mathrm{e}^{-\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}})\boldsymbol{\varepsilon}_n \tag{2.34c}$$

Field operators as a function of the creation and annihilation operators

where $\boldsymbol{\varepsilon}_n$ is the polarization vector of the electric field and $\mathscr{F}_n$ represents the *vacuum fluctuation field* given by:

$$\mathscr{F}_n = \sqrt{\frac{\hbar\omega_n}{2\varepsilon_0 L^3}} \tag{2.35}$$

Vacuum fluctuation field

We will return to the significance of this field later on. We note that the electric field operator may be broken down into the sum of two observables (which we can show to correspond to positive and negative frequencies):

$$\mathbf{E}_\perp(\mathbf{r}) = \mathbf{E}_\perp^{(+)}(\mathbf{r}) + \mathbf{E}_\perp^{(-)}(\mathbf{r}) \tag{2.36}$$

with, for example, the positive frequency electric field being:

$$\mathbf{E}_\perp^{(+)}(\mathbf{r}) = \mathrm{i}\sum_{n=1}^{\infty} \boldsymbol{\varepsilon}_n \mathscr{F}_n a_n \mathrm{e}^{\mathrm{i}\mathbf{k}_n\cdot\mathbf{r}} \tag{2.37}$$

These last two equations will allow us to move from a classical description of an electromagnetic wave to a quantum mechanical one.

2.5 The photon

We consider an optical cavity with only one electromagnetic mode n. The Hamiltonian for this mode n may then be written:

[1] Beware! In (2.34a) to (2.34c), the symbol $\mathbf{r}$ which appears in $\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$ is a real space variable and not an operator!

$$\hat{H}_{\perp n} = \hbar\omega_n\left(a_n^+ a_n + \frac{1}{2}\right) \tag{2.38}$$

The ith eigenstates, or stationary states of the electromagnetic mode n of frequency ω_n, are given by:

$$|i_n\rangle = \frac{(a_n^+)^{i_n}}{\sqrt{i_n!}}|0_n\rangle \tag{2.39}$$

where the state $|0_n\rangle$ is the empty state of the mode n. The energy of these stationary states is:

$$E_{n,i} = \hbar\omega_n\left(i_n + \frac{1}{2}\right) \tag{2.40}$$

Finally, the creation and destruction operators a_n^+ and a_n allow transitions to occur between the $|i_n\rangle$ state and $|i_n - 1\rangle$ or $|i_n + 1\rangle$ states:

$$a_n|i_n\rangle = \sqrt{i_n}|i_n - 1\rangle, \text{ if } i_n > 0 \tag{2.41a}$$

$$a_n|0_n\rangle = 0 \tag{2.41b}$$

$$a_n^+|i_n\rangle = \sqrt{i_n + 1}|i_n + 1\rangle \tag{2.41c}$$

Equations (2.41a) to (2.41c), may be interpreted in corpuscular terms. The electromagnetic mode is then considered to be formed by an ensemble of elementary excitations, called *photons*, so that:

$$|i_n\rangle = |\text{there are } i_n \text{ photons of the } n\text{th electromagnetic mode in the cavity}\rangle \tag{2.42}$$

where each photon of the mode n carries an energy E_n:

$$E_n = \hbar\omega_n \tag{2.43}$$

Assuming i_n photons of the nth mode to be in the cavity, the total energy is given by:

$$E_{n,i} = i_n\hbar\omega_n + \frac{1}{2}\hbar\omega_n = i_n E_n + \frac{1}{2}\hbar\omega_n \tag{2.44}$$

We note that, even if the cavity is unoccupied by photons, the energy of the nth mode is not zero, but given by $E_{\text{vac},n}$, corresponding to the vacuum energy of the nth mode (i.e. of the $|0_n\rangle$ state):

$$E_{\text{vac},n} = \frac{1}{2}\hbar\omega_n \tag{2.45}$$

This ground state energy may again be seen as resulting from the first of Heisenberg's uncertainty principles (given in Complement 1.D for the harmonic oscillator). The creation and annihilation operators create or destroy photons possessing momenta of $\mathbf{p}_n = \hbar\mathbf{k}_n$, which are eigenstates of the observable:

$$\hat{\mathbf{p}}_n = \hbar\mathbf{k}_n a_n^+ a_n \tag{2.46}$$

Let us now return to the optical cavity of arbitrary volume $V = L^3$ and populated by an arbitrarily large number of n electromagnetic modes. The Hamiltonian for the normal electromagnetic modes in the cavity is then written as:

$$\hat{H}_{\perp \mathrm{em}} = \sum_n \hat{H}_{\perp n} \tag{2.47}$$

The eigenstates for this Hamiltonian then result from the concatenation of the independent modes of each Hamiltonian (referred to as a *tensor product* space):

$$|i_1, i_2, \ldots, i_n, \ldots\rangle = \frac{(a_1^+)^{i_1}(a_2^+)^{i_2}\ldots(a_n^+)^{i_n}\ldots}{\sqrt{i_1! i_2! \cdots i_n! \cdots}}|0\rangle \tag{2.48}$$

where the state $|0\rangle$ refers to the empty cavity state $|0_1, 0_2, \ldots\rangle$ being devoid of photons. Each mode n may be populated by an arbitrary number of photons. These particles are therefore *bosons* as they are not subject to exclusion principles. The electromagnetic energy of the cavity is a sum of two terms. The first corresponds to the sum of the energies of the i_n photons, for each mode n:

$$E_{\mathrm{photon}} = \sum_n i_n \hbar\omega_n \tag{2.49}$$

and the second results from the vacuum energy E_{vac} associated with each mode:

$$E_{\mathrm{vac}} = \sum_n \frac{1}{2}\hbar\omega_n \tag{2.50}$$

Clearly, the latter term will diverge as the number of modes in the cavity is infinite. This observation therefore places before us a physical absurdity. This problem was solved by Feynman, Schwinger, and Tomonoga's tremendously beautiful and elegant renormalization theory. Unfortunately, it will not receive any further consideration from us.

We are now in a position, using the results in Complement 1.D, to calculate the expectation values and variances of the observables for the electric $\hat{\mathbf{E}}_\perp(\mathbf{r})$, magnetic $\hat{\mathbf{B}}_\perp(\mathbf{r})$, and vector potential $\hat{\mathbf{A}}_\perp(\mathbf{r})$ fields for the i_n photon state of mode n. Equations (1.D.32) and (1.D.33) show that:

$$\langle i_n|\hat{\mathbf{E}}_\perp|i_n\rangle = i\sum_n \mathscr{F}_n(\langle i_n|\hat{a}_n|i_n\rangle e^{i\mathbf{k}_n\cdot\mathbf{r}} - \langle i_n|\hat{a}_n^+|i_n\rangle e^{-i\mathbf{k}_n\cdot\mathbf{r}})\boldsymbol{\varepsilon}_n \tag{2.51a}$$

so that:

$$\langle i_n|\hat{\mathbf{E}}_\perp(\mathbf{r})|i_n\rangle = \langle i_n|\hat{\mathbf{B}}_\perp(\mathbf{r})|i_n\rangle = \langle i_n|\hat{\mathbf{A}}_\perp(\mathbf{r})|i_n\rangle = 0 \tag{2.51b}$$

In (2.51a), we left the terms $e^{i\mathbf{k}\cdot\mathbf{r}}$ outside of the brackets when evaluating $\langle i_n|\hat{a}_n|i_n\rangle$ as we recall that $\mathbf{r}$ is a variable here, and not an observable in this theory. Therefore, if the number i_n of photons in the nth mode is known exactly, the average value of the electromagnetic field will be zero in every location in the cavity at all times. This may seem strange given our conception of an 'electromagnetic wave', however, we will see in the following paragraph that this paradox is resolved by introducing the notion of a *coherent state*.

The variance of the observable of the electric field in the state $|i_n\rangle$ is found by following the same line of reasoning as that put forward in Complement 1.D:

$$(\hat{E}_{\perp n})^2 = -\mathscr{F}_n^2(\hat{a}_n^2 e^{2i\mathbf{k}_n\cdot\mathbf{r}} + \hat{a}_n^{+2} e^{-2i\mathbf{k}_n\cdot\mathbf{r}} - \hat{a}_n\hat{a}_n^+ - \hat{a}_n^+\hat{a}_n) \tag{2.52}$$

or by taking into account (1.D.32) to (1.D.33):

$$\overline{(\hat{E}_{\perp n})^2} = \langle i_n|(\hat{E}_{\perp n})^2|i_n\rangle = \mathscr{F}_n^2(2i_n + 1) \tag{2.53}$$

or again

$$\overline{(\hat{E}_{\perp n})^2} = \frac{\hbar\omega_n}{2\varepsilon_0 L^3}(2i_n + 1) \tag{2.54}$$

The same type of result is obtained for both the vector potential and the magnetic field. Equation (2.54) brings with it a certain number of insights. First, even when the cavity is devoid of all photons, the electric field has a variance different from zero, and found to be:

$$\overline{(\hat{E}_{\perp n})^2_{\text{vac}}} = \mathscr{F}_n^2 = \frac{\hbar\omega_n}{2\varepsilon_0 L^3} \tag{2.55}$$

We now understand the significance of $\mathscr{F}_n$, introduced as a normalization factor in the definitions for the creation and annihilation operators, as a *vacuum fluctuation field*. Following (2.40), Eq. (2.54) may be written as:

$$\overline{(\hat{E}_n)^2} = \frac{E_{n,i}}{\varepsilon_0 L^3} \tag{2.56}$$

which is the classical relationship between the electrical field and the energy contained by an electromagnetic mode in a cavity.

To summarize this somewhat complex section, we might say that photons

represent eigenstates of the Hamiltonian for an electromagnetic field in a cavity. These eigenstates then, also present a certain number of paradoxes: (1) the average values for the electric, magnetic, and vector potential fields are zero everywhere inside the cavity and do not oscillate with time; (2) even with no photons inside the cavity, there is a non-zero fluctuation in the variance of the electric field given by Eq. (2.55). The first paradox is resolved in Section 2.6. We will have to live with the second paradox, however, which has been confirmed by experiment (in, for instance, the Lamb shift) and which offers a means of understanding the phenomenon of spontaneous emission.

Example
The root mean square of the electromagnetic vacuum fluctuations for 'green light' photons ($\hbar\omega_n = 2\,\text{eV}$) in a cavity with a volume of $1\,\text{cm}^3$ ($L = 1\,\text{cm}$) is:

$$\mathscr{F}_n = (1.6 \times 10^{-19}\,\text{C} \times 2\,\text{eV}/(2 \times 8.85\,10^{-12}\,\text{Fd}\,\text{m}^{-1} \times 10^{-6}\,\text{m}^3))^{1/2} \approx 0.13\,\text{V}\,\text{m}^{-1}$$

Although this value is very weak in terms of field strength, its effect may be observed, for instance, in the minute displacements produced in atomic transition energies.

2.6 The coherent state

Glauber was the first to present a physical state that allowed reconciliation between oscillatory and corpuscular views of photonic states. We will not discuss Glauber's hypothesis in detail here, even though our following explanation may appear a little arbitrary. We consider a cavity possessing a single one-dimensional mode, with frequency ω (the generalization to three dimensions and multimodal excitations is immediate, but rather heavy in indices). The Glauber, or *coherent state*, $|\alpha\rangle$ is defined as:

$$|\alpha\rangle = \sum_m \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^m}{\sqrt{m!}} |m\rangle \tag{2.57}$$

Glauber's coherent state

The probabilistic interpretation of quantum mechanics allows us to understand the significance of this state in the following way: $|\alpha\rangle$ is the state in the cavity in which there is a probability p_m:

$$p_m = \mathrm{e}^{-|\alpha|^2} \frac{|\alpha|^{2m}}{m!} \tag{2.58}$$

of finding m photons in the cavity. We recognize in (2.58) *Poisson's law* from

probability theory. This classical law gives the probability of finding m photons in the cavity during a random sampling, knowing that there are on average $|\alpha|^2$ photons in the cavity.

Clearly, $|\alpha\rangle$ is a normalized physical state as:

$$\sum_m p_m = \mathrm{e}^{-|\alpha|^2} \sum_m \frac{|\alpha|^{2m}}{m!} = 1 \tag{2.59}$$

We seek to evaluate the evolution of such a state over time. We will assume the cavity to be in the state given by (2.57) at time $t = 0$ (i.e. the state in (2.57) corresponds to $|\alpha(0)\rangle$). Following Eq. (1.29), Glauber's state $|\alpha(t)\rangle$ evolves into:

$$|\alpha(t)\rangle = \sum_m \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^m}{\sqrt{m!}} \mathrm{e}^{-\mathrm{i}E_m t/\hbar} |m\rangle = \sum_m \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^m}{\sqrt{m!}} \mathrm{e}^{-\mathrm{i}\omega(m+\frac{1}{2})t} |m\rangle \tag{2.60}$$

which immediately takes the form:

$$|\alpha(t)\rangle = \mathrm{e}^{-\mathrm{i}\omega t/2} \sum_m \mathrm{e}^{-[|(\alpha \mathrm{e}^{-\mathrm{i}\omega t})|^2/2]} \frac{(\alpha \mathrm{e}^{-\mathrm{i}\omega t})^m}{\sqrt{m!}} |m\rangle \tag{2.61}$$

As every physical prediction concerning the state $|\alpha(t)\rangle$ is insensitive to the phase $\mathrm{e}^{-\mathrm{i}\omega t}$, we see that the time evolution for the Glauber state $|\alpha(t)\rangle$ may be written as:

$$|\alpha(t)\rangle = |\alpha \mathrm{e}^{-\mathrm{i}\omega t}\rangle \tag{2.62}$$

We now show a relationship that forms the basis of the quantum properties of the Glauber state:

$$a|\alpha\rangle = \alpha|\alpha\rangle \tag{2.63a}$$

or in its dual form:

$$\langle\alpha|a^+ = \alpha^*\langle\alpha| \tag{2.63b}$$

Indeed:

$$a|\alpha\rangle = \sum_m \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^m}{\sqrt{m!}} a|m\rangle = \sum_m \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^m}{\sqrt{m!}} \sqrt{m}|m-1\rangle \tag{2.64}$$

In this last sum, we change the summation index from $m - 1$ to n, so that:

$$a|\alpha\rangle = \sum_{n=1} \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^{n+1}}{\sqrt{n!}} |n\rangle = \alpha \sum_{n=1} \mathrm{e}^{-(|\alpha|^2/2)} \frac{\alpha^n}{\sqrt{n!}} |n\rangle = \alpha|\alpha\rangle \tag{2.65}$$

which is what we sought to demonstrate.

On the other hand, it is worth noting that $a^+|\alpha\rangle$ does not yield anything of use.

Equation (2.63) will easily allow us to evaluate the expectation values of the observables. Let us first seek the average number of photons N_α in a Glauber state $|\alpha\rangle$:

$$\bar{N}_\alpha = \langle\alpha|a^+ a|\alpha\rangle = \alpha^*\alpha = |\alpha|^2 \tag{2.66}$$

Also, the average value of the square of the number of photons in this state is:

$$\begin{aligned}\overline{N_\alpha^2} &= \langle\alpha|a^+ aa^+ a|\alpha\rangle = |\alpha|^2\langle\alpha|aa^+|\alpha\rangle \\ &= |\alpha|^2\langle\alpha|a^+ a + 1|\alpha\rangle = |\alpha|^2(|\alpha|^2 + 1)\end{aligned} \tag{2.67}$$

so that the standard deviation ΔN_α of the photon number is:

$$\Delta N_\alpha = \sqrt{\overline{N_\alpha^2} - \overline{N_\alpha^2}} = |\alpha| = \sqrt{\overline{N_\alpha}} \tag{2.68}$$

These equations are characteristic of Poisson's distribution and could have been obtained directly by application of probability theory. In addition, we see that the average number and its associated standard deviation are time independent. Lastly, the state respects the law of large numbers, whereby:

$$\frac{\Delta N_\alpha}{\overline{N_\alpha}} = \frac{1}{\sqrt{\overline{N_\alpha}}} \tag{2.69}$$

This last equation may also be written:

$$S/N = \frac{(\text{average})^2}{\text{variance}} = \frac{\overline{N_\alpha^2}}{\Delta N_\alpha^2} = \overline{N_\alpha} \tag{2.70}$$

where S/N refers to the signal-to-noise ratio associated with the Poissonian fluctuation in the photon number. This last relation forms the basis of detection theory, which will be given in Chapter 11.

Figure 2.1 shows the Poisson distribution for three values of average photon number N_α. We notice that the distribution narrows as the average number of photons is increased.

We now seek the average value of the observable $\mathbf{E}_{\perp,\alpha}$ for the electric field in the Glauber state $|\alpha\rangle$. Following Eqs. (2.34a) and (2.35):

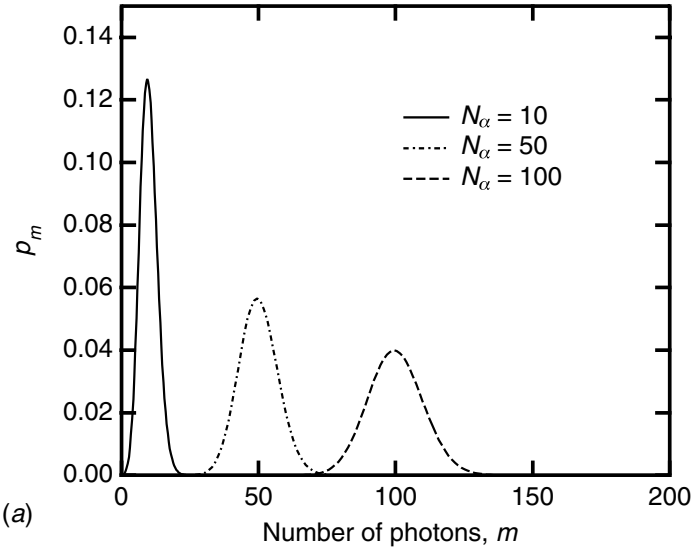
$$\bar{\mathbf{E}}_{\perp,\alpha} = \langle\alpha|\hat{\mathbf{E}}_\perp|\alpha\rangle = \mathrm{i}\mathscr{F}\langle\alpha|a\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} - a^+\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}|\alpha\rangle\boldsymbol{\varepsilon}_n \tag{2.71}$$

where $\mathscr{F}$ is the vacuum fluctuation field given by (2.35), so that given (2.63a) and (2.63b):

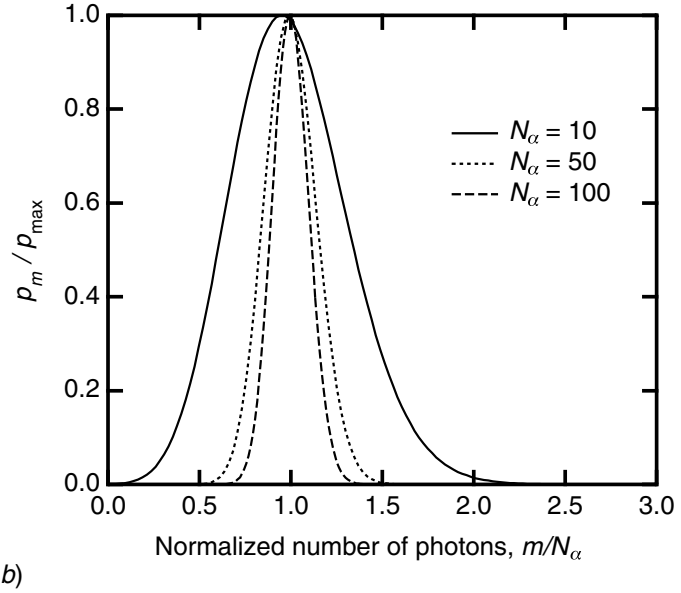
$$\bar{\mathbf{E}}_{\perp,\alpha} = \mathrm{i}\mathscr{F}(\alpha\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} - \alpha^*\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}})\boldsymbol{\varepsilon}_n \tag{2.72}$$

Similarly, the variance $\Delta E_{\perp,\alpha}$ of the observable for the electric field is given by:

$$\langle\alpha|(\hat{\mathbf{E}}_\perp)^2|\alpha\rangle = -\mathscr{F}^2\langle\alpha|(a^2\mathrm{e}^{2\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + a^{+2}\mathrm{e}^{-2\mathrm{i}\mathbf{k}\cdot\mathbf{r}} - a^+ a - a^+ a)|\alpha\rangle \tag{2.73}$$



(*a*)



(*b*)

Fig. 2.1. Probability of finding m photons in a cavity measured for three coherent states having average numbers of 10, 50, and 100 photons.

or

$$\Delta E_{\perp,\alpha} = \mathscr{F}^2[(\alpha e^{i\mathbf{k}\cdot\mathbf{r}} - \alpha^* e^{-i\mathbf{k}\cdot\mathbf{r}})^2 + 1] \tag{2.74}$$

Reintroducing relation (2.62), which describes the temporal evolution, and further supposing, without loss of generality, that α is a real number equal to α_0, we finally obtain for the time dependence of the mean electric field:

$$\bar{\mathbf{E}}_{\perp,\alpha}(t) = 2\mathscr{F}\alpha_0 \sin(\mathbf{k}\cdot\mathbf{r} - \omega t)\boldsymbol{\varepsilon}_n \tag{2.75}$$

which corresponds to the classical electric field:

$$\mathbf{E}_{\text{classical}}(t) = E_0 \sin(\mathbf{k}\cdot\mathbf{r} - \omega t)\boldsymbol{\varepsilon}_n \tag{2.76}$$

Using (2.35) and (2.66), the amplitude of the classical electric field E_0 of an electromagnetic wave of frequency ω and the average number N of photons in the cavity of energy equal to $\hbar\omega$ are then related by:

$$E_0 = \sqrt{\frac{2\hbar\omega}{\varepsilon_0 L^3}N} \tag{2.77}$$

Amplitude of the electric field in terms of the average number of photons with energy $\hbar\omega$

This last equation may also be interpreted in terms of the corpuscular model. The average number of photons N in the cavity is given by the ratio of the electromagnetic energy stored in the cavity $\varepsilon_0 V E_0^2/2$ divided by the energy carried by each photon $\hbar\omega$. We may also define a photon flux per unit area Φ either entering or exiting the cavity per unit time in terms of optical power per unit area P:

$$P = \hbar\omega\Phi \tag{2.78}$$

Example
The flux of incident 'green' photons ($\hbar\omega = 2\,\text{eV}$) in a $1\,\text{mW cm}^{-2}$ beam ($P = 10^{-3}\,\text{W cm}^{-2}$) is:

$$\Phi = 10^{-3}\,\text{W cm}^{-2}/(1.6\times10^{-19}\,\text{C}\times 2\,\text{eV}) = 3\times10^{15}\,\text{photons cm}^{-2}\,\text{s}^{-1}$$

The density of incident 'green' photons $\hbar\omega = 2\,\text{eV}$ corresponding to an electric field strength of $1\,\text{V cm}^{-1}$ ($E_0 = 100\,\text{V m}^{-1}$) is:

$$\begin{aligned} N/L^3 &= 10^4(\text{V m}^{-1})^2 \times 8.85\times10^{-12}\,\text{Fd}^{-1}\,\text{m}/(2\times1.6\times10^{-19}\,\text{C}\times 2\,\text{V}) \\ &= 1.4\times10^{11}\,\text{m}^{-3} \end{aligned}$$

Therefore, even for modest optical power levels and electric field strengths, the large number of photons involved make the corpuscular nature of these particles difficult to observe. Nonetheless, we will soon encounter devices capable of detecting single photons. Amazingly, the human eye, if given the opportunity to accommodate to the dark over the span of a few minutes, can detect a photon flux consisting of a few photons per second!

2.7 Blackbody radiation

It is by investigating the spectral distribution of the light emitted by heated objects

that Planck introduced the notion of a discrete energy quantum, and from which Einstein further elaborated the concept of the photon. We may define a blackbody as corresponding to a cavity in which the electromagnetic modes are in thermodynamic equilibrium with the walls at temperature T. In this case, all radiation is absorbed and re-emitted within the cavity without exchange with the outside. Without the introduction of the photon (i.e. without the postulate of minimum energy quanta), Rayleigh and Jeans found that energy contained within such a cavity tended towards infinity. This notable impasse in classical physics is referred to as the 'ultraviolet catastrophe'. We will not repeat this calculation here, but do point out that its implications are of historical significance to the early development of quantum mechanics.

Let us consider a cavity of volume L^3 and seek out the optical mode density for each frequency v (i.e. the number of modes $\mathrm{d}N$ contained in an interval of frequency $\mathrm{d}v$):

$$\mathrm{d}N = \rho_m(v)\mathrm{d}v \tag{2.79}$$

Inside the cavity, the waves consist of linear combinations of the modes described by $(\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \pm \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}})$. Requiring as boundary conditions that the electromagnetic fields be null at the cavity walls forces the wavevectors $\mathbf{k} = (k_x, k_y, k_z)$ to be equal to:

$$k_x = \frac{\pi}{L}n_x;\ k_y = \frac{\pi}{L}n_y;\ k_z = \frac{\pi}{L}n_z \tag{2.80}$$

where n_x, n_y, and n_z are *positive* integers. Therefore, we may associate with each mode (k_x, k_y, k_z) a volume element $\mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z = (\pi/L)^3$ in reciprocal space (see Fig. 2.2). The number of modes contained between k and $k + \mathrm{d}k$ is then one-eighth (n_x, n_y, and n_z being positive!) of the volume of a spherical shell of volume $4\pi k^2\mathrm{d}k$, multiplied by two to account for the two possible polarizations of each wavevector. There are therefore:

$$\mathrm{d}N = \frac{L^3k^2}{\pi^2}\mathrm{d}k \tag{2.81}$$

electromagnetic modes. Using the dispersion relation for radiation propagating in vacuum,

$$v = \frac{c}{2\pi}k \tag{2.82}$$

we deduce a blackbody mode density ρ_m per unit frequency and volume of:

$$\rho_m(v) = \frac{8\pi v^2}{c^3} \tag{2.83}$$

The above equation yields the density of electromagnetic (em) modes available

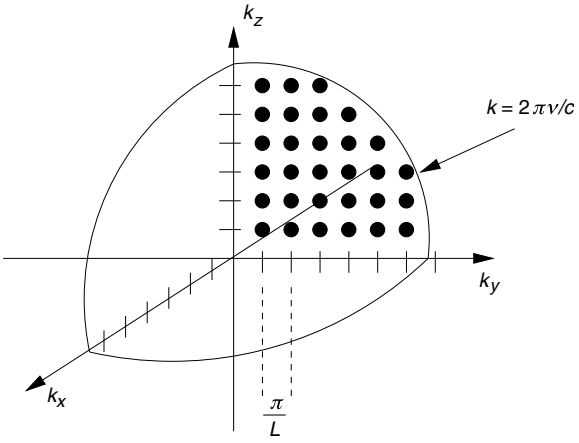


Fig. 2.2. The allowed electromagnetic modes in the cavity form a grid defined by $k_x = n_x\pi/L$, $k_y = n_y\pi/L$, and $k_z = n_z\pi/L$. The number of allowed modes with frequency less than ν in a cavity of volume L^3 is given by the number of points contained in one-eighth of the volume of a sphere of radius $k = 2\pi\nu/c$.

between ν and $\nu + d\nu$. Now, what is the probability of having this frequency range $(\nu, \nu + d\nu)$ actually filled with oscillator modes at thermal equilibrium?

The probability of finding the em oscillator in the energy state E_n is given by the Boltzmann law:

$$p_n \propto e^{-E_n/kT} \tag{2.84}$$

where k is the Boltzmann constant ($k = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ or $8.62 \times 10^{-5}\,\mathrm{eV\,K^{-1}}$) and $\propto$ in (2.84) is a symbol indicating proportionality. Since $E_n = h\nu(n + 1/2)$, this probability is then:

$$p_n = \kappa(e^{-(h\nu/kT)})^n \tag{2.85}$$

where κ is a proportionality constant given by the normalization condition:

$$\sum_{n=0}^{\infty} p_n = \kappa \sum_{n=0}^{\infty} (e^{-(h\nu/kT)})^n = \kappa \frac{1}{1 - e^{-(h\nu/kT)}} = 1 \tag{2.86a}$$

so that

$$p_n = (1 - e^{-(h\nu/kT)})(e^{-(h\nu/kT)})^n \tag{2.86b}$$

It is now interesting to determine the average number $\bar{n}_\nu$ of photons of frequency $h\nu$ in a cavity at temperature T:

$$\bar{n}_\nu = \sum_{n=0}^{\infty} np_n = (1 - e^{-(h\nu/kT)}) \sum_{n=0}^{\infty} n(e^{-(h\nu/kT)})^n = \frac{1}{e^{h\nu/kT} - 1} \tag{2.87}$$

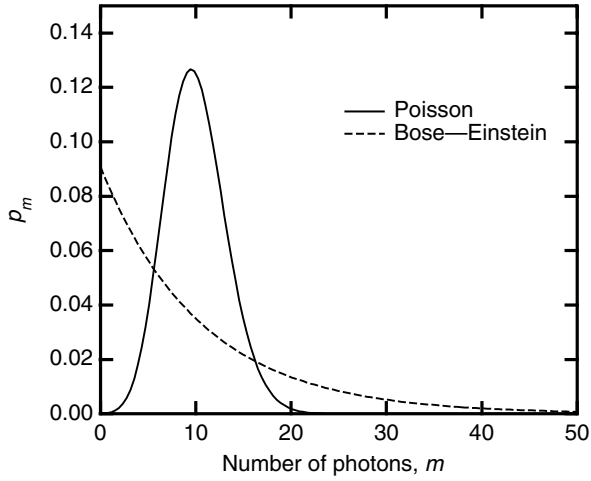


Fig. 2.3. Comparison between the Poisson and Bose–Einstein distributions for an average of ten photons in a cavity.

Substituting Eq. (2.87) into condition (2.86b), one obtains for the Bose–Einstein distribution:

$$p_n = \frac{1}{1+\bar{n}_v}\left(\frac{\bar{n}_v}{1+\bar{n}_v}\right)^n \tag{2.88}$$

A similar calculation (left as an exercise) allows one to find the variance σ_n for the number of photons:

$$\sigma_n = \bar{n}_v + \bar{n}_v^2 \tag{2.89}$$

Comparing the above expression with (2.70), obtained in the Glauber state approach, it is clear that the Bose–Einstein distribution is much larger than the Poisson distribution. Figure 2.3 shows this distribution for ten average photons in a cavity ($\bar{n}_v = 10$). The fluctuation in the number of photons in a blackbody is much greater than in a coherent state, which is not surprising. We are now ready to find the spectral distribution of the light emitted by a blackbody at temperature T.

The average energy distributed in a mode of frequency v is given by the average number of photons in this mode multiplied by the energy of the photons:

$$\bar{E}_v = \frac{hv}{e^{hv/kT} - 1} \tag{2.90}$$

As a result, the energy density $\rho_e(v)$ per unit frequency per unit volume in the cavity (in units of J s m^{-3}) is given by the product of (2.83) and (2.90):

$$\rho_e(\nu) = \frac{8\pi h\nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1} \tag{2.91}$$

Planck's law (blackbody spectrum)

This formula for the blackbody emission spectrum was discovered by Planck in 1900. In Complement 2.B, we will examine in more detail the peculiarities associated with thermal radiative emission and the use of the blackbody spectra in thermography and infrared imaging.

FURTHER READING

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).
W. H. Louisell, *Quantum Statistical Properties of Radiation*, Wiley, New York (1973).
P. Meystre and M. Sargent III, *Elements of Quantum Optics*, Springer-Verlag, Berlin (1989).
A. Aspect, C. Fabre, and G. Grynberg, *Introduction aux lasers et à l'optique quantique*, Ellipses, Paris (1997).

Complement to Chapter 2

2.A Radiation field for an oscillating charge: the Lorentz gauge

We will now show how the Maxwell–Lorentz equations allow the calculation of electromagnetic power radiated by an oscillating charge. This calculation is important for several reasons. First, from a historical perspective, this theory failed to explain the stability of the hydrogen atom and provided yet another path which led to the development of quantum mechanics. It also demonstrated an inherent link between the motion of charged particles and light, which helped to place Lorentz, Poincaré, and Einstein on the trail leading to relativity theory. Finally, the dependence of radiated power as a function of emission wavelength has technological implications wherever diffusion occurs, e.g. for optical fibres, atmospheric properties, etc.

We consider a charged particle situated at the origin **O** and subject to a small displacement $\mathbf{r}_e(t)$ about **O** (see Fig. 2.A.1). We recall Maxwell's and Lorentz's equations:

$$\boldsymbol{\nabla} \cdot \mathbf{E}(\mathbf{r}, t) = \frac{1}{\varepsilon_0} \rho(\mathbf{r}, t) \tag{2.A.1a}$$

$$\boldsymbol{\nabla} \cdot \mathbf{B}(\mathbf{r}, t) = 0 \tag{2.A.1b}$$

$$\boldsymbol{\nabla} \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t} \mathbf{B}(\mathbf{r}, t) \tag{2.A.1c}$$

$$\boldsymbol{\nabla} \times \mathbf{B}(\mathbf{r}, t) = \frac{1}{c^2} \frac{\partial}{\partial t} \mathbf{E}(\mathbf{r}, t) + \frac{1}{\varepsilon_0 c^2} \mathbf{j}(\mathbf{r}, t) \tag{2.A.1d}$$

with, in the case of an oscillating charge, the charge densities and currents given by:

$$\rho(\mathbf{r}, t) = q\delta[\mathbf{r} - \mathbf{r}_e(t)] \tag{2.A.2a}$$

$$\mathbf{j}(\mathbf{r}, t) = q\dot{\mathbf{r}}_e\delta[\mathbf{r} - \mathbf{r}_e(t)] \tag{2.A.2b}$$

We have seen that (2.A.1b) and (2.A.1c) allow the introduction of the vector and

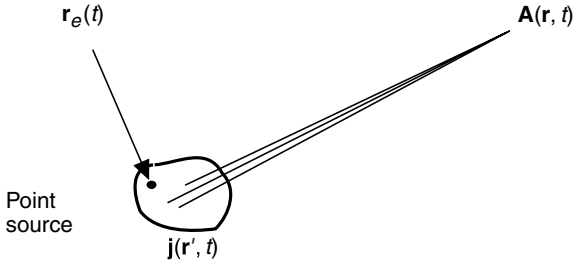


Fig. 2.A.1. Calculation of the vector potential due to a point charge.

scalar potentials $\mathbf{A}(\mathbf{r}, t)$ and $U(\mathbf{r}, t)$ defined to within a gauge transformation by:

$$\mathbf{B}(\mathbf{r}, t) = \nabla \times \mathbf{A}(\mathbf{r}, t) \tag{2.A.3a}$$

$$\mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r}, t) - \nabla U(\mathbf{r}, t) \tag{2.A.3b}$$

Instead of going the route of the Fourier transform, which is really best suited for plane waves, we will seek the differential equations which the scalar and vector potentials must satisfy. To do so, we substitute (2.A.3b) into Poisson's equation (2.A.1a), which gives:

$$\nabla^2 U + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\frac{\rho}{\varepsilon_0} \tag{2.A.4}$$

Similarly, introducing (2.A.3a) and (2.A.3b) into the Faraday–Ampère equation (2.A.1d), we obtain:

$$\nabla \times (\nabla \times \mathbf{A}) = \frac{1}{c^2}\left(-\frac{\partial^2}{\partial t^2}\mathbf{A} - \nabla\frac{\partial}{\partial t}U\right) + \frac{1}{\varepsilon_0 c^2}\mathbf{j} \tag{2.A.5}$$

We now apply the classic vector identity to $\mathbf{A}$ and ∇:

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c} \tag{2.A.6}$$

which allows (2.A.5) to be put into the form:

$$-\nabla^2\mathbf{A} + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{A} + \nabla\left(\nabla\mathbf{A} + \frac{1}{c^2}\frac{\partial}{\partial t}U\right) = \frac{1}{\varepsilon_0 c^2}\mathbf{j} \tag{2.A.7}$$

Since the vector potential is defined to within a potential gradient, it can be shown that we can profit from this degree of freedom to introduce a gauge (called the *Lorentz gauge*) in which the vector and scalar potentials, $\mathbf{A}_L$ and U_L, are related by:

$$\nabla \cdot \mathbf{A}_L + \frac{1}{c^2}\frac{\partial}{\partial t}U_L = 0 \tag{2.A.8}$$

Lorentz gauge

This gauge is particularly useful for radiation problems. In fact, (2.A.7) simplifies greatly as it directly relates the single vector potential **A** to the motion of the charge that created it, becoming:

$$-\nabla^2 \mathbf{A}_L + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{A}_L = \frac{1}{\varepsilon_0 c^2}\mathbf{j} \tag{2.A.9}$$

The solutions to this differential equation are familiar and correspond to the *retarded potential*, or explicitly:

$$\mathbf{A}_L(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0 c^2}\int_V \frac{\mathbf{j}[\mathbf{r}', t-(|\mathbf{r}-\mathbf{r}'|/c)]}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}^3\mathbf{r}' \tag{2.A.10}$$

We now make the following simplifying assumptions:

- the displacement $\mathbf{r}_e(t)$ of the particle is small relative to the observation distance. Therefore, the integration volume may be considered to be point-like relative to the distance r used to evaluate the vector potential $\mathbf{A}(\mathbf{r},t)$.
- the displacement of the source is small relative to the wavelength of the emitted light, which is a restatement of the condition that the particle's motion be non-relativistic.

Equation (2.A.10) then simplifies immediately to:

$$\mathbf{A}_L(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0 c^2 r}\int_V \mathbf{j}\left(\mathbf{r}', t-\frac{r}{c}\right)\mathrm{d}^3\mathbf{r}' \tag{2.A.11a}$$

where r is the magnitude of vector $\mathbf{r}$ and, given the definition of the current density due to the point source (2.A.2b), becomes:

$$\mathbf{A}_L(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0 c^2}\frac{\dot{\mathbf{D}}(t-r/c)}{r} \tag{2.A.11b}$$

Vector potential generated by a moving charge

where $\mathbf{D}(t)$ is the vector dipole $\mathbf{D}(t) = q\mathbf{r}_e(t)$. To determine the entire electromagnetic field, we must calculate the scalar potential given by the Lorentz condition (2.A.8):

$$\frac{\partial}{\partial t}U_L = -c^2\nabla\cdot\mathbf{A}_L$$

We must now calculate the divergence of $\dot{\mathbf{D}}(t-r/c)/r$. Simply, we have:

$$\nabla\left(\frac{1}{r}\right) = \begin{bmatrix} \frac{\partial}{\partial x}\left(\frac{1}{r}\right) = -\frac{x}{r^3} \\ \frac{\partial}{\partial y}\left(\frac{1}{r}\right) = -\frac{y}{r^3} \\ \frac{\partial}{\partial z}\left(\frac{1}{r}\right) = -\frac{z}{r^3} \end{bmatrix} = -\frac{\mathbf{r}}{r^3} \tag{2.A.12}$$

as $r^2 = x^2 + y^2 + z^2$. Similarly:

$$\nabla \cdot [\dot{\mathbf{D}}(t - r/c)] = -\ddot{\mathbf{D}}(t - r/c)\nabla\left(\frac{r}{c}\right) = -\ddot{\mathbf{D}}(t - r/c)\frac{\mathbf{r}}{rc}$$

We are now ready to calculate the divergence of the vector potential for (2.A.11b):

$$\frac{\partial}{\partial t} U_L = \frac{1}{4\pi\varepsilon_0}\left[\dot{\mathbf{D}}(t - r/c)\frac{\mathbf{r}}{r^3} + \ddot{\mathbf{D}}(t - r/c)\frac{\mathbf{r}}{r^2 c}\right] \tag{2.A.13}$$

The ratio of the first term in the parenthesis of (2.A.13) to the second is of the order of r_e/r and is therefore negligible. Equation (2.A.13) may then be easily integrated and equals (to within a constant):

$$U_L(\mathbf{r}, t) = \frac{\mathbf{r}}{4\pi\varepsilon_0 r^2 c}\dot{\mathbf{D}}(t - r/c) \tag{2.A.14}$$

From the expressions for the vector and scalar potential in the Lorentz gauge, we are now in a position to calculate the electric and magnetic fields.

The magnetic field is given by the curl of the vector potential:

$$\begin{aligned} \mathbf{B} = \nabla \times \mathbf{A}_L &= \frac{1}{4\pi\varepsilon_0 c^2}\nabla \times \left[\frac{\dot{\mathbf{D}}(t - r/c)}{r}\right] \\ &= \frac{1}{4\pi\varepsilon_0 c^2}\left[\nabla\left(\frac{1}{r}\right) \times \dot{\mathbf{D}}(t - r/c) + \frac{1}{r}\nabla \times \dot{\mathbf{D}}(t - r/c)\right] \\ &= \frac{1}{4\pi\varepsilon_0 c^2}\left[-\frac{\mathbf{r}}{r^3} \times \dot{\mathbf{D}}(t - r/c) - \frac{\mathbf{r}}{cr^2} \times \ddot{\mathbf{D}}(t - r/c)\right] \end{aligned} \tag{2.A.15}$$

where again, the first term is negligible relative to the second, thus allowing us to write:

$$\mathbf{B}(\mathbf{r}, t) = -\frac{1}{4\pi\varepsilon_0 c^3 r^2}\mathbf{r} \times \ddot{\mathbf{D}}(t - r/c) \tag{2.A.16}$$

Magnetic field radiated by a moving charge

Calculation of the electric field proceeds in the same fashion, but requires a little more effort:

$$\begin{aligned}\mathbf{E}(\mathbf{r},t) &= -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t) - \nabla U(\mathbf{r},t)\\ &= -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\ddot{\mathbf{D}}(t-r/c)}{r} - \frac{1}{4\pi\varepsilon_0 c}\nabla\left[\frac{\mathbf{r}\cdot\dot{\mathbf{D}}(t-r/c)}{r^2}\right]\\ &= -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\ddot{\mathbf{D}}(t-r/c)}{r} - \frac{1}{4\pi\varepsilon_0 c}\left[\frac{\mathbf{r}\cdot\dot{\mathbf{D}}(t-r/c)}{r^3} - \frac{\mathbf{r}}{cr^3}\ddot{\mathbf{D}}(t-r/c)\cdot\mathbf{r}\right]\\ &\cong -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\ddot{\mathbf{D}}(t-r/c)}{r} + \frac{\mathbf{r}}{4\pi\varepsilon_0 c}\frac{1}{cr^3}\ddot{\mathbf{D}}(t-r/c)\cdot\mathbf{r}\end{aligned} \tag{2.A.17}$$

which finally yields:

$$\mathbf{E}(\mathbf{r},t) = -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\mathbf{r}\times[\mathbf{r}\times\ddot{\mathbf{D}}(t-r/c)]}{r^3} \tag{2.A.18a}$$

or as written by Rayleigh:

$$\mathbf{E}(\mathbf{r},t) = -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\ddot{\mathbf{D}}_\perp(t-r/c)}{r} \tag{2.A.18b}$$

Electric field radiated by a moving point charge

where $\mathbf{D}_\perp$ is the component of the dipole vector perpendicular to the direction of the point source (Fig. 2.A.2). We note that the strength of the radiated electric field falls as $1/r$ and not as $1/r^2$ as in the case of the static field measured from a point charge. For this reason, radio waves have a much greater range than an electrostatic field – a result which fascinated Hertz and led to radio communication. We will return to the implications of this later. The electric and magnetic fields are both noted to be perpendicular to the propagation vector $\mathbf{r}$ and to each other. More precisely, they are related by:

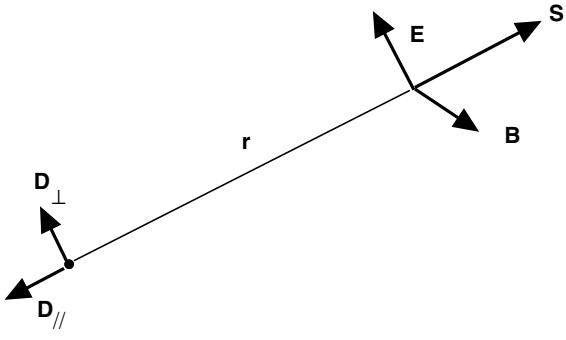


Fig. 2.A.2. Electromagnetic field radiated by a moving point charge.

$$\mathbf{B}(\mathbf{r},t) = \frac{1}{c}\frac{\mathbf{r}\times\mathbf{E}(\mathbf{r},t)}{r} \tag{2.A.19}$$

The flux of energy radiated per unit surface per unit time by the point charge is then given by the *Poynting vector* **S**:

$$\mathbf{S}(\mathbf{r},t) = \frac{1}{\mu_0}\mathbf{E}\times\mathbf{B} = \frac{1}{\mu_0 rc}\mathbf{E}\times(\mathbf{r}\times\mathbf{E}) = \frac{1}{\mu_0 rc}[E^2\mathbf{r} - (\mathbf{E}\cdot\mathbf{r})\mathbf{E}] \tag{2.A.20}$$

Since **E** and **r** are perpendicular, expression (2.A.20) may be written in the form:

$$\mathbf{S}(\mathbf{r},t) = \frac{[\ddot{\mathbf{D}}(t-r/c)]^2}{4\pi\varepsilon_0 c^3}\frac{\mathbf{u}_r}{4\pi r^2} \tag{2.A.21}$$

Energy per unit surface and unit time radiated by a moving point charge

where $\mathbf{u}_r$ is the unit vector in the direction of the point source. Expression (2.A.21) contains a wealth of information. The flux of radiated energy may be imagined as fleeing from the source at the speed of light, while preserving a constant integrated value over the surface of a sphere (of radius r) centred on the source.

We will now apply this general result to the particular case of a charge q, moving sinusoidally along the Oz axis. The dipole vector is given by:

$$\mathbf{D}_\perp(\mathbf{r},t) = qa\cos(\omega t)\mathbf{e}_z \tag{2.A.22}$$

and the Poynting vector becomes:

$$\mathbf{S}(\mathbf{r},t) = \frac{q^2a^2\omega^4\cos^2\omega t\sin^2\theta}{4\pi\varepsilon_0 c^3}\frac{\mathbf{u}_r}{4\pi r^2} \tag{2.A.23}$$

The flux of the radiated energy averaged over several cycles is given by:

$$S(\mathbf{r}) = \frac{q^2a^2\omega^4}{32\pi^2\varepsilon_0 c^3}\frac{\sin^2\theta}{r^2} \tag{2.A.24}$$

The ensembles of these constant energy points (such that $\sin\theta/r$ remains a constant) are located along circles tangent to the oscillating dipole (see Fig. 2.A.3). Dipole emission is thus rather directional. It is a direct result of this directionality that the field strength drops as $1/r$ and not as $1/r^2$.

The power radiated across a sphere centred on the oscillating charge is obtained by evaluating the surface integral of (2.A.24) over such a sphere giving:

$$P = \frac{q^2a^2\omega^4}{32\pi^2\varepsilon_0 c^3}\int_{r=\text{constant}}\frac{\sin^2\theta}{r^2}r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{2.A.25}$$

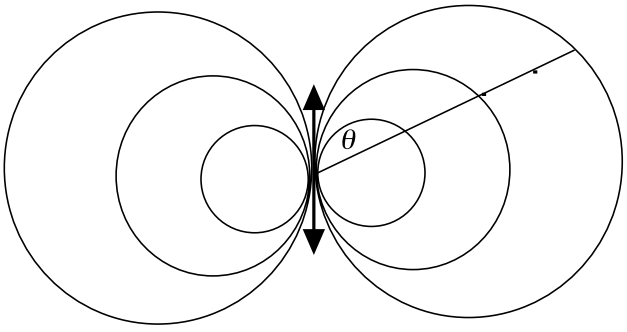


Fig. 2.A.3. Constant energy surfaces for a radiating dipole.

or after integration:

$$P = \frac{q^2 a^2 \omega^4}{12\pi\varepsilon_0 c^3} \tag{2.A.26}$$

Total power radiated by an oscillating dipole

This expression is rich in information. It indicates, for example, that the power emitted by an oscillating dipole is proportional to the fourth power of the frequency. Therefore, blue light ($\lambda \approx 0.4\,\mu m$) is scattered about a factor of 6 times more than red light ($\lambda \approx 0.65\,\mu m$) by particles in the atmosphere. This explains why the sky is blue, why sunset appears red, and why blue neon signs cast a bright halo on foggy nights. In more prosaic terms, it gives us a method of calculating the lifetime of electrons in atomic levels.

What (2.A.26) tells us is that as electrons lose their energy by electromagnetic radiation, their motions should eventually peter out. The origin of this effect is the work done by the moving charged particle on the electromagnetic field that it creates itself! To describe this we introduce in the expression for the dipole, a friction term γ_R (equal to $1/\tau_R$, where τ_R is the radiative lifetime) corresponding to the radiation loss:

$$m\ddot{z} + m\gamma_R \dot{z} + m\omega^2 z = 0 \tag{2.A.27}$$

We know that if $\tau_R \gg 1/\omega$, the solutions for the motion take the form $z = a\cos\omega t\, e^{-t/2\tau_R}$ and the particle's energy decreases as Ee^{-t/τ_R}. The lifetime τ_R is then given by:

$$\frac{dE}{dt} = P = -\frac{E}{\tau_R} \tag{2.A.28}$$

where P is given by (2.A.26) and the energy of the particle is obtained from:

$$E = \frac{1}{2}m\dot{z}^2 + \frac{1}{2}m\omega^2 z^2 \approx \frac{1}{2}m\omega^2 a^2 \tag{2.A.29}$$

We thereby obtain the radiative lifetime of an oscillating electron surrounding an atom:

$$\tau_R = \frac{6\pi\varepsilon_0 c^3 m}{q^2\omega^2} \tag{2.A.30}$$

Radiative lifetime of an electron oscillating with frequency ω

which in terms of wavelength becomes:

$$\tau_R = \frac{3}{2\pi}\frac{\varepsilon_0 cm\lambda^2}{q^2} = 4.5 \times 10^{-8}\,\lambda_{\mu m}^2 \tag{2.A.31}$$

Radiative lifetime of an electron emitting an electromagnetic wave of wavelength λ

Expression (2.A.31) allows us to understand the despair of physicists at the end of the nineteenth century. As electrons were thought of as typically orbiting atomic nuclei at frequencies of the order of 10^{15} Hz, (2.A.31) indicates that they all should screech to a halt and collapse onto the nuclei after a few nanoseconds. Quantum mechanics, as we shall see in Chapter 3, allows us to resolve this historical paradox. Equation (2.A.31) nonetheless holds reasonably true in predicting the lifetimes of radiative atomic transitions in lasers. Below are two typical examples:

$\lambda = 1\,\mu m, \quad \tau = 45\,ns$
$\lambda = 10\,nm, \quad \tau = 4.5\,ps$

It is therefore much easier to accumulate electrons in excited states subject to infrared transitions, than to populate transitions leading to X-ray emission. As a result, we shall see that the implementation of infrared lasers is much less difficult than is the case for X-ray lasers.

Example

Equation (2.A.26) is a very valuable tool for calculating the optical power emitted by a point source, i.e. when no propagation effect is taken into account. As an application, we shall determine the total power scattered at 2ω by a non-linear material illuminated by photons of pulsation ω.

Let ρ be the atomic density of non-linear scattering centres (m^{-3}), d the non-linear susceptibility (in $m\,V^{-1}$), and p_ω the incident light intensity ($W\,m^{-2}$). The non-linear dipole moment qa is related to coefficient d (see Chapter 12) through:

$$qa\rho = \varepsilon_0 dE_\omega^2 = 2Z_0\varepsilon_0 dp_\omega \tag{2.A.32}$$

The total power $P_{2\omega}$ radiated at 2ω in all directions by the ρV scattering centres (V is the focus volume) is given by (2.A.26), i.e:

$$P_{2\omega} = \rho V\frac{(d/\rho)^2\omega^4}{3\pi\varepsilon_0 c^5}p_\omega^2 \tag{2.A.33}$$

so that

$$\frac{P_{2\omega}}{V} = \frac{d^2\omega^4}{3\pi\varepsilon_0 c^5 \rho} p_\omega^2 \tag{2.A.34}$$

For lithium niobate, $d = 17\,\mathrm{pm\,V^{-1}}$, $\rho = 10^{22}\,\mathrm{cm^{-3}}$, and a pumping wavelength of 1.06 μm and power of $100\,\mathrm{MW\,cm^{-2}}$, one finds a total power of 1.4 μW emitted at 532 nm in all directions of space.

FURTHER READING

A. Aspect, C. Fabre, and G. Grynberg, *Laser et Atomes*, Ellipses, Paris (1997).
J. D. Jackson, *Classical Electrodynamics*, Wiley, New York (1975).
R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).

2.B Thermography

Equation (2.91) gives the explicit spectral distribution of the energy trapped inside a blackbody at temperature T. We still need to find the incident optical power per unit area per unit wavelength called the *spectral emittance* (i.e. a measurable quantity which allows *thermography* or thermal imaging).

Let us consider a surface area element $\mathrm{d}\mathbf{S}$ of the cavity within the blackbody (see Fig. 2.B.1) In the solid angle $\mathrm{d}\Omega$ subtended by this surface element, the optical power $\mathrm{d}R$ incident upon $\mathrm{d}\mathbf{S}$ in the spectral interval $\mathrm{d}\nu$ during time $\mathrm{d}t$ is the product of the mode density $\rho(\nu)$ with the volume occupied by the modes $(c/2\,\mathrm{d}t)(\mathrm{d}S\cos\theta)\mathrm{d}\Omega/2\pi$ or:

$$\mathrm{d}^4R = \frac{c}{2}\mathrm{d}t\rho(\nu)\mathrm{d}\nu\frac{\mathrm{d}S\cos\theta\mathrm{d}\Omega}{2\pi} \tag{2.B.1}$$

In this last equation, the factor of two in $c/2$ derives from the fact that Eq. (2.91) counts standing waves (comprising both incident and reflected waves) whereas only those waves *incident* upon $\mathrm{d}\mathbf{S}$ are relevant to the present tally. The factor $\mathrm{d}\Omega/2\pi$ presupposes that the emission is isotropic and the factor $\cos\theta$ takes into account the projection of the incident flux upon the surface element $\mathrm{d}\mathbf{S}$. Recalling that the solid angle element is given by $\mathrm{d}\Omega = 2\pi\sin\theta\,\mathrm{d}\theta$, Eq. (2.B.1) may be integrated to find:

$$\mathrm{d}^2R = \frac{c}{2}\rho(\nu)\,\mathrm{d}\nu\,\mathrm{d}S\int_0^{\pi/2} 2\pi\frac{\cos\theta\sin\theta}{2\pi}\mathrm{d}\theta = \frac{c}{4}\rho(\nu)\,\mathrm{d}\nu\,\mathrm{d}S \tag{2.B.2}$$

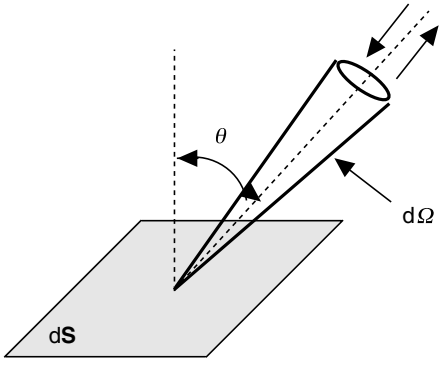


Fig. 2.B.1. Thermal electromagnetic flux on a blackbody surface.

The spectral emittance for the blackbody is obtained by substituting (2.91) into (2.B.2) to give:

$$\mathrm{d}R(v,T) = \frac{2\pi h}{c^2}\frac{v^3\,\mathrm{d}v}{\mathrm{e}^{hv/kT}-1} \tag{2.B.3}$$

The spectral emittance is most often represented in the wavelength rather than in the frequency space. We thus replace v in (2.B.3) by c/λ giving:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}R(\lambda,T) = \frac{2\pi hc^2}{\lambda^5}\frac{1}{\mathrm{e}^{hc/\lambda kT}-1} \tag{2.B.4}$$

Spectral emittance from a blackbody in $\mathrm{W\,m^{-1}\,m^{-2}}$

Different emittance spectra are represented in Fig. 2.B.2.
Considering the relative energy associated with a photon of wavelength λ (hc/λ) and the thermal energy kT of a source, two limiting cases of physical interest are worth distinguishing:

1. Short wavelength photons with energies greatly in excess of the source temperature $hc/\lambda \gg kT$.

   In this case, Eq. (2.B.4) becomes:

   $$\frac{\mathrm{d}}{\mathrm{d}\lambda}R(\lambda,T) \approx \frac{2\pi hc^2}{\lambda^5}\mathrm{e}^{-(hc/\lambda kT)} \tag{2.B.5}$$

   which is valid for $\lambda T < 5000\ \mu\mathrm{m\,K}$. This is a regime dominated by quantization of the photon energy attested by the presence of hv/kT in the equation.
2. Long wavelength photons carrying negligible energy relative to the thermal source energy $hc/\lambda \ll kT$.

   A first-order expansion of Eq. (2.B.4) immediately yields:

   $$\frac{\mathrm{d}}{\mathrm{d}\lambda}R(\lambda,T) \approx 2\pi ckT\lambda^{-4} \tag{2.B.6}$$

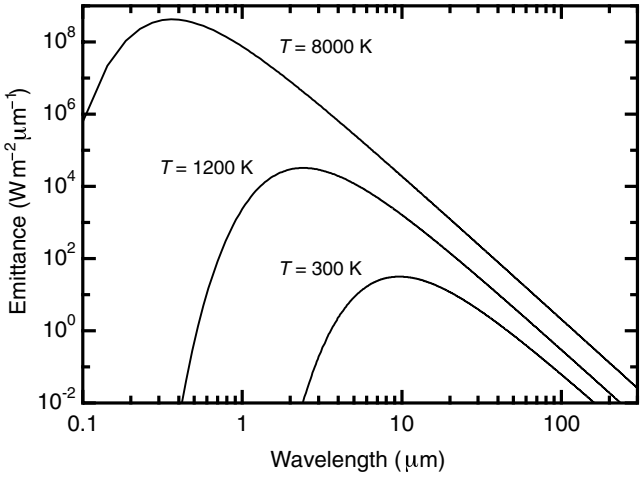


Fig. 2.B.2. Spectral emittance in $\mathrm{W\,m^{-2}\,\mu m^{-1}}$ for a blackbody at several different temperatures (ambient, 300 K; 'red hot', 1200 K; surface of the Sun, 8000 K).

in which we notice the disappearance of Planck's constant h from the expression for the emittance. In this regime, called the *Rayleigh–Jeans regime*, the quantized aspect of the photons is masked by the thermal agitation of the system. This expression could have been obtained directly (as done by Rayleigh and Jeans) starting from the classical expression (2.24) for electromagnetic energy in a cavity. We note as well, that (2.B.6) diverges as $\lambda \to 0$, which at the beginning of the twentieth century was referred to as the *ultraviolet catastrophe*. It is this failure of classical physics which led ultimately to the development of the theory of quanta.

Taking the integral of (2.B.4) over all wavelength space, leads to an expression for the total emittance of a blackbody:

$$R(\lambda, T) = 2\pi hc^2 \int_0^\infty \frac{1}{\lambda^5} \frac{1}{\mathrm{e}^{hc/\lambda kT} - 1} \mathrm{d}\lambda \tag{2.B.7}$$

Making use of Bernoulli's integral, $\int_0^\infty u^3/(\mathrm{e}^u - 1)\mathrm{d}u = \pi^4/15$, Eq. (2.B.7) immediately gives:

$$R(\lambda, T) = \sigma T^4 \tag{2.B.8a}$$

Stephan–Boltzmann law for total blackbody emittance

where σ is Stephan's constant:

$$\sigma = \frac{2\pi^5 k^4}{15c^2h^3} = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}} \tag{2.B.8b}$$

Example

Let us calculate the total power P emitted by a human body having a surface $S = 2\,\mathrm{m}^2$ and a temperature of 300 K.

$$P = R \times S = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}} \times (300\,\mathrm{K})^4 \approx 0.5\,\mathrm{kW}$$

where this quantity gives the expected loss for an unprotected body exposed to an ambient of 0 K – not a common occurrence!

As may be seen in Fig. 2.B.2, the emittance of a blackbody peaks at a maximum λ for a given temperature. This maximum may be obtained by differentiation of Eq. (2.B.4) with respect to λ, which gives the following implicit equation:

$$\mathrm{e}^{hc/\lambda kT} = \frac{5}{5 - (hc/\lambda kT)} \tag{2.B.9}$$

and admits as a solution:

$$\lambda_{\max} = \frac{2898}{T}, \text{ in } \mu\mathrm{m} \tag{2.B.10}$$

Wien's law

Thus, the emissivity maximum for a body at ambient temperature is located at a wavelength of roughly 2898/290 or 10 μm. Therefore, the imaging of objects at room temperature may be achieved by using detectors which are sensitive to wavelengths in the vicinity of 10 μm.

This range is also interesting by virtue of the fact that the atmosphere is effectively transparent at these wavelengths. Figure 2.B.3 shows a graph of atmospheric transparency as a function of wavelength. We notice in particular two wide bands situated between 3 and 5 μm (band II) and 8 and 12 μm (band III) where the atmosphere is particularly transparent.

These two detection windows each have their particular subscribers. A number of considerations will help determine which of the two bands are best exploited in a given context:

- Climatic conditions: moist air masses maintain better transparency in the range of band II, while scattering clouds, present less of an obstacle in band III.
- Thermal contrast: clearly, a greater portion of radiation will be emitted by objects at 300 K in the 8–12 μm range (12.2 mW cm^{-2} – see the example below), than in the 3–5 μm range (0.6 mW cm^{-2}). Nonetheless, sometimes we are interested in being able to detect a thermal source having a temperature T_s against a background of temperature T_b. The thermal contrast is therefore the more

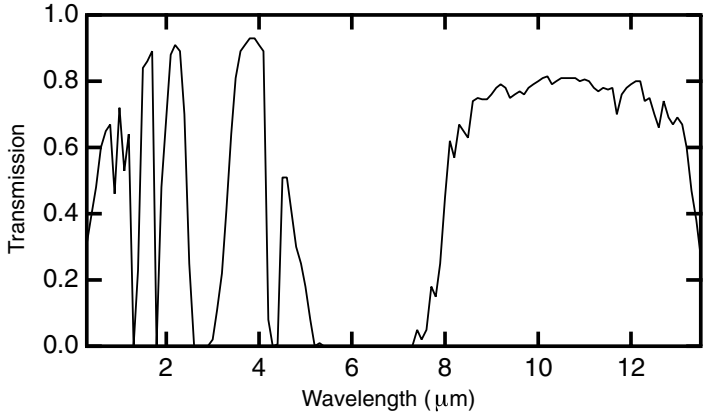


Fig. 2.B.3. Spectral dependence of the atmospheric transmission under typical conditions (2 km, sea level, 20 °C, 40% humidity).

physically pertinent quantity and corresponds to the normalized difference between the power emitted by the source and the background in a spectral range of $\Delta\lambda$ or:

$$C(\Delta\lambda) = \frac{\int_{\Delta\lambda} \frac{dR(\lambda, T_s)}{d\lambda} d\lambda - \int_{\Delta\lambda} \frac{dR(\lambda, T_b)}{d\lambda} d\lambda}{\int_{\Delta\lambda} \frac{dR(\lambda, T_s)}{d\lambda} d\lambda + \int_{\Delta\lambda} \frac{dR(\lambda, T_b)}{d\lambda} d\lambda} \tag{2.B.11}$$

Figure 2.B.4 compares the thermal contrast in bands II and III relative to a background of 280 K. The best contrast is seen to lie in the 3–5 μm range, an important consideration depending on the application.

Another important concept is the differential contrast C_d (in $W\,cm^{-2}\,K^{-1}$). By definition, $C_d\Delta T$ is the thermal emittance per degree K in a $\Delta\lambda$ band given by:

$$C_d(\Delta\lambda) = \frac{d}{dT} \int_{\Delta\lambda} \frac{dR(\lambda, T_s)}{d\lambda} d\lambda \tag{2.B.12}$$

This last notion is extremely useful for predicting the performance of infrared detectors. If the response of a detector with surface S is given by $\mathscr{R}$ (in $A\,W^{-1}$), then the detected current I (in A) due to a variation in the source temperature ΔT is given by:

$$I = S\mathscr{R}C_d(\Delta\lambda)\Delta T \tag{2.B.13}$$

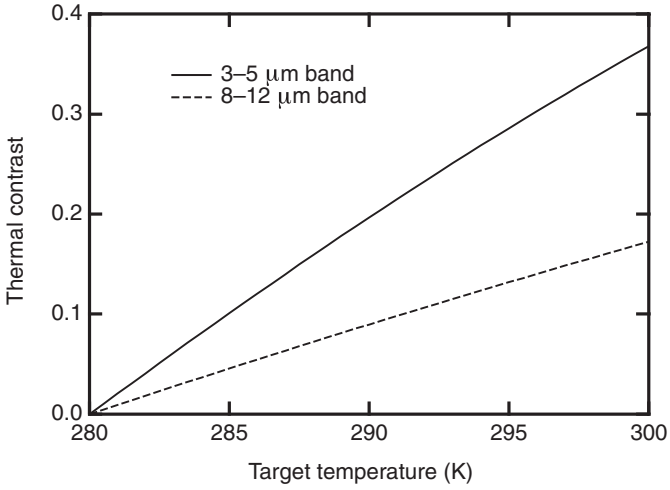


Fig. 2.B.4. Thermal contrast in the 3–5 μm (band II) and 8–12 μm (band III) bands. We note that although the photon count is lower in band II, the thermal contrast (or the ability in this case to discern a target with temperature T against a background at 280 K) is twice as large.

Example

The MATHEMATICA program below allows calculation of the emissivity of a perfect blackbody in the spectral range λ_1–λ_2 as well as the differential contrast.

```
(* this program calculates the blackbody emission as a function of the temperature ''temp'' between two
wavelengths, lambda1 and lambda2*)
    (*universal constants*)
c=2.988 10^8 (* m/s*);
k=1.38 10^-23 (*J/K*);
h=6.625 10^-34 (*J.s*);
hb=h/(2*3.1416);
m0=0.91 10^-30 (*kg*);
    (* Wavelength range in μm *)
lambda1= 8;lambda2= 12;
    (* Blackbody as a function of temperature*)
temp=.; lambda=.;
emm:= 2*3.1416*h*c^2*(lambda*10^-6)^-5/(Exp[h*c/((lambda*10^-6)*k*temp)]-1)
emm = emm * 10^-6 (* W/m^2/μm*);
contr=D[emm,temp];(*contrast used for the calculation of NETD*)
temp=300 (*temperature of the blackbody*);
Plot[emm,{lambda,.1,20}, Frame->True, RotateLabel->False, FrameLabel->
{''micron'',''W/m^2''} ]
Plot[contr,{lambda,.1,20}, Frame->True, RotateLabel->False, FrameLabel->
{''micron'',''W/m^2/K''} ]
    (* Blackbody power radiated over the spectral range*)
```

```
NIntegrate[emm,{lambda,lambda1,lambda2}] (*W/m^2*);
pp=% * 10^-4 (* W/cm^2*)
    (* differential contrast for the spectral range *)
NIntegrate[contr,{lambda,lambda1,lambda2}] (*W/m^2/K*);
cc=%*10^-4 (* W/cm^2/K*)
```

This program then gives for a source at 300 K, an emitted power density of 12.2 mW cm^{-2} between 8 and 12 μm and 0.6 mW cm^{-2} between 3 and 5 μm. The detector response will therefore be much larger in band III then in band II. We also find a differential contrast of 2.1 × 10^{-5} W cm^{-2} K^{-1} between 3 and 5 μm and of 1.98 × 10^{-4} W cm^{-2} K^{-1} between 8 and 12 μm. For a detector with a surface of 10^{-4} cm^2 and a sensitivity of 1 A W^{-1} in each of the two spectral bands, a variation of 10 mK in the source temperature will lead to a detector current of 20 pA from the 3–5 μm band and 200 pA from the 8–12 μm band. Values such as these are readily measurable and are suggestive of the great usefulness of infrared detectors. We will return to these important devices again in Chapter 11.

Lastly, it is worth pointing out that most ordinary objects are not in fact blackbodies. They only absorb a portion $\varepsilon_{se}(\lambda)$ called the *spectral emissivity*, with the remaining amount of radiation either being reflected and/or transmitted. The spectral emittance $dR/d\lambda$ is therefore related to the perfect blackbody emittance $dR_{BB}/d\lambda$ by:

$$\frac{d}{d\lambda}R(\lambda, T) = \varepsilon_{se}(\lambda)\frac{d}{d\lambda}R_{BB}(\lambda, T) \tag{2.B.14}$$

Typical emittances range from about 0.03 for polished aluminium (considered to be a good reflector) to 0.95 for black soot (a good absorber).

FURTHER READING

G. Gaussorgues and S. Chomet, *Infrared Thermography*, Kluwer, Boston (1993).

R. J. Keyes, ed., *Optical and Infrared Detectors*, Topics In Applied Physics Vol. 19, Springer-Verlag, Berlin (1980).

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).

3 Quantum mechanics of electron–photon interaction

3.1 Introduction

We begin this chapter by describing the interaction between a quantum system and an electromagnetic wave in a cavity. During this interaction, the electromagnetic field may be considered to be classical, taking on arbitrary values within the cavity or, alternately, to be quantized in the form of photons. The first case allows us to describe the phenomena of absorption and stimulated emission in a quantum system. The second case will afford a description of spontaneous emission from an excited state – an occurrence which may not be easily addressed in classical terms (see Complement 2.A). We then discuss the Einstein rate equations for a system subject to an ensemble of these processes, which form the basis for a description of laser operation.

3.2 Dipolar interaction Hamiltonian for electrons and photons

We will consider a particle of charge q and mass m subject to a static potential $V(\mathbf{r})$ resulting from an atom, a quantum well, etc. In the classical approach, the particle's Hamiltonian is given by:

$$H = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r}) \tag{3.1}$$

In the presence of a classical electromagnetic field, it is not very difficult (although perhaps somewhat tedious) to show that the classical Hamiltonian for the particle is:

$$H = \frac{[\mathbf{p} - q\mathbf{A}(\mathbf{r}, t)]^2}{2m} + V(\mathbf{r}) + qU(\mathbf{r}, t) \tag{3.2}$$

where $\mathbf{A}(\mathbf{r}, t)$ and $U(\mathbf{r}, t)$ are the vector and scalar potentials, respectively, for the electromagnetic wave (see Section 2.2). The correspondence principle *applied to a single particle* furnishes us with an expression for the quantum Hamiltonian:

$$\hat{H} = \frac{[\hat{\mathbf{p}} - q\mathbf{A}(\hat{\mathbf{r}}, t)]^2}{2m} + V(\hat{\mathbf{r}}) + qU(\hat{\mathbf{r}}, t) \tag{3.3}$$

We notice in this first approach that $\mathbf{A}$ is not an operator itself, but a vector function of the operator $\hat{\mathbf{r}}$. Before going any further, it is useful to decide upon which gauge to use (refer to Section 2.2). Here we will employ the Coulomb gauge as it leads to particularly simple expressions for the interaction Hamiltonian with a monochromatic plane wave.

Coulomb gauge

In the Coulomb gauge, the vector potential is assumed to have a null divergence:

$$\nabla \cdot \mathbf{A}_{\text{Coulomb}}(\mathbf{r}, t) = 0 \tag{3.4}$$

Referring to relation (2.13) in reciprocal space, this signifies that the vector potential has no parallel component and is determined uniquely by its normal component $\mathbf{A}_{\perp}$. Let us take, for example, a monochromatic electromagnetic plane wave having an electric field given by:

$$\mathbf{E} = \mathbf{E}_0 \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \tag{3.5a}$$

$$\mathbf{B} = \frac{\mathbf{k} \times \mathbf{E}_0}{\omega} \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \tag{3.5b}$$

$$\mathbf{E}_0 \cdot \mathbf{k} = 0 \tag{3.5c}$$

The vector potential is given by (2.23) or:

$$\mathbf{A}_{\perp} = \frac{\mathbf{E}_0}{\omega} \sin(\mathbf{k} \cdot \mathbf{r} - \omega t) \tag{3.6}$$

Applying (2.4b) and (2.12), we find that the scalar potential is null ($U(\mathbf{r}, t) = 0$), in which case the interaction Hamiltonian (3.3) takes the simple form:

$$\hat{H} = \frac{[\hat{\mathbf{p}} - q\mathbf{A}_{\perp}(\hat{\mathbf{r}}, t)]^2}{2m} + V(\hat{\mathbf{r}}) \tag{3.7}$$

Hamiltonian for a single particle subjected to an electromagnetic field

and may be expanded as:

$$\hat{H} = \frac{\hat{\mathbf{p}}^2}{2m} + V(\hat{\mathbf{r}}) - \frac{q}{2m}[\hat{\mathbf{p}} \cdot \mathbf{A}_{\perp}(\hat{\mathbf{r}}, t) + \mathbf{A}_{\perp}(\hat{\mathbf{r}}, t) \cdot \hat{\mathbf{p}}] + \frac{q^2}{2m} \cdot \mathbf{A}_{\perp}^2(\hat{\mathbf{r}}, t) \tag{3.8}$$

Two important observations allow simplification of this last expression. The second-order term in $\mathbf{A}_{\perp}^2$ may be neglected as it can be shown to be small in comparison with the linear term. Second, the operators $\hat{\mathbf{p}}$ and $\mathbf{A}_{\perp}(\hat{\mathbf{r}})$ commute. By evaluating the effect of the del operator ∇on the product $\mathbf{A}_{\perp}(\hat{\mathbf{r}}, t)\psi(\mathbf{r}, t)$ we immediately find:

$$[\mathrm{i}\hbar\nabla, \mathbf{A}_{\perp}(\hat{\mathbf{r}}, t)] = \mathrm{i}\hbar\nabla \cdot \mathbf{A}_{\perp}(\hat{\mathbf{r}}, t) \tag{3.9}$$

and given the definition of the normal component $\mathbf{A}_{\perp}$ (see relation (2.13)), this last term is null. Finally, thanks to the use of the Coulomb gauge, the Hamiltonian for the system takes the following simple form:

$$\hat{H} = \hat{H}_0 + \hat{W}(t) \tag{3.10}$$

where $\hat{H}_0$ is the Hamiltonian of the unperturbed system and $\hat{W}(t)$ is the perturbing Hamiltonian:

$$\hat{W}(t) = -\frac{q}{m}\mathbf{A}_{\perp}(\hat{\mathbf{r}}, t) \cdot \hat{\mathbf{p}} \tag{3.11}$$

$\mathbf{A} \cdot \mathbf{p}$ interaction Hamiltonian for an electron in an electromagnetic field

This last Hamiltonian carries the name $\mathbf{A} \cdot \mathbf{p}$. We end this paragraph by noting two important points. The first is that, generally, the mean particle displacement amplitude is much smaller than the wavelength of the electromagnetic excitation. The typical quantum systems studied in this book have physical dimensions between 0.1 and 10 nm, which are effectively very small in comparison with their resonant optical transition wavelengths (200 nm to 10 μm). As a result, the spatial variation of the vector potential will be neglected, so that if the system is centred at $\mathbf{r}_0$, the perturbing Hamiltonian is then:

$$\hat{W}(t) = -\frac{q}{m}\mathbf{A}_{\perp}(\mathbf{r}_0, t) \cdot \hat{\mathbf{p}} \tag{3.12}$$

The second point is that there is an equivalent derivation to (3.12) which is also commonly used (called the *Göppert–Mayer* gauge):

$$\hat{W}(t) = -q\hat{\mathbf{r}} \cdot \mathbf{E}(\mathbf{r}_0, t) = -\hat{\mathbf{D}} \cdot \mathbf{E}(\mathbf{r}_0, t) \tag{3.13}$$

$\hat{\mathbf{D}}$ is the dipole operator, and this Hamiltonian is referred to as the *electric dipole Hamiltonian* or the $\mathbf{D} \cdot \mathbf{E}$ *Hamiltonian*. The equivalence between these two Hamiltonians is investigated in Complement 3.D.

3.3 Linear optical susceptibility obtained by the density matrix

We now interest ourselves in the interaction between a two-level quantum system and a plane monochromatic electromagnetic wave. This two-level system is described by a Hamiltonian H_0 having stationary states $|1\rangle$ and $|2\rangle$ and energies E_1 and E_2 (i.e. $H_0|i\rangle = E_i|i\rangle$). No longer having to worry about confusing operators with classically measurable quantities, we will (temporarily) allow ourselves to

drop the '^' symbol to designate operators. In the stationary state basis, the Hamiltonian H_0 may be written:

$$H_0 = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \tag{3.14}$$

At $t = 0$, this system is subjected to a perturbation $W(t) = -e\mathbf{r} \cdot \mathbf{E}_0 \cos \omega t$ (the electric dipole Hamiltonian). With respect to the basis formed by $|1\rangle$ and $|2\rangle$, the matrix representation for W is then:

$$W = -E_0 \begin{bmatrix} 0 & D_{12} \\ D_{12} & 0 \end{bmatrix} \cos \omega t \tag{3.15}$$

where the D_{ij} terms are elements of the dipole matrix:

$$D_{ij} = q\langle i|\boldsymbol{\varepsilon} \cdot \hat{\mathbf{r}}|j\rangle \tag{3.16}$$

from which we recall that $\boldsymbol{\varepsilon}$ is the polarization vector of the wave. We have assumed the system to be symmetric with respect to the centre and, as a result, the elements D_{11} and D_{22} are null. Also, to streamline the notation, we have supposed the terms D_{12} and D_{21} to be real and therefore equal. We saw in Section 1.9 that a system may be described by its density matrix, in which case Schrödinger's equation (see (1.101)) may be written:

$$\begin{aligned} \frac{\mathrm{d}\rho_{11}}{\mathrm{d}t} &= \mathrm{i}\frac{D_{12}}{\hbar} E_0(\rho_{21} - \rho_{12})\cos \omega t - \Gamma_1(\rho_{11} - \rho_{11}^{\mathrm{eq}}) \\ \frac{\mathrm{d}\rho_{22}}{\mathrm{d}t} &= -\mathrm{i}\frac{D_{12}}{\hbar} E_0(\rho_{21} - \rho_{12})\cos \omega t - \Gamma_1(\rho_{22} - \rho_{22}^{\mathrm{eq}}) \\ \frac{\mathrm{d}\rho_{21}}{\mathrm{d}t} &= -\mathrm{i}(\omega_{21} - \mathrm{i}\Gamma_2)\rho_{21} - \mathrm{i}\frac{D_{12}}{\hbar} E_0(\rho_{22} - \rho_{11})\cos \omega t \end{aligned} \tag{3.17}$$

where Γ_1 $(= 1/T_1)$ is the inelastic population relaxation rate between levels, and Γ_2 $(= 1/T_2)$ is the phase relaxation rate. As described in Complement 1.E, we know that after a transitory period lasting a few T_2s, ρ_{12} will oscillate as $\cos \omega t$, whereas the terms ρ_{11} and ρ_{22} will tend towards their respective stationary state values. It is therefore useful to introduce the change of variables:

$$\begin{aligned} \rho_{21} &= \sigma_{21} \mathrm{e}^{-\mathrm{i}\omega t} \\ \rho_{12} &= \sigma_{12} \mathrm{e}^{\mathrm{i}\omega t} \end{aligned} \tag{3.18}$$

which, once introduced into (3.17), and after neglecting the non-resonant terms in $\omega + \omega_{21}$, leads to the coupled equations:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\sigma_{21} + \mathrm{i}\left[(\omega_{21} - \omega) - \mathrm{i}\frac{1}{T_2}\right]\sigma_{21} = \mathrm{i}\frac{D_{12}}{2\hbar}E_0\Delta\rho \\
&\frac{\mathrm{d}}{\mathrm{d}t}\Delta\rho + \frac{\Delta\rho - \Delta\rho^{\mathrm{eq}}}{T_1} = \mathrm{i}\frac{D_{12}}{\hbar}E_0(\sigma_{21} - \sigma_{12})
\end{aligned} \tag{3.19}$$

Once in a stationary state, the populations ρ_{22}, ρ_{11} and, therefore, $\Delta\rho = \rho_{11} - \rho_{22}$ are constant, as well as the amplitudes σ_{12} and σ_{21}. Equations (3.19) then become:

$$\begin{aligned}
&\sigma_{21} = \frac{\Omega_{12}}{2}\frac{1}{(\omega_{21} - \omega) - \mathrm{i}\dfrac{1}{T_2}}\Delta\rho \\
&\frac{\Delta\rho - \Delta\rho^{\mathrm{eq}}}{T_1} = -2\Omega_{12}\mathrm{Im}(\sigma_{21})
\end{aligned} \tag{3.20}$$

where we recall that $\Omega_{12} = qE_0|\langle 1|\mathbf{r}\cdot\boldsymbol{\varepsilon}|2\rangle/\hbar = E_0 D_{12}/\hbar$ is the *Rabi frequency* (see Complement 1.E). This system may be immediately solved to yield:

$$\Delta\rho = \frac{1 + (\omega_{21} - \omega)^2 T_2^2}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2}\Delta\rho^{\mathrm{eq}} \tag{3.21a}$$

$$\mathrm{Im}\sigma_{21} = \frac{\Omega_{12}T_2}{2}\frac{1}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2}\Delta\rho^{\mathrm{eq}} \tag{3.21b}$$

$$\mathrm{Re}\sigma_{21} = \frac{\Omega_{12}T_2^2}{2}\frac{\omega_{21} - \omega}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2}\Delta\rho^{\mathrm{eq}} \tag{3.21c}$$

To make the bridge to macroscopic scales, we must recall on one hand that the number of particles in the states $|1\rangle$ and $|2\rangle$ is given by $N_1 = N\rho_{11}$ and $N_2 = N\rho_{22}$, where N is the total number of particles. On the other hand, equation (1.92) gives us the average displacement of the particle in the two-level quantum system:

$$\langle\hat{\mathbf{r}}\rangle = \mathrm{Tr}(\rho\hat{\mathbf{r}}) = \mathbf{r}_{12}(\rho_{12} + \rho_{21}) \tag{3.22}$$

Given (3.16) and (3.18), the polarization of the medium in the direction of the polarization vector $\boldsymbol{\varepsilon}$ is given by:

$$P = N\langle\hat{\mathbf{D}}\cdot\boldsymbol{\varepsilon}\rangle = ND_{12}(\sigma_{12}\mathrm{e}^{\mathrm{i}\omega t} + \sigma_{21}\mathrm{e}^{-\mathrm{i}\omega t}) \tag{3.23a}$$

or again:

$$P = 2ND_{12}(\mathrm{Re}\,\sigma_{21}\cos\omega t + \mathrm{Im}\,\sigma_{21}\sin\omega t) \tag{3.23b}$$

We therefore introduce the linear optical susceptibility $\chi(\omega)$ defined as:

$$P(t) = \mathrm{Re}[\varepsilon_0 \chi(\omega) E_0 \mathrm{e}^{\mathrm{i}\omega t}] = \varepsilon_0 E_0 (\chi_{\mathrm{Re}} \cos \omega t - \mathrm{i}\chi_{\mathrm{Im}} \sin \omega t) \tag{3.24}$$

Optical polarization and linear susceptibility

where

$$\chi(\omega) = \chi_{\mathrm{Re}}(\omega) + \mathrm{i}\chi_{\mathrm{Im}}(\omega) \tag{3.25}$$

In atomic physics, the quantity $\varepsilon_0 \chi(\omega)$ is the polarizability of the quantum system. The term χ_{Re}, which is in phase with the wave $E_0 \cos \omega t$, describes the instantaneous response of the medium and gives rise to the index of refraction. On the other hand, the term χ_{Im} is 90° behind the phase of the exciting wave and, as we shall show in the next paragraph, describes the absorption (or the dissipation occurring in the medium). Proceeding from a term by term identification, we may find the three macroscopic values which describe the interaction of a two-level system with an electromagnetic wave (i.e. the population differences, and the real and imaginary portions of the linear optical susceptibility):

$$N_1 - N_2 = \frac{1 + (\omega_{21} - \omega)^2 T_2^2}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2} (N_1^{\mathrm{eq}} - N_2^{\mathrm{eq}}) \tag{3.26a}$$

$$\chi_{\mathrm{Re}} = \frac{D_{12}^2 T_2^2}{\varepsilon_0 \hbar} \frac{\omega_{21} - \omega}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2} (N_1^{\mathrm{eq}} - N_2^{\mathrm{eq}}) \tag{3.26b}$$

$$\chi_{\mathrm{Im}} = -\frac{D_{12}^2 T_2^2}{\varepsilon_0 \hbar} \frac{1}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2} (N_1^{\mathrm{eq}} - N_2^{\mathrm{eq}}) \tag{3.26c}$$

equations at times written more concisely as:

$$\Delta N = \frac{1 + \delta\omega^2 T_2^2}{1 + \delta\omega^2 T_2^2 + \Omega_{12}^2 T_1 T_2} \Delta N^{\mathrm{eq}} \tag{3.27a}$$

$$\chi_{\mathrm{Re}} = \frac{D_{12}^2 T_2^2}{\varepsilon_0 \hbar} \frac{\delta\omega}{1 + \delta\omega^2 T_2^2} \Delta N \tag{3.27b}$$

$$\chi_{\mathrm{Im}} = -\frac{D_{12}^2 T_2^2}{\varepsilon_0 \hbar} \frac{1}{1 + \delta\omega^2 T_2^2} \Delta N \tag{3.27c}$$

We will now describe the connection between these quantities and macroscopic optical properties such as the absorption and the gain of a quantum system.

3.4 Linear optical susceptibility: absorption and optical gain

Matter is made of negative particles (electrons) which shield the positively charged nucleus background. When an electromagnetic (em) field is applied to matter, the

motion of the electrons (light) relative to the nuclei (heavy) induces an oscillating dipole which radiates in tune with the em field. This is the atomic polarization. The radiated field interferes with the original applied field, leading to an apparent decrease of the speed of light. This effect can be taken into account by using a microscopic approach based on the oscillating dipole expression (Eq. (2.A.18)). The reader is encouraged to deepen this instructive aspect of refraction in Chapter 23 of the mythical *Feynman Lectures*.

However, though less illuminating, Maxwell's equations provide a far more powerful tool for taking into account this effect without resorting to complex expressions. For that purpose, the displacement current due to the motion of the bound charges is introduced in the $\partial E/\partial t$ term of Faraday–Ampère's equation. This leads to Maxwell's equations for an uncharged, non-magnetic, and polarizable medium:

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \tag{3.28}$$

$$\nabla \times \mathbf{B}(\mathbf{r},t) = \frac{1}{\varepsilon_0 c^2}\frac{\partial}{\partial t}\mathbf{D}(\mathbf{r},t) \tag{3.29}$$

where $\mathbf{D}$ is the displacement vector given by:

$$\mathbf{D}(\mathbf{r},t) = \varepsilon_0\mathbf{E} + \mathbf{P} + \mathbf{P}_{\text{res}} \tag{3.30}$$

and $\mathbf{P}_{\text{res}}$ is the polarization vector of the system near resonance. $\mathbf{P}$ is the polarization vector corresponding to all other off-resonant contributions (i.e. resulting from the host material in which the two level system is contained). The term $\varepsilon_0\mathbf{E} + \mathbf{P}$ is, by definition of the dielectric constant ε, replaced by $\varepsilon\mathbf{E}$. Similarly, following the definition of linear optical susceptibility (3.24), (3.30) may be rewritten as:

$$\mathbf{D} = \varepsilon\left[1 + \frac{\varepsilon_0}{\varepsilon}\chi(\omega)\right]\mathbf{E} \tag{3.31}$$

Maxwell's equations (3.28) and (3.29) admit as solutions travelling waves of the form:

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \mathrm{Re}[\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\mathbf{r}-\omega t)}] \tag{3.32}$$

where this time, the wavevector $\mathbf{k}$ and the radial frequency ω are related by:

$$k = |\mathbf{k}| = \frac{n_{\text{op}}\omega}{c}\left|1 + \frac{\varepsilon_0}{\varepsilon}\chi\right|^{1/2} \tag{3.33}$$

n_{op} is the *optical index of refraction* equal to $\sqrt{\varepsilon/\varepsilon_0}$. As the susceptibility may not be large compared with unity:

$$k = \frac{n_{op}\omega}{c}\left(1 + \frac{1}{2}\frac{\varepsilon_0}{\varepsilon}\chi_R\right) + i\frac{\omega}{c}\frac{\chi_{Im}}{2n_{op}} \tag{3.34}$$

The real part of the latter expression seems to indicate that the speed of light is reduced by a factor n_{op} in the medium. Once again, the reader is encouraged to deepen this aspect in the *Feynman Lectures*. Substituting (3.34) into (3.32), we see that the amplitude of the electromagnetic wave dampens exponentially as a function of distance:

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 e^{-\alpha/2 z}\mathrm{Re}[e^{-i(\mathbf{k}_R\cdot\mathbf{r}-\omega t)}] \tag{3.35}$$

where $\mathbf{k}_R$ is given by the real portion of (3.34) and the coefficient α is the absorption coefficient given by:

$$\alpha(\omega) = \frac{\omega}{c n_{op}}\chi_{Im} \tag{3.36}$$

Absorption coefficient and linear susceptibility

or:

$$\alpha(\omega) = \frac{D_{12}^2}{c n_{op}\varepsilon_0\hbar}\frac{\omega T_2}{1 + (\omega_{21} - \omega)^2 T_2^2 + \Omega_{12}^2 T_1 T_2}(N_1^{eq} - N_2^{eq}) \tag{3.37}$$

Absorption coefficient for a two-level system

The electromagnetic wave loses intensity as $I(z) = I_0 e^{-\alpha z}$ as a function of propagation distance and over which the absorbed electromagnetic energy is converted into heat. It is a fundamental observation to note that it is the 90° phase lag of the χ_{Im} term with respect to the propagating electromagnetic wave that is at the origin of the thermal dissipation of the energy (and the irreversibility of the process). This is a general result applicable to all linear processes.

Figure 3.1 shows the variation of the form of the normalized absorption as a function of frequency ω for different values of the product $\Omega_{12}^2 T_1 T_2$. The numerical constants in (3.37) have been chosen so that $\alpha(\omega_{21}) = 1$ for $\Omega_{12} = 0$. We notice that the absorption decreases as the intensity of the electromagnetic wave increases and that the peak width increases as:

$$\Delta\nu(E_0) = \frac{1}{\pi T_2}\sqrt{1 + \frac{D_{12}^2 E_0^2 T_1 T_2}{\hbar^2}} \tag{3.38}$$

This phenomenon is referred to as *absorption saturation*, and we will return to it later in Chapter 4. For low wave intensities, $N_1 - N_2 \approx N_1$, and Eq. (3.37) takes on the form of a Lorentzian, which may be derived from classical theory for the elastically bound electron:

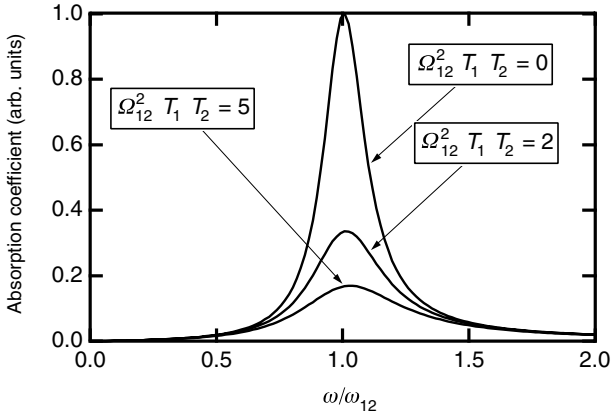


Fig. 3.1. Absorption saturation for $\omega_{21}T_2 = 10$ and different values of the product $\Omega_{12}^2 T_1 T_2$.

$$\alpha(\omega) \approx N_1 \frac{q^2\pi}{2\varepsilon_0 mcn_{\text{op}}} f_{21}\mathscr{L}(\omega - \omega_{21}) = f_{21}\alpha_{\text{clas}}(\omega) \tag{3.39a}$$

where $\mathscr{L}(\omega)$ is a Lorentzian of width $1/T_2$ centred at 0:

$$\mathscr{L}(\omega) = \frac{1/\pi T_2}{\omega^2 + (1/T_2)^2}$$
$$\mathscr{L}(\nu) = \frac{\Delta\nu/2\pi}{\nu^2 + (\Delta\nu/2)^2} \tag{3.39b}$$

Lorentzian lineshape

where $\Delta\nu$ is the full width at half maximum (FWHM) of the Lorentzian ($\Delta\nu = 2/T_2 = 2\Gamma_2$). α_{clas} is the result from Lorentz's classical theory and f_{21} is a dimensionless physical quantity called the *oscillator strength*, which describes the strength of the quantum coupling between the two levels $|1\rangle$ and $|2\rangle$:

$$f_{21} = \frac{2m}{\hbar^2} E_{21} |\langle 1|\mathbf{r}\cdot\varepsilon|2\rangle|^2 \tag{3.40a}$$

Oscillator strength for a transition from $|1\rangle \rightarrow |2\rangle$

In a system with several levels, we may show the Thomas–Reiche–Kuhn sum rule (Complement 3.D):

$$\sum_{i\neq j} f_{ji} = 1 \tag{3.40b}$$

This rule stipulates that although quantum mechanics may distribute the oscillator strengths among several levels, the sum of these absorptions must equal the classically derived quantity. The absorption in a system resulting from optical transitions between the fundamental level $|1\rangle$ and an arbitrary number of levels $|j\rangle$

results from a generalization of (3.39)[1]:

$$\alpha(\omega) \approx N_1 \sum_{j=1}^{\infty} \frac{q^2\pi}{2\varepsilon_0 mcn_{\text{op}}} f_{j1}\mathscr{L}(\omega-\omega_{j1}) \tag{3.41}$$

3.5 The rate equations

3.5.1 Adiabatic approximation and corpuscular interpretation

We may now give a corpuscular account of the absorption equations. To begin, we shall make the adiabatic approximation which assumes that the phase relaxation time T_2 is very short in comparison with the relaxation time for population T_1. This hypothesis is supported by the fact that the elastic mechanisms that dominate T_2 are much more numerous than the inelastic processes (also included in T_2) that control T_1. Consequently, we shall take for σ_{12} the expression obtained in (3.20), which relates the instantaneous value of ρ_{12} to that of $\Delta\rho$. We introduce this expression into the second differential equation of (3.19), by supposing that *for a few moments* the lifetime of the population T_1 is very long and effectively infinite ($\Gamma_1 = 0$). We may then write the variation of the population in level $|1\rangle$ as:

$$\dot{N}_1 = R_{12}N_2 - R_{21}N_1 \tag{3.42}$$

where the terms R_{12} and R_{21} are the optical transition rates (in s^{-1}) given by:

$$R_{12} = R_{21} = \frac{D_{12}^2 E_0^2}{2\hbar^2} \frac{1/T_2}{(\omega_{12}-\omega)^2 + (1/T_2)^2} \tag{3.43}$$

We find, to within a Lorentzian, the result obtained by time-dependent perturbation theory (Eq. (1.85)). Exposed to a flux of photons, electrons in a starting level have a transition rate towards a second level given by (3.43). The equality $R_{12} = R_{21}$ is a demonstration of the principle of *microreversibility*. We may rewrite (3.43) more recognizably as:

$$R_{12} = \sigma_{\text{op}}\Phi \tag{3.44}$$

Transition rate and optical cross-section

where Φ is the photon flux (in photons s^{-1} m^{-2}) related to the electric field strength E_0 by:

[1] In the context of this section, absorption cannot be negative. The only way for the oscillator strength (f_{j1}) to be negative is if $|1\rangle$ is not the fundamental level (i.e. $E_f < E_1$), in which case Eq. (3.41) must keep track of the populations, N_f.

$$\Phi = \frac{1}{2}\frac{n_{op} c \varepsilon_0 E_0^2}{\hbar\omega} \tag{3.45}$$

Photon flux and electromagnetic field strength

and σ_{op} is the optical cross-section (expressed generally in cm^2) given by:

$$\sigma_{op} = \frac{D_{12}^2 \omega}{\varepsilon_0 \hbar c n_{op}} \frac{1/T_2}{(\omega_{12} - \omega)^2 + (1/T_2)^2} \tag{3.46}$$

Optical cross-section for a two-level system

3.5.2 Stimulated emission

Equation (3.42) may be rewritten as:

$$\dot{N}_1 = \sigma_{op}\Phi N_2 - \sigma_{op}\Phi N_1 \tag{3.47}$$

We may also write this last equation in terms of the number of photons. The number of photons n in a cavity of volume V, traversed by a flux Φ of photons travelling at the speed of light in that material $c' = c/n_{op}$ is then trivially:

$$n = \frac{\Phi}{c/n_{op}} V = \frac{\Phi}{c'} V \tag{3.48}$$

and (3.42) becomes:

$$\dot{N}_1 = -\dot{N}_2 = \frac{c'\sigma_{op}}{V} n N_2 - \frac{c'\sigma_{op}}{V} n N_1 \tag{3.49}$$

Each term may be easily interpreted. The electrons leave level $|1\rangle$ at a rate proportional to the flux or the photon density, which corresponds well with our understanding of absorption. Alternately, the electrons in level $|2\rangle$ relax to level $|1\rangle$ accompanied by photon emission with a rate proportional to the photon flux – this case corresponds to the phenomenon of *stimulated emission* by photons in a cavity. We may thus conclude that the number of photons in the cavity decreases by 1 with each absorption event, but that each of these photons is eventually re-emitted as a clone during subsequent stimulated emission by a cavity photon. Working from (3.49), we may also find the equation describing the variation in the number of photons in the cavity. For each transition from level $|2\rangle$ to level $|1\rangle$, an additional photon is released into the cavity. Conversely, each electron transition from $|1\rangle$ to $|2\rangle$ leads to the removal of one photon from the cavity. Consequently, $dn/dt = +dN_1/dt = -dN_2/dt$, or:

$$\dot{n} = -\frac{c'\sigma_{op}}{V} n(N_1 - N_2) \tag{3.50}$$

Another expression for the variation in the number of photons in the cavity, useful in quantum optics, is obtained starting from (3.19):

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\rho + \frac{\Delta\rho - \Delta\rho^{\mathrm{eq}}}{T_1} = \mathrm{i}\frac{D_{12}}{\hbar}E_0(\sigma_{21} - \sigma_{12}) \tag{3.51}$$

We note that the term in $1/T_1$ is non-radiative (i.e. it neither creates nor absorbs photons). Consequently, the difference between the variation rates of the two populations $(N_1 - N_2)$ (and which lies at the origin of the disappearance of the photons from the cavity) is due only to differences between σ_{12} and σ_{21} and may be expressed as:

$$\dot{n} = -2\Omega_{12}\mathrm{Im}(\sigma_{21}) \tag{3.52}$$

Absorption and off-diagonal elements of the density matrix

Utilizing expression (3.21b) for σ_{21}, we effectively obtain (3.50) again.

Let us now study the behaviour of a two-level system at resonance. At resonance, the optical cross-section is at a maximum and may be written in the form:

$$\sigma_{\mathrm{op}} = \frac{r_{12}^2}{n_{\mathrm{op}}}(\omega T_2)\left(\frac{q^2}{\hbar c\varepsilon_0}\right) = 9.2 \times 10^{-2}(\omega T_2)\frac{r_{12}^2}{n_{\mathrm{op}}} \tag{3.53}$$

As ωT_2 lies generally in the range 10^{-2}–10^2 and the displacement dipole r_{12} is of the order of Ångströms, we expect typical optical cross-sections to lie between 10^{-20} and $10^{-14}\,\mathrm{cm}^2$.

Example
Calculate the typical optical cross-section at resonance for a quantum well where:

$T_2 = 0.1\,\mathrm{ps}$
$\omega = 10^{14}\,\mathrm{s}^{-1}$
$D_{12} = q \times 1\,\mathrm{C\,nm}$
$n_{\mathrm{op}} = 3$

We obtain a typical value of $\sigma_{\mathrm{op}} = 3 \times 10^{-15}\,\mathrm{cm}^2$.

3.5.3 Absorption saturation

Let us return to expressions (3.26) which yield the population difference between the two levels as a function of the Rabi frequency Ω_{12}, which in turn relates to the electric field strength E_0 by $\Omega_{12} = E_0 D_{12}/\hbar$. Recalling that the number of particles in the system is constant, we have the following system of two equations:

$$N_1 + N_2 = N$$
$$N_1 - N_2 = \frac{1}{1 + \Omega_{12}^2 T_1 T_2}(N_1^{eq} - N_2^{eq}) \tag{3.54}$$

where N is the total number of electrons in the system. We therefore see that the populations in each of the two levels tend to become equal as the photon flux increases ($\Omega_{12} \to \infty$). This may be understood intuitively as follows: as the photon flux increases, the absorption tends to populate the excited level up to the point when the rate of stimulated emission exactly cancels out the absorption rate.

More precisely, if we consider the two levels to be sufficiently separated in energy, so that at thermodynamic equilibrium, all the electrons are in level $|1\rangle$ ($\Delta N^{eq} = N$), then the number of particles in level $|2\rangle$ becomes:

$$N_2 = \frac{N}{2}\left(1 - \frac{1}{1 + \Phi/\Phi_{sat}}\right) \tag{3.55}$$

where Φ_{sat} is the *saturation flux* for the two-level system given by the condition that $\Omega_{sat} = 1/\sqrt{T_1 T_2}$ in (3.54) and by taking into account (3.45):

$$\Phi_{sat} = \frac{\hbar c \varepsilon_0 n_{op}}{2\omega D_{12}^2 T_1 T_2} \tag{3.56a}$$

Saturation flux of a two-level system

Once the photon flux attains this value, the absorption is reduced by a factor of 2, and tends towards zero at sufficiently high power. Note that the inequality, $N_1 > N_2$, holds regardless of the excitation intensity level.

If this was not the case, the immediate consequence of $N_2 > N_1$ would be that the system would exhibit negative absorption and, therefore, optical gain. The condition whereby $N_2 > N_1$ is referred to as *population inversion* and we have just demonstrated that this condition cannot be attained in a two-level system.

Example

Starting from the saturation power level P_{sat}, Eq. (3.56a) allows one to determine the lifetime of the excited state T_1 experimentally once time T_2 has been obtained (from the width of the absorption peak) using:

$$P_{sat} = \frac{c\varepsilon_0 n_{op}}{2(D_{12}/\hbar)^2 T_1 T_2} \tag{3.56b}$$

For quantum wells having dipolar matrix elements r_{12} of 1 nm and absorption line widths of 10 meV (i.e. $T_2 = 0.1$ ps), a saturation power density of 2 MW cm^{-2} is obtained. Figure 3.2 shows the absorption saturation for a two-level quantum well. The value measured for P_{sat} is actually much higher than that predicted here and is a consequence of selection rules which will be dealt with later in Chapter 8.

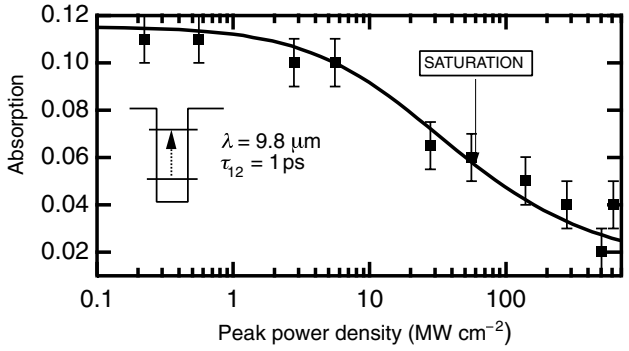


Fig. 3.2. Absorption saturation for a transition between two levels in a quantum well. The saturation threshold refers to the value at which the absorption drops by a factor of 2 corresponding here to an optical power density of 50 MW cm^2 (after J. Y. Duboz et al., *J. Appl. Phys.* **77**, 6492 (1995)).

3.6 Spontaneous emission and radiative lifetime

3.6.1 Spontaneous emission

When the second level is occupied, it tends to relax under the effect of a photon flux, as indicated by the term $-\sigma_{op}c'nN_2/V$ in (3.49). Therefore, if the number of photons in the cavity is zero (i.e. if there is no light in the cavity) the rate of stimulated emission is equally null. As a result, in the absence of inelastic collisions ($\Gamma_1 = 0$), the particle should remain in the excited state indefinitely. This, however, runs counter to experience. For instance, when atoms are excited by an electric discharge in a rarefied atmosphere to keep atomic collisions to a minimum (as is intentionally the case in a neon lamp) the excited atoms relax in a few microseconds. How then can we reconcile this paradox between the impeccable theory presented so far and this most mundane observation from the realm of day-to-day experience?

The answer is that we have only done half the task. While we have allowed for quantization of the behaviour of the particle, we have not done so for the light. To complete our work, we must introduce in our expression for the interaction Hamiltonian (3.12) or (3.13) the quantized form of the electric field corresponding to the lth mode given by equation (2.34a):

$$\hat{W} = \mathrm{i}q\mathscr{F}_l(a_l \mathrm{e}^{\mathrm{i}\mathbf{k}_l\cdot\mathbf{r}} - a_l^\dagger \mathrm{e}^{-\mathrm{i}\mathbf{k}_l\cdot\mathbf{r}})\boldsymbol{\varepsilon}_l \cdot \hat{\mathbf{r}} \tag{3.57}$$

where a_l and $a_l^\dagger$ are the creation and annihilation operators for the photons in the

lth mode and $\mathscr{F}_l$ is the field strength associated with the quantum vacuum fluctuations for the same mode (Eq. (2.35)). The Hamiltonian $\hat{W}$ now acts in the two Hilbert spaces (or more exactly on their *tensor product*) corresponding to that of the photons and the charged particles, respectively. We should therefore speak of states labelled as $|i, n_l\rangle$, where i refers to the state of the charged particle ($i = 1, 2$ for a two-level system) and n_l designates the number of photons in mode l (see Fig. 3.3). The generalization of this problem to an arbitrary number of modes is straightforward, but heavy in terms of the required notation (e.g. $|i, n_1, \ldots, n_l, \ldots\rangle$). We will therefore set aside the mode indices for the time-being.

Optical transitions may occur between the states $|1, n\rangle$ and $|2, n'\rangle$ and vice versa. We will interest ourselves for the present in the absorption mechanism. The initial level is then $|1, n\rangle$ and the final level is $|2, n-1\rangle$ (see Fig. 3.3) as a photon is lost during the absorption process. From (1.85b), the absorption probability is:

$$P_{1,2} = \frac{2\pi}{\hbar} |\langle 1, n|W|2, n-1\rangle|^2 \delta(E_{\text{final}} = E_{\text{initial}}) \tag{3.58}$$

where $E_{\text{initial}} = E_1 + \hbar\omega(n + 1/2)$ and $E_{\text{final}} = E_2 + \hbar\omega((n-1) + 1/2)$. The energy conservation condition therefore corresponds to $\hbar\omega = E_2 - E_1$, which is in accord with the theory of Chapter 1 (Eq. (1.80)). Equation (3.58) may then be written:

$$P_{1,2} = \frac{2\pi}{\hbar} q^2 \mathscr{F}_l^2 |\langle 1, n|(a_l e^{i\mathbf{k}_l \cdot \mathbf{r}} - a_l^+ e^{-i\mathbf{k}_l \cdot \mathbf{r}}) \boldsymbol{\varepsilon}_l \cdot \hat{\mathbf{r}}|2, n-1\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \tag{3.59}$$

As the state of the particle and that of the photon are independent of one another, this expression may be separated as:

$$P_{1,2} = \frac{2\pi}{\hbar} q^2 \mathscr{F}_l^2 |\langle n-1|(a_l e^{i\mathbf{k}_l \cdot \mathbf{r}} - a_l^+ e^{-i\mathbf{k}_l \cdot \mathbf{r}})|n\rangle|^2 |\langle 2|\boldsymbol{\varepsilon}_l \cdot \hat{\mathbf{r}}|1\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \tag{3.60}$$

Following the properties of the creation and annihilation operators ((1.D.23) and (1.D.24)), the only non-null photon related term is:

$$\langle n-1|a_l|n\rangle = \sqrt{n} \tag{3.61}$$

and the absorption probability per unit time becomes:

$$P_{1,2} = \frac{\pi}{\hbar} \frac{\hbar\omega_l n}{\varepsilon_0 L^3} |\langle 1|q\boldsymbol{\varepsilon}_l \cdot \mathbf{r}|2\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \tag{3.62}$$

or again, using expression (2.77) for the electric field of a coherent state:

$$P_{\text{abs}} = \frac{\pi}{2\hbar} E_0^2 |\langle 1|q\boldsymbol{\varepsilon}_l \cdot \hat{\mathbf{r}}|2\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \tag{3.63}$$

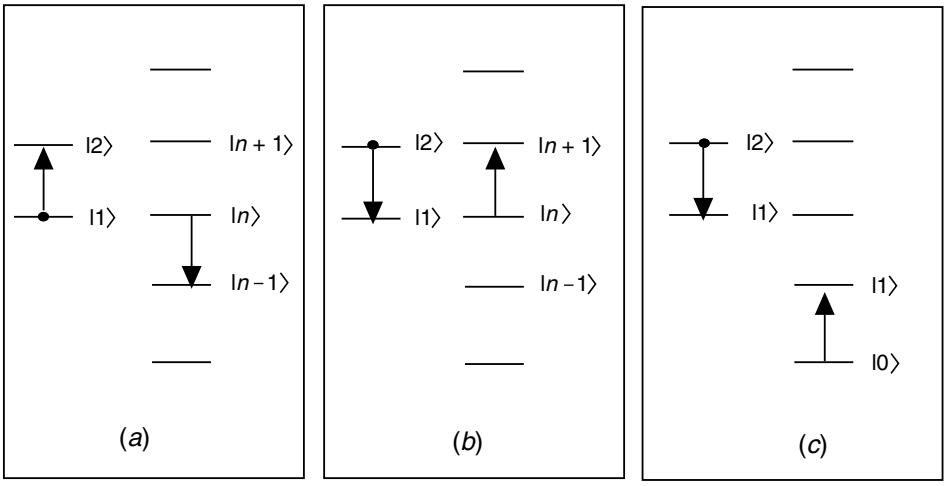


Fig. 3.3. Interaction between a two-level system and a quantized electromagnetic field in the form of a harmonic oscillator. The figure shows different mechanisms for the following optical transitions: (*a*) absorption, (*b*) stimulated or induced emission, and (*c*) spontaneous emission.

We thereby exactly recover expression (1.85b) for a classical electric field (comforting and suggestive of the overall coherence of these theories).

We now consider the cavity containing a population of n photons in the lth mode, but this time with the particle lying initially in the $|2\rangle$ state. Emission then corresponds to a transition from the $|2, n\rangle$ state, to the $|1, n+1\rangle$ state (Fig. 3.3). The probability for this transition is:

$$P_{1,2} = \frac{2\pi}{\hbar} q^2 \mathscr{F}_l^2 |\langle n+1|(a_l e^{i\mathbf{k}_l\cdot\mathbf{r}} - a_l^+ e^{-i\mathbf{k}_l\cdot\mathbf{r}})|n\rangle|^2 |\langle 1|\boldsymbol{\varepsilon}_l\cdot\hat{\mathbf{r}}|2\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \quad (3.64)$$

The only non-zero photon related term is this time:

$$\langle n+1|a^+|n\rangle = \sqrt{n+1} \quad (3.65)$$

and the emission probability per unit time then becomes:

$$P_{\text{em}} = \frac{\pi\omega_l}{\varepsilon_0 L^3}(n+1)|\langle 1|q\boldsymbol{\varepsilon}_l\cdot\hat{\mathbf{r}}|2\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \quad (3.66)$$

This last expression shows that two distinct physical effects are involved in emission:

1. An emission mechanism whereby the transition rate P_{stim} is proportional to the number of photons already present in the cavity – this is the phenomenon of *stimulated emission* described earlier.
2. A new emission mechanism, which is present even when the cavity is devoid of photons – this mechanism is referred to as *spontaneous emission*. The sponta-

neous emission rate P^l_{spon} in the lth mode is given by:

$$P^l_{\text{spon}} = \frac{\pi\omega_l}{\varepsilon_0 L^3}|\langle 1|q\boldsymbol{\varepsilon}_l \cdot \mathbf{r}|2\rangle|^2 \delta(\hbar\omega = E_2 - E_1) \tag{3.67}$$

We may also interpret this term as corresponding to the emission rate stimulated by the *vacuum fluctuations* $\mathscr{F}_l$. Under the influence of the vacuum fluctuations, the particle may relax by releasing a photon in the lth mode. Clearly, the particle may emit this photon into any one of the possible cavity modes. The total spontaneous emission rate is obtained by summing the rates (3.67) for each of the different modes in the cavity with frequency ω_{12}:

$$\Gamma_{\text{spon}} = \iiint P^l_{\text{spon}} \mathrm{d}^3\mathbf{k}_l \tag{3.68}$$

The calculation of the integral makes use of expression (2.83) for the density of electromagnetic cavity modes and presents no particular challenge, as we may draw upon the integral over all space evaluated in (2.A.25). In this manner, we obtain the spontaneous emission rate over all spatial directions for a quantum transition with energy $\hbar\omega_0$ in an optical medium with index of refraction n_{op}:

$$\Gamma_{\text{spon}} = \frac{q^2 r_{12}^2 \omega^3 n_{\text{op}}}{3\pi c^3 \hbar \varepsilon_0} = \frac{1}{t_{\text{spon}}} \tag{3.69a}$$

Spontaneous emission rate and radiative lifetime for a two-level system

where

$$r_{12}^2 = |x_{12}|^2 + |y_{12}|^2 + |z_{12}|^2 \tag{3.69b}$$

We recall, however, that this expression only holds if the photons are emitted isotropically.

Several points are worth noting:

1. If a particle is in its fundamental level $|1\rangle$, the spontaneous transition rate is null, i.e. the lifetime of an electron in its fundamental level is infinite. We thereby resolve one of the greatest worries of theoreticians at the beginning of the century – the reason as to why all the electrons in all the atomic orbitals in the universe do not simply collapse onto their respective nuclei in a matter of nanoseconds by emitting photons.
2. The ω^3 behaviour of the spontaneous emission rate is as predicted in the classical theory of Rayleigh. Assuming a constant dipolar element, the spontaneous lifetime becomes shorter as the transition energy increases. This is one of the reasons why X-ray lasers are intrinsically more difficult to fabricate than infrared lasers. Of course, the dipolar matrix element will also depend on the transition energy as shown in the example below.

3. The absence of Planck's constant $\hbar$ in Eq. (3.67) shows why the classical approach employed in Complement 2.A could be successfully applied in certain situations.

Example: radiative lifetime in an infinite quantum well

We consider a GaAs quantum well of width a, with infinite barriers in which:

$$m^* = 0.067 m_0$$
$$n_{op} = 3.3$$

We saw in Chapter 1 that the energy level separation $\hbar\omega_{21}$, and the dipolar element z_{12} are given by:

$$\hbar\omega_{21} = E_2 - E_1 = \frac{3}{2}\frac{\hbar^2\pi^2}{m^* a^2}$$
$$z_{12} = \frac{2^4}{3^2}\frac{a}{\pi^2} \tag{3.70}$$

We thereby obtain an expression for the radiative lifetime as a function of the transition energy:

$$t_{spon} = \frac{3^4\pi^3}{2^7}\frac{m^* c^3 \hbar^2 \varepsilon_0}{q^2 n_{op}}\frac{1}{E^2} \tag{3.71a}$$

The variation of the lifetime as a function of transition energy in the simple model is presented in Fig. 3.4. For typical transition energies (lying between 50 and 250 meV), the corresponding lifetimes lie between 10 ns and 1 μs. We see that the radiative lifetime in this case is proportional to ω^{-2} and not ω^{-3} (as indicated above) as a result of the dependence of the dipolar matrix element r_{12} on the transition energy.

The above example illustrates the possibility of a classical approach to describing the spontaneous emission. For this, we introduce the ratio between the classical radiation lifetime t_{clas} (given by (2.A.30)) and the quantum lifetime t_{spon}:

$$\frac{t_{clas}}{t_{spon}} = \frac{2mr_{12}^2\omega}{\hbar} \tag{3.71b}$$

Now, all the quantum structures (ions, atoms, semiconductor quantum wells, . . .) may be crudely described by a model infinite quantum well of width a with transition energies $\hbar\omega$ and dipole r_{12} given by:

$$\hbar\omega = E_2 - E_1 = N\frac{\hbar^2\pi^2}{2m_0 a^2}$$

and

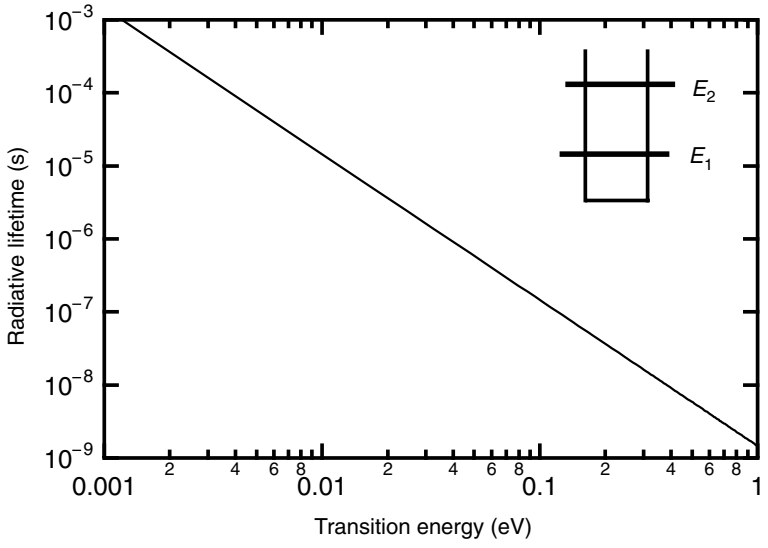


Fig. 3.4. Radiative lifetime as a function of transition energy between two levels in a GaAs quantum well with infinite barriers.

$$r_{12} = \alpha a$$

where N is an integer number in the 1–10 range and α a real number in the 10^{-1} range. The ratio $t_{\text{clas}}/t_{\text{spon}}$ may thus be written as:

$$\frac{t_{\text{clas}}}{t_{\text{spon}}} = \alpha^2 N \pi^2 \approx 1$$

The classical expression for radiative lifetime is therefore reasonable, but fails to describe effects such as the stability of the ground state of quantum structures, symmetry forbidden transitions, etc.

3.6.2 The rate equations including spontaneous emission

We shall now include spontaneous emission into the rate equations in (3.49) and (3.50). To simplify the required notation temporarily, we will suppose until otherwise indicated that only one photon mode exists in the cavity. The spontaneous emission and the level lifetimes T_1 are taken care of by introducing the terms in $\pm N_2/t_{\text{spon}}$, $\pm N_1/T_1$, and $\pm N_2/T_1$ into the rate equations for the electrons in (3.49). Regarding (3.50), only a fraction β of the spontaneously emitted photons will be released into the cavity mode. Correspondingly, we add a term in $\beta N_2/t_{\text{spon}}$ into the rate equation of the photons in (3.50), yielding:

$$\dot{N}_1 = \frac{c'\sigma_{\text{op}}}{V} n N_2 - \frac{c'\sigma_{\text{op}}}{V} n N_1 - \frac{(N_1 - N_1^{\text{eq}})}{T_1}$$

$$\dot{N}_2 = \frac{c'\sigma_{\text{op}}}{V} n N_1 - \frac{c'\sigma_{\text{op}}}{V} n N_2 - \frac{(N_2 - N_2^{\text{eq}})}{\tau} \tag{3.72a}$$

$$\dot{n} = -\frac{c'\sigma_{\text{op}}}{V} n(N_1 - N_2) + \beta \frac{N_2}{t_{\text{spon}}}$$

Rate equations for the cavity

where τ is the lifetime in $|2\rangle$ and takes into account spontaneous emission:

$$\frac{1}{\tau} = \frac{1}{T_1} + \frac{1}{t_{\text{spon}}} \tag{3.72b}$$

β is referred to as the *spontaneous emission factor*. The introduction of the terms N^{eq} in (3.72a) guarantees that the system returns to thermodynamic equilibrium once the optical excitation ceases.

We have now only to relate the expression for the radiative lifetime to the optical cross-section and absorption, as this will be of use to us in several upcoming occasions. Supposing the spontaneous emission is isotropic ($x_{12} = y_{12} = z_{12}$), (3.46) and (3.69) lead to the following well known equations:

$$\sigma_{\text{op}} = \frac{\lambda^2}{4n_{\text{op}}^2 t_{\text{spon}}} \mathscr{L}(\omega - \omega_{21}) = \frac{\lambda^2}{8\pi n_{\text{op}}^2 t_{\text{spon}}} \mathscr{L}(\nu - \nu_{21}) \tag{3.73}$$

Optical cross-section and radiative lifetime

and

$$\alpha(\nu) = \frac{\lambda^2}{8\pi n_{\text{op}}^2 t_{\text{spon}}} \mathscr{L}(\nu - \nu_{21})(N_1 - N_2) \tag{3.74}$$

Absorption coefficient and radiative lifetime

where λ is the vacuum wavelength and we recall that $\mathscr{L}$ is the Lorentzian function.

3.7 Polychromatic transitions and Einstein's equations

Equation (3.63) gives the absorption probability for an electromagnetic wave in a two-level system. This equation may be as readily obtained within the framework of time-dependent perturbation theory, as well as that of electromagnetic field quantization. Replacing the Dirac function by a Lorentzian brings us back to (3.37), obtained using density matrix formalism. Another use for (3.63) is when the transition results from an arbitrary distribution of electromagnetic energy $\rho_e(\nu)$. This situation is of particular concern in solids where the optical transitions occur between energy bands.

We therefore consider a volume V, in which a *spectral energy distribution* $\rho_e(\nu)$ is

defined (the width of the distribution being assumed to be larger than that resulting from isolated atomic transitions). Between ν and $\nu + \mathrm{d}\nu$ there are $\rho_e(\nu)V\mathrm{d}\nu/h\nu$ photons (where ρ_e is in units of energy per unit frequency per unit volume). Each photon has a probability $\sigma_{\mathrm{op}}c/Vn_{\mathrm{op}}$ (see (3.49)) of inducing a transition (i.e. either by absorption or by stimulated emission). The total absorption and stimulated emission rates are then:

$$P_{\mathrm{abs}} = P_{\mathrm{stim}} = \int_0^\infty \frac{\rho_e(\nu)V}{h\nu}\frac{c}{n_{\mathrm{op}}V}\sigma_{\mathrm{op}}(\nu)\mathrm{d}\nu \tag{3.75}$$

Using (3.73), this equation may be immediately integrated to find:

$$P_{\mathrm{abs}} = P_{\mathrm{stim}} = B_{12}\rho_e(\nu_{12}) \tag{3.76}$$

where the Einstein coefficient B_{12} is given by:

$$B_{12} = \frac{\lambda^3}{8\pi h n_{\mathrm{op}}^3 t_{\mathrm{spon}}} \tag{3.77a}$$

Einstein coefficient for absorption and stimulated emission

The other coefficient introduced by Einstein is the spontaneous emission rate:

$$A_{12} = \frac{1}{t_{\mathrm{spon}}} \tag{3.77b}$$

Einstein coefficient for spontaneous emission

the rate equations (3.49) may then be generalized to deal with polychromatic transitions between bands of width $\mathrm{d}\nu$:

$$\mathrm{d}\dot{N}_1 = -\mathrm{d}\dot{N}_2 = A_{12}\mathrm{d}N_2 + (\mathrm{d}N_2 - \mathrm{d}N_1)B_{12}\rho_e(\nu) \tag{3.78}$$

Einstein rate equation

where $\mathrm{d}N_1$ and $\mathrm{d}N_2$ are now the differential quantities $\mathrm{d}N_1 = \rho_1(\nu)\mathrm{d}\nu$ and $\mathrm{d}N_2 = \rho_2(\nu)\mathrm{d}\nu$, and where the ρ_is correspond to the population densities per unit transition frequency (in s cm^{-3}). Einstein did not have at his disposal the concept of a quantized electromagnetic field from which to establish Eq. (3.78). He rather proceeded heuristically starting from the rate equations (3.49), which allowed him to obtain Planck's distribution for $\rho_e(\nu)$ at thermal equilibrium. This line of reasoning is investigated in more detail in Complement 3.C.

3.8 Rate equations revisited

As may be easily concluded from our work so far, the physics of electron–photon

interaction is as rich in concepts as in definitions. It is therefore worthwhile in this section to summarize some of the principal concepts we have introduced while giving some indication of their inter-relatedness. All these concepts emerge from an amalgam of the density matrix approach, time-dependent perturbation theory, and quantization of the electromagnetic field.

We consider a cavity of volume V filled by an ensemble of two-level ($|1\rangle$ and $|2\rangle$) systems with a density of N per cm^3, having a transition energy of $\hbar\omega_{12}$, a dipole moment D_{12}, a non-radiative lifetime T_1, and a linewidth of $1/\pi T_2$. The interaction of this system with electromagnetic waves is described below.

3.8.1 Monochromatic single-mode waves

We will suppose that the electromagnetic wave is monochromatic (i.e. possessing only one frequency ω). The rate of change of the number of particles in levels $|1\rangle$ and $|2\rangle$ is given by:

$$\dot{N}_1 = \sigma_{op}\Phi N_2 - \sigma_{op}\Phi N_1 - \frac{(N_1 - N_1^{eq})}{T_1}$$
$$\dot{N}_2 = -\sigma_{op}\Phi N_2 + \sigma_{op}\Phi N_1 - \frac{(N_2 - N_2^{eq})}{\tau} \tag{3.79}$$

where Φ is the photon flux in the lth mode (in cm^{-2} s^{-1}). The lifetime τ results from inelastic non-radiative recombination ($1/T_1$) and radiative recombination ($1/t_{spon}$):

$$\frac{1}{\tau} = \frac{1}{T_1} + \frac{1}{t_{spon}} \tag{3.80}$$

These different transition mechanisms are represented schematically in Fig. 3.5.

The optical cross-section σ_{op} is given by:

$$\sigma_{op} = \frac{D_{12}^2\omega}{\varepsilon_0\hbar c n_{op}} \frac{1/T_2}{(\omega_{21} - \omega)^2 + (1/T_2)^2} \tag{3.81a}$$

or again as:

$$\sigma_{op} = \frac{\lambda^2}{4n_{op}^2 t_{spon}} \mathscr{L}(\omega - \omega_{21}) = \frac{\lambda^2}{8\pi n_{op}^2 t_{spon}} \mathscr{L}(\nu - \nu_{21}) \tag{3.81b}$$

where t_{spon} is the radiative lifetime. This lifetime is the probability per unit time of spontaneously emitting a photon *in any mode* and is given by:

$$\Gamma_{spon} = \frac{q^2 r_{12}^2 \omega^3 n_{op}}{3\pi c^3 \hbar \varepsilon_0} = \frac{1}{t_{spon}} \tag{3.82}$$

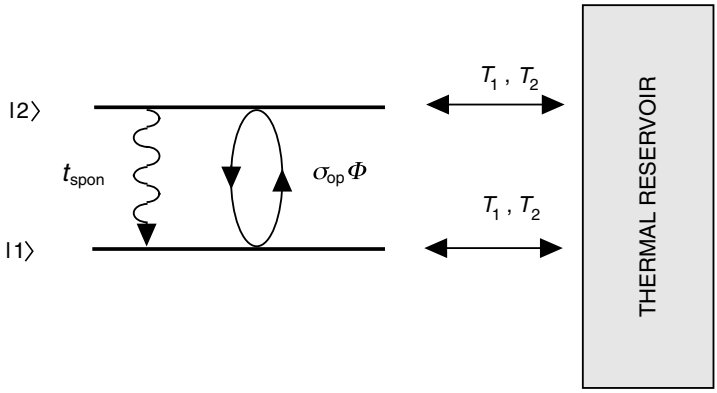


Fig. 3.5. Various transition rates defined in a two-level system interacting with an optical field. The relaxation mechanisms are shown figuratively as interacting with a thermal reservoir.

This gives rise to an absorption (cm^{-1}):

$$\alpha(\omega) = \sigma_{op}(\omega)(N_1 - N_2) \tag{3.83}$$

The rate equations may also be written in the form:

$$\begin{aligned}
\dot{N}_1 &= \frac{c'\sigma_{op}}{V} nN_2 - \frac{c'\sigma_{op}}{V} nN_1 - \frac{(N_1 - N_1^{eq})}{T_1} \\
\dot{N}_2 &= \frac{c'\sigma_{op}}{V} nN_1 - \frac{c'\sigma_{op}}{V} nN_2 - \frac{(N_2 - N_2^{eq})}{\tau} \\
\dot{n} &= -\frac{c'\sigma_{op}}{V} n(N_1 - N_2) + \beta \frac{N_2}{t_{spon}}
\end{aligned} \tag{3.84a}$$

where n is the number of photons in the lth cavity mode and c' is the speed of light in the medium ($c' = c/n_{op}$). β is the portion of spontaneously emitted photons released into the lth mode. These coupled differential equations in N_1, N_2, and n, are highly non-linear and lead to the multifaceted behaviour of laser emission examined in the next chapter.

3.8.2 Multimode monochromatic waves

This case is a simple generalization of (3.84a) to all the l allowed modes in a cavity:

$$\begin{aligned}
\dot{N}_1 &= \sum_l \left(\frac{c'\sigma_{op}^l}{V} n_l N_2 - \frac{c'\sigma_{op}^l}{V} n_l N_1 \right) - \frac{(N_1 - N_1^{eq})}{T_1} \\
\dot{N}_2 &= \sum_l \left(\frac{c'\sigma_{op}^l}{V} n_l N_1 - \frac{c'\sigma_{op}^l}{V} n_l N_2 \right) - \frac{(N_2 - N_2^{eq})}{\tau}
\end{aligned} \tag{3.84b}$$

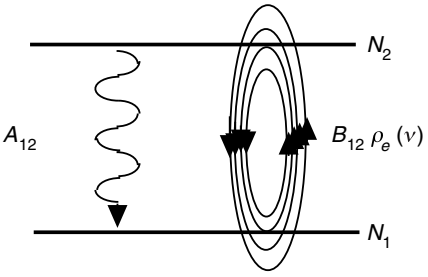


Fig. 3.6. Optical transitions implicated in Einstein's rate equation.

$$\dot{n}_l = -\frac{c'\sigma_{\text{op}}^l}{V} n_l(N_1 - N_2) + \beta_l \frac{N_2}{t_{\text{spon}}}$$

where we have indicated that naturally the cross-sections and spontaneous emission factors depend on mode l.

3.8.3 Polychromatic waves

If the electromagnetic wave has a spectral energy density which is much wider than the spectral width of the interlevel transition, then the probability density for stimulated emission or absorption is:

$$P_{\text{abs}} = P_{\text{stim}} = B_{12}\rho_e(\nu_{12}) \tag{3.85}$$

where B_{12} is the Einstein coefficient:

$$B_{12} = \frac{\lambda^3}{8\pi h n_{\text{op}}^3 t_{\text{spon}}} \tag{3.86}$$

The rate equation may then be written for each band $\mathrm{d}\nu$ of frequencies:

$$\mathrm{d}\dot{N}_1 = -\mathrm{d}\dot{N}_2 = A_{12}\mathrm{d}N_2 + (\mathrm{d}N_2 - \mathrm{d}N_1)B_{12}\rho_e(\nu) \tag{3.87}$$

(see Fig. 3.6) where $\mathrm{d}N_1$ and $\mathrm{d}N_2$ are the infinitesimal populations involved in the optical transitions by the modes $\rho_e(\nu)\mathrm{d}\nu$.

FURTHER READING

A. Aspect, C. Fabre, and G. Grynberg, *Introduction aux lasers et à l'optique quantique*, Ellipses, Paris (1997).

C. Cohen-Tannoudji, J. Dupont-Roc, and G. Grynberg, *Atom–Photon Interactions*: *Basic Processes and Applications*, Wiley, New York (1998).

R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Mass (1964).

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).

W. H. Louisell, *Quantum Statistical Properties of Radiation*, Wiley, New York (1973).

P. Meystre and M. Sargent III, *Elements of Quantum Optics*, Springer-Verlag, Berlin (1989).

Complement to Chapter 3

3.A Homogeneous and inhomogeneous broadening: coherence of light

We saw in several places that two approaches may be used in describing the optical transitions in a two-level quantum system. One approach is founded upon time-dependent perturbation theory (Section 1.6), while the other utilizes the density matrix to keep track of the finite lifetimes (Eq. (3.43)). We will fill the gap between the two descriptions by widening the Dirac function used in the first approach into a Lorentzian function $\mathscr{L}(v)$:

Time-dependent perturbation:

$$P_{12}(\omega) = \frac{\pi}{2\hbar^2}|W_{12}|^2\delta(\omega - \omega_{21}) \tag{3.A.1}$$

Density matrix:

$$P_{12}(\omega) = \frac{\pi}{2\hbar^2}|W_{12}|^2\mathscr{L}(\omega - \omega_{21})$$

Within the framework of the $\mathbf{D}\cdot\mathbf{E}$ perturbation Hamiltonian, $|W_{12}| = qE_0r_{12}$. In fact, to rewrite (3.A.1) in a more general fashion:

$$P_{12}(\omega) = \frac{\pi}{2\hbar^2}|W_{12}|^2 g(\omega - \omega_{21}) \tag{3.A.2a}$$

where g is a lineshape function describing the broadening of the transition and which must satisfy the normalization condition:

$$\int_{-\infty}^{+\infty} g(\omega)\mathrm{d}\omega = 1 \tag{3.A.2b}$$

This broadening may have several origins, which we will explore below. These mechanisms may be grouped into two very distinct categories – *homogeneous* and

inhomogeneous broadening. In both cases, we must no longer consider a single two-level system, but rather an ensemble of such systems.

3.A.1 Homogeneous broadening

In this case, each of the quantum systems forming the medium are identical. They each share the same Bohr oscillator frequency and are subject to the same broadening mechanisms. This broadening may have two origins.

1a Broadening due to finite lifetime (inelastic collisions)
A particularly elegant method to take into account the finite lifetime of the population in each of the levels $|1\rangle$ and $|2\rangle$ is to consider each of the eigenenergies E_n $(n = 1, 2)$ to possess imaginary parts:

$$E_n = \hbar\omega_n + \frac{\mathrm{i}\hbar}{2\tau_n} \tag{3.A.3a}$$

Then, the probability of finding a particle in level $|n\rangle$ becomes:

$$|\mathrm{e}^{-\mathrm{i}E_n t/\hbar}|^2 = \mathrm{e}^{-t/\tau_n} \tag{3.A.3b}$$

The form of the transition lineshape is proportional to the amplitude of the Fourier transform of the wavefunction describing the state of the particle in $|n\rangle$. The form of the lineshape in this case is easily found to be:

$$g(\nu) = \frac{\Delta\nu/2\pi}{(\nu - \nu_{21})^2 + (\Delta\nu/2)^2} \tag{3.A.4}$$

where the full width at half maximum is:

$$\Delta\nu = \frac{1}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) \tag{3.A.5}$$

1b Broadening due to elastic collisions: temporal coherence
In this case, we consider each quantum system within the ensemble to be subjected to random jolts, which cause it to lose its phase. Therefore the electromagnetic field due to the system m is:

$$E_m(t) = E\cos[\omega_{21}t + \phi_m(t)] \tag{3.A.6}$$

where $\phi_m(t)$ is the product of stochastic processes. Figure 3.A.1 represents such a process. The resulting electric field from the contributions of each of these identical systems having no particular phase relation between them is then:

$$E(t) = \sum_m E_m(t) = E\sum_m \cos[\omega_{21}t + \phi_m(t)] \tag{3.A.7}$$

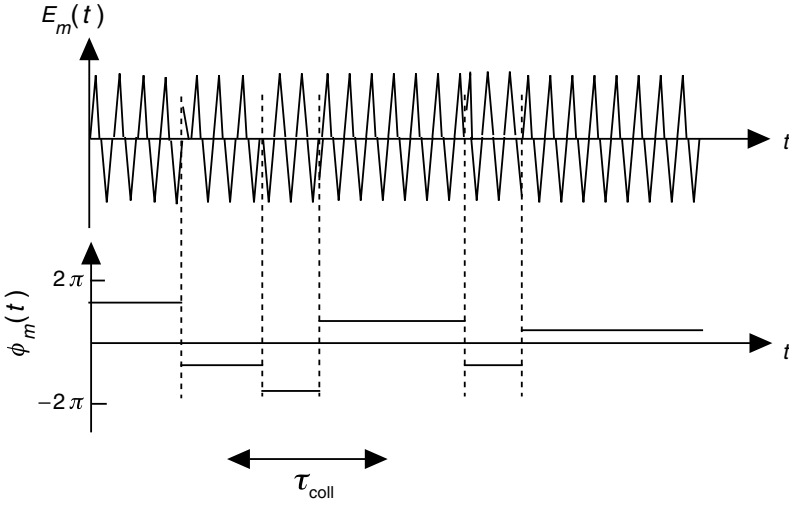


Fig. 3.A.1. Amplitude and phase of the electric field emitted by a quantum system m subjected to random 'Poisson' jolts. The mean time between jolts (i.e. the collision or dephasing time) is τ_{coll}.

From the theory of *stochastic processes*, we know that the frequency spectrum obtained for such a process by evaluating the Fourier transform of the electric field:

$$E(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} E(t) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \tag{3.A.8}$$

is of no significance. This notion must be replaced by the concept of the *autocorrelation spectrum*:

$$\begin{aligned} G(\omega) &= \frac{1}{2\pi} \int\int_{-\infty}^{+\infty} E^*(t) E(t') \mathrm{e}^{\mathrm{i}\omega(t'-t)} \mathrm{d}t \, \mathrm{d}t' \\ &= \frac{1}{2\pi} \int\int_{-\infty}^{+\infty} E^*(t) E(t+\tau) \mathrm{e}^{\mathrm{i}\omega\tau} \mathrm{d}t \, \mathrm{d}\tau \end{aligned} \tag{3.A.9}$$

where we have allowed for the possibility of the electric field having a complex component. We therefore introduce the first-order autocorrelation function of the electric field $G_1(\tau)$ given by:

$$G_1(\tau) = \frac{1}{T}\int_{-T/2}^{T/2} E^*(t)E(t+\tau)\mathrm{d}t = \overline{E^*(t)E(t+\tau)} \tag{3.A.10a}$$

First-order correlation function

This formula only applies for sufficiently large values of T which enable the statistical sum to converge. The existence of this limit (which we will not demonstrate) is one of the major issues of the study of stochastic processes and is referred to as the *ergodic principle*. This principle states that there is an equivalence between temporal and statistical averaging. The 'bar' sign then signifies the statistical average of a random process.

The autocorrelation spectrum $G(\omega)$ and the autocorrelation function $G_1(\tau)$ are related by:

$$G(\omega) = \frac{T}{2\pi}\int_{-\infty}^{+\infty} G_1(\tau)\mathrm{e}^{\mathrm{i}\omega\tau}\mathrm{d}\tau \tag{3.A.10b}$$

The autocorrelation spectrum therefore has a form close to that of the Fourier transform of the autocorrelation function. To allow for the normalization of (3.A.10b) we note that for $\tau = 0$, Eq. (3.A.10a) gives:

$$G_1(0) = \frac{1}{T}\int_{-T/2}^{T/2} E^*(t)E(t)\mathrm{d}t = \overline{|E(t)|^2} \tag{3.A.11}$$

which is the average intensity. Additionally, by integrating (3.A.10b), we find:

$$\begin{aligned}\frac{1}{T}\int_{-\infty}^{+\infty} G(\omega)\mathrm{d}\omega &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\mathrm{d}\tau G_1(\tau)\left[\int_{-\omega}^{+\omega}\mathrm{e}^{\mathrm{i}\omega\tau}\mathrm{d}\omega\right] \\ &= \int_{-\infty}^{+\infty}\mathrm{d}\tau G_1(\tau)\delta(\tau) = G_1(0)\end{aligned} \tag{3.A.12}$$

We introduce the lineshape $g(\omega)$ as the normalized version of (3.A.10b) (i.e. with no explicit reference to averaging time T):

$$g(\omega) = \frac{G(\omega)}{\int_{-\infty}^{+\infty} G(\omega)\mathrm{d}\omega}$$

which may be written, given (3.A.12) and (3.A.11):

$$g(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} g_1(\tau) \mathrm{e}^{\mathrm{i}\omega\tau} \mathrm{d}\tau \tag{3.A.13a}$$

Lineshape and first-order coherence

where $g_1(\tau)$ is the first-order temporal coherence of light:

$$g_1(\tau) = \frac{\int_{-T/2}^{+T/2} E^*(t)E(t+\tau)\mathrm{d}t}{\int_{-T/2}^{+T/2} E^*(t)E(t)\mathrm{d}t} = \frac{G_1(\tau)}{G_1(0)} \tag{3.A.13b}$$

Equation (3.A.13a) is a major result of the statistical theory of light. The lineshape, or spectral distribution of light, is the Fourier transform of the first-order coherence of the electric field.[2] This is known as the *Wiener–Kintchine* theorem.

We still have to specify the process describing the jolts of the particles having an average time between impulses of τ_{coll}. This most common process is a Poisson process with probability $p(\tau)\mathrm{d}\tau$ that an atom does not experience a collision during a time τ given by:

$$p(\tau)\mathrm{d}\tau = \frac{1}{\tau_{\mathrm{coll}}} \mathrm{e}^{-\tau/\tau_{\mathrm{coll}}} \mathrm{d}\tau \tag{3.A.14}$$

The autocorrelation function of the field emitted by m is:

$$G_1(\tau) = E_0^2 \mathrm{e}^{-\mathrm{i}\omega_{21}\tau} \overline{\mathrm{e}^{\mathrm{i}(\phi_m(t+\tau) - \phi_m(t))}} \tag{3.A.15}$$

In this last equation, the only non-zero contribution results from atoms which have a period between collisions longer than τ for which $\phi_m(t+\tau) = \phi_m(t)$, the other terms have a null average when evaluated in the summation over m, so that:

$$G_1(\tau) = E_0^2 \mathrm{e}^{-\mathrm{i}\omega_{21}\tau} \frac{1}{\tau_{\mathrm{coll}}} \int_{\tau}^{+\infty} \mathrm{e}^{-t/\tau_{\mathrm{coll}}} \mathrm{d}t = E_0^2 \mathrm{e}^{-\mathrm{i}\omega_{21}\tau - \tau/\tau_{\mathrm{coll}}} \tag{3.A.16}$$

The lineshape function, shown in Fig. 3.A.2, is then given by the Fourier transform (3.A.13a) of the above field's autocorrelation function:

[2] There are about as many different ways of normalizing the Fourier transform as there are authors. Some use $1/\sqrt{2\pi}$, while others use $1/2\pi$ for the transformation from ω to t. Others use these same factors when transforming from t to ω!

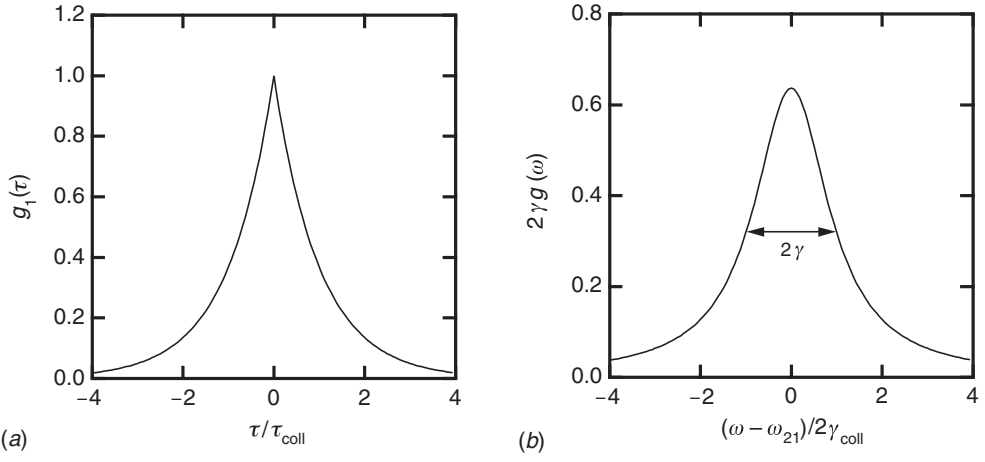


Fig. 3.A.2. (*a*) Autocorrelation function $g_1(\tau)$ and (*b*) spectral distribution $g(\omega)$ of the Poisson process shown in Fig. 3.A.1.

$$g(\omega) = \frac{1/\pi\tau_{coll}}{(\omega - \omega_{21})^2 + (1/\tau_{coll})^2} \tag{3.A.17}$$

Returning to the expression obtained using the density matrix, we see the coherence between density matrix formalism and the statistical theory of light (compare Eqs. (3.43) and (3.A.17)).

Taking into account finite lifetime broadening (3.A.5) and elastic collisions (3.A.17), homogeneous broadening Δv is then given by:

$$\Delta v = \frac{1}{2\pi}\left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + \frac{2}{\tau_{coll}}\right) \tag{3.A.18}$$

3.A.2 Inhomogeneous broadening

In this case, the medium results from an inhomogeneous ensemble of m distinct systems possessing different Bohr oscillation frequencies, ω_{ij}^m. The overall lineshape is then an average of all the individual lineshapes (see Fig. 3.A.3):

$$\bar{g}(\omega) = \int_{\omega_m} g_m(\omega)\rho(\omega_m)\mathrm{d}\omega_m \tag{3.A.19}$$

where $\rho(\omega_m)\mathrm{d}\omega_m$ is the number of states in the system having a Bohr frequency within the range $\omega_m - \mathrm{d}\omega_m/2$ to $\omega_m + \mathrm{d}\omega_m/2$. In atomic systems, there are many possible causes for inhomogeneous line broadening, such as the Doppler effect. In optoelectronic materials, inhomogeneous broadening typically results from fluctuations in the growth parameters of a given sample.

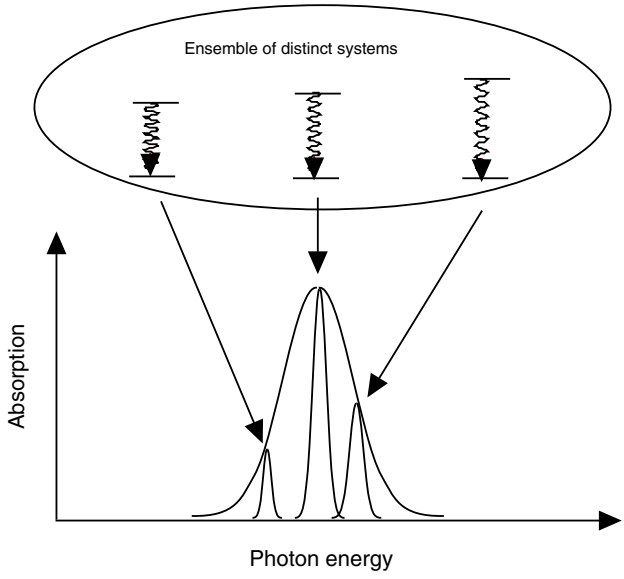


Fig. 3.A.3. Inhomogeneous line broadening due to the cumulative contributions of several distinct and independent systems. Each of these systems is responsible for optical transitions at different characteristic photon energies.

Example: inhomogeneous broadening of an infinite quantum well

We consider an infinite quantum well where the width is taken to vary across the sample. Supposing the occurrence of these fluctuations to be described by a Gaussian distribution:

$$\rho(a) = \frac{1}{\sqrt{2\pi\delta a^2}} \mathrm{e}^{-((a-a_0)^2/2\delta a^2)} \tag{3.A.20}$$

where δa is the standard deviation of the distribution of well widths. This distribution of well widths may be expected to lead to a similar Gaussian distribution of associated Bohr frequencies, written as:

$$\rho(\omega_m) = \frac{1}{\sqrt{2\pi\delta\omega^2}} \mathrm{e}^{-((\omega_m-\omega_0)^2/2\delta\omega^2)} \tag{3.A.21}$$

Equation (1.49) allows us to calculate the resulting width of the distribution of Bohr frequencies:

$$\frac{\delta\omega}{\omega_{21}} = -\frac{2\delta a}{a_0} \tag{3.A.22}$$

Assuming a 6-nm wide quantum well and a width fluctuation of 0.5 nm, the transition energy E_{21} is 300 meV and the induced energy fluctuation is $\delta E_{21} =$ 300 meV/6 or 50 meV. This value is smaller in a finite quantum well, but the

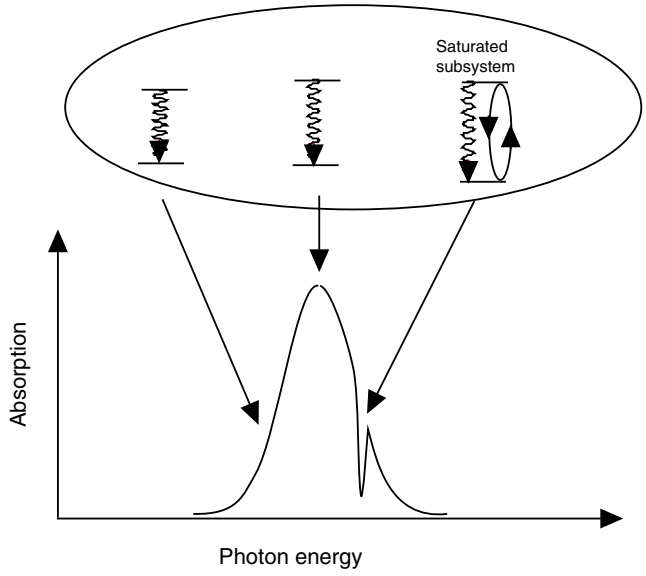


Fig. 3.A.4. Illumination of an inhomogeneous quantum system by an intense monochromatic beam of light which saturates the absorption in the subsystem having a transition energy equal to the photon energy of the pump beam. This opens a hole in the absorption spectrum of the inhomogeneous system as measured by a secondary probe beam of lesser intensity. This experiment is referred to as *spectral hole burning* and allows the differentiation between inhomogeneously and homogeneously broadened systems.

fluctuations in the well width must nonetheless be kept down to the scale of a single monolayer of atoms if the inhomogeneous broadening is not to exceed the homogeneous line broadening (typically a value of 10 meV).

A simple way to discriminate between homogeneous and inhomogeneous broadening in a system is to employ the technique known as *spectral hole burning*. In this case, a high power monochromatic source (referred to as a *pump beam*) of frequency v_0 lying within the absorption peak of the sample is mixed in with a low power broadband, or tuneable source (the *probe beam*), and directed onto the sample (see Fig. 3.A.4). If the broadening is inhomogeneous, the only element which will be optically saturated by the pump beam will be the mth element in the ensemble such that $v_0 = v_{21}^m$. As a result, the absorption spectrum of the probe beam will show a discrete dip at a frequency of v_0. If, however, the broadening is purely homogeneous, then all the elements in the system will be saturated by the pump beam and the entire absorption peak will drop in amplitude.

FURTHER READING

M. Born and E. Wolf, *Principles of Optics*, 6th Edn, Pergamon Press, New York (1980).

R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford (1973).
A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 4th Edn, Mc-Graw Hill, New York (1984).

3.B Second-order time-dependent perturbations

In Chapter 1, we saw how a time-dependent Hamiltonian may induce transitions between different states in a quantum system. In doing so, we wrote the time-dependent Schrödinger equation in the basis formed by the eigenstates of the system and proceeded to identify term by term, each of the elements in the differential equation (Eq. (1.70)). We now present an alternate and more powerful approach, which allows graphical interpretation of the time evolution of a quantum system under the influence of external perturbations. This will enable us to generalize the study of higher-order quantum transitions, such as two-photon absorption or optical absorption by free carriers.

We begin with Schrödinger's time-dependent equation which describes the evolution of the state $|\Psi(t)\rangle$ of a system under the influence of a time-dependent Hamiltonian $H(t)$:

$$H(t)|\Psi(t)\rangle = \mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\Psi(t)\rangle \tag{3.B.1}$$

As this last equation is linear, the state of a system at time t_b is related to that of the system at time t_a by a linear operator $u(t_b, t_a)$ defined by:

$$|\Psi(t_b)\rangle = u(t_b, t_a)|\Psi(t_a)\rangle \tag{3.B.2}$$

This *propagation operator* must then satisfy:

$$H(t_b)u(t_b, t_a) = \mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t_b}u(t_b, t_a) \tag{3.B.3}$$

If the Hamiltonian H is time independent ($H(t) = H_0$), then this last equation may be easily integrated giving:

$$u(t_b, t_a) = \mathrm{e}^{-\mathrm{i}(H_0/\hbar)(t_b - t_a)} \tag{3.B.4}$$

(in this last equation, the *exponential operator* e^H corresponds to $\sum H^n/n!$).

We introduce the basis formed by the eigenstates $|m\rangle$ of H_0 as defined by $H_0|m\rangle = E_m|m\rangle$. Making use of the *closure relation*:

$$I = \sum_m |m\rangle\langle m| \tag{3.B.5}$$

where I is the identity operator, the operator u may be written as:

$$u(t_b, t_a) = Iu(t_b, t_a) = \sum_m |m\rangle\langle m| e^{-i\omega_m(t_b - t_a)} \tag{3.B.6}$$

with ω_m corresponding to the Bohr frequency $E_m/\hbar$.

We now suppose that the system's Hamiltonian H_0 is perturbed by a time-dependent interaction $V(t)$ such that $H = H_0 + V(t)$. The propagation operator is then the solution to the differential equation:

$$[H_0 + V(t_b)]u(t_b, t_a) = i\hbar \frac{d}{dt_b} u(t_b, t_a) \tag{3.B.7}$$

Taking as inspiration the change of variables performed earlier in (1.67), we introduce the operator U given by:

$$u(t_b, t_a) = e^{-i(H_0/\hbar)(t_b - t_a)} U(t_b, t_a) \tag{3.B.8}$$

This change of variables is at times referred to as the *interaction picture*. Substituting (3.B.8) into (3.B.7) then leads to:

$$i\hbar \dot{U}(t_b, t_a) = e^{-i(H_0/\hbar)(t_b - t_a)} V(t_b) u(t_b, t_a) \tag{3.B.9}$$

This equation may then be formally integrated to obtain:

$$U(t_b, t_a) = I - \frac{i}{\hbar} \int_{t_a}^{t_b} e^{i(H_0/\hbar)(t - t_a)} V(t) u(t, t_a) dt \tag{3.B.10}$$

or alternately:

$$u(t_b, t_a) = u^{(0)}(t_b, t_a) - \frac{i}{\hbar} \int_{t_a}^{t_b} u^{(0)}(t_b, t) V(t) u(t, t_a) dt \tag{3.B.11}$$

Schrödinger equation for the propagation operator

where $u^{(0)}$ is the propagation operator of the unperturbed system given by (3.B.6). Equation (3.B.11) is of no practical use in the sense that it is little more than an integral form of Schrödinger's differential equation. It does, however, readily loan itself to graphical interpretation and allows a basis for iterative calculations which we will now describe.

Equation (3.B.11) may be interpreted as a limited expansion in perturbations of increasing order, with $u^{(0)}$ corresponding to the first term in the series. The propagation operator may then be written as:

$$u(t_b, t_a) = u^{(0)}(t_b, t_a) + u^{(1)}(t_b, t_a) + u^{(2)}(t_b, t_a) + \cdots \tag{3.B.12}$$

where:

$$u^{(0)}(t_b, t_a) = \mathrm{e}^{-\mathrm{i}(H_0/\hbar)(t_b - t_a)} \tag{3.B.13a}$$

$$u^{(1)}(t_b, t_a) = -\frac{\mathrm{i}}{\hbar}\int_{t_a}^{t_b} u^{(0)}(t_b, t_1) V(\mathrm{t}_1) u^{(0)}(t_1, t_a)\mathrm{d}t_1 \tag{3.B.13b}$$

$$u^{(2)}(t_b, t_a) = \left(-\frac{\mathrm{i}}{\hbar}\right)^2 \int_{t_a}^{t_b}\int_{t_a}^{t_1} u^{(0)}(t_b, t_1) V(t_1) u^{(0)}(t_1, t_2) V(t_2) u^{(0)}(t_2, t_a)\mathrm{d}t_1\mathrm{d}t_2 \tag{3.B.13c}$$

We are now in a position to interpret these equations graphically:

(a) To first order $u^{(1)}$: the system evolves from t_a to t_1 without interaction ($u^{(0)}(t_1, t_a)$ term), interacts with the perturbation at $t = t_1$ ($V(t_1)$ term) and continues to evolve without interaction until its final destination at t_b ($u^{(0)}(t_b, t_1)$ term). Figure 3.B.1 shows a graphical representation of this evolution, known as a *Feynman diagram*. The first-order correction to the propagation operator is then the integral of these elementary contributions over all intermediate times t_1. For a detailed and enlightening discussion on this topic, we refer the reader to Richard Feynman's masterpiece *QED*: *The Strange Theory of Light and Matter* (1985).

Starting from expression (3.B.13b), we may calculate the first-order correction to the state of the system:

$$|\Psi^{(1)}(t)\rangle = u^{(1)}(t, t_a)|\Psi^{(0)}(t_a)\rangle$$

where $|i\rangle = |\Psi^{(0)}(t = 0)\rangle$ is the initial state of the system, and calculate the transition rate between the states $|i\rangle$ and $|f\rangle$. We would then return to the earlier results of Section 1.6 (this will be left as an exercise to the reader). We now turn our attention to the second-order expansion. The second-order interaction plays a fundamental role when the first-order interaction yields a null transition rate (as in the case of a forbidden transition).

(b) To second order: the integral (3.B.13c) may be read as follows. The system evolves from t_a to t_1, interacts at t_1, evolves again between t_1 and t_2, interacts again at $t = t_2$, and continues on its course without interaction until t_b (Fig. 3.B.2).

In (3.B.13c), we replace the expression for the unperturbed operator $u^{(0)}$ by its explicit definition in (3.B.6) giving:

$$u^{(2)}(t, 0) = \frac{1}{\hbar^2}\int_{t_1=0}^{t_1=t}\int_{t_2=0}^{t_2=t_1} \sum_f |f\rangle\langle f|\mathrm{e}^{-\mathrm{i}\omega_f(t-t_1)} V(t_1) \times \sum_n |n\rangle\langle n|\mathrm{e}^{-\mathrm{i}\omega_n(t_1-t_2)} V(t_2) \sum_m |m\rangle\langle m|\mathrm{e}^{-\mathrm{i}\omega_m t_2}\mathrm{d}t_1\mathrm{d}t_2 \tag{3.B.14}$$

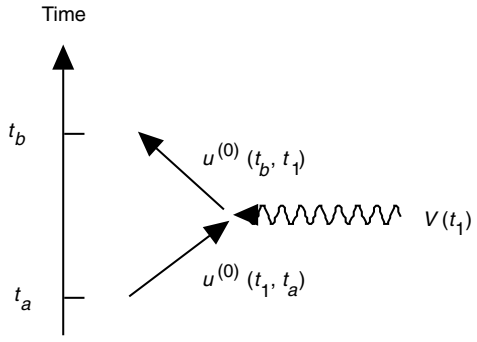


Fig. 3.B.1. Feynman diagram for a first-order perturbation.

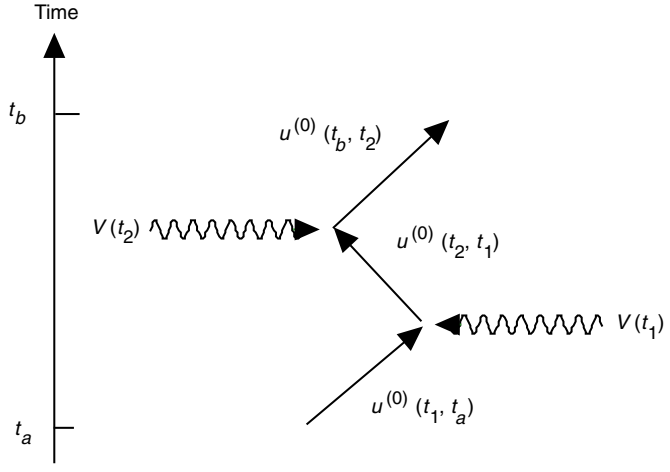


Fig. 3.B.2. Feynman diagram for a second-order perturbation.

The second-order correction to the initial state $|i\rangle = |\Psi^{(0)}(t=0)\rangle$ is therefore $|\Psi^{(2)}(t)\rangle = u^{(2)}(t,0)|i\rangle$ or:

$$|\Psi^{(2)}(t)\rangle = \frac{1}{\hbar^2}\int_{t_1=0}^{t_1=t}\int_{t_2=0}^{t_2=t_1}\sum_f |f\rangle\langle f|\mathrm{e}^{-\mathrm{i}\omega_f(t-t_1)}V(t_1) \times \sum_n |n\rangle\langle n|\mathrm{e}^{-\mathrm{i}\omega_n(t_1-t_2)}V(t_2)\mathrm{e}^{-\mathrm{i}\omega_i t_2}|i\rangle \mathrm{d}t_1\mathrm{d}t_2 \tag{3.B.15}$$

which after regrouping term by term may be written as:

$$|\Psi^{(2)}(t)\rangle = \frac{1}{\hbar^2}\sum_f |f\rangle e^{-i\omega_f t} \int_{t_1=0}^{t_1=t}\int_{t_2=0}^{t_2=t_1}\left[\sum_n \langle f|V(t_1)|n\rangle e^{i\omega_{fn}t_1}\langle n|V(t_2)|i\rangle e^{i\omega_{ni}t_2}\right] dt_1\, dt_2 \tag{3.B.16}$$

We recognize in this last expression, the decomposition of the state $|\Psi^{(2)}(t)\rangle$ into all of the possible final states $|f\rangle$, i.e:

$$|\Psi^{(2)}(t)\rangle = \sum_f c_f(t)|f\rangle e^{-i\omega_f t}$$
$$c_f(t) = \frac{1}{\hbar^2}\int_{t_1=0}^{t_1=t}\int_{t_2=0}^{t_2=t_1}\left[\sum_n \langle f|V(t_1)|n\rangle e^{i\omega_{fn}t_1}\langle n|V(t_2)|i\rangle e^{i\omega_{ni}t_2}\right] dt_1\, dt_2 \tag{3.B.17}$$

It is now high time to specify the interaction $V(t)$. We will suppose that $V(t)$ is due to the contribution of two oscillating sources:

$$V(t) = \sum_{I=1}^{2} V_I e^{-i\omega_I t} + \text{c.c.} \tag{3.B.18}$$

Substituting this expression for $V(t)$ into (3.B.17), we notice that $c_f(t)$ will be the sum of the different contributions of $2\omega_1$, $2\omega_2$, $\omega_1 + \omega_2$, $\omega_1 - \omega_2$.

We will first take interest in the sum of the frequencies $\omega_1 + \omega_2$ (corresponding to a two-photon transition). Expression (3.B.17) then takes the somewhat unwieldy form:

$$c_f(t) = \frac{1}{\hbar^2}\sum_I^2\sum_J^2\sum_n C_{I,J,n} \tag{3.B.19}$$

where each element $C_{I,J,n}$ is obtained by integration of (3.B.17) or:

$$C_{I,J,n} = \frac{\langle f|V_J|n\rangle\langle n|V_I|i\rangle}{i(\omega_{ni} - \omega_I)}\left[\frac{e^{i(\omega_{fi}-\omega_I-\omega_J)t} - 1}{i(\omega_{fi} - \omega_I - \omega_J)} - \frac{e^{i(\omega_{fn}-\omega_J)t} - 1}{i(\omega_{fn} - \omega_J)}\right] \tag{3.B.20}$$

We will assume that the one-photon transitions are non-resonant ($\omega_{ni} \neq \omega_I$ and $\omega_{fn} \neq \omega_J$), whereas the two-photon transitions are close to resonance, i.e:

$$E_f - E_i = \hbar\omega_1 + \hbar\omega_2 \tag{3.B.21}$$

In this case, the first term in parentheses dominates the expression (3.B.20). The final amplitude $c_f(t)$ takes the form:

$$c_f(t) = \frac{1}{\hbar}\sum_n\sum_J\sum_I \frac{\langle f|V_J|n\rangle\langle n|V_I|i\rangle}{E_n - E_i - E_I}\left[\frac{\sin(\omega_{fi} - \omega_I - \omega_J)t/2}{(\omega_{fi} - \omega_I - \omega_J)/2}\right] \tag{3.B.22}$$

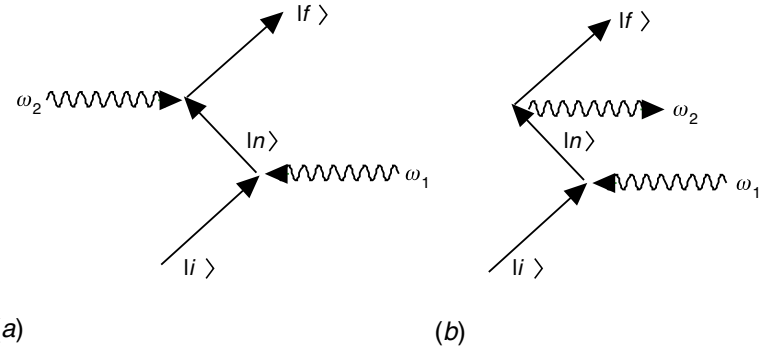


Fig. 3.B.3. Terms intervening in the two-photon transition rate. Only the mechanism shown in (*a*) is resonant with this process and contributes to two-photon absorption.

We recognize in this expression the issues discussed in Section 1.6. For long times, the probability $|c_f(t)|^2$ of finding the system (initially in state $|i\rangle$) in the state $|f\rangle$ tends towards (see (1.77) and (1.78)):

$$|c_f(t)|^2 = \frac{1}{\hbar^2}\left|\sum_n \sum_J \sum_I \frac{\langle f|V_J|n\rangle\langle n|V_I|i\rangle}{E_n - E_i - E_I}\right|^2 2\pi t\delta(\omega_{fi} = \omega_I + \omega_J) \tag{3.B.23}$$

We find, as in the case of a first-order time-dependent perturbation, that the second-order perturbation induces a constant transition rate G_{fi} given in this case by:

$$G_{fi} = \frac{2\pi}{\hbar}\left|\sum_n \sum_J \sum_I \frac{\langle f|V_J|n\rangle\langle n|V_I|i\rangle}{E_n - E_i - E_I}\right|^2 \delta(E_{fi} = \hbar\omega_I + \hbar\omega_J) \tag{3.B.24}$$

Fermi's second golden rule

This last expression is referred to as *Fermi's second golden rule*. Each term appearing in the summation of (3.B.24) is interpreted as follows: under the influence of the oscillations at ω_I, the system evolves from state $|i\rangle$ to state $|n\rangle$ with probability amplitude proportional to $\langle n|V_I|i\rangle$. As energy is not conserved, the transition can only occur over a time interval Δt, set by Heisenberg's second uncertainty relation ($\Delta t \approx \hbar/\Delta E$ or $\hbar/(E_{ni} - E_I)$), and which appears as a weighting factor in (3.B.24). Figure 3.B.3*a* illustrates such a two-photon absorption process.

We will now apply Fermi's second golden rule to the problem of two-photon (2ω) absorption. We consider a two-level system ($|1\rangle$ and $|2\rangle$) having an eigenenergy separation of $\hbar\omega_{21}$. This system is subjected to an electromagnetic wave of frequency ω. The perturbation Hamiltonian is then:

$$V(t) = \frac{qEz}{2}e^{i\omega t} + \text{c.c.} \tag{3.B.25}$$

Noting that there is no summation over the I and J indices, the two-photon transition rate is then given by (3.B.24) to be:

$$G_{21} = \frac{2\pi}{\hbar^2}\left|\sum_n \frac{\langle f|V|n\rangle\langle n|V|i\rangle}{E_n - E_i - \hbar\omega}\right|^2 \delta(\omega_{21} = 2\omega) \tag{3.B.26}$$

As the intermediate state can only be $|1\rangle$ or $|2\rangle$, this gives:

$$G_{21} = \frac{2\pi}{\hbar^2}\left|\frac{\langle 2|V|2\rangle\langle 2|V|1\rangle}{E_2 - E_1 - \hbar\omega} + \frac{\langle 2|V|1\rangle\langle 1|V|1\rangle}{-\hbar\omega}\right|^2 \delta(\omega_{21} = 2\omega) \tag{3.B.27}$$

and as $\omega_{21} \approx 2\omega$:

$$G_{21} = \frac{2\pi}{\hbar^4}\left|\frac{(\langle 2|V|2\rangle - \langle 1|V|1\rangle)\langle 2|V|1\rangle}{\omega_{21} - \omega}\right|^2 \delta(\omega_{21} = 2\omega) \tag{3.B.28}$$

and posing:

$$\langle 2|V|1\rangle = \frac{qEz_{12}}{2}$$

$$\langle 2|V|2\rangle - \langle 1|V|1\rangle = \frac{qE\delta_{12}}{2} \tag{3.B.29}$$

$$\langle 2|z|2\rangle - \langle 1|z|1\rangle = \delta_{12}$$

the transition rate between the two states in the system is then:

$$G_{21} = \frac{\pi}{8\hbar^4}\frac{q^4E^4z_{12}^2\delta_{12}^2}{(\omega_{21} - \omega)^2}\delta(\omega_{21} = 2\omega) \tag{3.B.30}$$

Therefore, even if the transition $|1\rangle \rightarrow |2\rangle$ is not permitted to first order, the two-photon transition is allowed as long as the element δ_{12} is non-zero (i.e. as long as the system is asymmetric). The oscillating wave with frequency ω then experiences a decrease in its intensity I_ω ($=\varepsilon cE^2/2n_{\text{op}}$) during propagation equal to:

$$\frac{\partial}{\partial z}I_\omega = 2\hbar\omega G_{21}(N_1 - N_2) \tag{3.B.31}$$

where N_1 and N_2 correspond to the population densities in each of their respective levels. Substituting (3.B.30) into (3.B.31), we see that the absorption may be written as:

$$\frac{\partial}{\partial z}I_\omega = -\beta I_\omega^2 \tag{3.B.32}$$

where β is the *two-photon absorption* coefficient given by:

$$\beta = \frac{\pi q^4 \omega (N_1 - N_2) z_{12}^2 \delta_{12}^2}{\varepsilon^2 c^2 \hbar^3 (\omega_{21} - \omega)^2} \delta(\omega_{21} = 2\omega) \tag{3.B.33}$$

Two-photon absorption coefficient

This theory may be generalized without difficulty to treating two-photon transitions in semiconductors. In this case, (3.B.33) must be integrated over the entire band structure following a procedure outlined in another context in Chapter 5. Table 3.B.1 shows the two-photon absorption coefficients for different materials. Two-photon absorption is very important as it limits the optical power density that may readily propagate in optoelectronic semiconductor components (e.g. lasers, modulators, . . .). Two-photon absorption may also be utilized for optical protection from excessive laser fluxes.

Example

(a) We consider a GaAs laser having a guiding layer thickness of 1 μm and a width of 5 μm. For an internal power of 500 mW, the power density I_0 is $10^7\,\mathrm{W\,cm^{-2}}$ resulting in a two-photon absorption of βI_0 ($\beta = 25\,\mathrm{cm\,GW^{-1}}$), or a parasitic loss of $0.25\,\mathrm{cm^{-1}}$. This is not a negligible quantity and it limits the power levels which may easily propagate in GaAs waveguides.

(b) A $d = 1\,\mathrm{mm}$ thick slab of InSb is placed before the focal point of a lens (resulting spot size diameter of 0.5 mm or surface area of $2 \times 10^{-3}\,\mathrm{cm^2}$). The output power at the exit of the slab is then given by the integral of the differential equation (3.B.32):

$$\frac{1}{I} - \frac{1}{I_0} = \beta d \tag{3.B.34}$$

where I_0 is the entrance power. For a power of 2 MW (or power density of $1 \times 10^9\,\mathrm{W\,cm^{-2}}$), the exit power density is equal to $1/\beta d$ or $1/(5 \times 10^3\,\mathrm{cm\,GW^{-1}} \times 10^{-1}\,\mathrm{cm})$, which equals $2 \times 10^{-3}\,\mathrm{GW\,cm^{-2}}$. This corresponds to an attenuation factor I/I_0 of 2×10^{-3}.

Table 3.B.1. *Two-photon absorption coefficients for various semiconductor materials. The uncertainty indicated in the values for* β *corresponds to the great experimental difficulties involved in these measurements.*

Material	Wavelength (μm)	$\beta\,(\mathrm{cm\,GW^{-1}})$
ZnSe	0.532	5–6
GaAs	1.064	23–26
CdTe	1.064	15–25
ZnTe	1.064	4–5
InSb	10.6	5000

FURTHER READING

L. I. Schiff, *Quantum Mechanics*, 2nd Edn, Mc-Graw Hill, New York (1955).
A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

An illuminating explanation of Feynman diagrams may be found in:
R. P. Feynman, *QED*: *The Strange Theory of Light and Matter*, Princeton University Press, NJ (1985).

3.C Einstein coefficients in two limiting cases: quasi-monochromatic and broadband optical transitions

We will now present the treatment of induced and spontaneous emission given by Albert Einstein in 1905. This approach is of significant historical interest, as it demonstrates the power of well directed *heuristic* reasoning.

We consider a two-level system in thermodynamic equilibrium with a cavity forming a blackbody. The energy of the electromagnetic modes in the cavity has a spectral density distribution $\rho_e(\nu)$ (in J s m^{-3}). As the photons within the cavity are in a state of thermodynamic equilibrium, their energy density is given by the Planck distribution (2.79). We designate by $G_{12}\mathrm{d}\nu$ the transition rate from level $|1\rangle$ to level $|2\rangle$ due to electromagnetic energy contained in the frequency interval dν and $G_{21}\mathrm{d}\nu$ the inverse transition rate. The transition rate equations (1.85) indicate that, in an interval dν, there is an available energy $\rho_e(\nu)\mathrm{d}\nu$ which will lead to a transition rate $G_{12}\mathrm{d}\nu$ (we recall that, in G_{12}, the term $|W_{12}^2|$ is proportional to E^2, which is itself proportional to the electromagnetic field energy). *By definition*, the Einstein coefficients B_{12} and B_{21} are therefore the proportionality coefficients between the transition rates and the energy density $\rho_e(\nu)$;

$$\begin{aligned} G_{12} &= B_{12}\rho_e(\nu) \\ G_{21} &= B_{21}\rho_e(\nu) \end{aligned} \tag{3.C.1}$$

Actually, Eq. (3.C.1) is incomplete. We must add the term which figures in the spontaneous emission:

$$\begin{aligned} G_{12} &= B_{12}\rho_e(\nu) \\ G_{21} &= B_{21}\rho_e(\nu) + A_{12} \end{aligned} \tag{3.C.2}$$

At thermodynamic equilibrium, the occupation densities in the two levels remain constant, the transitions from $|1\rangle \rightarrow |2\rangle$ must equal those from $|2\rangle \rightarrow |1\rangle$ or:

$$N_1 B_{12}\rho_e(\nu) = N_2(B_{21}\rho_e(\nu) + A_{12}) \tag{3.C.3}$$

As already noted, the energy density for the modes in the cavity at thermodynamic

equilibrium $\rho_e(\nu)$ is given by Planck's law (2.91), so that the above equation yields:

$$N_1 B_{12} \frac{8\pi n_{\text{op}}^3 h\nu^3}{c^3} \frac{1}{e^{h\nu/kT}-1} = N_2 \left(B_{21} \frac{8\pi n_{\text{op}}^3 h\nu^3}{c^3} \frac{1}{e^{h\nu/kT}-1} + A_{12} \right) \tag{3.C.4}$$

(where we have introduced the index of refraction n_{op} to take into account the optical dispersion occurring in the cavity medium). Now, at thermodynamic equilibrium, the ratio of the populations distributed in each of the levels is given by:

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} e^{-h\nu/kT} \tag{3.C.5a}$$

where we have temporarily introduced the possibility that the degree of degeneracy in the two levels g_1 and g_2 may be different from 1. Substituting (3.C.5a) into (3.C.4), we obtain the self-consistent condition:

$$\frac{8\pi n_{\text{op}}^3 h\nu^3}{c^3(e^{h\nu/kT}-1)} = \frac{A_{12}(g_2/g_1)}{B_{12}e^{h\nu/kT} - B_{21}(g_2/g_1)} \tag{3.C.5b}$$

This equation must be obeyed at any given temperature. This is only possible if the two following relationships hold:

$$\begin{aligned} \frac{B_{12}}{B_{21}} &= \frac{g_2}{g_1} \\ \frac{A_{12}}{B_{12}} &= \frac{8\pi n_{\text{op}}^3 h\nu^3}{c^3} \end{aligned} \tag{3.C.6}$$

From which we recover Eq. (3.77) obtained within the framework of electromagnetic field quantization, as well as the expression for spontaneous emission. Einstein's unique insight (or genius) allowed him to derive the relationship between the A and B coefficients by using the conditions imposed by the particular case of thermodynamic equilibrium of a blackbody while realizing the universal applicability of these findings outside this state of equilibrium (and constituting a characteristic of the quantum system itself). The transition rates (3.C.2) are then given by:

$$G_{12} = \frac{c^3}{8\pi n_{\text{op}}^3 h\nu^3 t_{\text{spon}}} \rho_e(\nu) \tag{3.C.7}$$

Transition rate (per second) between two levels due to an electromagnetic wave having a large spectral distribution $\rho_e(\nu)$

where $\rho_e(\nu)$ is the spectral energy distribution of the incident electromagnetic wave.

The above approach deals with an entirely monochromatic transition, i.e. a single frequency transition induced by an electromagnetic wave with a broad

spectral distribution. We may wonder whether this approach would allow us nonetheless to obtain the results given by (3.73) or (3.74), which describe the absorption and the optical cross-section due to transitions induced by a quasi-monochromatic wave between levels and having a Lorentzian lineshape $\mathscr{L}(v)$. To do so, we need only employ the same argument which allowed us to pass from (3.75) to (3.76), i.e. by introducing the spectral dependence of the transition rate G'_{ij}:

$$G'_{12} = \frac{c^3 \rho_v}{8\pi n_{\text{op}}^3 h v^3 t_{\text{spon}}} \mathscr{L}(v) \tag{3.C.8}$$

Spectral dependence of the transition rate (dimensionless!) between two levels due to a quasi-monochromatic electromagnetic wave having an energy density of ρ_v. The transition is broadened by the Lorentzian function $\mathscr{L}(v)$

where ρ_v (joules per cubic metre) is the energy density of the photons at frequency v given by $I_v = c\rho_v / n_{\text{op}}$ (watts per square metre). Equation (3.C.7) is obtained by integrating (3.C.8) over the broadened spectrum, i.e.:

$$G_{12} = \int_{-\infty}^{+\infty} G'_{12}(v) \mathrm{d}v \tag{3.C.9}$$

Thus (3.C.7) and (3.C.8) represent both extreme situations which may arise in the application of the rate equations.

3.D Equivalence of the A · p and D · E Hamiltonians and the Thomas–Reiche–Kuhn sum rule

A certain number of general conclusions on the different quantities intervening in the optical properties (oscillator strengths, . . .) may be drawn from the commutation properties of the **p** and **r** observables. In this section, **p** and **r** will refer strictly to operators and as such we will allow ourselves to drop the hat '^' used otherwise to differentiate them from their use as variables. We will begin by showing the equivalence of the $\mathbf{A} \cdot \mathbf{p}$ and $\mathbf{D} \cdot \mathbf{E}$ Hamiltonians.

We recall that two types of Hamiltonians may be used to describe the interaction between a quantum system (described by its Hamiltonian $H_0 = \mathbf{p}^2/2m + V(\mathbf{r})$) and an electromagnetic wave (see (3.12) and (3.13)):

$$\begin{aligned} &\text{Hamiltonian } \mathbf{A} \cdot \mathbf{p} \qquad W_{Ap}(t) = -\frac{q}{m} \mathbf{A}_\perp(\mathbf{r}_0, t)\mathbf{p} \\ &\text{Hamiltonian } \mathbf{D} \cdot \mathbf{E} \qquad W_{DE}(t) = -q\mathbf{E}(\mathbf{r}_0, t)\mathbf{r} \end{aligned} \tag{3.D.1}$$

For a plane wave, the vector potential and the electric field are given by (3.5a) and (3.6):

$$\begin{aligned}\mathbf{E}(\mathbf{r}_0, t) &= \mathbf{E}_0 \cos(\mathbf{k}\cdot\mathbf{r}_0 - \omega t)\\ \mathbf{A}_\perp(\mathbf{r}_0, t) &= \frac{\mathbf{E}_0}{\omega}\sin(\mathbf{k}\cdot\mathbf{r}_0 - \omega t)\end{aligned} \tag{3.D.2}$$

We recall that $\mathbf{r}_0$ designates the position of the quantum system and may be taken to be the origin. To lighten the required notation, we may suppose that $\mathbf{A}_\perp$ and $\mathbf{E}$ are oriented along the Oz axis. Equations (3.D.1) may then be written:

$$\begin{aligned}W_{Ap}(t) &= W_{Ap}\cos\omega t = -\frac{q}{m}\frac{E_0}{\omega}p_z\cos\omega t\\ W_{DE}(t) &= W_{DE}\sin\omega t = -qE_0 z\sin\omega t\end{aligned} \tag{3.D.3}$$

We saw that the transition rates which enter into the absorption formulas use elements of type $\langle f|z|i\rangle$ or $\langle f|p_z|i\rangle$. We will therefore seek the relationship between these two terms. To do so, we will make use of the commutator of z and H_0:

$$\begin{aligned}[z, H_0] &= \left[z, \frac{p_z^2}{2m}\right] = \frac{1}{2m}(zp_z^2 - p_z^2 z)\\ &= \frac{1}{m}(zp_z^2 - p_z z p_z + p_z z p_z - p_z^2 z)\\ &= \frac{1}{2m}([z, p_z]p_z + p_z[z, p_z])\\ &= \frac{i\hbar p_z}{m}\end{aligned} \tag{3.D.4}$$

Projected onto states $|i\rangle$ and $|f\rangle$, Eq. (3.D.4) becomes:

$$\frac{i\hbar}{m}\langle f|p_z|i\rangle = \langle f|[z, H_0]|i\rangle = (E_i - E_f)\langle f|z|i\rangle \tag{3.D.5}$$

from which the relationship between the two matrix elements is found to be:

$$\langle f|p_z|i\rangle = \mathrm{i}m\omega_{fi}\langle f|z|i\rangle \tag{3.D.6}$$

Relationship between the matrix elements z and p_z

Substituting (3.D.6) into (3.D.3), gives:

$$\langle f|W_{Ap}|i\rangle = -\frac{q}{m}\frac{E_0}{\omega}\langle f|p_z|i\rangle = -\mathrm{i}\frac{\omega_{fi}}{\omega}\langle f|W_{DE}|i\rangle \tag{3.D.7}$$

which establishes the equivalence between the two Hamiltonians only at resonance. The *oscillator strength* within the $\mathbf{A}\cdot\mathbf{p}$ approach is written:

$$f_{21} = \frac{2}{m\hbar\omega_{21}} |\langle 1|\mathbf{p}\cdot\boldsymbol{\varepsilon}|2\rangle|^2 \tag{3.D.8}$$

Oscillator strength for a transition $1 \rightarrow 2$ using the $\mathbf{A}\cdot\mathbf{p}$ Hamiltonian

The preferential use of matrix elements taken from either the $\mathbf{D}\cdot\mathbf{E}$ (3.40) or $\mathbf{A}\cdot\mathbf{p}$ (3.D.8) Hamiltonians depends above all on the context. We may say in general, that the $\mathbf{A}\cdot\mathbf{p}$ Hamiltonian is best used for transitions involving delocalized states (i.e. momentum eigenstates), as encountered in the description of absorption between bands in semiconductors. The $\mathbf{D}\cdot\mathbf{E}$ Hamiltonian, however, is better suited to describing transitions between localized states (deep defect states, quantum wells, etc.).

We will show yet another consequence of these commutation rules. We will calculate the sum of the oscillator strengths for the transitions between $|i\rangle$ and all possible final states $|f\rangle$:

$$\sum_{f\neq i} f_{fi} = \frac{2m}{\hbar^2}\left[\sum_{f\neq i} E_f |\langle f|z|i\rangle|^2 - \sum_{f\neq i} E_i |\langle f|z|i\rangle|^2\right] \tag{3.D.9}$$

This last equation may be simplified in a very significant manner. On one hand:

$$\begin{aligned}\sum_f E_f |\langle f|z|i\rangle|^2 &= \sum_f E_f \langle i|z|f\rangle\langle f|z|i\rangle \\ &= \langle i|z\left[\sum_f E_f |f\rangle\langle f|\right] z|i\rangle\end{aligned} \tag{3.D.10}$$

We will suppose that the basis of eigenvectors $|f\rangle$ is complete, which is to say that any state in the system $|n\rangle$ may be decomposed in unique fashion onto this basis. We then have:

$$H_0 = \sum_f E_f |f\rangle\langle f| \tag{3.D.11}$$

Expression for a Hamiltonian in terms of its eigenenergies and eigenstates

This last equation may be explained by noting that if (3.D.11) holds for all eigenstates $|f\rangle$, it must also hold for any arbitrary state $|n\rangle$. We then have:

$$\sum_f E_f |\langle f|z|i\rangle|^2 = \langle i|zH_0 z|i\rangle \tag{3.D.12}$$

The second right-hand term of (3.D.9) reads:

$$\sum_f E_f |\langle f|z|i\rangle|^2 = E_i \sum_f \langle i|z|f\rangle\langle f|z|i\rangle$$
$$= E_i \left\langle i \left| z \left[\sum_f |f\rangle\langle f| \right] z \right| i \right\rangle \tag{3.D.13}$$

The following *closure relation* may be easily shown to yield:

$$\sum_f |f\rangle\langle f| = 1 \tag{3.D.14}$$

Closure relation for a complete basis of eigenvectors of an observable

where 1 is the identity operator. We therefore have:

$$\sum_f E_i |\langle f|z|i\rangle|^2 = E_i \langle i|z^2|i\rangle = \langle i|H_0 z^2|i\rangle \tag{3.D.15}$$

Using (3.D.12) and (3.D.15), Eq. (3.D.9) then becomes:

$$\sum_{f\neq i} f_{fi} = \frac{2m}{\hbar^2}\langle i|zH_0 z - H_0 z^2|i\rangle \tag{3.D.16}$$

Now, from (3.D.4) one may write:

$$[[z, H_0], z] = \frac{\mathrm{i}\hbar}{m}[p_z, z] = \frac{\hbar^2}{m} \tag{3.D.17}$$

but also:

$$\langle i|[[z, H_0], z]|i\rangle = \langle i|(2zH_0 z - H_0 z^2 - z^2 H_0)|i\rangle$$
$$= 2\langle i|zH_0 z - H_0 z^2|i\rangle \tag{3.D.18}$$

as $\langle i|H_0 z^2|i\rangle$ and $\langle i|z^2 H_0|i\rangle$ are both equal to $E_i\langle i|z^2|i\rangle$. Substituting (3.D.17) and (3.D.18) into (3.D.16), we obtain:

$$\sum_{f\neq i} f_{fi} = 1 \tag{3.D.19}$$

Thomas–Reiche–Kuhn sum rule

Coupled to the formula for the absorption of a multi-level system (3.41):

$$\alpha(\omega) \approx N_1 \sum_{j=1}^{\infty} \frac{q^2\pi}{2\varepsilon_0 mcn} f_{j1}\mathscr{L}(\omega - \omega_{j1}) \tag{3.D.20}$$

the sum rule in (3.D.19) may be interpreted as saying that the classical absorption $\alpha = N_1 q^2\pi/(2\varepsilon_0 mcn)\mathscr{L}(\nu)$ which may be obtained within the context of the elasti-

cally bound electron model remains valid in quantum mechanics. Only, this absorption is distributed between different oscillators corresponding to transitions between different quantized levels in the system. This distribution is weighted by oscillator strengths which gauge the strength of each of these transitions. *In spite of this, the sum of all these oscillator strengths will be found to equal unity*. In more empirical terms, if we consider the complete absorption curve for a quantum system, the ratio of the integral of the absorption to the population of the fundamental level N_1:

$$\frac{\int_0^\infty \alpha(\omega)\mathrm{d}\omega}{N_1} \approx \frac{q^2\pi}{2\varepsilon_0 mcn} \tag{3.D.21}$$

is the same for all physical systems. This concept reveals its true strength when one realizes that oscillator strengths can be negative (refer to the footnote accompanying (3.41)). We shall return to the point later in our study of intersubband transitions in quantum wells.

Example: oscillator strengths in an infinite quantum well

We will illustrate the concept of oscillator strength for the easily managed case of the infinite square well. We saw in Section 1.4.2 that the eigenenergies and eigenfunctions of the stationary states for a well of width a, with infinitely high potential barriers at 0 and a are given by:

$$\begin{aligned} E_n &= n^2 \frac{\hbar^2\pi^2}{2m_e a^2} \\ |n\rangle &= \sqrt{\frac{2}{a}} \sin n\frac{\pi}{a} z \end{aligned} \tag{3.D.22}$$

The matrix elements $\langle i|z|j\rangle$ are then given by the integrals:

$$\begin{aligned} \langle i|z|j\rangle &= \frac{2}{a}\int_0^a z \sin i\frac{\pi}{a} z \sin j\frac{\pi}{a} z \,\mathrm{d}z \\ &= \frac{a}{\pi^2}\left[\frac{1}{(j-i)^2} - \frac{1}{(j+i)^2}\right][\cos(i-j)\pi - 1] \end{aligned} \tag{3.D.23}$$

and:

$$|\langle i|z|j\rangle|^2 = \frac{2^6 a^2}{\pi^4}\frac{(ij)^2}{(j^2-i^2)^4} F(i-j) \tag{3.D.24}$$

where $F(i-j)$ is equal to 0 if $i-j$ is even, and 1 if $i-j$ is odd. The oscillator strength for the transition $i \to j$ is then:

$$f_{ij} = \frac{2m}{\hbar^2}\left[(j^2 - i^2)\frac{\hbar^2\pi^2}{2ma^2}\right]\left[\frac{2^6 a^2}{\pi^4}\frac{(ij)^2}{(j^2-i^2)^4}F(i-j)\right] \tag{3.D.25}$$

so that:

$$f_{ij} = \frac{2^6}{\pi^2}\frac{(ij)^2}{(j^2-i^2)^3}F(i-j) \tag{3.D.26}$$

Oscillator strength for a transition $i \to j$ in an infinite quantum well

We may verify the sum rule by utilizing the series expansion for π^2. We notice that the oscillator strength is concentrated in the transitions between the lowest energy levels. More precisely, the oscillator strengths f_{j1} starting from the fundamental level $|1\rangle$ decrease as $1/j^4$ with the first two values corresponding to:

$$f_{21} = \frac{2^8}{\pi^2 3^3} = 0.960$$
$$f_{41} = \frac{2^{10}}{\pi^2 15^3} = 0.030 \tag{3.D.27}$$

4 Laser oscillations

4.1 Introduction

We were able to show in Chapter 3 that a medium in which we can obtain a population inversion (i.e. a situation in which the population density in the excited state is greater than that in the fundamental level) allows for optical gain of an electromagnetic wave having a frequency near to the resonant frequency of the system. By introducing feedback of the amplified signal into the medium, the system can be made to oscillate naturally, resulting in *laser oscillations*. To obtain this population inversion, we must introduce at least a third (and perhaps even a fourth) energy level into the system. (We saw how a two-level system under the influence of an intense pump beam will saturate with no resulting population inversion.) The aim then of this chapter is to introduce the concepts necessary to extend our two-level system into a working model capable of illustrating the phenomenon of laser oscillation. We will not spend too much time discussing atomic transition lasers as they do not figure readily in our treatment of quantum electronic properties of semiconductors. An exception will be made, however; we brush upon the particular topics of a diode pumped laser in Complement 4.E and a quantum cascade laser in Complement 13.H.

4.2 Population inversion and optical amplification

4.2.1 Population inversion

We will show how population inversion can be achieved by carrier transfer from higher lying levels to the upper level of a two-level subsystem of interest. To do so, we consider a cavity of volume V filled with a density n (in cm^{-3}) of identical quantum systems each possessing an arbitrary number of levels E_i, $i = 0, 1, 2,$ etc. We are interested in the absorption and emission effects near resonance $|1\rangle \rightarrow |2\rangle$ (i.e. for transitions involving photon energies $h\nu \approx E_{21} = h\nu_{21}$). We will suppose that the second level is filled by some as yet undefined mechanism described by a filling rate R_2 ($\text{cm}^{-3}\,\text{s}^{-1}$). This filling occurs either by pumping electrons from inferior levels directly into $|2\rangle$ or via recombination from superior levels (see Fig. 4.1). Electrons in level $|1\rangle$ in turn relax to lower levels with a lifetime of τ_1 and electrons in level $|2\rangle$ relax to $|1\rangle$ by spontaneous emission (rate $= 1/t_{\text{spon}}$), by

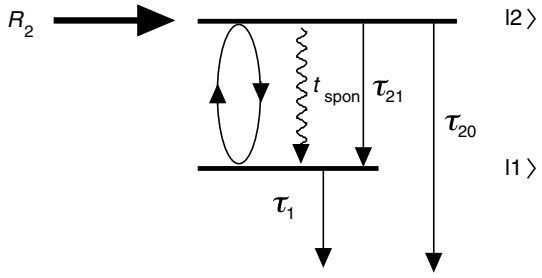


Fig. 4.1. Different dynamic mechanisms for a two-level system in interaction with other levels.

non-radiative recombination (rate $= 1/\tau_{21}$), and may also relax to other lower energy levels (rate $= 1/\tau_{20}$). Let us suppose for a moment that $t_{spon} \gg \tau_{21}$; the rate equations describing the population densities $n_1 = N_1/V$ and $n_2 = N_2/V$ were given in Eq. (3.84):

$$\frac{dn_1}{dt} = -\sigma_{op}\Phi n_1 + \sigma_{op}\Phi n_2 - \frac{n_1}{\tau_1} + \frac{n_2}{\tau_{21}}$$

$$\frac{dn_2}{dt} = -\sigma_{op}\Phi n_2 + \sigma_{op}\Phi n_1 - \frac{n_2}{\tau_2} + R_2 \tag{4.1}$$

We note certain changes, however, between (3.84) and (4.1). First, Eq. (3.84) was obtained within the context of a uniquely two-level system. In that case, n_1 and n_2 could not tend towards any other possible values than those imposed by thermodynamic equilibrium between the two levels. No limitations of the sort appear in (4.1), as the occurrence of other levels relaxes this constraint. We further recall that σ_{op} is the optical cross-section (in cm^2) given by (3.81a) or (3.81b), related to the absorption α (cm^{-1}) by $\alpha = \sigma_{op}(n_1 - n_2)$ and to the gain γ by $\gamma = \sigma_{op}(n_2 - n_1)$, and where Φ is the photon flux (in $cm^{-2}\,s^{-1}$), and τ_2 is the net lifetime in level $|2\rangle$ given by:

$$\frac{1}{\tau_2} = \frac{1}{\tau_{21}} + \frac{1}{\tau_{20}} + \frac{1}{t_{spon}} \tag{4.2}$$

As already indicated, we supposed that levels $|1\rangle$ and $|2\rangle$ are sufficiently far apart in energy from lower lying levels to be able to neglect thermal population of these levels under unpumped conditions (i.e. $n_1^{eq} = n_2^{eq} = 0$). In the absence of any photon flux $h\nu_{21}$ ($\Phi = 0$ whereby we speak of a *cold cavity*), we obtain the stationary solution for Eq. (4.1) by setting $dn_1/dt = dn_2/dt = 0$. We then find the population difference between the upper and lower levels n_{d0} in the absence of a photon flux to be:

$$n_{d0} = n_2 - n_1 = R_2\tau_2\left(1 - \frac{\tau_1}{\tau_{21}}\right) \tag{4.3}$$

We see that if the lifetime of the lower level τ_1 is shorter than the recombination time τ_{21} we may achieve *population inversion*. This term designates that pumping leads, in this case, to a situation far from that allowed under thermodynamic equilibrium, where $n_2/n_1 = \exp(-E_{12}/kT) < 1$.

Some authors speak of *negative temperature*; however, this concept is of little practical use. We may additionally note, that the shorter τ_1 becomes in comparison with the recombination time τ_{21}, the greater the density of inverted carriers will be. Correspondingly, a greater degree of optical amplification in the system will be possible and may be attained with greater ease.

4.2.2 Optical amplification and gain saturation

We saw in Eqs. (3.50) and (3.74) that, under circumstances of population inversion, an optical signal close to resonance is amplified. In this case, its intensity increases as it travels through the medium and varies as:

$$I(z) = I_0 e^{\gamma(\nu)z} \tag{4.4}$$

where $\gamma(\nu)$ is the amplification coefficient (in cm^{-1}) given by Eq. (3.74):

$$\gamma(\nu) = \sigma_{\text{op}} n_d = \frac{\lambda^2}{8\pi\varepsilon_R t_{\text{spon}}} g(\nu) n_d \tag{4.5}$$

Optical gain and population inversion

where $n_d = n_2 - n_1$ is now the population difference in the presence of a photon flux and $g(\nu)$ is the lineshape described in Complement 3.A. We recall that $\varepsilon_R = n_{\text{op}}^2$ is the relative permittivity of the host medium, λ is the vacuum wavelength, and t_{spon} is the spontaneous lifetime given by (3.82). Equation (4.4) is only valid if the amplification factor γ is independent of position (this is generally not the situation and is discussed further in Complement 4.B). We obtain the stationary state value of n_d with the help of Eq. (4.1), this time by considering that $\Phi \neq 0$. The inversion population is then found to be:

$$n_d = \frac{n_{d0}}{1 + \tau_{\text{sat}}\sigma_{\text{op}}\Phi} \tag{4.6}$$

where τ_{sat} is the optical saturation time constant given by:

$$\tau_{\text{sat}} = \tau_2 + \tau_1\left(1 - \frac{\tau_2}{\tau_{21}}\right) \tag{4.7}$$

Equation (4.6) may also be written as:

$$n_d = \frac{n_{d0}}{1 + \Phi/\Phi_{\text{sat}}} \tag{4.8}$$

Population inversion and saturation flux

where Φ_{sat} is the saturation photon flux given by:

$$\Phi_{\text{sat}} = \frac{1}{\sigma_{\text{op}}(\nu)\tau_{\text{sat}}} = \frac{8\pi\varepsilon_R}{\lambda^2}\frac{t_{\text{spon}}}{\tau_{\text{sat}}}\frac{1}{g(\nu)} \tag{4.9}$$

Saturation flux of a laser medium

This quantity is very similar to the one introduced in (3.56). We therefore immediately deduce the variation of gain in the medium as a function of photon flux to be:

$$\gamma(\nu) = \frac{\gamma_0(\nu)}{1 + \Phi/\Phi_{\text{sat}}(\nu)} \tag{4.10}$$

Saturation of laser gain

where γ_0 is the low flux or *cold cavity* amplification given by:

$$\gamma_0(\nu) = n_{d0}\sigma_{\text{op}}(\nu) = n_{d0}\frac{\lambda^2}{8\pi\varepsilon_R t_{\text{spon}}}g(\nu) \tag{4.11}$$

Cold cavity gain for a laser medium

Figure 4.2 shows the relative variation γ/γ_0 of the amplification coefficient in the medium as a function of the normalized optical flux Φ/Φ_{sat}. We note that the gain decreases as the photon flux increases. This mechanism is called, for obvious reasons, *optical gain saturation* and results from progressive equalization of the populations between the levels providing the amplification, once the absorption and stimulated emission processes become dominant over all others.

Example

Table 4.1 gives the characteristics for the principal transitions used in solid state lasers. The neodymium–YAG laser operates chiefly using an available transition at 1.064 μm. The measured spontaneous lifetime $t_{\text{spon}} = 1.2$ ms and the transition width is 120 GHz. At resonance, the optical cross-section is given by (3.73):

$$\sigma_{\text{op}} = \frac{\lambda^2}{8\pi n^2 t_{\text{spon}}}\frac{1}{\pi\Gamma} = 4 \times 10^{-19}\,\text{cm}^2$$

If we suppose that the inversion population density at low flux n_{d0} is of the order of $10^{16}\,\text{cm}^{-3}$, we obtain a laser gain amplification γ_0 of:

$$\gamma_0 = n_{d0}\sigma_{\text{op}} = 4 \times 10^{-3}\,\text{cm}^{-1}$$

The corresponding gain along a 20 cm rod appears to be deceivingly small (0.08). Multiple passes of the amplified beam through the same medium, however, will

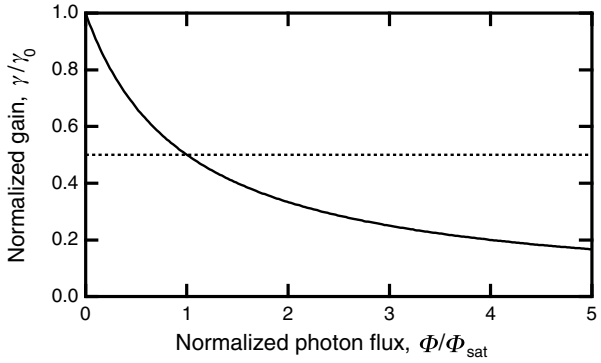


Fig. 4.2. Optical gain saturation for a two-level system as a function of normalized photon flux Φ/Φ_{sat}.

allow for significant overall levels of amplification to be achieved. Assuming $\tau_1 = 30\,\text{ns}$, $\tau_2 = t_{spon}$, and $\tau_{sat} \approx t_{spon}$ we obtain a saturation flux using (4.9) of:

$$\Phi_{sat} = \frac{8\pi n^2}{\lambda^2}\pi\Gamma = 2.1 \times 10^{21}\,\text{cm}^{-2}\,\text{s}^{-1}$$

This value corresponds to an optical power density of $390\,\text{W cm}^{-2}$. For such a photon flux (relatively low), the gain is reduced by a factor of 2.

Table 4.1. *Emission wavelengths* (λ_{em}), *optical cross-sections* (σ_{op}), *spontaneous lifetimes* (t_{spon}), *linewidths* ($\Delta\nu$), *and optical indices* (n_{op}) *for technologically significant laser media.*

Laser medium	λ_{em} (μm)	σ_{op} (cm^2)	t_{spon}	$\Delta\nu$	n_{op}
Nd^{3+}:YAG	1.064	4×10^{-19}	1.2 ms	120 GHz	1.82
Er^{3+}:SiO_2	1.55	5×10^{-21}	10 ms	4 THz	1.45
Ti^{3+}:Al_2O_3	0.66–1.18	3×10^{-19}	3 μs	100 THz	1.8
Ar^+	0.515	3×10^{-12}	10 ns	3.5 GHz	1

4.3 Three- and four-level systems

We will now refine our discussion of the pumping mechanisms employed in Section 4.2 to create the conditions for optical gain between two levels. Explicitly, we shall consider systems comprising three and four levels altogether. Equation (4.1) needs to be slightly modified to take into account the different recombination mechanisms at work in such systems. We seek in these two cases, expressions for *cold cavity inversion populations*, i.e. when the density of amplified photons $h\nu_{12}$ is

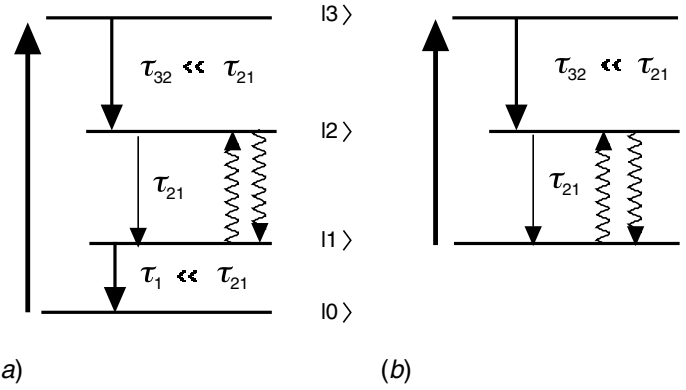


Fig. 4.3. Dynamic populations in a four-level (*a*) and three-level (*b*) system.

zero in the cavity ($\Phi_{12} = 0$). Expressions (4.7) and (4.10) then remain valid once the inversion populations n_{a0} are known.

Figure 4.3*a* represents, schematically, the effect of pumping and subsequent carrier absorption and relaxation mechanisms occurring in a four-level system. Level $|3\rangle$ is populated by electrons excited from level $|0\rangle$ and it is the ensuing relaxation of carriers $|3\rangle \rightarrow |2\rangle$ which is responsible for feeding the population inversion in level $|2\rangle$. The population in level $|3\rangle$ is given by:

$$\frac{\mathrm{d}n_3}{\mathrm{d}t} = R_3 - \frac{n_3}{\tau_{32}} \tag{4.12}$$

where R_3 results from the optical pumping of carriers from level $|0\rangle$:

$$R_3 = W_{03}(n_0 - n_3) = \sigma_{\mathrm{op}}^{03}\Phi_{03}(n_0 - n_3) \tag{4.13}$$

Φ_{03} is the photon flux associated with the pump beam. At the stationary state, the carrier density in level $|3\rangle$ is:

$$n_3 = \frac{\sigma_{\mathrm{op}}^{03}\Phi_{03}\tau_{32}}{1 + \sigma_{\mathrm{op}}^{03}\Phi_{03}\tau_{32}} n_0 \tag{4.14}$$

As τ_{32} is very small, we expect absorption saturation only under very high pump conditions. The stationary-state occupation of states $|2\rangle$ and $|1\rangle$ is given by:

$$\begin{aligned} -\frac{n_2}{\tau_{21}} + \frac{n_3}{\tau_{32}} &= 0 \\ -\frac{n_1}{\tau_1} + \frac{n_2}{\tau_{21}} &= 0 \end{aligned} \tag{4.15a}$$

in which we have neglected transitions of the type $|3\rangle \rightarrow |1\rangle$ or $|2\rangle \rightarrow |0\rangle$. (Contributions from such processes could clearly be taken into account, but only at the

cost of obscuring our illustration of the more relevant physical concepts.) To these three equations (i.e. (4.14) and (4.15a)), we add the conservation law for the total number of particles:

$$n_0 + n_1 + n_2 + n_3 = n \tag{4.15b}$$

This leads in simple fashion to an expression for the inversion population:

$$n_{d0} = \frac{\sigma_{op}^{03}\Phi_{03}\tau_{21}}{1 + \sigma_{op}^{03}\Phi_{03}(\tau_{21} + 2\tau_{32})}n \tag{4.16}$$

assuming the depopulation mechanism from level $|1\rangle$ to be very rapid in comparison with others (and reflecting the desired situation). We note, as expected, that as the photon flux of the pump beam increases, the inversion saturates at $\tau_{21}/(\tau_{21} + 2\tau_{32})n$, which is close to the total population n of the system ($\tau_{32} \ll \tau_{21}$).

A three-level system is also represented in Fig. 4.3*b*. In this case, the ground state of the system also corresponds to the lower transition level associated with the amplification mechanism. The calculation proceeds as before, allowing us to find:

$$n_{d0} = \frac{\sigma_{op}^{13}\Phi_{13}(\tau_{21} - \tau_{32}) - 1}{1 + \sigma_{op}^{13}\Phi_{13}(\tau_{21} + 2\tau_{32})}n \approx \frac{\sigma_{op}^{13}\Phi_{13}\tau_{21} - 1}{1 + \sigma_{op}^{13}\Phi_{13}\tau_{21}}n \tag{4.17}$$

where Φ_{13} now refers to the pump photon flux. Comparing (4.16) with (4.17), we note that three-level systems are systematically more difficult to invert than four-level systems. This is because in a three-level system, level $|1\rangle$ cannot empty itself by spontaneous carrier relaxation to a lower energy level, but rather must rely upon photon absorption from the pump beam. Figure 4.4 shows the achieved inversion densities n_{d0}/n in the three- and four-level systems as a function of the normalized pump photon flux Φ/Φ_{sat} (where Φ_{sat} refers to the saturation photon flux $1/\sigma_{op}\tau_{sat}$ for transitions between the base level $|0\rangle$ or $|1\rangle$ and the pump level $|3\rangle$). Figure 4.4 also shows how, in a four-level system, population inversion is obtained for all pump photon flux values. In a three-level system, inversion only results beyond a certain pump photon flux threshold corresponding to the *transparency condition.*

Example

Of particular importance to telecommunications technologies interested in light signal propagation at 1.55 μm is the erbium-doped fibre optic laser. In this case, Er atoms are introduced into a silicon dioxide optical fibre ($Er^{3+}:SiO_2$) forming a three-level laser gain medium. Characteristics for the $Er^{3+}:SiO_2$ system are presented in Table 4.1, to which we add the optical cross-section for the relevant pump transition $\sigma_{op}^{13} = 2 \times 10^{-21}\,cm^2$ and the pump photon wavelength $\hbar\omega_{13} = 1.26\,eV$ ($\lambda_{pump} = 0.98\,\mu m$). The transparency condition that must be reached before achieving optical gain occurs at:

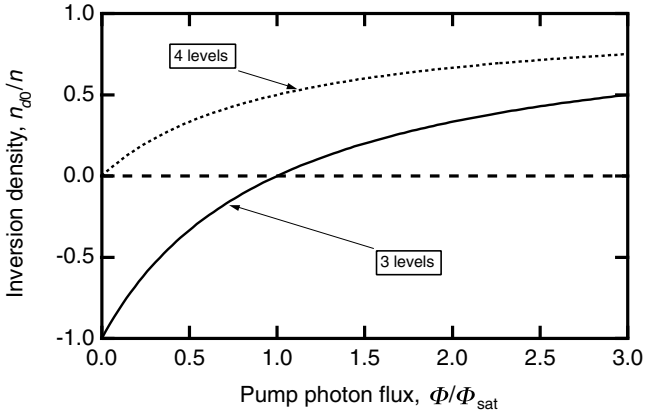


Fig. 4.4. Comparison of induced population inversion densities as a function of pump photon flux in three- and four-level systems.

$$\Phi_{\text{sat}}^{13} = 1/(2 \times 10^{-21}\,\text{cm}^2 \times 10^{-2}\,\text{s}) = 5 \times 10^{22}\,\text{cm}^2\,\text{s}^{-1}$$

or

$$P_{\text{transparency}} = 5 \times 10^{22}\,\text{cm}^2\,\text{s}^{-1} \times 1.26\,\text{eV} \times 1.6 \times 10^{-19} \approx 10^4\,\text{W}\,\text{cm}^{-2}$$

For a 10-μm diameter optical fibre, this represents a required input power of 8 mW to obtain population inversion and achieve the transparency condition.

4.4 Optical resonators and laser threshold

A laser material contains an inverted population and as such may act as an amplifier. Therefore, we need only reinject the amplified optical signal (referred to as *optical feedback*) to induce laser oscillation. We now describe the conditions that must be satisfied to obtain this laser oscillation. We consider a laser medium of length d, with an entrance surface M_e, and an exit surface M_s covered by reflecting mirror surfaces having transmission and reflection coefficients t_e and r_e for the entrance mirror, and t_s and r_s for the exit mirror (see Fig. 4.5). We note that all the reflection coefficients r may be written as $\sqrt{R}\,e^{i\phi}$ where R is the reflectance and ϕ the optical phase shift induced by the mirror. Additionally, we may suppose that this medium 'loses' photons by parasitic absorption, diffusion, diffraction, etc. These mechanisms are then taken into account by a single parasitic attenuation α_p (in cm^{-1}) which opposes the optical gain. We will now follow the plight of an electromagnetic wave $E = E_0 \exp i(\omega t - kx)$ leaving the mirror M_e and heading towards the exit mirror M_s. Leaving M_e, the wave is described by:

$$E_1 = E_0 t_e e^{i\omega t} \tag{4.18a}$$

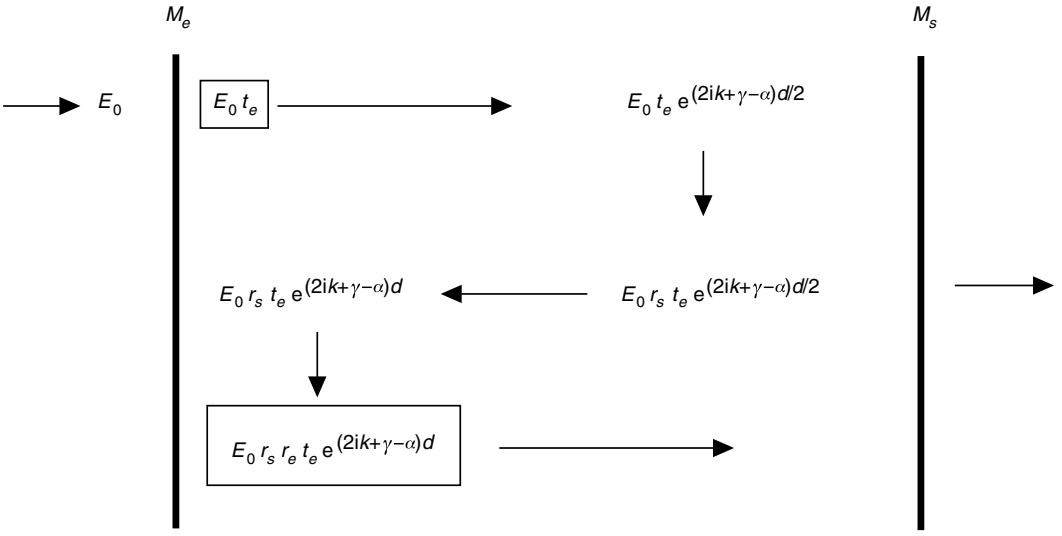


Fig. 4.5. Mechanisms leading to laser oscillation in a cavity. The electric field at the entrance mirror, corresponding to the wave moving towards the right, results from the sum of all the contributions of the electromagnetic fields depicted in the figure. The phenomenon of laser oscillation is seen to involve both a dependence on the phase and the amplitude of the waves.

Arriving at M_s, the power of the signal has been amplified by a factor of $\exp(\gamma - \alpha_p)d$, whereas the amplitude has increased by a factor of $\exp(\gamma - \alpha_p)d/2$ and been multiplied by a phase shift of $\exp(ikd)$. A fraction (r_s) of the wave amplitude is reflected and retransmitted through the medium, receiving a further amplification of $\exp(\gamma - \alpha_p)d/2$ and a phase shift of $\exp(ikd)$, with fraction r_e being re-reflected by the entrance mirror. The component of the electromagnetic field after this single round-trip between both mirrors may be written:

$$E_2 = E_0 t_e r_e r_s e^{i\omega t} e^{(\gamma-\alpha_p)d} e^{i2kd} \tag{4.18b}$$

The signal at the entrance mirror is the superposition of the fields of all the waves going back and forth. It is then given by the sum of $E = E_1 + E_2 + E_3 + \cdots$ or:

$$E = E_0 t_e e^{i\omega t}[1 + r_e r_s e^{(\gamma-\alpha_p)d} e^{i2kd} + (r_e r_s e^{(\gamma-\alpha_p)d} e^{i2kd})^2 + \cdots] \tag{4.19}$$

The summation is trivial and leads to:

$$E = E_0 t_e \frac{e^{i\omega t}}{1 - r_e r_s e^{(\gamma-\alpha_p)d} e^{i2kd}} \tag{4.20}$$

The electromagnetic field diverges when the denominator in (4.20) goes to zero. When achieved, this condition refers to the phenomenon of *laser oscillation*. This

condition could also have been obtained by requiring that in the stationary state, the two expressions for the field at $x = 0$ ($E_0 t_e$ and $E_0 t_e r_s r_e e^{(2ik+\gamma-\alpha)d}$) be equal.

We see, therefore, that two conditions must be met before laser oscillations may be established in a cavity.

Gain condition

The gain in the amplification medium must exceed the various cavity losses (e.g. light transmission by mirrors, diffusion, . . .). This condition is represented by the following inequality:

$$|r_e r_s| e^{(\gamma-\alpha_p)d} > 1$$

There is thus a corresponding gain threshold above which the medium begins to oscillate spontaneously. This threshold is given by:

$$\gamma_{\text{threshold}} = \alpha_p - \frac{1}{d}\ln|r_e r_s| = \alpha_t \tag{4.21a}$$

where α_t is the total attenuation coefficient. Alternatively, we may write (4.21a) as:

$$\gamma_{\text{threshold}} = \alpha_p - \frac{1}{2d}\ln R_e R_s \tag{4.21b}$$

Optical gain and laser oscillation threshold

where R_e and R_s refer to the mirror reflectances. We note that the gain is *clamped* to this value, as (4.20) shows that this value cannot be surpassed. Equation (4.21b) may be restated in terms of a corresponding population inversion density with the help of (4.11):

$$n_{\text{threshold}} = \frac{8\pi\varepsilon_R t_{\text{spon}}}{\lambda^2 g(\nu)}\left(\alpha_p - \frac{1}{2d}\ln R_e R_s\right) \tag{4.22a}$$

or at the peak of the Lorentzian:

$$n_{\text{threshold}} = \frac{4\pi^2\varepsilon_R t_{\text{spon}}\Delta\nu}{\lambda^2}\left(\alpha_p - \frac{1}{2d}\ln R_e R_s\right) \tag{4.22b}$$

This equation may also be written in a more illustrative fashion by introducing the concept of *photon lifetime*. Let us imagine a photon moving to and fro within a cavity at speed $c' = c/n_{\text{op}}$. The exit probability for the photon at any particular cycle through the cavity is then given by $d\alpha_p - \ln(r_e r_s)$. This probability may also be considered as a ratio between the cavity length d *and the mean path* λ_c *of the photon before escaping from the cavity*. This mean path is the product of the photon propagation speed $c' = c/n_{\text{op}}$ and the photon lifetime in the cavity τ_c, which is then given by:

$$\tau_c = \frac{1}{\left(\alpha_p - \frac{1}{2d}\ln R_e R_s\right)c'} \tag{4.23a}$$

Photon lifetime in a cavity

or again supposing the parasitic absorption α_p is weak, $R_e = 1$ and $R_s = 1 - T_s$ with the transmittance of the exit mirror $T_s \ll 1$:

$$\tau_c = \frac{2d}{T_s c'} = \frac{2}{T_s}\tau_{\text{light}} \tag{4.23b}$$

τ_{light} is the time required for the photon to travel the distance between the two mirrors. Equation (4.23b) is a very illustrative and useful relation for the laser engineer. It directly shows how the photon round-trip time in the cavity $2\tau_{\text{light}}$ is enhanced by the low exit mirror transmittance T_s. Equation (4.22) may then be written:

$$n_{\text{threshold}} = \frac{1}{c'\sigma_{\text{op}}\tau_c} = \frac{4\pi^2\varepsilon_R\Delta\nu}{\lambda^2 c'}\frac{t_{\text{spon}}}{\tau_c} \tag{4.24}$$

Inversion population density at laser oscillation threshold

In the case of optical pumping, the required threshold power for the pump beam is that necessary to achieve population inversion (Eqs. (4.16) and (4.17)).

Phase condition

The phase condition that leads to a value of zero in the denominator of (4.20) is:

$$kd + \phi = q\pi,\ q = 1, 2, \ldots \tag{4.25a}$$

where ϕ is the average of the phase shifts introduced by the two mirrors. If we suppose both mirrors to be metallic ($\phi = \pi$), the amplified modes will be given by those modes:

$$\nu_q = q\frac{c}{2n_{\text{op}}d} \tag{4.25b}$$

which fall within the gain spectrum of the amplification medium (see Fig. 4.9). We shall return to this topic in Section 4.6.

Example

1. In the case of a Nd^{3+}:YAG laser, the emission wavelength is 1.064 μm, the width of the gain spectrum is 120 GHz, the index of refraction is 1.82, the spontaneous lifetime is 1.2 ms, and we consider a photon lifetime in the cavity of 1 ns (Table 4.1). The inversion density at threshold is then given by

(4.24) and is $n_{\text{threshold}} = 1/(4 \times 10^{-19}\,\text{cm}^2 \times 10^{-9}\,\text{s} \times 1.65 \times 10^{10}\,\text{cm s}^{-1})$ or $1.5 \times 10^{17}\,\text{cm}^{-3}$.

2. We consider a 10 m section of an erbium-doped silica fibre with mirrors at both ends. One mirror has zero transmittance, while the other has a 1% transmittance. The photon lifetime in the fibre (neglecting parasitic absorption) is $\tau_c = 2 \times 10\,\text{m}/(10^{-2} \times 3 \times 10^8\,\text{m s}^{-1}/1.45)$ or 9.6 μs. The inversion population density is then obtained by using the optical cross-section given in Table 4.1 ($\sigma_{\text{op}} = 5 \times 10^{-21}\,\text{cm}^2$):

$$n_{\text{threshold}} = 1/[(3 \times 10^{10}\,\text{cm s}^{-1}/1.45) \times 5 \times 10^{-21}\,\text{cm}^2 \times 9.6 \times 10^{-6}\,\text{s}]$$

or

$$= 1.0 \times 10^{15}\,\text{cm}^{-3}$$

4.5 Laser characteristics

4.5.1 Internal laser characteristics and gain clamping

We consider a laser system undergoing optical pumping at a growing rate (i.e. pumping rate R_2 increasing). In the stationary state (i.e. before the onset of laser oscillation), the cold cavity inversion population density is given by n_{d0}, Eqs. (4.16) or (4.17), depending on whether we are dealing with a three- or four-level system. The gain in the medium $\gamma_0(\nu)$ is that given by (4.11). As long as the system remains below the oscillation threshold, the inversion density is proportional to the pumping rate R_2, as indicated by (4.3). Let us now suppose that the system is pumped beyond laser threshold. The system then begins to oscillate and the amplified photon flux within the cavity increases, saturating the inversion density and reducing the gain (see Section 4.1 and Fig. 4.2). Figure 4.6 shows the progression of this effect. The decrease in gain stops once the gain in the medium $\gamma(\nu)$ exactly balances out the cavity losses α_t resulting from the parasitic losses and the mirrors (Eq. (4.22)). The stationary state of the cavity is then given by the *clamped gain*, i.e. the condition for which $\gamma(\nu) = \alpha_t$. As a result, the inversion density is also clamped to its threshold value $n_{\text{threshold}}$. Evolution of the inversion density as a function of pump rate is shown in Fig. 4.7.

Taking the expression for saturated gain $\gamma(\nu) = \gamma_0(\nu)/(1 + \Phi/\Phi_{\text{sat}})$, we obtain as an expression for laser photon flux in the cavity at the stationary state:

$$\Phi = \begin{cases} 0, \text{ if } \gamma_0(\nu) < \alpha_t \\ \Phi_{\text{sat}}(\nu)\left(\dfrac{\gamma_0(\nu)}{\alpha_t} - 1\right), \text{ if } \gamma_0(\nu) > \alpha_t \end{cases} \tag{4.26a}$$

Note that the flux Φ refers *to photons propagating in both directions along the*

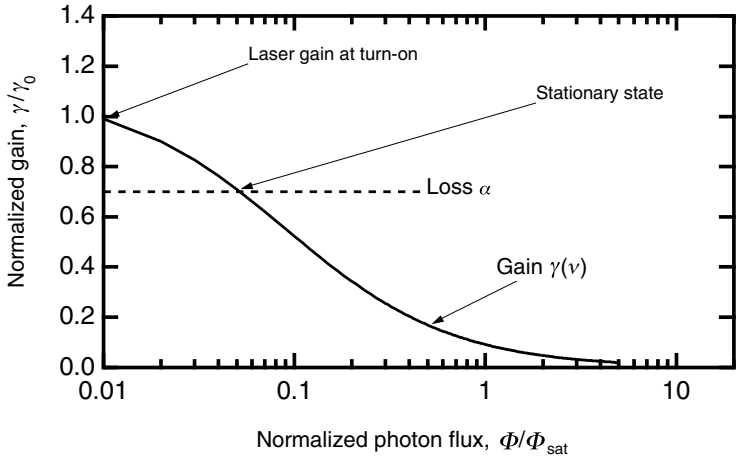


Fig. 4.6. *Clamping* mechanism for lasers. At the onset of pumping, the number of photons in the cavity is low and the gain medium is unsaturated. The photon density in the cavity gradually increases until the gain in the medium reaches the optical saturation level, which exactly balances out the cavity losses.

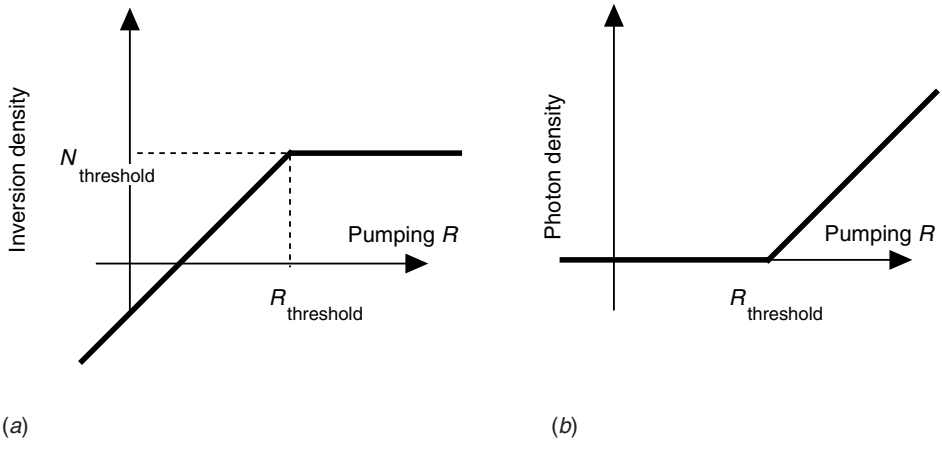


Fig. 4.7. Variation in the inversion and photon densities in a cavity as a function of pump rate R. We note that once laser threshold is achieved, the inversion population *clamps* to $N_{\text{threshold}}$, while the photon density increases linearly with optical pumping. Above threshold, all additional pump energy is ideally converted into an amplified optical signal.

cavity. As $\gamma_0 = n_{d0}\sigma_{op}$, and given the definition of the laser oscillation threshold $\alpha_t = \gamma_{\text{threshold}} = n_{\text{threshold}}\sigma_{op}$, (4.26a) may also be written:

$$\Phi = \begin{cases} 0, \text{ if } n_{d0} < n_{\text{threshold}} \\ \Phi_{sat}(\nu)\left(\dfrac{n_{d0}}{n_{\text{threshold}}} - 1\right), \text{ if } n_{d0} > n_{\text{threshold}} \end{cases} \tag{4.26b}$$

It is interesting to write this last equation in terms of the photon density and optical pumping power. To do so, we introduce the quantities:

$$p_{\text{sat}} = \frac{\Phi_{\text{sat}}}{c'} = \frac{1}{c'\sigma_{\text{op}}\tau_{\text{sat}}} \text{ and } R_{\text{threshold}} = \frac{n_{\text{threshold}}}{\tau_2} \tag{4.27}$$

which allow us to rewrite (4.26) as:

$$p = \begin{cases} 0, \text{ if } R_2 < R_{\text{threshold}} \\ p_{\text{sat}}\left(\dfrac{R_2}{R_{\text{threshold}}} - 1\right), \text{ if } R_2 > R_{\text{threshold}} \end{cases} \tag{4.28a}$$

Photon density as a function of pump rate

The variation in photon flux as a function of pump rate is also shown schematically in Fig. 4.7. This last figure illustrates several of the particularities of laser emission: the existence of a threshold, clamping of inversion population density, gain, and linear increase in the photon density above threshold. This approach is greatly simplified as we have notably neglected spontaneous emission. A more complete approach will be presented in Complement 4.A.

Finally, let us note that far above threshold, the number of photons in the cavity is given by:

$$p \approx p_{\text{sat}} \frac{n_{d0}}{n_{\text{threshold}}} = \frac{\tau_c}{\tau_{\text{sat}}} n_{d0} \approx \frac{\tau_c}{\tau_2} n_{d0} = R_2 \tau_c \tag{4.28b}$$

This last equation clearly shows the role played by the cavity. As a result of the optical resonator, the cavity is able to store electrons on excited states, eventually yielding for each electron a photon quantity τ_c/τ_2. These various concepts are illustrated again in Complement 4.E in discussing the operation of diode pumped lasers.

4.5.2 Output power

For the moment, we are interested in internal densities and fluxes in the laser medium. The output flux is the light intensity emitted through the output mirror M_s with transmittance T_s. The smaller the value of T_s, the greater the reflectance ($R_s \approx 1 - T_s$) and the lower the laser threshold. On the other hand, if the transmittance is null, no light can emerge from the laser cavity! Therefore, given both these considerations, there is an optimal value for the reflectance. The output flux from the laser Φ_{ext} is found simply by taking *half* the internal flux (4.26) (to take into account only that flux instantaneously directed towards the output mirror), multiplied by the mirror transmittance yielding:

$$\Phi_{\text{ext}} = \frac{1}{2}\Phi_{\text{sat}} T_s \left(\frac{2d\gamma_0}{2d\alpha_p - \ln R_e - \ln(1 - T_s)} - 1\right) \tag{4.29a}$$

Figure 4.8 shows the variation of the output flux Φ_{ext} as a function of T_s. This plot

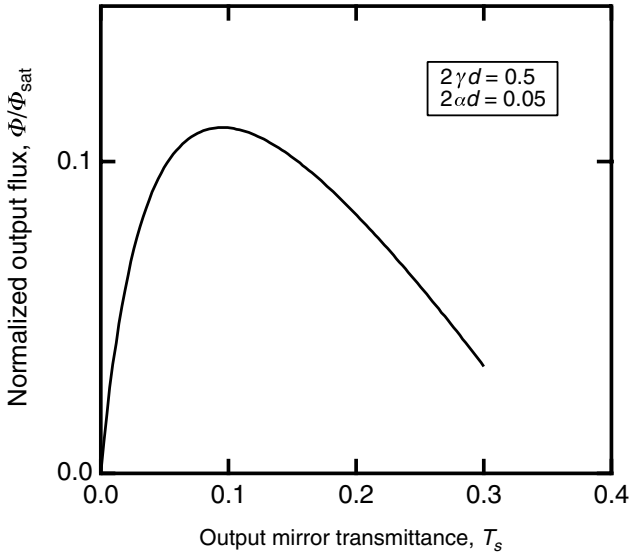


Fig. 4.8. External photon flux as a function of output mirror transmittance for a laser cavity with a total gain $2\gamma d$ of 0.5 and a loss $2\alpha d$ of 0.05. We note the existence of an optimal value for the output mirror transmittance.

has a maximum value for T_{op} given by ($\ln(1 - T) \approx -T$):

$$T_s = \sqrt{2\gamma_0 d}\sqrt{2\alpha_p d - \ln R_e} - (2\alpha_p d - \ln R_e) \tag{4.29b}$$

Another useful result is the relationship between the number of photons in the cavity P, and the external power $\mathscr{P}$ in the case where (4.23b) applies ($R_e = 1$, $T_s \ll 1$, $\alpha_p = 0$). For this, let us consider a time interval Δt. During this time a portion $c'\Delta t/d$ of the total photon population P in the cavity is incident on the exit mirror (d is the cavity length). Only half of the photons in the population have the correct propagation direction. Moreover, only a fraction T_s of these photons, each carrying an energy $\hbar\omega$, will succeed in leaving the cavity, thus emitting a laser energy of $\mathscr{P}\Delta t$, so that:

$$\mathscr{P}\Delta t = \frac{1}{2}\frac{c'\Delta t}{d} P T_s \hbar\omega \tag{4.30a}$$

or

$$\mathscr{P} = \frac{1}{2} T_s c' \frac{P}{d}\hbar\omega = P\frac{\hbar\omega}{\tau_c} \tag{4.30b}$$

Output power and number of photons in a laser cavity

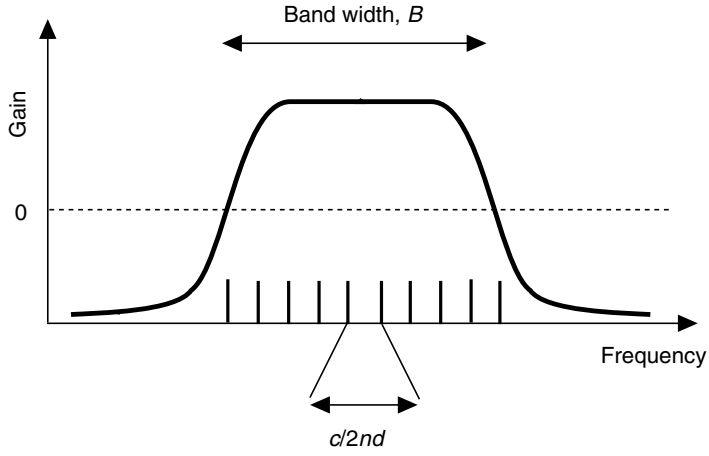


Fig. 4.9. The maximum number of modes which may be supported in a cavity is given by the ratio of the spectral bandwidth to the modal frequency spacing.

Example

1. For a total medium gain $2\gamma d$ of 0.5, neglecting any transmission from the entrance mirror ($R_e = 1$) and assuming a total parasitic loss $2\alpha d$ of $0.05\,\text{cm}^{-1}$, using (4.29b) we find an optimal transmittance of 10%.
2. We consider a gas cavity laser 1 m in length, fitted with an output mirror with a transmittance of 0.5%. The photon lifetime, given by (4.23b), is found to be $\tau_c = 2 \times 10^2\,\text{cm}/(5 \times 10^{-3} \times 3 \times 10^{10}\,\text{cm s}^{-1}) = 1.3\,\mu\text{s}$. The number of photons with an energy of 1 eV in this cavity and corresponding to an output power of 1 W is given by (4.30), or $P = 1\,\text{W} \times 1.3 \times 10^{-6}\,\text{s}/1.6 \times 10^{-19}\,\text{J} = 8 \times 10^{12}$ photons.

4.5.3 Spectral characteristics

Only those cavity eigenfrequencies satisfying the phase condition (4.25) and which have a cold cavity gain $\gamma_0(v_q)$ greater than the threshold gain ($\gamma_0(v_q) > \gamma_{\text{threshold}}$) are oscillating. Figure 4.9 shows the oscillating modes allowed within the cavity. As the frequency spacing between each mode is $c/2nd$, the maximum number of laser modes is then given by the ratio of the gain bandwidth divided by the mode spacing (see Fig. 4.9), yielding:

$$N_{\text{mode}} = \frac{B}{c/2nd} \tag{4.31}$$

The actual number of observed modes depends, additionally, on the nature of the broadening of the absorption transition lineshape in any given medium (see Complement 3.A).

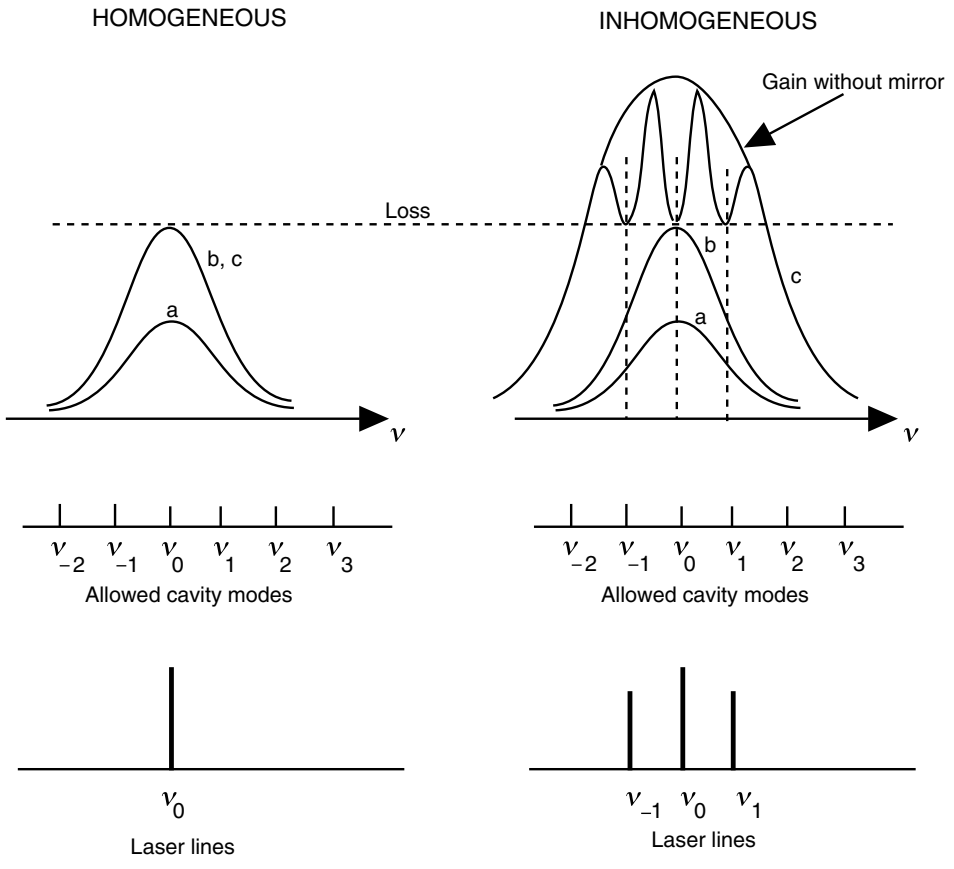


Fig. 4.10. Gain spectra and corresponding laser lines resulting from amplification in homogeneous and inhomogeneous media below threshold (a), at threshold (b), and above threshold (c).

Homogeneous gain spectrum (see Complement 3.A)

In this case, broadening is due to phase relaxation mechanisms (represented by $\Delta\nu$ in the Lorentzian). The lineshape is monolithic, in the sense that the saturation mechanisms act upon the entire absorption spectrum and may be written from (4.10) as:

$$\gamma(\nu) = \frac{\lambda^2}{8\pi\varepsilon_R t_{\text{spon}}} \frac{1}{1 + \Phi/\Phi_{\text{sat}}(\nu)} \frac{\Delta\nu/2\pi}{(\nu - \nu_0)^2 + (\Delta\nu/2)^2} \tag{4.32}$$

The spectral response of the laser amplification is depicted in Fig. 4.10. We suppose that at $t = 0$, the system is pumped far above laser threshold. All allowed modes ν_q having a gain greater than the threshold requirement begin to lase. As all the modes begin to be amplified, saturation effects tend to decrease the gain

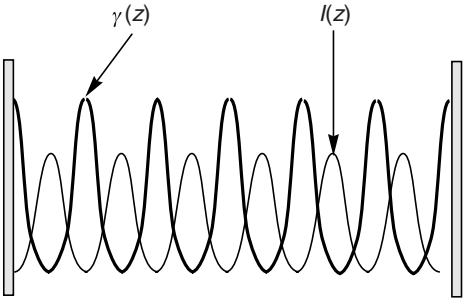


Fig. 4.11. The gain experienced by a stationary mode depends on its position across the cavity. The overall gain curve is thereby broadened in an inhomogeneous fashion leading to multimode amplification even if the fundamental quantum transitions are themselves homogeneous in nature. The periodic minima induced in the gain curve across the cavity corresponds to the phenomenon of *spatial hole burning*.

available for each mode; the amplification envelope decreases uniformly in amplitude, while preserving its original spectral distribution. The overall amplitude continues to drop until only a single mode positioned nearest to the gain maximum may sustain amplification. Thus, laser oscillations established in a homogeneous medium will tend to result in *single mode* laser operation.

In practice, homogeneous mediums may often exhibit multimode laser amplification. Several parasitic phenomena may be responsible for this. Most often, the reason is that the amplitudes of the stationary waves in the cavity vary from one location to the next (as $\cos kz$ or $\sin kz$). Therefore, even a single mode experiences different levels of gain across the cavity even if the gain spectrum is homogeneous in nature (see Fig. 4.11). This situation refers to the phenomenon of *spatial hole burning*.

Inhomogeneous gain spectrum (see Complement 3.A)
This situation is relatively easy to understand (Fig. 4.10). In this case, the system is composed of an assembly of independent subsystems with each subsystem possessing its own set of associated rate equations and saturation characteristics. Each mode allowed by (4.25) will then correspond to a maximum in the gain spectrum over that frequency range, leading to independent amplification of each of these individual modes. Thus, laser oscillation in an inhomogeneous medium is fundamentally multimode in nature (see Fig. 4.10).

4.6 Cavity rate equations and the dynamic behaviour of lasers

The rate equations (4.1), introduced at the start of this chapter, are time-dependent differential equations. Up to this point, we have been chiefly interested in the

stationary state behaviour of laser emission. We now turn our attention to the dynamic behaviour of lasers. With this goal in mind, we return to the four-level system described in Fig. 4.3. We recall that level $|2\rangle$ is inverted by the optical pumping of carriers from level $|0\rangle$ to level $|3\rangle$. The recombination of electrons from $|3\rangle$ to $|2\rangle$ was assumed to be instantaneous, so that the feed rate of carriers into $|2\rangle$ was taken to be given by the pump rate of level $|3\rangle$. The dynamic equation describing the occupation of level $|2\rangle$ is then:

$$\frac{dn_2}{dt} = R - \frac{n_2}{\tau_2} - \sigma_{op}\Phi(n_2 - n_1) \tag{4.33}$$

We suppose the carrier lifetime in level $|1\rangle$ is sufficiently short so that the occupation of level $|1\rangle$ is effectively empty at all times, i.e. $n_d = n_2 - n_1 \approx n_2$. Equation (4.33) may be written in a more illustrative manner by introducing the cold cavity ($\Phi = 0$) inversion density given by $n_{d0} = R\tau_2$. The photon density p relates to the flux Φ by $p = \Phi/c'$, and the optical cross-section is given by (4.24), i.e. $\sigma_{op} = 1/c' n_{threshold}\tau_c$, so that:

$$\frac{dn_d}{dt} = \frac{n_{d0} - n_d}{\tau_2} - \frac{p}{\tau_c}\frac{n_d}{n_{threshold}} \tag{4.34}$$

The dynamic behaviour of the photon density p results from contributions from the stimulated emission in the system ($+n_d\sigma_{op}\Phi$) and the photon losses (diffusion, residual absorption, loss to mirrors, . . .) described by the photon lifetime (p/τ_c), so that taking into account the same change of variables as in (4.34):

$$\frac{dp}{dt} = -\frac{p}{\tau_c} + \frac{p}{\tau_c}\frac{n_d}{n_{threshold}} \tag{4.35}$$

Equations (4.34) and (4.35) form a pair of coupled non-linear differential equations (via the products pn_d in (4.34) and (4.35)). Although they do not take into account spontaneous emission (see Complement 4.A), they are of such importance that we shall rewrite them again side by side:

$$\frac{dn_d}{dt} = \frac{n_{d0} - n_d}{\tau_2} - \frac{p}{\tau_c}\frac{n_d}{n_{threshold}} \tag{4.34}$$

$$\frac{dp}{dt} = -\frac{p}{\tau_c} + \frac{p}{\tau_c}\frac{n_d}{n_{threshold}} \tag{4.35}$$

Dynamic coupled equations for a laser cavity

These types of non-linear equations lead to extremely complex and chaotic behaviour. As such, we will restrict ourselves to examining some of the simpler situations.

First, stationary solutions of these equations occur for:

- either $p = 0$ and $n_d = n_{d0}$, corresponding to the situation below threshold;
- or $p \neq 0$ and $n_d = n_{\text{threshold}}$ and $p = (n_{d0} - n_{\text{threshold}})\tau_2/\tau_c$, which is the situation above threshold with clamping.

4.6.1 Damped oscillations

We suppose that at time $t = 0$, the system is in an inverted state with n_{d0} being in excess of $n_{\text{threshold}}$ (i.e. before clamping – we shall see in the following section exactly how this occurs). We rewrite (4.34) and (4.35) and introduce the following dimensionless constants:

$$X = \frac{n_d}{n_{\text{threshold}}},\ Y = \frac{p}{n_{\text{threshold}}},\ X_0 = \frac{n_{d0}}{n_{\text{threshold}}},\ T = \frac{t}{\tau_c},\ u = \frac{\tau_c}{\tau_2} \tag{4.36}$$

These equations then become:

$$\begin{cases} \dfrac{\mathrm{d}Y}{\mathrm{d}T} = Y(X - 1) \\ \dfrac{\mathrm{d}X}{\mathrm{d}T} = u(X_0 - X) - XY \end{cases} \tag{4.37}$$

In order to describe the behaviour of the laser near equilibrium, we define the variables x and y which describe the excursion of the system from stationary behaviour:

$$\begin{cases} X = 1 + x \\ Y = u(X_0 - 1) + y \end{cases} \tag{4.38}$$

Neglecting the terms in xy (being of second order), (4.37) becomes:

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}T} = u(X_0 - 1)x \\ \dfrac{\mathrm{d}x}{\mathrm{d}T} = -uX_0x - y \end{cases} \tag{4.39}$$

This system of coupled linear differential equations is easily integrated. The system then behaves as a damped oscillator with an eigenfrequency $\omega_r = 2\pi/T_r$, and a relaxation time τ_r given by:

$$\tau_r = 2\frac{n_{\text{threshold}}}{n_{d0}}\tau_2$$

$$T_r = \frac{2\pi\sqrt{\tau_2\tau_c}}{\sqrt{(n_{d0}/n_{\text{threshold}}) - 1}} \tag{4.40}$$

Relaxation time and oscillation period for a laser cavity

Figure 4.12 shows the damped oscillations resulting from a starting carrier density

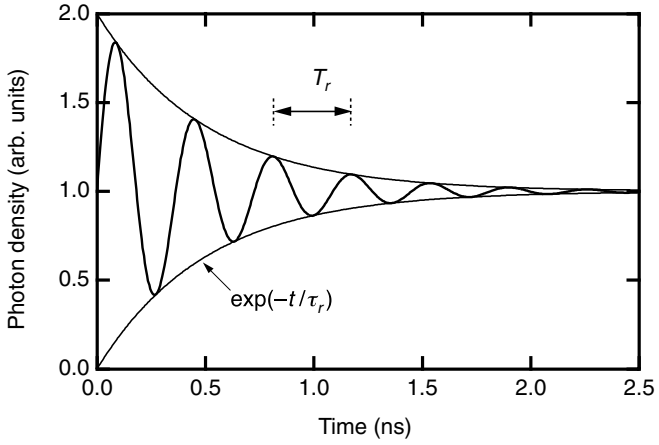


Fig. 4.12. Time evolution of the photon density in a cavity assuming a photon lifetime of 3 ps, and a carrier lifetime τ_2 in level $|2\rangle$ of 1 ns. At time $t = 0$, an inversion density n_{d0} twice that of the threshold value $n_{\text{threshold}}$ is assumed.

twice that of the threshold value, i.e. $n_{d0}/n_{\text{threshold}} = 2$. These oscillations result from the alternating energy exchange between the photons in the cavity and the carriers in the two-level system.

Example

We consider a semiconductor laser with a typical τ_2 of 1 ns and a photon lifetime of 3 ps. Assuming an initial inversion density of $n_{d0} = 2n_{\text{threshold}}$, we obtain an oscillation frequency of $T_r = 2\pi\sqrt{(3 \times 10^{-12} \times 1 \times 10^{-9})}/\sqrt{(2-1)}$ or 0.34 ns. The relaxation time is then $\tau_r = 2 \times 10^{-9}/2$ or 1 ns, as shown in Fig. 4.12.

4.6.2 Laser cavity dumping by loss modulation (*Q*-switching)

We will now describe a method enabling one to exceed the threshold carrier density in a laser (allowing $n_{d0} \gg n_{\text{threshold}}$). To do so, during an initial period, the cavity losses α_c are artificially increased to maintain cold cavity conditions. The density of electrons occupying level $|2\rangle$ is then given by $n_2(t) = R\tau_2(1 - e^{-t/\tau_2})$. The length of the period T_a is made sufficiently large so that the majority of available states $|2\rangle$ become occupied (i.e. $T_a \approx n/R$, where n is the total density of emission centres). To introduce these large losses temporarily, the reflectivity of the exit mirror can be considerably reduced by means of a Pockels cell (see Chapter 12). As a result, the quality factor Q of the cavity can be modulated or *Q-switched*. The inversion population density $n_i \approx n$ (i for initial) in level $|2\rangle$ at the end of the first low-Q cycle is very large in comparison to the clamped inversion density $n_{\text{threshold}}$ during laser emission (see Fig. 4.13). The same is true for the energy stored by the occupied states. The cavity is then switched to a high-Q value (by restoring the

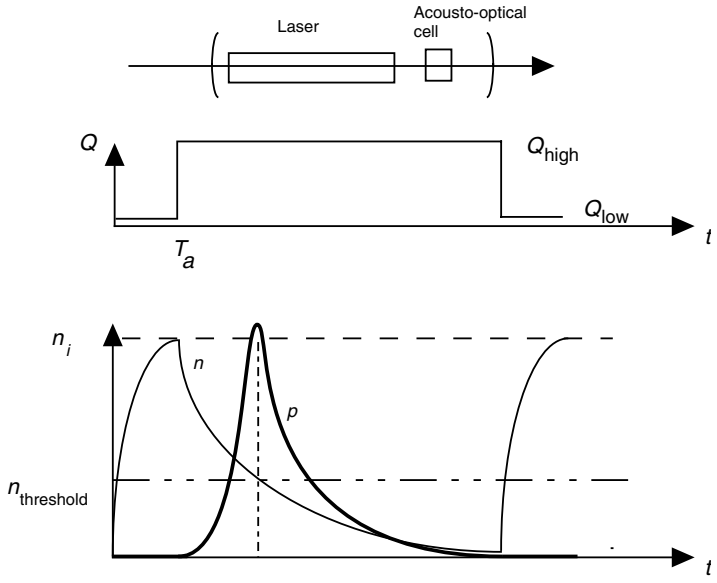


Fig. 4.13. Principle of laser cavity dumping by loss modulation (*Q-switching*). Between 0 and T_a significant cavity losses are introduced as a means of decreasing the quality factor Q of the cavity (by using, for example, an acousto-optic modulator). While the Q of the cavity is artificially maintained at a value Q_{low}, the inversion population is allowed to build up to a value n_i exceeding the value of $n_{threshold}$ for the otherwise high-Q cavity. The Q factor of the cavity is then restored to its original value of Q_{high} and the additional energy stored during the low-Q charging phase is released into a single giant pulse.

high reflectivity of the mirrors, for example). The energy stored in the cavity is then released as described in (4.37). To simplify the notation, we will assume that the lifetime of level $|2\rangle$ remains very large in comparison with the photon lifetime (i.e. $u \ll 1$). Equation (4.37) may then be written as:

$$\begin{cases} \dfrac{\mathrm{d}Y}{\mathrm{d}T} = Y(X-1) \\ \dfrac{\mathrm{d}X}{\mathrm{d}T} = -XY \end{cases} \tag{4.41}$$

Dividing one of the equations by the other, we obtain:

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = \frac{1}{X} - 1 \tag{4.42}$$

which may be easily integrated and put into the following form, taking into account earlier definitions:

$$p = n_{\text{threshold}} \ln \frac{n_d}{n_i} - (n_d - n_i) \tag{4.43}$$

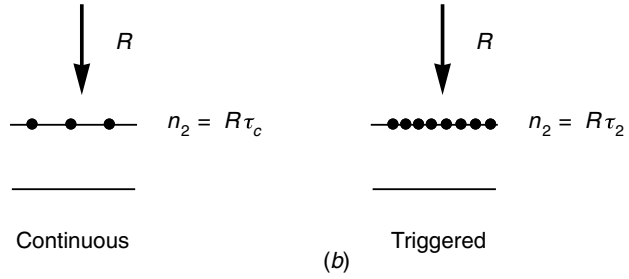


Fig. 4.14. In a laser operating in continuous mode (*a*), the maximum density of electrons in the excited state is given by the threshold density value determined by photon lifetime in the cavity τ_c. In Q-switched mode (*b*), the maximum electron density is determined by carrier lifetime in the excited state τ_2, which can be considerably larger than τ_c.

The photon density p may also take on *transitory* values greater than those allowed by the stationary number of photons in the high-Q cavity. The photon density is at a maximum when $\mathrm{d}p/\mathrm{d}t = 0$ in Eq. (4.41), i.e. when $n = n_{\text{threshold}}$:

$$p_{\max} = n_i\left(1 + \frac{n_{\text{threshold}}}{n_i}\ln\frac{n_{\text{threshold}}}{n_i} - \frac{n_{\text{threshold}}}{n_i}\right) \approx n_i \tag{4.44}$$

when $n_{\text{threshold}} \ll n_i$, which is often the case. We are now ready to calculate the increase in *maximum power* of the laser relative to its stationary operation. Without Q-switching, the number of photons available (p_{CW} where CW refers to *continuous wave*) in the cavity is given by (4.28b), or $p_{\text{CW}} = R\tau_c$, where we recall that R is the pump rate and τ_c the photon lifetime in the high-Q cavity. The maximum photon density (p_{QS}) achieved by Q-switching is given by (4.44) or $p_{\text{QS}} = n_i = RT_a$, where we recall that $n_i \approx n$ is the total density of the emission centres. Generally, the time T_a is fairly close to the lifetime τ_2. As a result, the ratio of the CW exit power to the Q-switch power (see Fig. 4.14) is:

$$\frac{P_{\text{QS}}}{P_{\text{CW}}} = \frac{n_i}{n_{\text{threshold}}} \approx \frac{\tau_2}{\tau_c} \gg 1 \tag{4.45}$$

As lifetimes τ_2 and τ_c are generally of the order of 1 ms and 1 ns, respectively, we see that the ratio given in (4.45) may become considerable. Following the same reasoning as for (4.30), we see that the peak output power of the pulse is precisely given by the product of half the internal photon density (we divide by two as only those photons moving in the direction of the exit mirror are to be considered) with the transmittance T_s of the output mirror and the cross-sectional area A of the laser cavity:

$$P_{\text{QS}} = h\nu\frac{c}{2}T_s A n_i \tag{4.46}$$

Peak output power of a Q-switched laser

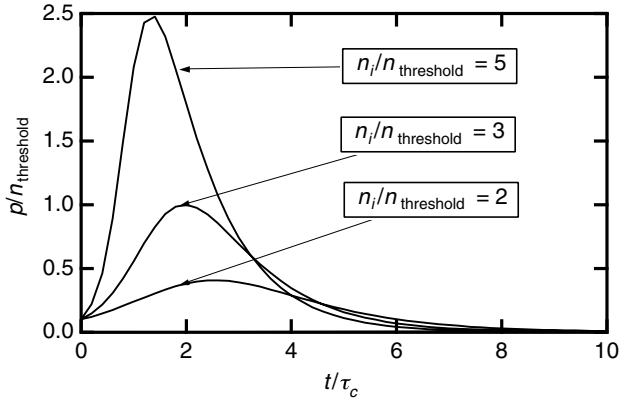


Fig. 4.15. Time dependence of photon density p in a Q-switched laser mode for different values of $n_i/n_{\text{threshold}}$ above threshold. These curves are solutions to the differential Eq. (4.41).

Finally, the length of the laser pulse may be obtained by numerically integrating the non-linear differential Eq. (4.41). In the next example, we give a short program written for MATHEMATICA which allows one to obtain the dynamic behaviour of a Q-switched laser.

Figure 4.15 gives a few examples from which we conclude that the typical pulse lengths are of the order of a few τ_cs (photon lifetimes), i.e. tens of nanoseconds for laser cavities measuring a few centimetres in length. We note that loss modulation in the cavity can be achieved passively (i.e. without external application of electrical pulses to an optical shutter, for instance). This can be achieved by placing a saturable absorber in the laser cavity. This material then absorbs electromagnetic radiation up to a certain intensity level beyond which the medium becomes transparent. Thus, the introduction of such a passive element will result in spontaneous Q-switching of the cavity. At the beginning of the cycle, the absorbing characteristics of the material lower the Q factor of the cavity, during which time electrical energy is stored in the cavity by increasing the electron occupation density in the excited level. Once the medium saturates, the cavity losses are reduced and stimulated emission proceeds to liberate the stored energy in the carriers into the form of a high-power optical pulse.

Example

As (4.41) has a zero photon density as an initial condition ($Y(0) = 0$), it possesses as acceptable solutions the time-independent inversion population $X(T) = X(0)$ and zero photon density $Y(T) = 0$. As indicated by Fig. 4.15, we have had to introduce artificial initial conditions ($Y(0) = 0.1$) to simulate the effect of Q-switching. This underlines the *physical necessity* of spontaneous emission in triggering laser oscillations. The MATHEMATICA program below allows calculation of the temporal response of a laser to modulated cavity loss. It is written assuming an initial

condition of $X(0) = 2$.

```
eq1= y′[t]==y[t]*(x[t]-1)
eq2=x′[t]==-x[t]*y[t]
sol=NDSolve[{eq1,eq2,x[0]==2,y[0]==0.1},{x[t],y[t]},{t,0,10}]
plot=Plot[Evaluate[y[t] /.sol,{t,0,10}]]
```

4.6.3 Mode locking

We now consider the particular case of dynamic response of an inhomogeneous laser medium. We saw how an inhomogeneous medium may be described by an ensemble of independent oscillators which may lase over the entire range of modes allowed by the phase constraints given in (4.25). The complex electromagnetic field set up inside the cavity is then given by the sum of all contributions $S = 2N + 1$ of the independent modes q:

$$E(x,t) = \sum_{q=-N}^{+N} A_q \exp\left[2\pi i \nu_q \left(t - \frac{x}{c}\right)\right] \tag{4.47}$$

We suppose that the gain spectrum is centred at ν_0 and possesses a width $\Delta\nu = (2N + 1)\delta\nu$, where $\delta\nu$ is the frequency separation between the modes permitted by the cavity. We saw in (4.25) that $\delta\nu = c/2nd$, so that $\delta\nu = (2N + 1)/T_c = S/T_c$ (and where T_c is the photon round-trip time for the cavity, i.e. $T_c = 2d/(c/n) = 2\tau_{\text{light}}$). As we are only interested in time-dependent variations of the electromagnetic field (4.47) may be written:

$$E(t) = A(t)\exp[2\pi i \nu_0 t] \tag{4.48a}$$

where the amplitude $A(t)$ is given by:

$$A(t) = \sum_{q=-N}^{+N} A_q \exp\left(2i\pi q \frac{t}{T_c}\right) \tag{4.48b}$$

The amplitudes A_q have no a priori phase relations between each other – their phases ϕ_q are random. We will see in Complement 4.D that, as a consequence, intensity and phase fluctuations occurring at the output of multimode lasers can be significant. Furthermore, if all the amplitudes are identical ($A_q = A$), then the average power of the laser is $(2N + 1)A^2 = SA^2$.

In this section we are interested in what happens to the behaviour of a laser when the amplitudes of the various modes have a specific phase relation imposed upon them, e.g. $\phi_q = 0$. In this case, we speak of *mode locking*. We will see further on, how this may be achieved in practice. We note that the envelope of the time dependence for the electromagnetic field given by (4.48b) is a well defined periodic function with period T_c. More precisely, if we assume that the gain curve is constant over the interval $\Delta\nu$, (i.e. that $A_q = A$) then the wave intensity is given by:

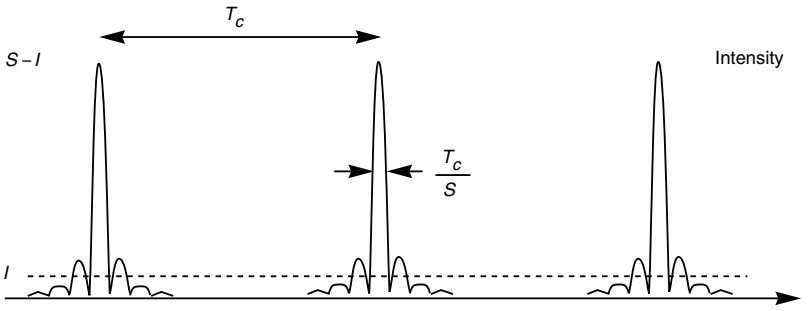


Fig. 4.16. Time dependence of the sum of S sinusoidal contributions separated in frequency by $1/T_c$. The result is a succession of pulses of width T_c/S separated in time by T_c, and having a peak intensity S-times greater than the time averaged output value.

$$I(t) = |A|^2 \left| \sum_{q=-N}^{+N} \exp\left(2i\pi q \frac{t}{T_c}\right) \right|^2 = |A|^2 \left[\frac{\sin(2N+1)\pi t/T_c}{\sin \pi t/T_c} \right]^2 \tag{4.49}$$

The light output from a mode-locked laser consists of a train of pulses of intensity $2(N+1)^2A^2$, separated in time by a period T_c, with pulse widths T_{pulse} given by:

$$T_{\text{pulse}} = \frac{T_c}{2N+1} = \frac{1}{\Delta v} \tag{4.50}$$

(see Fig. 4.16). Therefore, the wider the laser gain spectrum, the greater the number of supported modes and the shorter the resulting pulse width. To obtain ultrashort pulses, titanium-doped sapphire crystals having gain bandwidths of 500 meV allow the generation of pulses 10 fs in duration. We will see that semiconductors with gain curves of several hundred meV are also good candidates for the implementation of mode synchronized lasers. Table 4.2 summarizes the key characteristics of light output from mode-locked lasers.

Practically (see Fig. 4.17), mode synchronization can be accomplished by introducing a fast response electro-optic shutter between the two cavity mirrors. In this way, light transmission between the two mirrors is possible only during short periods of time as determined by an externally applied electrical signal. The only portion of the overall wave train that will be able to propagate in the cavity (and experience amplification by the laser medium) will be that comprised of modes which are in phase with the modulated transmittance of the electro-optic shutter. This situation refers to a form of *survival of the fittest*, with those modes which do not satisfy the requisite phase conditions being eliminated from amplification. A succinct description of this mechanism, using electromagnetic laser equations is given in Complement 4.C. We note in closing, that mode synchronization may also be achieved using a saturable absorber. This latter method is employed to achieve mode locking in semiconductor lasers.

Table 4.2. *Key characteristics of light output from mode-locked lasers*

Period	$T_c = \dfrac{2d}{c'}$
Pulse width	$T_{\text{pulse}} = \dfrac{T_c}{S} = \dfrac{1}{\Delta v}$
Average intensity	$\bar{I} = SA^2$
Peak intensity	$I_{\text{peak}} = S^2A^2 = S\bar{I}$

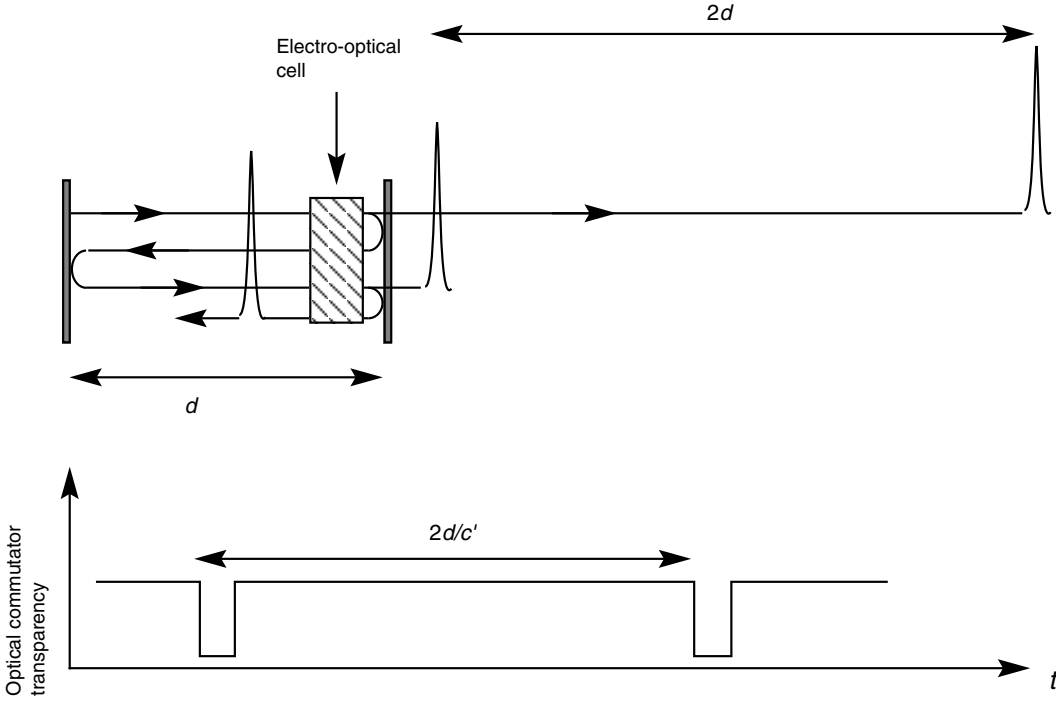


Fig. 4.17. In mode locking, only pulses that are synchronous with the opening of the electro-optic shutter every $2d/c'$ will be able to experience amplification by the laser medium. This is equivalent to a form of natural selection being imposed upon all possible modes in the cavity, with those modes not satisfying the requisite phase restrictions being eliminated from amplification.

Example

We consider a 1 m long Nd^{3+} : YAG glass laser. The index of refraction of glass n_{op} is 1.5, and the transition width $\Delta v = 3 \times 10^{12}$ Hz. The period T_c between the pulses is $2\,\text{m}/2 \times 10^8\,\text{m s}^{-1}$ or 10 ns, the width of the mode-locked pulses is

$T_{\text{pulse}} = 0.33\,\text{ps}$, and the number of synchronized modes is $S = 10^{-8}\,\text{s}/3.3 \times 10^{-13}\,\text{s}$ or 30 000 modes. If the average laser output power is 1 W, the peak power of a single pulse is 30 kW.

FURTHER READING

G. Grynberg, A. Aspect, and C. Fabre, *Introduction aux lasers et à l'optique quantique*, Ellipses, Paris (1997).

P. W. Milony and J. H. Eberly, *Lasers*, Wiley, New York (1988).

B. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley Interscience, New York (1991).

A. E. Siegman, *Lasers*, University Science Book, Mill Valley, CA (1986).

J. T. Verdeyen, *Laser Electronics*, Prentice-Hall, Englewood Cliffs, NJ (1989).

A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

Complement to Chapter 4

4.A The effect of spontaneous emission and photon condensation

As mentioned previously, in Chapter 4 we have neglected the influence of spontaneous emission on laser effect. One of the consequences of this omission, which we flagged in Section 4.6, is that stimulated emission cannot commence if the initial number of photons in the cavity is zero. In other words, a number of photons $p = 0$ in mode l with an inversion density $n = n_d = \text{constant}$ constitutes an acceptable solution to (4.34) and (4.35). We will now show how spontaneous emission may be included in laser equations and discuss the resulting effects on laser characteristics. To begin, we return to the two coupled equations (4.34) and (4.35) that describe the dynamic relationship between the number of photons p in a given mode l and the inversion density $n_d = n_2 - n_1$ for a four-level system in a cavity:

$$\frac{\mathrm{d}n_d}{\mathrm{d}t} = R - \frac{n_d}{\tau_2} - \frac{p}{\tau_c}\frac{n_d}{n_{\text{threshold}}} \tag{4.A.1}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{p}{\tau_c} + \frac{p}{\tau_c}\frac{n_d}{n_{\text{threshold}}} \tag{4.A.2}$$

We recall that R is the pumping rate of level $|2\rangle$, τ_c is the photon lifetime in the cavity, and $n_{\text{threshold}} = 1/c'\tau_c\sigma_{\text{op}}$ is the inversion density required to reach laser threshold. We also suppose that the lifetime τ_1 in level $|1\rangle$ is short enough to allow us to neglect the population density n_1 (where $n_d \approx n_2$), so consequently $\tau_{\text{sat}} = \tau_2$ (Eq. (4.6)). Spontaneous emission is introduced into (4.A.1) by means of including the rate $1/t_{\text{spon}}$ in the calculation of $1/\tau_2$, as may be seen in Eq. (4.2). On the other hand, in (4.A.2), we must incorporate the fact that, even in the absence of any photons in the cavity ($p = 0$), photons may be released through spontaneous emission. We saw in Chapter 3 that the emission rate for a two-level system may be written in the form $W(P + 1)$ where this time P is the number of photons (not the photon density), WP is the stimulated emission, and W is the spontaneous emission (Eq. (3.66)). To include the effects of spontaneous emission, it is therefore sufficient to write (4.A.1) and (4.A.2) in terms of the number and not the density (by

multiplying by the cavity volume V), and by replacing P by $P + 1$ in (4.A.2), from which we obtain for the coupled system:

$$\begin{cases} \dfrac{\mathrm{d}N_d}{\mathrm{d}t} = RV - \dfrac{N_d}{\tau_2} - \dfrac{P}{\tau_c}\dfrac{N_d}{N_{\text{threshold}}} \\ \dfrac{\mathrm{d}P}{\mathrm{d}t} = -\dfrac{P}{\tau_c} + \dfrac{P+1}{\tau_c}\dfrac{N_d}{N_{\text{threshold}}} \end{cases} \tag{4.A.3}$$

Dynamic equations for a laser cavity including spontaneous emission

In (4.A.3), $N_d = n_d V$ and $P = pV$. Furthermore, these equations suppose that there is a single mode in the cavity into which all spontaneously emitted photons are released (i.e. the spontaneous emission factor β introduced in (3.72) is equal to 1). We now introduce the following dimensionless variable:

$$r = \frac{R}{R_{\text{threshold}}} \tag{4.A.4}$$

and recall that:

$$R_{\text{threshold}} = \frac{n_{\text{threshold}}}{\tau_2};\ n_{\text{threshold}} = \frac{1}{c'\sigma_{\text{op}}\tau_c};\ P_{\text{sat}} = \frac{V}{c'\sigma_{\text{op}}\tau_2};\ X = \frac{n_d}{n_{\text{threshold}}} \tag{4.A.5}$$

With all the changes of variables, for a stationary state (4.A.3) may be written as:

$$\begin{cases} r = X\left(1 + \dfrac{P}{P_{\text{sat}}}\right) \\ X(P+1) = P \end{cases} \tag{4.A.6}$$

By solving the resulting second-degree equation, the stationary number of photons P is trivially found to be:

$$P = \left[(r-1) + \sqrt{(r-1)^2 + 4\frac{r}{P_{\text{sat}}}}\right]\frac{P_{\text{sat}}}{2} \tag{4.A.7}$$

Number of photons in the cavity as a function of pump rate (dimensionless)

This seemingly innocent equation actually implies the existence of a rather surprising physical effect, which we shall now examine. Let us begin by noting that P_{sat} is a very large number, typically $\sim 10^{10}$, which gives the equation a rather distinct behaviour.

For a pump rate slightly above threshold ($r = 1 + \delta$), the term $4r/P_{\text{sat}}$ under the square root immediately becomes negligible in comparison to δ, leading to $P = P_{\text{sat}}(r-1)$, which is the behaviour given by (4.28). In the case of a pump rate slightly below threshold ($r = 1 - \delta$), (4.A.7) becomes:

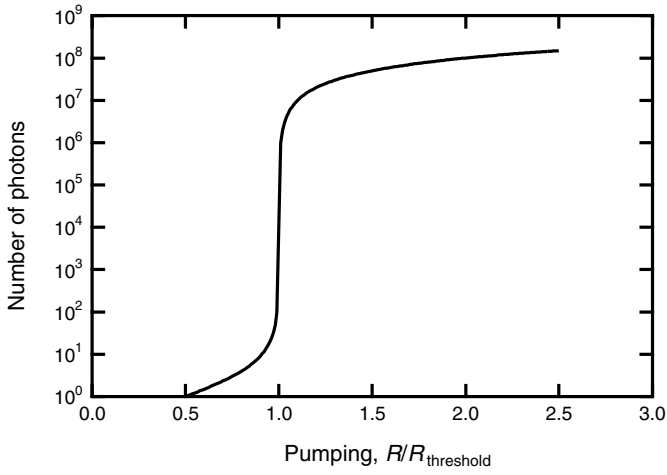


Fig. 4.A.1. Number of photons in a cavity as a function of pump rate (for a cavity with a saturation flux P_{sat} of 10^8 photons).

$$P = \left(-\delta + \delta \sqrt{1 + 4 \frac{1-\delta}{\delta^2} \frac{1}{P_{\text{sat}}}} \right) \frac{P_{\text{sat}}}{2} \tag{4.A.8}$$

or

$$P \approx \left[-\delta + \delta \left(1 + 2 \frac{1-\delta}{\delta^2} \frac{1}{P_{\text{sat}}} \right) \right] \frac{P_{\text{sat}}}{2} = \frac{1-\delta}{\delta} \tag{4.A.9}$$

In other words, even if the system is ever so slightly below threshold, the number of photons is well approximated by:

$$P = \frac{r}{1-r} \tag{4.A.10}$$

Therefore, by passing from a pump value 1% below to 1% above threshold, the number of photons in the cavity changes from 10^2 to $P_{\text{sat}} \times 10^{-2} \sim 10^8$. This corresponds to a veritable phase condensation which results from the Boson nature of light being taken into account by the $P + 1$ term in (4.A.3). Figure 4.A.1 shows the behaviour of the number of photons in a cavity as a function of reduced pump rate r for $P_{\text{sat}} = 10^8$ photons.

We now underline an important consequence of spontaneous emission on laser cavity dynamics which will be of use to us in Complement 4.D. The second dynamic equation in (4.A.3) may be rewritten as:

$$\frac{\mathrm{d}P}{\mathrm{d}t} + \frac{P}{\tau_c} = (P+1) \frac{N_d}{\tau_c N_{\text{threshold}}} \tag{4.A.11}$$

or again, taking into account the definitions in (4.A.5),

$$\frac{\mathrm{d}P}{\mathrm{d}t} + \frac{P}{\tau_c} = (P+1)\frac{G}{\tau_{\text{light}}} \tag{4.A.12}$$

where G is the optical gain resulting from a single pass of light across the length of the cavity ($G \approx gd$), g is the optical gain per unit length (in cm^{-1}) given by $g = n_d\sigma_{\text{op}}$, and $\tau_{\text{light}} = d/c'$ is the light transit time between the two cavity mirrors. Equation (4.A.12) may be easily integrated to yield:

$$P = \frac{\mathrm{e}^{(G/\tau_{\text{light}} - 1/\tau_c)t} - 1}{1 - (\tau_{\text{light}}/\tau_c G)} \tag{4.A.13}$$

Thus, even if there are no photons in the cavity at $t = 0$, laser oscillations will be triggered by the spontaneous emission of a single photon into the mode as long as the condition $G > \tau_{\text{light}}/\tau_c = 2T$ (gain > mirror loss) is satisfied.

Equation (4.A.13) shows, moreover, that the photon population builds up from noise with a characteristic build-up time τ_b, which may be easily calculated assuming that $G_0/\tau_{\text{light}} \gg 1/\tau_c$ (G_0 is the unsaturated initial gain), i.e when the mirror transmission T is small. Indeed, in that case (4.A.13) reads:

$$P \approx \mathrm{e}^{t/\tau_b} \tag{4.A.14}$$

with the *characterictic laser build-up time* τ_b given by:

$$\tau_b = \frac{\tau_{\text{light}}}{G_0} = \frac{L/c'}{\gamma_0 L} = \frac{1}{\gamma_0 c'} \tag{4.A.15}$$

Characteristic laser build-up time

Let us recall that γ_0 is the unsaturated laser medium gain.

With many authors, a different (and somewhat more arbitrary) notion is introduced, which we will call the *laser build-up time*. It is the time t_b necessary for the laser to deliver a measurable power P_m starting from a noise power P_0. From (4.A.14), this build-up time is given by:

$$t_b = \log(P_m/P_0)\tau_b = \log(P_m/P_0)\frac{1}{\gamma_0 c'} \tag{4.A.16}$$

Laser build-up time

The usual value taken for $\log(P_m/P_0)$ is 30, corresponding to an arbitrary detection threshold power of 1 mW and a photon noise power of 10^{-16} W.

Example

When the laser gain is small, build-up time can be fairly long compared with other characteristic times of the cavity. If we assume a laser medium of unsaturated gain of $0.01\ \text{cm}^{-1}$ and a light velocity of $c' = 2 \times 10^{10}\ \text{cm s}^{-1}$, the characteristic build-up time τ_b is 5 ns and the build-up time t_b is $30\,\tau_b$, i.e. 150 ns.

4.B Saturation in laser amplifiers

We saw in Chapter 3, and again in Chapter 4, how the optical amplification resulting from stimulated emission saturates rapidly. This saturation was described by (4.10) and shall be briefly rederived here. We consider a four-level system (Fig. 4.3) – in which carriers in the ground state $|0\rangle$ are promoted by optical pumping (R) to an excited state $|3\rangle$. Carriers in level $|3\rangle$ are assumed to relax instantaneously to level $|2\rangle$, which in turn possesses a long lifetime. Stimulated emission then proceeds involving radiative carrier relaxation from level $|2\rangle$ to $|1\rangle$, with a characteristic time τ_2. Subsequent relaxation processes between levels $|1\rangle$ and $|0\rangle$ are finally assumed to proceed instantaneously. This situation is described by the following system of differential equations:

$$\begin{cases} \dfrac{\mathrm{d}N_2}{\mathrm{d}t} = R - \sigma_{\mathrm{op}} N_2 \Phi - \dfrac{N_2}{\tau_2} \\ N_1 = 0 \end{cases} \tag{4.B.1}$$

where Φ is the photon flux and σ_{op} is the optical cross-section for the transition $|2\rangle \rightarrow |1\rangle$. The stationary electron density N_2 in level $|2\rangle$ is then given by:

$$N_2 = \frac{R\tau_2}{1 + \Phi/\Phi_{\mathrm{sat}}} = \frac{N_{d0}}{1 + \Phi/\Phi_{\mathrm{sat}}} \tag{4.B.2}$$

where N_{d0} is the low signal inversion population density and Φ_{sat} is the saturation flux $1/\sigma_{\mathrm{op}}\tau_2$. The gain of the amplification medium is given by $\gamma = \sigma_{\mathrm{op}}(N_2 - N_1)$ or:

$$\gamma = \frac{\gamma_0}{1 + I/I_{\mathrm{sat}}} \tag{4.B.3}$$

with $\gamma_0 = \sigma_{\mathrm{op}} R \tau_2$ being the cold cavity gain for the medium and the saturation intensity I_{sat} being given by:

$$I_{\mathrm{sat}} = \frac{\hbar\omega}{\sigma_{\mathrm{op}}\tau_2} \tag{4.B.4}$$

We recall that this value is generally quite small. For instance, for an optical cross-section of $4 \times 10^{-19}\,\mathrm{cm}^{-2}$, a lifetime of 1.2 ms, and a transition energy $\hbar\omega$ of 1.16 eV, $I_{\mathrm{sat}} \approx 500\,\mathrm{W\,cm}^2$, which corresponds to a fairly low optical power density.

We now consider a laser medium of length L and surface A, which we shall use as an amplifier. At one end, we inject a light beam with an intensity I_e and seek its output intensity I_s at the exit of the amplification stage. There are no mirrors in

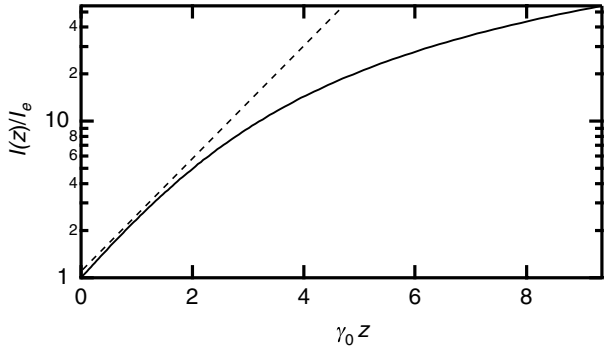


Fig. 4.B.1. Intensity variation I as a function of z position in a laser medium. The intensity values have been normalized to the light intensity at the amplifier entrance I_e, γ_0 is the cold cavity gain for the system, and a ratio of 0.1 is assumed for I_e/I_{sat}.

this amplifier to enable optical feedback. We recall that if the gain of a propagation medium is γ, then the electric field strength is proportional to $e^{\gamma z/2}$ and the intensity to $e^{\gamma z}$. As the gain has a dependence on the intensity resulting from saturation effects (Eq. (4.B.3)) and since the intensity is a function of the position z, the light intensity in the medium is therefore a solution of the differential equation:

$$\frac{1}{I(z)}\frac{\mathrm{d}}{\mathrm{d}z}I(z) = \frac{\gamma_0}{1 + I(z)/I_{\text{sat}}} \tag{4.B.5}$$

This equation may be easily integrated to obtain the dependence of intensity I on position z in the medium through the implicit relation:

$$\gamma_0 z = \ln\left(\frac{I(z)}{I_e}\right) + \frac{1}{I_{\text{sat}}}[I(z) - I_e] \tag{4.B.6}$$

Figure 4.B.1 shows this variation in intensity as a function of z position.

We notice that as the intensity increases as a function of position, the medium saturates along with the amplification. Taking G as the total amplifier gain ($G = I_s/I_e$) and G_0 as the small signal gain ($G_0 = e^{\gamma_0 L}$), Eq. (4.B.6) may be written as:

$$\frac{I_{\text{s}}}{I_{\text{sat}}} = \frac{G}{G-1}\ln\left(\frac{G_0}{G}\right) \tag{4.B.7}$$

We see from this last equation that G tends rapidly towards 1 when the light intensity exceeds its saturation value in the medium. The following example will look at the variation in gain G of a laser medium having a small signal gain of 800. We notice that the available gain decreases rapidly with I_s. Thus, for an entrance power of $882\,\text{W cm}^2$, the gain is only 4, and the output power $3530\,\text{W cm}^{-2}$. Saturation is a negative and unavoidable effect, as shall now be explained.

We label as I_{disp} the optical power dispensed by the amplification medium, i.e. $I_{\text{disp}} = I_s - I_e$. Expression (4.B.7) then gives:

$$I_{\text{disp}} = \ln\left(\frac{G_0}{G}\right) I_{\text{sat}} \tag{4.B.8a}$$

If we neglect the logarithmic term in front of the linear term in (4.B.6), I_{disp} may then be written:

$$I_{\text{disp}} = I_s - I_e = \gamma_0 L I_{\text{sat}} \tag{4.B.8b}$$

or again, using an alternative set of definitions:

$$\frac{I_{\text{disp}}}{L} = \frac{N_{d0}\hbar\omega}{\tau_2} \tag{4.B.9}$$

Maximum available power in an optical amplifier limited by optical saturation

This last equation states that the maximum power that an amplifier can give to an incident light beam is given by the energy $\hbar\omega$ stored by each inverted atom (having a density of N_{d0}) over time τ_2.

Example

For an amplification medium of Nd: YAG doped $10^{19}\,\text{cm}^{-3}$, the maximum stored energy density is then given by $10^{19}\,\text{cm}^{-3} \times 1.16 \times 1.6 \times 10^{-19}\,\text{J}$ or $1.85\,\text{J cm}^{-3}$. The maximum power which may be transferred to an incident beam is then $1.85\,\text{J cm}^{-3}/1.2\,\text{ms}$, or $1.54\,\text{kW cm}^{-3}$. For a 1-m long amplification stage, we therefore obtain a maximum available power of $154\,\text{kW cm}^{-2}$. The small signal gain is $\gamma_0 = \sigma_{\text{op}} N_{d0} = 4 \times 10^{-19}\,\text{cm}^2 \times 10^{19}\,\text{cm}^{-3} = 4\,\text{cm}^{-1}$. The amplifier gain as a function of input and output power levels, given by (4.B.7) (and by taking $G_0 = 2 \times 1\,\text{m} \times 4\,\text{cm}^{-1} = 800$), is shown in Fig. 4.B.2.

Another aspect of gain saturation in laser amplifiers is *pulse shape distortion*. Indeed, since the pulse amplitude non-linearly affects the amplification factor, different parts of the optical pulses will be subject to very different amplifications, leading to strong pulse shape distortions. This effect is of great importance in optical fibre amplifiers, which have immense practical applications in telecommunications – we will discuss this now.

We consider an optical pulse of flux $\Phi_{\text{in}}(t)$ entering a *fibre amplifier*. The amplifier population is totally inverted before time $t = 0$, so that the pumping rate R is identically zero during pulse amplification in the fibre. Moreover, we neglect spontaneous emission – too slow – during the pulse duration. Let us concentrate on the portion of the fibre situated between x and $x + \Delta x$ during time interval t and $t + \text{d}t$. The rate of population decrease $N_2(x, t)$ due to stimulated emission under photon flux $\Phi(x, t)$ is given by (4.B.1), i.e. with the above assumptions:

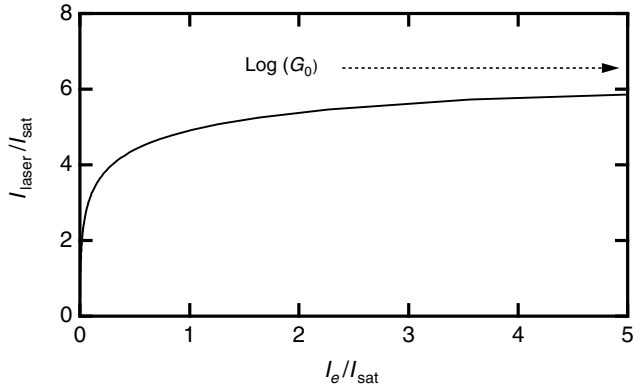


Fig. 4.B.2 Power supplied by an amplifier medium as a function of entrance power I_e normalized to the saturation power level. The cold cavity gain G_0 is 800.

$$\frac{\partial}{\partial t}N_2(x,t) = -\sigma_{op}N_2(x,t)\Phi(x,t) \tag{4.B.10}$$

Spatio-temporal variation of the photon flux is given by the *continuity equation*. We will re-derive it here, because it is always a stimulating exercise! In the Δx interval, the photon surface density ρ_{ph} is related to photon flux by $\Phi(x,t) = c'\rho_{ph}(x,t)$ where c' is the speed of light in the fibre. The rate of change in photon number in the Δx interval is given by the input photon flux minus the output photon flux, plus the rate of stimulated emission in the interval Δx, i.e:

$$\frac{\partial}{\partial t}[\rho_{ph}(x,t)\Delta x] = \Phi(x) - \Phi(x+\Delta x) + \sigma_{op}N_2(x,t)\Delta x\rho_{ph}(x,t) \tag{4.B.11a}$$

which leads to the continuity equation:

$$\frac{\partial}{\partial t}\Phi(x,t) + c'\frac{\partial}{\partial x}\Phi(x,t) = \sigma_{op}N_2(x,t)\Phi(x,t) \tag{4.B.11b}$$

$$\frac{\partial}{\partial t}N_2(x,t) = -\sigma_{op}N_2(x,t)\Phi(x,t) \tag{i.e. 4.B.10}$$

Pulse propagation equations in laser amplifier

We have repeated Eq. (4.B.10) above because these last two non-linear coupled equations are all we need to solve any pulse shape distortion problems due to saturation in amplifying fibres.

Like all the propagation equations, (4.B.11) can be solved in *Eulerian co-ordinates*. This means that we may 'follow' propagation of the pulse – riding on the photons so to say – by introducing a new system of co-ordinates:

$$\tilde{t} \equiv t$$

$$\tilde{x} \equiv x + c't \tag{4.B.12a}$$

By definition, the photon flux and inversion population in these new co-ordinates are:

$$\tilde{\Phi}(\tilde{x}, \tilde{t}) \equiv \Phi(x, \mathrm{t})$$

$$\tilde{N}_2(\tilde{x}, \tilde{t}) \equiv N_2(x, t) \tag{4.B.12b}$$

The $\tilde{\Phi}(\text{constant}, \tilde{t})$ follows the fate of the pulse during its propagation in the fibre. The Jacobian of this transformation can be derived in the following mnemotechnical way:

$$\frac{\partial}{\partial \tilde{t}} \equiv \frac{\partial}{\partial t}$$

$$\frac{\partial}{\partial \tilde{x}} \equiv \frac{\partial}{\partial x}\frac{\partial x}{\partial \tilde{x}} + \frac{\partial}{\partial t}\frac{\partial t}{\partial \tilde{x}} = \frac{\partial}{\partial x} + c'\frac{\partial}{\partial t} \tag{4.B.12c}$$

so that in the new co-ordinate system (4.B.10) and (4.B.11) read:

$$\frac{\partial}{\partial \tilde{x}}\tilde{\Phi}(\tilde{x}, \tilde{t}) = \sigma_{\mathrm{op}}\tilde{N}_2(\tilde{x}, \tilde{t})\tilde{\Phi}(\tilde{x}, \tilde{t}) \tag{4.B.13}$$

$$\frac{\partial}{\partial \tilde{t}}\tilde{N}_2(\tilde{x}, \tilde{t}) = -\sigma_{\mathrm{op}}\tilde{N}_2(\tilde{x}, \tilde{t})\tilde{\Phi}(\tilde{x}, \tilde{t}) \tag{4.B.14}$$

Pulse propagation equations in a laser amplifier in moving co-ordinates

Equations (4.B.13) and (4.B.14) seem nicer than their counterparts in Lagrangian co-ordinates. In fact, it is a *trompe-l'œil*: it is the boundary conditions which are now complex, but we will neglect this aspect in the rest of the discussion. We shall now derive integral formulations of (4.B.13) and (4.B.14) – which are in fact all we really need to know – known as the *Frantz–Nodvik model*. We integrate (4.B.13) over the fibre length, with the small cheat of using boundaries at $\tilde{x} = 0$ and $\tilde{x} = L$ (leaving to the reader the task of deriving the criteria for validity of these boundaries):

$$\int_{\Phi_{\mathrm{in}}(t)}^{\Phi_{\mathrm{out}}(t)} \frac{\mathrm{d}\tilde{\Phi}(\tilde{x}, \tilde{t})}{\tilde{\Phi}(\tilde{x}, \tilde{t})} = \sigma_{\mathrm{op}} \int_{\tilde{x}=0}^{\tilde{x}=L} \tilde{N}_2(\tilde{x}, \tilde{t})\mathrm{d}\tilde{x} \equiv \sigma_{\mathrm{op}} N_T(t) \tag{4.B.15}$$

where, by definition, $N_T(t)$ is the integrated inversion population over the fibre length (in cm^{-2}). From (4.B.15), one immediately obtains:

$$\Phi_{\mathrm{out}}(t) = \Phi_{\mathrm{in}}(t)\mathrm{e}^{\sigma_{\mathrm{op}}N_T(t)} \equiv G(t)\Phi_{\mathrm{in}}(t) \tag{4.B.16}$$

where $G(t)$ is the instantaneous gain of the laser amplifier. Now, (4.B.13) and (4.B.14) can be merged to form the second integral relation:

$$\frac{\partial}{\partial \tilde{t}} \int_{\tilde{x}=0}^{\tilde{x}=L} \tilde{N}_2(\tilde{x},\tilde{t})\mathrm{d}\tilde{x} = -\int_{\tilde{x}=0}^{\tilde{x}=L} \frac{\partial}{\partial \tilde{x}} \tilde{\Phi}(\tilde{x},\tilde{t})\mathrm{d}\tilde{x} = -[\Phi_{\mathrm{out}} - \Phi_{\mathrm{in}}(t)] \tag{4.B.17a}$$

so that

$$\frac{\partial}{\partial \tilde{t}} N_T(t) = -[\Phi_{\mathrm{out}}(t) - \Phi_{\mathrm{in}}(t)] \tag{4.B.17b}$$

The latter equation expresses that flux enhancement in the fibre amplifier equals the integrated population inversion decrease, which is reasonable. Now we can solve the system formed by (4.B.16) and (4.B.17b) to obtain, in the first place, the evolution of the integrated population inversion:

$$\frac{\mathrm{d}}{\mathrm{d}t} N_T(t) = -\Phi_{\mathrm{in}}(t)(\mathrm{e}^{\sigma_{\mathrm{op}} N_T(t)} - 1) \tag{4.B.18}$$

Introducing the *photon fluence* as the time integral of the photon flux:

$$\Psi(t) \equiv \int_0^t \Phi(t')\mathrm{d}t' \tag{4.B.19}$$

we obtain the time evolution of the integrated population inversion, instantaneous gain, and output photon flux:

$$N_T(t) = N_{\mathrm{sat}} \log\left[\frac{G_0}{G_0 - (G_0 - 1)\mathrm{e}^{-\Psi_{\mathrm{in}}(t)/N_{\mathrm{sat}}}}\right] \tag{4.B.20}$$

$$G(t) = \frac{G_0}{G_0 - (G_0 - 1)\mathrm{e}^{-\Psi_{\mathrm{in}}(t)/N_{\mathrm{sat}}}} \tag{4.B.21}$$

$$\Phi_{\mathrm{out}}(t) = \Phi_{\mathrm{in}}(t)\frac{G_0}{G_0 - (G_0 - 1)\mathrm{e}^{-\Psi_{\mathrm{in}}(t)/N_{\mathrm{sat}}}} \tag{4.B.22}$$

Frantz–Nodvik equations

where $\mathrm{N}_{\mathrm{sat}} = 1/\sigma_{\mathrm{op}}$ is the saturation fluence, G_0 is the unsaturated amplifier gain ($G_0 = \sigma_{\mathrm{op}} N_{T0}$), and N_{T0} is the initial integrated population inversion or total available fluence in the fibre (in cm^{-2}). These three equations describe the pulse shape distortion in a laser amplifier. For instance, if one injects a square pulse $\Phi_{\mathrm{in}}(t) = \Phi_{\mathrm{in}}$ at the fibre input, (4.B.22) shows that the output will be stretched with an exponential time constant $\mathrm{e}^{-\sigma_{\mathrm{op}}\Phi_{\mathrm{in}}t}$. Another case (Gaussian input) is numerically analysed in the example below and the result shown in Fig. 4.B.3. The effect of

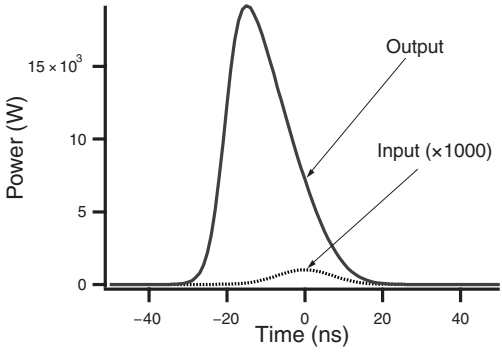


Fig. 4.B.3. Pulse distortion in a fibre amplifier due to saturation (see the example above).

pulse shape distortion is clearly seen in the figure.

Finally, as in the CW case derived at the beginning of this complement, it is easy to find the maximum available energy E_{disp} that one can extract from the fibre amplifier:

$$E_{\text{disp}} = N_{T0}\hbar\omega \tag{4.B.23}$$

This calculation is left as an exercise for the reader.

Example

Let us investigate the behaviour of a 2-m long erbium-doped fibre, with an inner core diameter of 8 μm and doped to $7 \times 10^{19}\,\text{cm}^{-3}$. The wavelength of the emitted photons is 1.55 μm (0.8 eV) and the optical cross-section of Er^{3+} in the silica fibre is taken as $10^{-21}\,\text{cm}^{-2}$. The saturation fluence is thus:

$$N_{\text{sat}} = 1/(10^{-21}\,\text{cm}^{-2}) = 10^{21}\,\text{cm}^{-2}$$

The total available fluence N_{T0} available in the fibre (4.B.15) is:

$$N_{T0} = 7 \times 10^{19}\,\text{cm}^{-3} \times 2 \times 10^{2}\,\text{cm} = 1.4 \times 10^{22}\,\text{cm}^{-2}$$

i.e. 14 times the saturation fluence. Large pulse distortion is thus expected. The total available energy from the fibre (4.B.23) is:

$$E_{\text{disp}} = 1.4 \times 10^{22}\,\text{cm}^{-2} \times 1.6 \times 10^{-19}\,\text{J} \times 0.8\,\text{eV} \times [3.14 \times (8 \times 10^{-4})^2\,\text{cm}^2/4] = 0.9\,\text{mJ}$$

The MATHEMATICA program below gives an example of Gaussian pulse distortion.

$$\text{N0} = 7.10^{19}\ (\text{*cm-3*});\ \sigma = 1.10^{-21}\ (\text{*cm-2*})\ ;\ \text{h}\nu = \frac{1.24}{1.55}\,1.6\,10^{-19};$$

$$\text{Usat} = \frac{\text{h}\nu}{\sigma}\ (\text{*J/cm}^2\text{*})\ ;\ \text{Dfibre} = 8\,10^{-4}\ (\text{*8 μm*})\ ;\ \text{S} = \text{N}[\pi\ \text{Dfibre}^2 / 4]\ ;\ \tau = 10\ (\text{*ns*})\ ;$$

```
lfibre= 200 (*2 m*) ; G0 = Exp[σ N0 lfibre] ; I0= 1/S (* 1 W input*) ;
Iin[t_] := I0 Exp[-(t/τ)^2] (*W/cm ^2*) ; Uin= 10^-9 I0 ∫_-∞^+∞ Exp[-(t/τ)^2] dt (* J/cm ^2*) ;
θ[t_] := 10^-9/Usat ∫_-∞^t Iin[t] dt; Nrel = 1/(σ N0 1fibre) Log[G0/(G0 - (G0- 1) Exp[-θ[t]])] ;
Iinp= S I0 Exp[-(t/τ)^2] ; Iout = S G0/(G0 - (G0 - 1) Exp[-θ[t]]) Iin[t] ;
p1 = Plot[Iinp, {t, -50, 50},
  PlotRange → All, PlotStyle → RGBColor[1, 0, 0], DisplayFunction → Identity] ;
p2 = Plot[Iout, {t, -50, 50},
  PlotRange → All, PlotStyle → RGBColor[0, 1, 0] , DisplayFunction → Identity] ;
Show[p1, p2, DisplayFunction → $DisplayFunction, AxesLabel → {"t (ns) ", "Pout (W) "}]
```

FURTHER READING

L. M. Frantz and J. S. Nodvik, *J. App. Phys.* **24**, 2346 (1963).

4.C Electrodynamic laser equations: electromagnetic foundations for mode locking

Up until now, we have been chiefly interested in the density of photons contained in a laser cavity, without paying much attention to the (electromagnetic) wave nature of laser light. We recall that the photon results from quantization of the electromagnetic mode of a cavity. This, however, does not free us from further having to impose upon the photons an appropriate set of boundary conditions relating to the cavity itself. An exact description of a laser cavity, therefore, requires that we solve Maxwell's equations, appropriately modified to include the laser effect. Here, we will present elements of the theory of *Slater modes*, which will allow us to understand the calculation method and its application to the phenomenon of mode locking. We consider a cavity of *arbitrary shape*, supposed to be completely empty for the time being. Slater was able to show that Maxwell's equations for such a hollow cavity admit an ensemble of solutions which form a complete and orthonormal basis. Furthermore, these solutions $\{\mathbf{E}_a, \mathbf{H}_a\}$ are time independent and must respect any and all other boundary conditions set by the cavity (e.g. in the case of metallic boundaries, $\mathbf{E}_a$ must be perpendicular to the surface; whereas $\mathbf{H}_a$ can only have a parallel component). Of course, the notion of orthonormality should be understood in the sense of distributions, e.g.

$\int_V \mathbf{E}_a(\mathbf{r})\mathbf{E}_b(\mathbf{r})\mathrm{d}^3\mathbf{r} = \delta_{ab}$, where δ_{ab} is the Kronecker delta. Furthermore, the elements forming this Slater double basis must satisfy the following relationships:

$$k_a\mathbf{E}_a(\mathbf{r}) = \nabla \times \mathbf{H}_a(\mathbf{r})$$

$$k_a\mathbf{H}_a(\mathbf{r}) = \nabla \times \mathbf{E}_a(\mathbf{r}) \tag{4.C.1}$$

(We recall that $\mathbf{H}$ is the magnetic induction related to the magnetic field by $\mathbf{H} = \mathbf{B}/\mu_0$). Note that with these definitions, the vectors $\mathbf{E}_a$ and $\mathbf{H}_a$ automatically satisfy two of Maxwell's equations in a vacuum, i.e. $\mathrm{div}\,\mathbf{E}_a = 0$ and $\mathrm{div}\,\mathbf{H}_a = 0$. The electric field $\mathbf{E}_a(\mathbf{r})$ must additionally meet the surface boundary requirement that:

$$\mathbf{n} \times \mathbf{E}_a = 0 \tag{4.C.2}$$

where $\mathbf{n}$ is the normal vector to the surface boundary. It is easy to show that condition (4.C.2) and the definition of $\mathbf{H}_a$ given by (4.C.1) further require the magnetic induction at the boundary surface to satisfy:

$$\mathbf{n} \cdot \mathbf{H}_a = 0 \tag{4.C.3}$$

From (4.C.1) and remembering that $\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}$, we obtain:

$$\nabla^2\mathbf{E}_a(\mathbf{r}) + k_a^2\mathbf{E}_a(\mathbf{r}) = 0$$

$$\nabla^2\mathbf{H}_a(\mathbf{r}) + k_a^2\mathbf{H}_a(\mathbf{r}) = 0 \tag{4.C.4}$$

Every electromagnetic wave, being subject to Maxwell's equations in any arbitrarily shaped cavity, may therefore be decomposed in the Slater basis, *thus automatically simultaneously satisfying the boundary conditions.* The Slater basis is therefore a more generalized notion for photons in complex geometry cavities in which plane waves may no longer constitute valid cavity solutions (and as such no longer offer a viable eigenfunction basis). It is customary (following Slater) to write such a decomposition in the form:

$$\mathbf{E}(\mathbf{r},t) = -\sum_a \frac{1}{\sqrt{\varepsilon}} p_a(t)\mathbf{E}_a(\mathbf{r})$$

$$\mathbf{H}(\mathbf{r},t) = -\sum_a \frac{1}{\sqrt{\mu_0}} \omega_a q_a(t)\mathbf{H}_a(\mathbf{r}) \tag{4.C.5}$$

where $\omega_a = k_a c'$. The last two of Maxwell's equations act on the decomposition given in (4.C.5) and their effect may be summarized as:

$$p_a = \dot{q}_a$$

$$\dot{p}_a = -\omega_a^2 q_a \tag{4.C.6a}$$

which is in fact quite compact in terms of the required notation! Eliminating q_a from these last two equations, we find:

$$\hat{p}_a + \omega_a^2 p_a = 0 \tag{4.C.6b}$$

This tells us that the time dependence is constructed of functions which oscillate with frequencies satisfying the requirement that $\omega_a = k_a c'$. Also, (4.C.6) reminds us that p_a and q_a are *conjugate variables* and their quantization will give rise to observables that will not commute. The Slater bases lead to a very concise formulation of Maxwell's equations in an empty cavity of arbitrary shape. It remains, however, a somewhat formal notation as considerable work resides in finding the basis $\{\mathbf{E}_a, \mathbf{H}_a\}$. We will see that in spite of this drawback, this approach considerably simplifies the treatment of oscillation problems in cavities.

We will now interest ourselves in the electromagnetic fields which reside in a cavity filled with a laser medium characterized by a susceptibility $\chi = \chi_R + i\chi_{Im}$. The electromagnetic waves in this laser cavity are solutions to Maxwell's equations:

$$\begin{aligned} &\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r}, t) \\ &\frac{1}{\mu_0}\nabla \times \mathbf{B}(\mathbf{r}, t) = \mathbf{i}(\mathbf{r}, t) + \frac{\partial}{\partial t}\mathbf{D}(\mathbf{r}, t) \end{aligned} \tag{4.C.7a}$$

where the current density of free charge carriers $\mathbf{i}(\mathbf{r}, t)$ is given by Ohm's law to be:

$$\mathbf{i}(\mathbf{r}, t) = \sigma \mathbf{E}(\mathbf{r}, t) \tag{4.C.7b}$$

(Beware: here σ is the electrical conductivity and not a cross-section!) The displacement current $\mathbf{D}(\mathbf{r}, t)$ is given by:

$$\mathbf{D}(\mathbf{r}, t) = \frac{\partial}{\partial t}[\varepsilon \mathbf{E}(\mathbf{r}, t) + \mathbf{P}_{\text{laser}}(\mathbf{r}, t)] \tag{4.C.7c}$$

In this last equation, $\mathbf{P}_{\text{laser}}$ is the polarization vector associated with the resonant susceptibility of the laser medium, whereas the polarization due to the host medium (for example, the YAG crystal in a Nd: YAG laser) is included in the permittivity ε.

Substituting (4.C.7b) and (4.C.7c) into (4.C.7a), and projecting equation (4.C.7a) onto the Slater basis, we find immediately that:

$$\omega_a^2 q_a + \frac{\sigma}{\varepsilon} p_a + \dot{p}_a = \frac{1}{\sqrt{\varepsilon}} \frac{\partial}{\partial t} \int_V \mathbf{P}_{\text{laser}}(\mathbf{r}, t) \mathbf{E}_a(\mathbf{r}) \mathrm{d}^3\mathbf{r} \tag{4.C.8}$$

Therefore, every Slater mode is decoupled from the others. We then differentiate

(4.C.8) and use the conjugation relation (4.C.6), which leads us to:

$$\omega_a^2 p_a + \frac{\sigma}{\varepsilon}\dot{p}_a + \ddot{p}_a = \frac{1}{\sqrt{\varepsilon}}\frac{\partial^2}{\mathrm{L}\partial t^2}\int_V \mathbf{P}_{\text{laser}}(\mathbf{r},t)\mathbf{E}_a(\mathbf{r})\mathrm{d}^3\mathbf{r} \tag{4.C.9}$$

If $\sigma = 0$ and $\mathbf{P}_{\text{laser}} = 0$, we recover the dispersion relation $\omega_a = k_a c'$ with a new group velocity being given by $c' = c/n_{\text{op}}$. The term in $\mathrm{d}p_a/\mathrm{d}t$ is a damping term. The photons in the cavity escape with a time constant which is none other than the photon lifetime in the cavity:

$$\tau_c = \frac{\sigma}{\varepsilon} = \frac{Q}{\omega_a} \tag{4.C.10}$$

where, by definition, Q is the *quality factor* of the cavity. In addition to ohmic losses (4.C.7b), we may incorporate the effects of other dissipation mechanisms in σ such as mirror transmission. In this instance, it is required that the losses at the cavity extremities be averaged over the entire volume of the medium.

We will suppose that the photon lifetime is very large in comparison to the mode frequency ω_a, or again that the quality factor Q is very large. The solutions to (4.C.9) without source terms are clearly sinusoidal functions with frequencies ω_a damped slowly by a factor of e^{-t/τ_c}. It is therefore tempting to express the solutions to (4.C.9) in the form:

$$p_a(t) = p_a^{\text{slow}}(t)\mathrm{e}^{\mathrm{i}\omega t} \tag{4.C.11}$$

where $\omega_a \approx \omega$ and $p_a^{\text{slow}}(t)$ are slowly varying functions (i.e. $\mathrm{d}p_a^{\text{slow}}(t)/\mathrm{d}t \ll \omega_a p_a^{\text{slow}}(t)$). We then substitute (4.C.11) into (4.C.9) to obtain the differential equation to which $p_a^{\text{slow}}(t)$ is subjected:

$$\left[(\omega_a^2 - \omega^2) + \mathrm{i}\frac{\sigma\omega}{\varepsilon}\right]p_a^{\text{slow}} + \left(2\mathrm{i}\omega + \frac{\sigma}{\varepsilon}\right)\dot{p}_a^{\text{slow}}$$
$$= \frac{\mathrm{e}^{-\mathrm{i}\omega t}}{\sqrt{\varepsilon}}\frac{\partial^2}{\partial t^2}\int_V \mathbf{P}_{\text{laser}}(\mathbf{r},t)\mathbf{E}_a(\mathbf{r})\mathrm{d}^3\mathbf{r} \tag{4.C.12}$$

Equation (4.C.12) is then the most general equation for calculating the behaviour of a real laser cavity of arbitrary shape and mode distribution.

To facilitate our physical interpretation of this result, we will temporarily assume the existence of a single mode l, in the cavity. In (4.C.12), field polarization in the laser is then given by:

$$\mathbf{P}_{\text{laser}}(\mathbf{r},t) = \varepsilon_0\chi_{\text{laser}}\mathbf{E}(\mathbf{r},t) \tag{4.C.13a}$$

further, taking (4.C.5) and (4.C.11) into account leads to:

$$\mathbf{P}_{\text{laser}}(\mathbf{r},t) = -\frac{\varepsilon_0}{\sqrt{\varepsilon}}(\chi_{\text{R}} + \text{i}\chi_{\text{Im}})p_a^{\text{slow}}\text{e}^{\text{i}\omega t}\mathbf{E}_a(\mathbf{r}) \tag{4.C.13b}$$

where χ_{laser} is the susceptibility of the laser medium as introduced in Section 3.3. Substituting (4.C.13b) into (4.C.12) and supposing the laser to be in a stationary state, we find the self-consistent equation to which ω and χ are subject:

$$(\omega_a^2 - \omega^2) + \text{i}\frac{\sigma\omega}{\varepsilon} = \frac{\omega^2\varepsilon_0}{\varepsilon}(\chi_{\text{R}} + \text{i}\chi_{\text{Im}}) \tag{4.C.14}$$

Condition (4.C.14) is none other than the laser oscillation conditions, with the real part constituting the condition on the phase and the imaginary portion setting the condition on the gain. For example, the latter condition may be written as $\sigma = \omega\varepsilon_0\chi_{\text{Im}}$. Recall that (3.36) relates the imaginary part of the optical susceptibility to the gain by:

$$\chi_{\text{Im}} = \frac{\gamma n^2}{k} \tag{4.C.15a}$$

where n is the index of refraction and k is the wavenumber. The condition on the imaginary portion of (4.C.14) may then be written:

$$\gamma = \frac{1}{c\tau_c} \tag{4.C.15b}$$

which is nothing other than the oscillation condition (4.21b). We may similarly develop a condition on the real part of (4.C.14). We find that the resonant frequency of the cavity is not its eigenvalue, but rather is slightly offset by dephasing resulting from dispersion in the laser medium. This phenomenon is referred to as *frequency pulling*.

We will now use the Slater basis to explore the mode-locking mechanism in a laser cavity in greater detail. We will suppose that the cavity is equipped with a device allowing one to modulate the cavity loss (e.g. an acousto-optical cell, a Pockels cell, etc., . . .) as shown in Fig. 4.17. The mechanism is described by a modulated loss $\sigma(\mathbf{r},t)$ given by:

$$\sigma(\mathbf{r},t) = \sigma_{\text{ml}}\cos(\omega_{\text{ml}}t)f(\mathbf{r}) \tag{4.C.16}$$

The function $f(\mathbf{r})$ is arbitrary – the function, for instance, could be something like $f(\mathbf{r}) = \delta(z - z_0)$, which would represent a shutter placed in the $z = z_0$ plane. To simplify the analysis, we will study the cavity response without taking into account laser amplification, i.e. without introducing the source term $\mathbf{P}_{\text{laser}}$. Including this effect would complicate the required notation unnecessarily and grant us little in terms of additional insights into the problem. We will concentrate on the mechan-

ism of *mode locking*, which is independent from the mechanism of laser oscillation. Clearly, the $\mathbf{P}_{\text{laser}}$ term will have to be reintroduced to obtain a more comprehensive description of mode locking.

The equations describing the cavity are then given by Maxwell's equations:

$$\begin{aligned}\nabla \times \mathbf{E}(\mathbf{r},t) &= -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t)\\ \frac{1}{\mu_0}\nabla \times \mathbf{B}(\mathbf{r},t) &= \sigma(\mathbf{r},t)\mathbf{E}(\mathbf{r},t) + \varepsilon\frac{\partial}{\partial t}\mathbf{E}(\mathbf{r},t)\end{aligned} \tag{4.C.17}$$

Again, using the fact that $\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}$, and supposing $E\partial\sigma/\partial t \ll \sigma\partial E/\partial t$ (which is reasonable in practice), (4.C.17) implies:

$$\nabla^2\mathbf{E} - \mu_0\sigma(\mathbf{r},t)\frac{\partial}{\partial t}\mathbf{E} - \mu_0\varepsilon\frac{\partial^2}{\partial t^2}\mathbf{E} = 0 \tag{4.C.18}$$

We will now use a Slater basis taking into account the expression for $p_a(t)$ in (4.C.11), i.e:

$$\mathbf{E}(\mathbf{r},t) = \sum_a \frac{1}{\sqrt{\varepsilon_0}} p_a^{\text{slow}}(t)\mathrm{e}^{\mathrm{i}\omega_a t}\mathbf{E}_a(\mathbf{r}) \tag{4.C.19}$$

Expressing (4.C.18) in terms of this Slater basis gives:

$$\begin{aligned}&\sum_a \{p_a^{\text{slow}}(\nabla^2\mathbf{E}_a + \omega_a^2\mu_0\varepsilon\mathbf{E}_a) - (\mu_0\sigma + 2\mathrm{i}\omega_a\mu_0\varepsilon)\dot{p}_a^{\text{slow}}\mathbf{E}_a\\ &- \mathrm{i}\mu_0\omega_a\sigma_a^{\text{slow}}\mathbf{E}_a - \mu_0\varepsilon\ddot{p}_a^{\text{slow}}\mathbf{E}_a\}\mathrm{e}^{\mathrm{i}\omega_a t} = 0\end{aligned} \tag{4.C.20a}$$

Using (4.C.3) and assuming a high-Q factor for the cavity ($\sigma/\varepsilon < \omega_a$) and a slow variation of $p^{\text{slow}}(\mathrm{d}t^{\text{slow}}/\mathrm{d}t \ll \omega_a p^{\text{slow}})$, allows us to neglect the term in $\ddot{p}_a^{\text{slow}}$. Equation (4.C.20a) then becomes:

$$\sum_a \mathbf{E}_a(\mathbf{r})\omega_a\dot{p}_a^{\text{slow}}(t)\mathrm{e}^{\mathrm{i}\omega_a t} = -\sum_a \mathbf{E}_a(\mathbf{r})\frac{\sigma(\mathbf{r},t)}{2\varepsilon}\omega_a p_a^{\text{slow}}(t)\mathrm{e}^{\mathrm{i}\omega_a t} \tag{4.C.20b}$$

We can then use expression (4.C.16) for the loss modulation and project the resulting equation onto a Slater mode $\mathbf{E}_q(\mathbf{r})$, which gives:

$$\begin{aligned}\dot{p}_q^{\text{slow}} &= -\sum_a S_{aq}\frac{\sigma_{\text{ml}}\cos\omega_{\text{ml}}t}{2\varepsilon}p_a^{\text{slow}}\mathrm{e}^{\mathrm{i}(\omega_a-\omega_q)t}\\ &= -\sum_a S_{aq}\frac{\sigma_{\text{ml}}}{4\varepsilon}p_a^{\text{slow}}\mathrm{e}^{\mathrm{i}(\omega_a-\omega_q\pm\omega_{\text{ml}})t}\end{aligned} \tag{4.C.21}$$

where S_{aq} is the overlap integral:

$$S_{aq} = \frac{\omega_a}{\omega_q} \int_V f(\mathbf{r}) \mathbf{E}_a(\mathbf{r}) \mathbf{E}_q(\mathbf{r}) \mathrm{d}^3\mathbf{r} \tag{4.C.22}$$

While perhaps not being entirely obvious, (4.C.21) is in fact the equation we were looking for. It expresses how the perturbation σ_{ml} couples the initially independent modes to one another. As integration of (4.C.21) leads to terms in $1/(\omega_q - \omega_a \pm \omega_{\mathrm{ml}})$, sizeable contributions will only result for $\omega_q \approx \omega_a \pm \omega_{\mathrm{ml}}$. Therefore, the only modes to survive (in analogy to the 'survival of the fittest' mechanism mentioned earlier) will be those having a frequency separation from neighbouring modes close to the modulation frequency ω_{ml}, i.e:

$$\delta = \omega_q - \omega_{q\pm1} \pm \omega_{\mathrm{ml}} \approx 0 \tag{4.C.23}$$

As in a cavity, the frequency separation between two adjacent modes is simply the inverse of the round-trip time in the cavity (4.25), Eq. (4.C.23) gives the synchronization condition for the modes found earlier in Section 4.7.3. Taking (4.C.23) into account, (4.C.21) becomes:

$$\dot{p}_q^{\mathrm{slow}} = \kappa_{q-1}^{\mathrm{slow}} \mathrm{e}^{\mathrm{i}\delta t} + \kappa p_{q+1}^{\mathrm{slow}} \mathrm{e}^{-\mathrm{i}\delta t} \tag{4.C.24a}$$

where the coupling constant κ is given by:

$$\kappa = \frac{S_{q,q\pm1}\sigma_{\mathrm{ml}}}{4\varepsilon} \approx \frac{\sigma_{\mathrm{ml}}}{4\varepsilon} \int_V f(\mathbf{r}) \mathbf{E}_q(\mathbf{r}) \mathbf{E}_{q\pm1}(\mathbf{r}) \mathrm{d}^3\mathbf{r} \tag{4.C.24b}$$

In order to solve this system of equations (4.C.24), we write the coefficients p_q^{slow} in the form:

$$p_q^{\mathrm{slow}}(t) = \mathrm{i}c_q \mathrm{e}^{\mathrm{i}q\delta t} \tag{4.C.25}$$

Substitution of this into (4.C.24) then gives:

$$-qc_q = \frac{\kappa}{\delta} c_{q+1} + \frac{\kappa}{\delta} c_{q-1} \tag{4.C.26}$$

This recurrence relation admits as a solution:

$$c_q = I_q\left(\frac{\kappa}{\delta}\right) \tag{4.C.27}$$

where I_q is the modified Bessel function of order q. Finally, the electromagnetic field in the cavity is given by (4.C.19) with the result (4.C.26) yielding:

$$\mathbf{E}(\mathbf{r}, t) = \sum_q I_q\left(\frac{\kappa}{\delta}\right) \mathrm{e}^{\mathrm{i}(\omega_0 + q\omega_{\mathrm{ml}})t} \mathbf{E}_q(\mathbf{r})$$

$$\cong \sum_q \frac{1}{\sqrt{\left(2\pi\frac{\kappa}{\delta}\right)}} e^{iq\pi/2} e^{i(\omega_0+q\omega_{ml})t} \mathbf{E}_q(\mathbf{r}) \tag{4.C.28}$$

where we have reintroduced ω_0 as the central oscillation frequency of the laser medium (which is somewhat arbitrary in relation to the derivation given above). Equation (4.C.28) shows how the phases, frequencies, and amplitudes of the different modes are strictly interrelated by the loss modulation.

Thus, use of the Slater mode formalism has allowed us to show that a modulated loss in a cavity will synchronize modes whose frequency separation approaches the modulation frequency. Equation (4.C.25) also shows that if the losses are distributed uniformly throughout the cavity ($f(\mathbf{r}) = \text{constant}$), then mode locking becomes ineffective. This formalism can be applied to many situations and remains an infallible approach for solving any and all laser problems. The most common Slater basis employed by far is that consisting of Gaussian functions.

Finally, it is worth noting that this formalism only describes a medium subject to inhomogeneous broadening where each mode may oscillate independently of the others. In a homogeneous gain medium, mode locking is still possible. In this case, the lateral modes $\omega_0 \pm q\omega_{ml}$ are created by non-linear interactions between the laser medium and the modulator oscillations ω_{ml}. The formalism required to describe the ensuing mode locking in the cavity then becomes very different from that developed in this section.

FURTHER READING

W. H. Louisell, *Quantum Statistical Properties of Radiation*, Wiley, New York (1973).
A. E. Siegman, *Lasers*, University Science Book, Mill Valley, CA (1986).
J. C. Slater, *Microwave Electronics*, Van Nostrand, Princeton, NJ (1964).
A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

4.D The Schawlow–Townes limit and Langevin-noise force

One of the most cited particularities of laser emission (outside of its brilliance) is its highly monochromatic nature. The question may then be asked, 'What limits the spectral width of laser emission?' We might reasonably think the finesse of the cavity resonance or the width of the atomic transition, which participates in the stimulated emission, as being the causes. Both of these answers, however, are very wrong! In fact, taking these considerations into account leads to expressions several orders of magnitude larger than what is observed! For example, for a

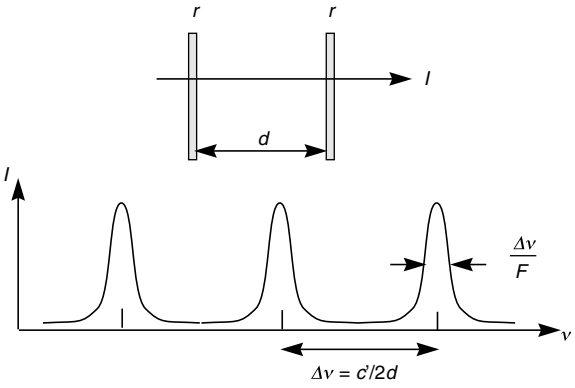


Fig. 4.D.1. The bandwidth of the transmission peaks in a Fabry–Pérot cavity is equal to the mode spacing $\Delta\nu = c'/2d$ divided by the cavity finesse.

Fabry–Pérot type cavity (of length d, with mirror reflectance r, see Fig. 4.D.1), the cavity finesse $\mathscr{F}$ and the width of the Fabry–Pérot resonance frequency $\Delta\nu_c$ are given (consult Complement 9.D) by:

$$\mathscr{F} = \pi\frac{\sqrt{r}}{1-r} \approx \frac{2\pi}{T}$$

$$\Delta\nu_c = \frac{1}{\mathscr{F}}\frac{c}{2d} = \frac{1}{2\pi\tau_c} \tag{4.D.1}$$

where τ_c is the photon lifetime in the cavity given by (4.24b). For a cavity 10 cm in length and a transmission coefficient T of 10^{-2}, we obtain a finesse of 628 and a bandwidth of 2.4×10^6 Hz, in stark contrast with experimental laser bandwidths found to be of the order of 1 Hz for certain lasers. Possible contributions from the widths of atomic transitions are often of the order of gigahertz (see Table 4.1). To understand the origin of the extremely narrow bandwidths associated with laser emission, we need to return to and, to some extent, modify the dynamic equations given for photons in a cavity (4.A.12). To begin, we recall that the electric field $\mathscr{E}(x,t)$ at any point x in the laser cavity is given by $\mathscr{E}(x,t) = \mathrm{Re}(E(x,t)\mathrm{e}^{-\mathrm{i}\omega t})$, where time dependence of the envelope function $E(x,t)$ describes the time evolution of the electromagnetic field due to laser amplification. This dependence is much slower than that described by the electromagnetic field oscillations given by $\mathrm{e}^{-\mathrm{i}\omega t}$. In (4.A.12), we replace the number of photons in the cavity P, by the envelope of the electric field $E(t)$ defined at any given point in the cavity (supposing the entrance mirror to be at $x = 0$) by recalling the relationship between these two quantities (see (2.77)) to be:

$$P = \frac{\varepsilon|E|^2 V}{2\hbar\omega} \tag{4.D.2}$$

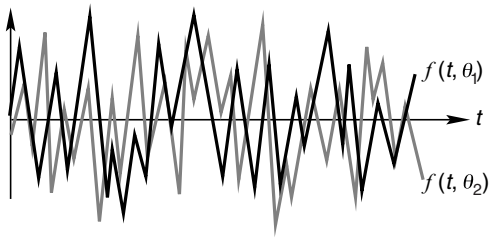


Fig. 4.D.2. A stochastic process is a random function $f(t, \theta)$ where each realization (or draw) θ leads to a function $f(t)$. At any given time t, $f(t, \theta)$ is a random variable.

where ε is the dielectric constant of the medium ($\varepsilon = \varepsilon_0 n_{\text{op}}^2$). We note that the spontaneous emission term (the $+1$ term in (4.A.12)) does not tell us anything about the moment when the photon is spontaneously released. We therefore replace it by a force $F(t)$ which will be used to describe the random character of this emission. We then obtain:

$$\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{1}{2}\left(\frac{1}{\tau_c} - \frac{G}{\tau_{\text{light}}}\right)E = F(t) \tag{4.D.3}$$

Langevin's equation

where $F(t)$ is the *Langevin force* and has the following characteristics:

- $F(t)$ is a stochastic process or a random variable with time-dependent behaviour. Figure 4.D.2 depicts such a stochastic process. It is a function $F(t, \theta)$ which is randomly selected at $t = -\infty$, and where θ is an event belonging to an ensemble of possible results Ω. For each draw (i.e. for each value of θ), $F(t, \theta)$ is a function of time t, and for each time value t, $F(t, \theta)$ is a random variable.
- $F(t)$ is an ergodic process (see Complement 3.A), i.e. implying that the temporal and statistical averages are identical at all times t:

$$\bar{F} = \int_{\Omega} F(t, \theta)\mathrm{d}\theta = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} F(t, \theta)\mathrm{d}t \tag{4.D.4}$$

- The process is without memory, i.e. the random variables $F(t)$ and $F^*(t + \tau)$ are uncorrelated for any $\tau \neq 0$:

$$\overline{F(t)F^*(t + \tau)} = A\delta(\tau) \tag{4.D.5a}$$

where A is a constant to be determined later on. We further suppose that:

$$\overline{F(t)F(t + \tau)} = 0 \tag{4.D.5b}$$

This is not absolutely necessary (it could be written in the form of $B\delta(\tau)$), but it greatly simplifies calculations.

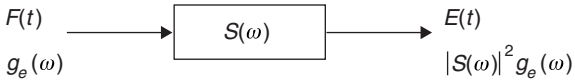


Fig. 4.D.3. A system with linear gain $S(\omega)$ transforms a stochastic process of spectrum $g_e(\omega)$ into a stochastic process with $g_s(\omega) = |S(\omega)|^2 g_e(\omega)$.

- Finally, $F(t)$ is a process with null average:

$$\overline{F(t)} = 0 \tag{4.D.6}$$

Let us now write the solution to (4.D.3). It can easily be shown that this solution is given by:

$$E(t) = \int_0^t F(t')\mathrm{e}^{\upsilon(t-t')}\mathrm{d}t' \tag{4.D.7}$$

where we have temporarily set:

$$\upsilon = \frac{1}{2}\left(\frac{1}{\tau_c} - \frac{G}{\tau_{\text{light}}}\right) \tag{4.D.8}$$

Equation (4.D.7) may be interpreted as the response of a system with linear gain (where the response to the impulse varies as $\mathrm{e}^{\upsilon t}$) to a random excitation $F(t)$. This is a classic problem in electronics (see Fig. 4.D.3). We first calculate the average value of the electric field $E(t)$:

$$\overline{E(t)} = \int_0^t \overline{F(t')}\mathrm{e}^{\upsilon(t-t')}\mathrm{d}t' = 0 \tag{4.D.9}$$

from (4.D.6). The variance of the field $E(t)$ is given by:

$$\begin{aligned}\overline{|E(t)|^2} &= \overline{\int_0^t F(t')\mathrm{e}^{\upsilon(t-t')}\mathrm{d}t' \int_0^t F^*(t'')\mathrm{e}^{\upsilon(t-t'')}\mathrm{d}t''} \\ &= \int_0^t \mathrm{d}t' \int_0^t \mathrm{d}t'' \overline{F(t')F^*(t'')}\mathrm{e}^{\upsilon(2t-t'-t'')}\end{aligned} \tag{4.D.10}$$

Using the fact that the process is without memory (4.D.5), this last expression takes the form:

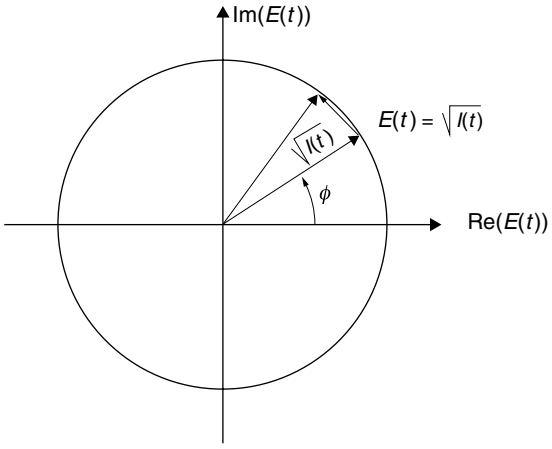


Fig. 4.D.4. Fresnel representation of the amplitude and phase of an electric field. Under the effect of the Langevin force, the vector $E(t)$ rotates in the plane (the phase is said to diffuse) while the amplitude remains largely unaffected.

$$\overline{|E(t)|^2} = A\int_0^t \mathrm{d}t' \int_0^t \mathrm{d}t''\delta(t'=t'')\mathrm{e}^{\upsilon(2t-t'-t'')} = A\int_0^t \mathrm{d}t'\mathrm{e}^{2\upsilon(t-t')} \tag{4.D.11}$$

or replacing υ by its value (4.D.8):

$$\overline{|E(t)|^2} = \frac{A}{(G/\tau_{\text{light}} - 1/\tau_c)}[\mathrm{e}^{(G/\tau_{\text{light}}-1/\tau_c)t} - 1] \tag{4.D.12a}$$

or again, by introducing the number of photons P with the help of (4.D.2):

$$P = \frac{\varepsilon VA}{2\hbar\omega}\frac{1}{(G/\tau_{\text{light}} - 1/\tau_c)}[\mathrm{e}^{(G/\tau_{\text{light}}-1/\tau_c)t} - 1] \tag{4.D.12b}$$

Comparing this expression to that obtained within the framework of Complement 4.A (see (4.A.13)), we see that the two approaches lead to an identical result by identifying A as:

$$A = \frac{2G\hbar\omega}{\varepsilon\tau_{\text{light}}V} \tag{4.D.13}$$

The approach of the Langevin equation is better, however, than that of the 'excess photon' introduced in Complement 4.A, in the sense that it describes the random dynamics associated with laser emission. We will now introduce the amplitude $I(t)$ and the phase $\phi(t)$ of the electric field, see Fig. 4.D.4:

$$E(t) = \sqrt{I(t)}\mathrm{e}^{\mathrm{i}\phi(t)} \tag{4.D.14}$$

Evidently, $I(t)$ and $\phi(t)$ are now stochastic processes. Substituting (4.D.14) into

(4.D.3), we obtain the following system of differential equations:

$$\frac{\mathrm{d}I}{\mathrm{d}t} + 2\upsilon I = \sqrt{I}\,\mathrm{Re}[F(t)\mathrm{e}^{-\mathrm{i}\phi}] \tag{4.D.15a}$$

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{1}{\mathrm{i}\sqrt{I}}\,\mathrm{Im}[F(t)\mathrm{e}^{-\mathrm{i}\phi}] \tag{4.D.15b}$$

We see that in (4.D.15a), I is damped by the term $2\upsilon I$. Consequently, the amplitude fluctuations of lasers are very small and often neglected. We will perform this calculation in Complement 13.G for the case of semiconductor lasers. Equation (4.D.15b), however, has no such damping term on the phase. Therefore, the phase fluctuations will be considerable. This effect results from the random temporal characteristics associated with spontaneous light emission. We rewrite (4.D.15b) in the form:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{1}{2\mathrm{i}\sqrt{\bar{I}}}[F(t)\mathrm{e}^{-\mathrm{i}\phi} - F^*(t)\mathrm{e}^{+\mathrm{i}\phi}] \tag{4.D.16}$$

where, considering the weak fluctuations in amplitude I, we have replaced it by its average value $\bar{I}$. Equation (4.D.16) may be formally integrated as:

$$\phi(t) - \phi(t_0) = \frac{1}{2\mathrm{i}\sqrt{\bar{I}}}\int_{t_0}^{t}[F(t')\mathrm{e}^{-\mathrm{i}\phi} - F^*(t')\mathrm{e}^{+\mathrm{i}\phi}]\mathrm{d}t' \tag{4.D.17}$$

Clearly, the average value of ϕ is zero. The variance, however, is given by:

$$\begin{aligned}\overline{\Delta\phi^2} &= \overline{[\phi(t) - \phi(t_0)]^2} \\ &= -\frac{1}{4\bar{I}}\int_{t_0}^{t}\mathrm{d}t'\int_{t_0}^{t}\mathrm{d}t''\,\overline{[F(t')\mathrm{e}^{-\mathrm{i}\phi} - F^*(t')\mathrm{e}^{+\mathrm{i}\phi}][F^*(t'')\mathrm{e}^{\mathrm{i}\phi} - F(t'')\mathrm{e}^{-\mathrm{i}\phi}]}\end{aligned} \tag{4.D.18a}$$

This last expression may be considerably simplified if we utilize the relationship in (4.D.5) and further suppose that:

$$\overline{F(t)\mathrm{e}^{-\mathrm{i}\phi}} = \overline{F(t)}\;\overline{\mathrm{e}^{-\mathrm{i}\phi}} \tag{4.D.18b}$$

This last expression generally holds, in spite of the fact that the processes F and ϕ are clearly correlated. Equation (4.D.18) then becomes:

$$\overline{\Delta\phi^2} = -\frac{1}{4\bar{I}}\int_{t_0}^{t}\mathrm{d}t'\int_{t_0}^{t}\mathrm{d}t''\,\overline{[F(t')F^*(t'')} + \overline{F(t'')F^*(t')]} \tag{4.D.19}$$

$$= \frac{A}{2\bar{I}} \int_{t_0}^{t} \mathrm{d}t' \int_{t_0}^{t} \mathrm{d}t'' \delta(t' = t'') = \frac{A}{2\bar{I}} \int_{t_0}^{t} 1 \, \mathrm{d}t' = \frac{A}{2\bar{I}}(t - t_0)$$

We find therefore that the standard deviation of the phase changes with time as described by:

$$\overline{\Delta\phi} = \sqrt{\overline{\Delta\phi^2}} = \sqrt{\frac{A}{2\bar{I}}(t - t_0)} \tag{4.D.20}$$

We recognize (4.20) takes the form of a *diffusion equation* where the diffusion constant D is given by $D = A/2\bar{I}$, or given the expressions for A in (4.D.13) and I from (4.D.2) and (4.D.24a):

$$D = \frac{A}{2\bar{I}} = \frac{1}{2} \frac{2\hbar\omega G}{\varepsilon\tau_{\text{light}} V} \frac{\varepsilon V}{2\hbar\omega\bar{P}} = \frac{1}{2} \frac{G}{\tau_{\text{light}}} \frac{1}{\bar{P}} \tag{4.D.21}$$

Above laser threshold, the gain G is *clamped* to its threshold value given by $G/\tau_{\text{light}} = 1/\tau_c$. Moreover, (4.D.1) tells us that the bandwidth of the cold cavity (i.e. the bandwidth of the Fabry–Pérot resonance) is given by $\Delta\nu_c = 1/2\pi\tau_c$. The diffusion coefficient for the phase is then finally found to be:

$$D = \frac{\pi\Delta\nu_c}{\bar{P}} \tag{4.D.22}$$

Phase diffusion constant (in s^{-1})

We now need only calculate the emission bandwidth for a laser. To do so, we write the electric field corresponding to the laser emission in the canonical form described at the beginning of this complement:

$$\mathscr{E}(t) = \mathrm{Re}[E(t)\mathrm{e}^{-\mathrm{i}\omega_0 t}] = \frac{1}{2}[E(t)\mathrm{e}^{-\mathrm{i}\omega_0 t} + E^*(t)\mathrm{e}^{+\mathrm{i}\omega_0 t}] \tag{4.D.23}$$

where ω_0 is the emission frequency of the laser. From Complement 3.A, we know that the spectral distribution of the signal $\mathscr{E}(t)$ is given by the Fourier transform of the autocorrelation function (Eq. (3.A.13)). This is the Wiener–Kintchine theorem (see Fig. 4.D.3). This autocorrelation function may be calculated by writing:

$$\overline{\mathscr{E}(t)\mathscr{E}(t+\tau)} =$$

$$\frac{1}{4}\overline{[E(t+\tau)\mathrm{e}^{-i\omega_0(t+\tau)} + E^*(t+\tau)\mathrm{e}^{\mathrm{i}\omega_0(t+\tau)}][E(t)\mathrm{e}^{-\mathrm{i}\omega_0 t} + E^*(t)\mathrm{e}^{\mathrm{i}\omega_0 t}]} \tag{4.D.24a}$$

or again

$$\overline{\mathscr{E}(t)\mathscr{E}(t+\tau)} = \frac{1}{4}\overline{[E(t+\tau)E^*(t)}\mathrm{e}^{-\mathrm{i}\omega_0\tau} + \text{c.c.}] \tag{4.D.24b}$$

We replace the electric field $E(t)$ by its expression in (4.D.14) and by remembering that the amplitude of the fluctuations in I are small enough to warrant its substitution by its average value:

$$\overline{E(t)E(t+\tau)} = \sqrt{\overline{I(t+\tau)I(t)}}\,\mathrm{e}^{\mathrm{i}\overline{(\phi(t+\tau)-\phi(t))}} \tag{4.D.25}$$

We have already encountered the average value for $\mathrm{e}^{\mathrm{i}(\phi(t+\tau)-\phi(t))}$ in (3.A.16) assuming a Poisson-type process. Making the same assumption about $\phi(t)$:

$$\mathrm{e}^{\mathrm{i}\overline{(\phi(t+\tau)-\phi(t))}} = \mathrm{e}^{-\frac{1}{2}\overline{\delta\phi^2}} = \mathrm{e}^{-\frac{1}{2}D\tau} \tag{4.D.26}$$

where $\delta\phi^2$ is the variance of the phase fluctuations. Consequently, the expression from the autocorrelation function (4.D.24) now becomes:

$$G_1(\tau) = \overline{\mathscr{E}(t)\mathscr{E}(t+\tau)} = \frac{1}{4}(\bar{I}\mathrm{e}^{-\frac{1}{2}D\tau}\mathrm{e}^{-\mathrm{i}\omega_0\tau} + \text{c.c.}) = \frac{\bar{I}}{2}\mathrm{e}^{-\frac{1}{2}D\tau}\cos(\omega_0\tau) \tag{4.D.27}$$

The spectral distribution for $E(\omega)$ is then given by the Wiener–Kintchine theorem (3.A.13), i.e. by taking the Fourier transform of (4.D.27), or:

$$\mathrm{i}(\omega) = \overline{|E(\omega)|^2} = \frac{\bar{I}}{\pi}\frac{D/2\pi}{(\omega-\omega_0)^2 + D^2/4} \tag{4.D.28}$$

Using the expression for the diffusion coefficient in (4.D.22), we find that the bandwidth of the spectral distribution for emitted laser light (as might be measured, for instance, by a dispersive spectrometer) is given by:

$$\Delta\nu_{\mathrm{ST}} = \frac{\Delta\nu_c}{\bar{P}} \tag{4.D.29}$$

The Schawlow–Townes linewidth

Thus, we conclude that the laser linewidth is the cavity bandwidth $\Delta\nu_c$ divided by the number of photons in the cavity (a considerable quantity). The bandwidth of a laser may therefore theoretically be extremely fine. Equation (4.D.29) can be put in a more accessible form for calculation by using (4.D.1) for $\Delta\nu_c$ and (4.30), which relates the number of photons P in the cavity to the laser output power $\mathscr{P}$:

$$\mathscr{P} = P\frac{h\nu}{\tau_c} \tag{4.D.30}$$

The Schawlow–Townes equation then takes the form:

$$\Delta\nu_{\mathrm{ST}} = \frac{h\nu}{2\pi\tau_c^2\mathscr{P}} \tag{4.D.31}$$

The Schawlow–Townes equation

Example

(a) The linewidth of a Nd^{3+}:YAG laser cavity ($h\nu = 1.17\,eV$, $n_{op} = 1.82$) 1 cm in length, having a mirror transparency of 0.5% (or a photon lifetime of 24 ns), and emitting an output power of 1 mW is $\Delta\nu_{ST} = 1.17 \times 1.6 \times 10^{-19}\,J/(2\pi \times (24 \times 10^{-9}\,s)^2 \times 10^{-3}\,W)$ or 5×10^{-2} Hz.

(b) The linewidth of a semiconductor laser ($h\nu = 1.4\,eV$, $n_{op} = 3$) 100 μm in length, possessing mirrors with a transparency of 30% (corresponding to a photon lifetime of 6 ps), and emitting an output power of 1 mW is $\Delta\nu_{ST} = 1.4 \times 1.6 \times 10^{-19}\,J/(2\pi \times (6 \times 10^{-12}\,s)^2 \times 10^{-3}\,W)$ or 1 MHz.

FURTHER READING

G. Grynberg, A. Aspect, and C. Fabre, *Introduction aux lasers et à l'optique quantique*, Ellipses, Paris (1997).

A. E. Siegman, *Lasers*, University Science Book, Mill Valley, CA (1986).

4.E A case study: diode pumped lasers

In this section, we will interest ourselves in an aspect of considerable importance in terms of the technological application of lasers – *laser efficiency*. This quantity is defined as the ratio of the output power delivered by a laser to the optical power required to pump it. We will take as an example a fairly common laser used in industry – the neodymium-doped YAG laser (or Nd^{3+}:YAG). The ensemble of atomic levels which come into play in the operation of the Nd^{3+}:YAG laser is shown schematically in Fig. 4.E.1*a*. As one may notice from the absorption spectra of this ionic system (Fig. 4.E.1*b*), the wavelength of the pump photon transition is 0.81 μm and its associated bandwidth is of the order of $\Delta\lambda = 30$ nm. Let us begin by calculating the efficiency of this laser used in conjunction with a xenon flash lamp as a pump source. A Xe lamp may be considered to be a blackbody raised to a temperature T; in which case, the spectral emittance in terms of energy is given by Eq. (2.B.4). At best, the power absorbed ΔP by the pump transition is given by the integral of the spectral emittance (2.B.4) over the absorption window $\Delta\lambda$ of the laser material (see Fig. 4.E.2) or:

$$\Delta P = \frac{\mathrm{d}}{\mathrm{d}\lambda} R(\lambda, T)\Delta\lambda = \frac{2\pi hc^2}{\lambda^5} \frac{1}{\mathrm{e}^{(hc/\lambda kT)} - 1} \Delta\lambda \tag{4.E.1a}$$

or

$$\Delta P \approx 1.14 \times 10^9\, \mathrm{e}^{-(1.79 \times 10^4/T)} \Delta\lambda\, \mathrm{W\,\mu m^{-1}\,m^{-2}} \tag{4.E.1b}$$

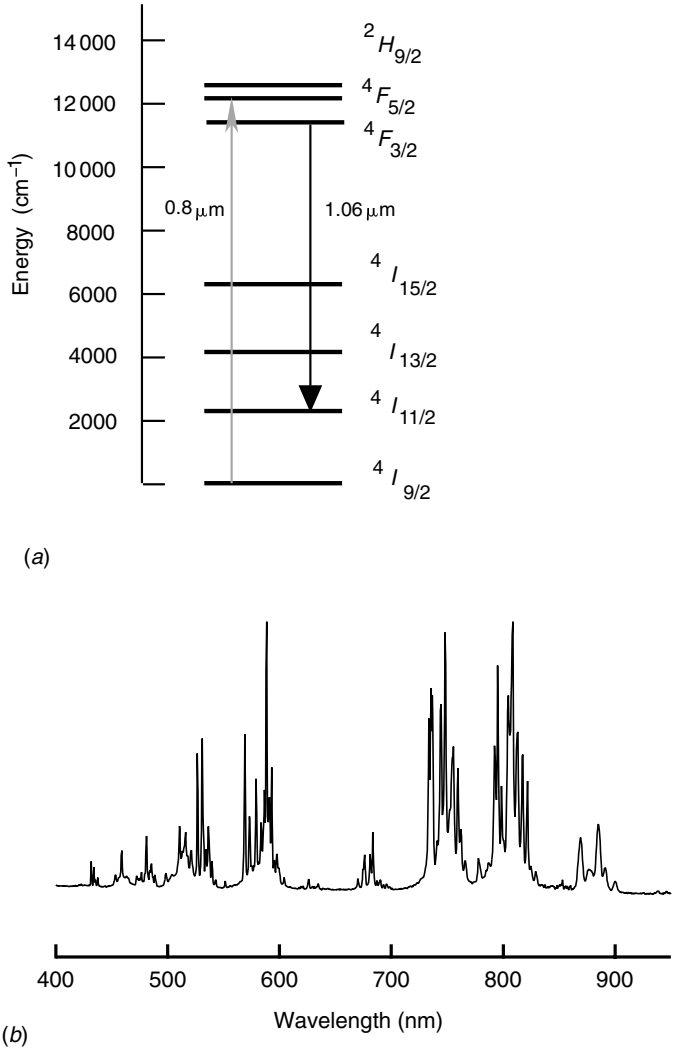


Fig. 4.E.1. (*a*) Levels involved in absorption and emission in a Nd^{3+} : YAG laser. (*b*) Absorption spectrum of Nd^{3+} : YAG (THALES).

The total energy emitted is the integral of (2.B.4) over the entire spectrum, which using Stephan's law (2.B.8) gives:

$$P = \sigma T^4 = 5.67 \times 10^{-8} T^4 \,\mathrm{W\,m^{-2}} \tag{4.E2}$$

The maximum possible efficiency of the laser is then $\Delta P/P$ or:

$$\frac{\Delta P}{P} \approx 2 \times 10^{16} T^{-4} \mathrm{e}^{-(1.79\times10^4/T)} \Delta\lambda\,(\mu\mathrm{m}) \tag{4.E.3}$$

Assuming a typical blackbody temperature of ~8000 K and a spectral absorption

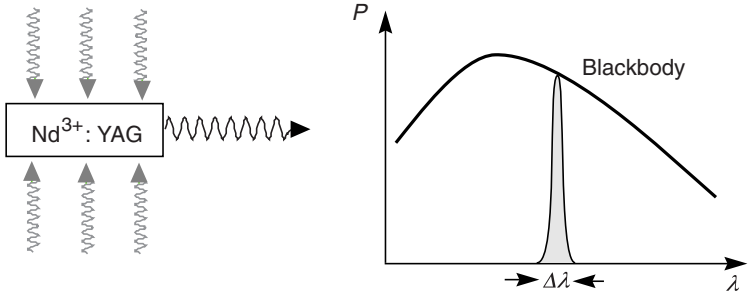


Fig. 4.E.2. In a laser pumped by a flash lamp (*a*), the overlap between the blackbody spectral distribution (very large) and the laser absorption band (very narrow) leads to poor conversion efficiencies (*b*).

width of 0.03 μm, we find an efficiency of roughly 1.5%. Therefore, intrinsically, the efficiency of lamp pumped lasers is very low as the bulk of the light emission falls outside of the absorption band of the laser (see Fig. 4.E.2).

It is therefore natural to seek a highly monochromatic and brilliant pump source as afforded by semiconductor laser diodes. At the end of this complement we will see the reason for using diode pumped lasers, as opposed to making direct use of the laser diodes themselves.

Figure 4.E.3 represents a diode pumped $Nd^{3+}:YAG$ laser assembly, as first conceived of by Robert Byer and his research team at Stanford. A lens is used to couple the otherwise divergent light beam of the GaAs laser diode to the entrance mirror of the laser rod. The laser rod shown in the figure was fashioned into the shape of a Gaussian resonator with length $d = 2L = 0.5\,\text{cm}$.

We recall that a *Gaussian beam* perfectly describes the propagation of diffraction limited electromagnetic waves. If a Gaussian wave propagates along the Oz axis, the amplitude of the electromagnetic field $U(\rho, z)$ as a function of the distance ρ from the Oz axis is given (see Fig. 4.E.4) by:

$$U(\rho,z) = A\frac{W_0}{W(z)}\mathrm{e}^{-(\rho^2/W(z)^2)}\mathrm{e}^{-\mathrm{i}kz-\mathrm{i}k\,\rho^2/2R(z)+\mathrm{i}\xi(z)} \tag{4.E.4}$$

where various parameters of interest are:

$$R(z) = z\left[1+\left(\frac{z_0}{z}\right)^2\right] \qquad \text{Radius of curvature}$$

$$W(z) = W_0\left[1+\left(\frac{z}{z_0}\right)^2\right]^{1/2} \qquad \text{Spatial extent}$$

$$z_0 = \frac{\pi W_0^2 n_{\text{op}}}{\lambda} \qquad \text{Rayleigh length} \tag{4.E.5}$$

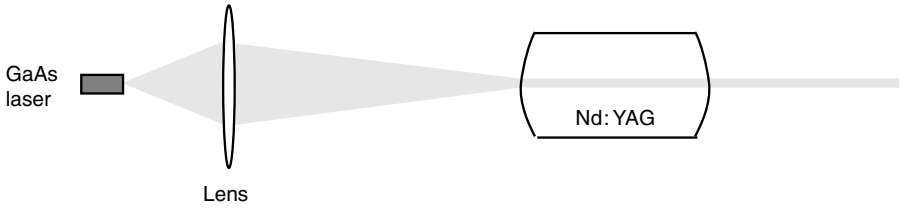


Fig. 4.E.3. Schematic of a Nd: YAG laser pumped by a laser diode.

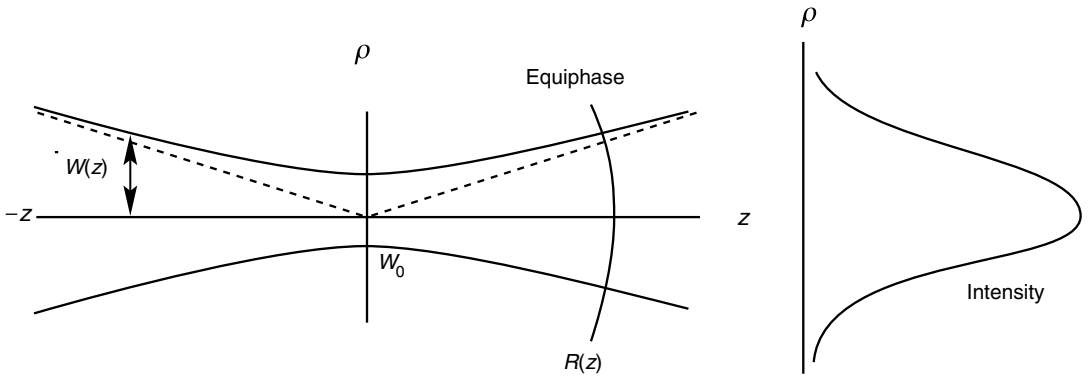


Fig. 4.E.4. Behaviour of a Gaussian beam.

$$\xi(z) = \arctan\frac{z}{z_0} \qquad \text{Guoy phase}$$

$$\theta = \frac{1}{\pi}\frac{\lambda/n_{\text{op}}}{W_0} \qquad \text{Divergence}$$

To minimize the threshold power of the pump, we are generally interested in diminishing the overall volume occupied by the laser mode. Let us suppose we have a beam width W_0 of 40 μm (this is fairly difficult to achieve with a laser diode). Since $\lambda = 1.06\,\mu\text{m}$ and $n_{\text{op}} = 1.82$, the third equation in (4.E.5) gives us the *Rayleigh length* of the beam, i.e. $z_0 = 8.6\,\text{mm}$. The radius of curvature of the entrance mirror is given by the first equation in (4.E.5), i.e. $R = 3.2\,\text{cm}$. As $z_0 \gg d/2$, we may consider $W(z) = \text{constant}$, so that the volume V occupied by the Gaussian mode is then $\pi d W_0^2$ or $2 \times 10^{-5}\,\text{cm}^{-3}$.

We are now interested in the photon lifetime in the cavity neglecting parasitic absorption α_p. The entrance mirror has 100% reflectivity for photons at 1.06 μm and the exit mirror has reflectivity $R_s = 99.7\%$, which yields a photon lifetime (see (4.24b)) of:

$$\tau_c = \frac{2d}{T_s c/n_{\text{op}}} = \frac{2 \times 0.5\,\text{cm}}{3 \times 10^{-3} \times 3 \times 10^{10}\,\text{cm s}^{-1}/1.82} = 20\,\text{ns}$$

The threshold inversion density is then (see Table 4.1 and Eq. (4.23c)):

$$n_{\text{threshold}} = \frac{1}{\sigma_{\text{op}}\tau_c c/n_{\text{op}}} = \frac{1.82}{4 \times 10^{-19}\,\text{cm}^2 \times 20\,\text{ns} \times 3 \times 10^{10}\,\text{cm s}^{-1}}$$
$$= 7.5 \times 10^{15}\,\text{cm}^{-3}$$

and the power density at threshold is given by:

$$p_{\text{threshold}} = \frac{n_{\text{threshold}}\hbar\omega_{\text{pump}}}{\tau_2} = \frac{7.5 \times 10^{15}\,\text{cm}^{-3} \times 1.5\,\text{V} \times 1.6 \times 10^{-19}\,\text{C}}{1.2 \times 10^{-3}\,\text{s}}$$
$$= 1.5\,\text{W cm}^{-3}$$

The pump power at threshold is then given by:

$$P_{\text{threshold}} = p_{\text{threshold}} V = 0.03\,\text{mW}$$

The required pump power at threshold is particularly low for two reasons:

- the volume V is very small (enhanced by the Gaussian resonator);
- all the pump photons have energies that are resonant with the pump transition in the laser medium (i.e. all photons are useful in pumping the laser medium).

Clearly, parasitic loss mechanisms (absorption by impurities, . . .) will tend to increase the threshold pump power requirement. Nonetheless, diode pumped lasers maintain extremely low threshold pump power requirements.

The output power P_s far above threshold is given by (4.28b), (4.29), and (4.24b), or in our case ($\alpha_p = 0$, $R_e = 1$, and $T_s \ll 1$):

$$P_s \approx \frac{1}{2} R_2 \tau_c c/n_{\text{op}} T_s \hbar\omega_{\text{laser}} = dR_2\hbar\omega_{\text{laser}} \tag{4.E.6}$$

We recall that the pump rate R_2 is in $\text{cm}^{-3}\,\text{s}^{-1}$. Supposing all the pump photons are absorbed along the total length d of the laser rod, the pump rate is then related to pump power by:

$$P_{\text{pump}} = dR_2\hbar\omega_{\text{pump}} \tag{4.E.7}$$

The efficiency is then found to be:

$$\frac{P_s}{P_{\text{pump}}} \approx \frac{\hbar\omega_{\text{laser}}}{\hbar\omega_{\text{pump}}} = 1 - \eta_q \tag{4.E.8}$$

which is easy to understand conceptually: η_q is the *laser quantum defect*. In this case, the excess energy in the pump photons (i.e. the quantum defect) is transformed into heat.

A threefold advantage is gained in using diode pumped lasers over the direct use of laser diodes. First, the Gaussian resonator in the YAG rod plays the role of a mode transformer. As a result, the optical power of several GaAs laser diodes

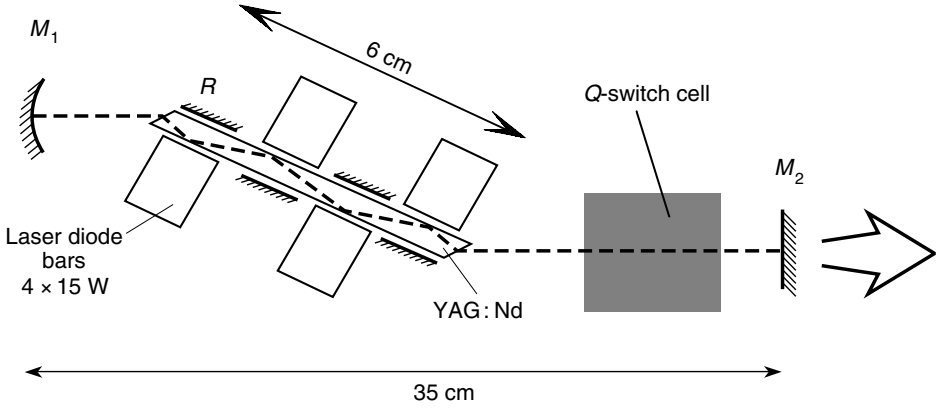


Fig. 4.E.5. Experimental set-up for a *Q*-switched Nd: YAG laser pumped by laser diodes (courtesy of J. P. Pocholle, THALES).

possessing relatively poor beam characteristics can be efficiently converted into a single diffraction limited beam by the resonator. Also, as explained in Complement 4.D, the bandwidth of a Nd: YAG laser is significantly less than that of a semiconductor laser (the principal reason being the large differences in the photon lifetimes between the two mediums). Last, thanks to the long lifetime of its excited state ($\tau_2 = 1.2\,\text{ms}$), the Nd: YAG laser allows for *Q-switched* operation, resulting in the production of laser pulses with peak power levels 10^4 times greater than the average power levels supplied directly by the GaAs diodes. Figure 4.E.5 shows the configuration of an actual diode pumped Nd: YAG laser.

FURTHER READING

W. Koechner, *Solid-State Laser Engineering*, 4th Edn, Springer, Berlin (1996).
T. Y. Pan and R. L. Byer, *IEEE J. Quantum Electron* **QE 24**, 895 (1988).

5 Semiconductor band structure

5.1 Introduction

In this chapter, for the first time we touch upon condensed matter physics *per se*. Until now, we have studied quantum systems comprising two, three, and up to four quantized levels situated either in a vacuum or in a medium characterized by a given index of refraction. We have not, however, considered the problem of interaction of many such systems present at elevated concentrations and leading to chemical bonding on a mass scale. This problem, a priori insoluble given that there are as many coupled equations needing to be solved as there are electrons in matter (of the order of $10^{22}\,\mathrm{cm}^{-3}$), becomes greatly simplified in the case of crystalline or semicrystalline materials.

5.2 Crystal structures, Bloch functions, and the Brillouin zone

The most stable form of matter at zero kelvin results from the periodic arrangement of atoms into a crystalline structure. The number of possible crystalline structures is immense and forms the subject of a science in itself – *crystallography*. A crystal structure is characterized by the periodic repetition (*ad infinitum* or close to in practice!) from point to point (with the ensuing pattern being referred to as a *lattice*) of a basis consisting of a single atom or assembly of atoms. The term 'monoatomic lattice' is used to describe the lattice that results when a basis comprising a single atom is used to fill the lattice. The most common lattice type for elemental semiconductor crystals (Si, Ge, C) is that of diamond. For binary semiconductors, such as GaAs, the most common structure is that of zincblende, shown in Fig. 5.1.

Of immediate concern is the fact that the crystal lattice may be spanned by linear combinations of three fundamental vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, i.e. all atomic bases are distributed at points $\mathbf{r}_i$ such that:

$$\mathbf{r}_i = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \tag{5.1}$$

where the n_is correspond to arbitrary integers. The Hamiltonian which describes the interaction between the electrons and the atoms in the crystal is extremely complex as it must take into account all the electron interactions, the nuclear

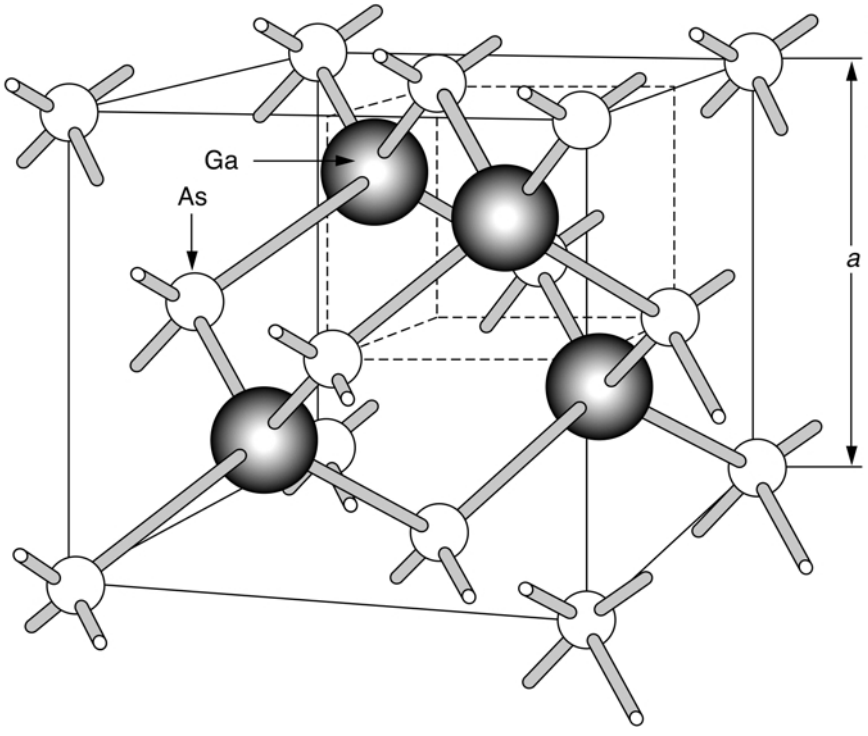


Fig. 5.1. The zincblende crystal structure.

motions, etc. A primary aim of solid state physics then is to show that many of these interactions average out across the crystal to the extent by which they may be successfully represented by an effective potential $U(\mathbf{r})$. Given the periodic nature of the lattice, the Hamiltonian used to describe an electron in a crystal takes the form:

$$H_{\text{cryst}} = \frac{\mathbf{p}^2}{2m} + \sum_i U(\mathbf{r} - \mathbf{r}_i) \tag{5.2}$$

or alternately:

$$H_{\text{cryst}} = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r}) \tag{5.3}$$

where $V(\mathbf{r})$ possesses the periodicity of the crystal:

$$V(\mathbf{r} + \mathbf{r}_i) = V(\mathbf{r}) \tag{5.4}$$

Given the periodicity condition (5.4), the periodic potential may be expanded in terms of a Fourier series:

$$V(\mathbf{r}) = \sum_{\mathbf{G}\in\mathbf{RL}} \tilde{V}(\mathbf{G}) \mathrm{e}^{-\mathrm{i}\mathbf{G}\mathbf{r}} \tag{5.5}$$

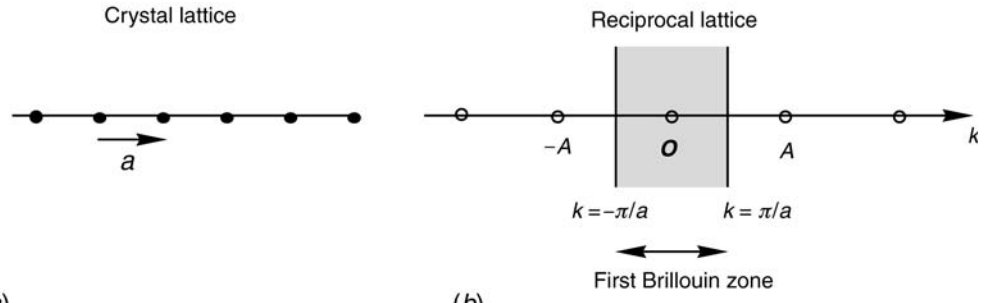


Fig. 5.2. Crystal lattice (*a*), and its associated reciprocal lattice (*b*). The translation vector for the reciprocal lattice is the vector **A** located at $2\pi/a$. The two perpendicular lines at the midpoints of OA and $-OA$ define the first Brillouin zone of the reciprocal lattice.

where the points $\mathbf{G}(OG = \mathbf{G})$ are distributed in a reciprocal lattice (RL) space spanned by the unitary vectors:

$$\mathbf{A}_1 = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}; \mathbf{A}_2 = 2\pi \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_2 \cdot \mathbf{a}_3 \times \mathbf{a}_1}; A_3 = 2\pi \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_3 \cdot \mathbf{a}_1 \times \mathbf{a}_2} \tag{5.6}$$

The construction of reciprocal lattices of crystalline structures lies within the venerable domain of geometry. Figure 5.2 shows a one-dimensional, monoatomic lattice (*a*) alongside its associated reciprocal lattice (*b*).

We recall that in a vacuum, where $H = \mathbf{p}^2/2m$, the stationary wavefunctions of free electrons are represented by the wavevectors **k** and may be written in the form of $\exp(-i\mathbf{k} \cdot \mathbf{r})$. It can easily be shown (Complement 5.A), that the solutions to Schrödinger's equation $H_{\text{crystal}}\Psi(\mathbf{r}) = E\Psi(\mathbf{r})$ are Bloch–Floquet functions, and that they may be indexed according to their wavevector **k**:

$$\Psi(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}} \tag{5.7}$$

Bloch–Floquet functions

where the functions $u_{\mathbf{k}}(\mathbf{r})$ possess the periodicity of the crystal:

$$u_{\mathbf{k}}(\mathbf{r} + \mathbf{r}_i) = u_{\mathbf{k}}(\mathbf{r}) \tag{5.8}$$

The real (or imaginary) portion of a Bloch–Floquet function is shown in Fig. 5.3. We note the general form of the functions, which consists of rapidly varying atomic wavefunctions under a more slowly modulated envelope function $\exp(-i\mathbf{k} \cdot \mathbf{r})$. It is these modulations that are transported within the crystal and which lie behind the phenomenon of *effective mass* described later on.

As usual, when dealing with an infinite medium, the wavevectors **k** are *pseudo-quantized* for reasons of convenience (see Complement 1.A). This can be achieved by introducing infinite potential barriers along the six planes defined by:

$$n_1 = 0, n_1 = n_{\max}$$

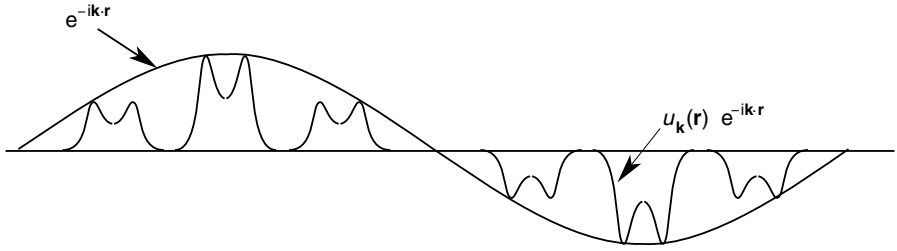


Fig. 5.3. Bloch–Floquet functions (stationary electron states in the periodically varying potential of atoms distributed throughout a crystalline lattice) are the periodic modulation of atomic wavefunctions by a travelling wave of the form $e^{-i\mathbf{k}\cdot\mathbf{r}}$.

$$n_2 = 0, n_2 = n_{\max} \tag{5.9}$$

$$n_3 = 0, n_3 = n_{\max}$$

In this case, pseudo-quantization would have led arbitrarily to positive discrete **k** vectors. The Bloch wavefunctions would therefore have been stationary as they result from waves propagating back and forth in both directions between the limiting surfaces. It is customary in solid state physics to employ another pseudo-quantization procedure: the *Born–von Karman cyclic boundary conditions*. We will suppose that at all $\mathbf{L}_i = N\mathbf{a}_i$ $(i = 1, 2, 3)$, the crystal repeats itself identically, or in other terms, that all space within the crystal is completely filled by adjacent boxes having dimensions of $L_1 \times L_2 \times L_3$. The Born–von Karman cyclic conditions require that the wavefunction $\Psi(\mathbf{r})$ be the same within each of these boxes, i.e. that $\Psi(\mathbf{r} + n_1\mathbf{L}_1 + n_2\mathbf{L}_2 + n_3\mathbf{L}_3) = \Psi(\mathbf{r})$. The wavevectors **k** then belong to a reciprocal lattice described by the points:

$$k_1 = n_1 \frac{2\pi}{Na_1}; k_2 = n_2 \frac{2\pi}{Na_2}; k_3 = n_3 \frac{2\pi}{Na_3} \tag{5.10}$$

where n_1, n_2, and n_3 are now positive or negative integers. Clearly, the quantity N may be made arbitrarily large, in such a manner as to keep unobservable any effects relating to the pseudo-quantification procedure.

Together with the Born–von Karman boundary conditions, the Schrödinger equation, (5.3), has solutions of Bloch form (5.7), i.e. $\psi_{\mathbf{k}}(\mathbf{r})$. For each **k** there is a whole family of solutions whose eigenenergies we designate as $E_m(\mathbf{k})$, m being an integer. We can conceive that these families form continuous *bands* in the sense that for $\mathbf{k}'$ close to **k**, $E_m(\mathbf{k}')$ will be close to $E_m(\mathbf{k})$. Therefore, m is referred to as a *band index*.

Over larger scales, however, we must realize that an eigenstate of Bloch form (5.7) corresponding to the band m, with **G** a reciprocal lattice vector, can also be written as:

$$\psi_{m\mathbf{k}}(\mathbf{r}) = u_{m\mathbf{k}}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}} = [u_{m\mathbf{k}}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\mathbf{G}\cdot\mathbf{r}}]\mathrm{e}^{-\mathrm{i}(\mathbf{k}-\mathbf{G})\cdot\mathbf{r}} \tag{5.11}$$

which is again of Bloch form since $\mathrm{e}^{-\mathrm{i}\mathbf{G}\cdot\mathbf{r}}$ is a periodic function of the crystal lattice. It is therefore *identical* to one of the solutions of the Schrödinger equation for the wavevector $\mathbf{k} - \mathbf{G}$. For the eigenenergies this means that we have:

$$E_{m'}(\mathbf{k} - \mathbf{G}) = E_m(\mathbf{k}) \tag{5.12}$$

for some m'. The energy bands are consequently *periodic in* $\mathbf{k}$, with the period given by the *reciprocal lattice*.

A further consequence of the fact that $\psi_{m',\mathbf{k}-\mathbf{G}}(\mathbf{r})$ is identical to and the same solution as $\psi_{m,\mathbf{k}}(\mathbf{r})$ is that to present all the solutions we *do not need all of the reciprocal space*.

A common practice (as explained in Complement 5.A) is to restrict the integer values n_i in (5.10) to the interval $-N/2 < n_i \leq N/2$ (i.e. $n_{1,2,3} = 0, \pm 1, \pm 2, \ldots, \pm N/2$). Thus, the wavevectors $\mathbf{k}$ belong to a zone defined by:

$$-\frac{\pi}{a_1} \leq k_1 \leq \frac{\pi}{a_1}; -\frac{\pi}{a_2} \leq k_2 \leq \frac{\pi}{a_2}; -\frac{\pi}{a_3} \leq k_3 \leq \frac{\pi}{a_3} \tag{5.13}$$

This zone is defined by a set of perpendicularly bisecting planes which cross the reciprocal lattice (RL) translation vectors (connecting the nearest neighbouring RL points) at their respective mid-points. The region of space in the RL delimited in such a fashion is referred to as the first Brillouin zone. Figure 5.2 represents the first Brillouin zone in a one-dimensional structure. Figure 5.4*a* and *b* shows the first Brillouin zone for a two-dimensional square lattice and for a three-dimensional zincblende structure. The most important points (and corresponding direction vectors), in the case of the latter, have been labelled according to convention.

In some cases it is more convenient to exploit the periodicity of the bands to count the states, by associating one eigenenergy with each point in reciprocal space. A particular case is the nearly free electron model studied in Complement 5.A. To return to the standard band structure in the first Brillouin zone from this representation of solutions, one should perform a translation of the band, as in Eq. (5.12), by the reciprocal lattice wavevector $\mathbf{G}$, which makes $\mathbf{k} - \mathbf{G}$ lie in the first Brillouin zone. This procedure is known as *band folding* and is shown explicitly in Fig. 5.A.2.

One of the principal aims of solid state physics is to solve Schrödinger's equation, obtained by the introduction of (5.11) into (5.2). This is generally carried out using heavy numerical methods and allows one to obtain a relationship $E_m(\mathbf{k})$ between the electron energies in a structure, their corresponding wavevectors $\mathbf{k}$ in the first Brillouin zone, and the band index m. The ensemble of resulting curves $E_m(\mathbf{k})$ defines the *band structure* of a material. We will content ourselves here with summarizing the principal aspects of this important area of solid state physics.

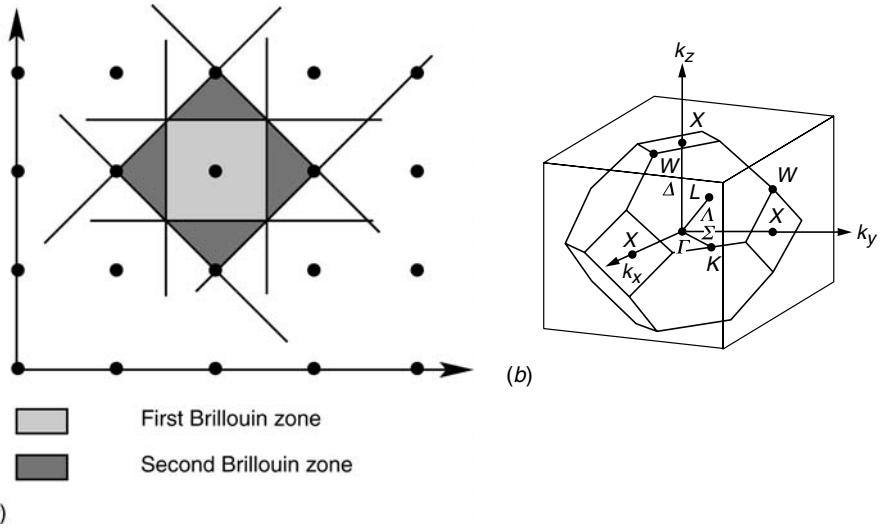


Fig. 5.4. First and second Brillouin zones for a two-dimensional square lattice (*a*) and the first Brillouin zone for a three-dimensional zincblende structure (*b*). In the latter case, the more significant points of symmetry have been labelled according to convention.

5.3 Energy bands

One of the most spectacular consequences of interaction between electrons in a crystalline lattice and the periodically varying crystal potential is the existence of forbidden energy bands for the electrons. There are several ways of understanding the origin of these forbidden bands and, of course, all of them are interrelated.

From the point of view of Chapter 1, the wavelength of an electron of energy E in vacuum is given by $\lambda = 2\pi\hbar/\sqrt{2mE}$. The typical ionization energy for electrons in an atom is $\sim$5 eV, corresponding to a wavelength of 5 Å. As atoms in a crystal are typically separated by 3 to 5 Å, the electronic matter-waves will be diffracted by the periodic crystal potential (see Fig. 5.5). More precisely, those electronic waves for which $2\pi/\lambda = k = \pi/a_i$, i.e. those which are located at the edge of the Brillouin zone, are diffracted and cannot propagate through the crystal thus forming a zone of forbidden energies. The theoretical approach founded upon this conceptual image corresponds to the *nearly free electron model*. This method is examined more closely in Complement 5.A.

A second viewpoint relies upon the chemical model. We may recall from Complement 1.B, that when two atoms approach one another, their orbitals become hybridized. The degeneracy between the energy levels is lifted and two distinct levels result (the bonding and anti-bonding configurations). Generalization of this phenomenon to an infinite number of atoms leads to the appearance

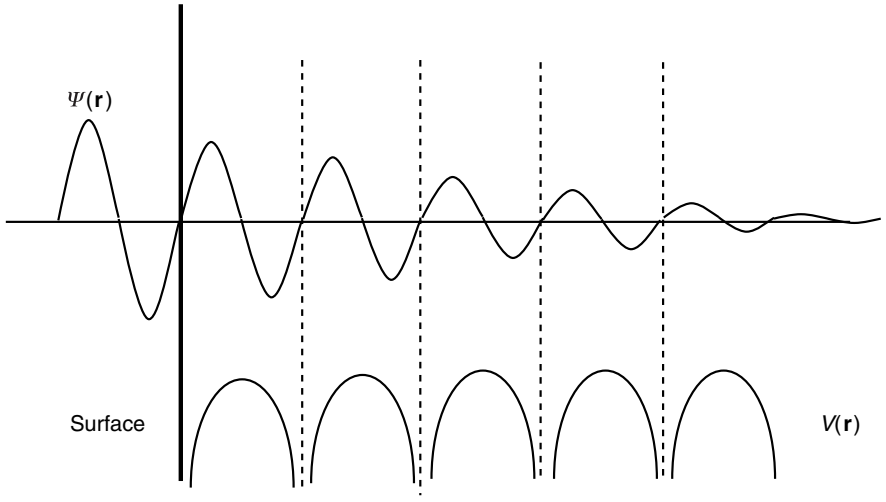


Fig. 5.5. Bloch functions equivalent in periodicity to the crystal lattice cannot propagate within the crystal. This effect is at the origin of forbidden energy bands. Thus an externally impinging electron with a matter-wave periodicity commensurate with atomic lattice spacing will not be able to penetrate the crystal.

of an ensemble of compact levels, with the bonding levels leading to the formation of a valence band and anti-bonding levels leading to a conduction band (Fig. 5.6). These two bands constitute two distinct continuums of allowed electron energies. A forbidden band of electron energies may or may not emerge depending on whether the energy separation between the two bands is large enough to keep them from overlapping. This approach, referred to as the *tight binding model*, is examined in Complement 5.B.

Without entering into unnecessary details in this book, we may nonetheless distinguish between two types of forbidden bands. One type emerges when the valence band's energy maximum and the conduction band's energy minimum are situated at the same location within the Brillouin zone (i.e. they share the same $\mathbf{k}$). We will see that optical transitions between inferior levels in the conduction band and superior levels in the valence band occur at constant $\mathbf{k}$. Such optical transitions are then possible between the two band extrema; the corresponding energy gap is referred to as being *direct*. In GaAs, the two extrema are situated at $\mathbf{k} = 0$ – this is the Γ point of the Brillouin zone (Fig. 5.7*b*).

A second type of band gap emerges when the band extrema of the valence and conduction bands are each located at different points within the Brillouin zone. Transitions between the band extrema must then take place with non-conserving wavevectors $\mathbf{k}_0$, the band gap is referred to as being *indirect* (Fig. 5.7*a*). This is the case for silicon, where the valence band extremum is situated at the Γ point and the conduction band minimum lies 85% of the way along ΓX.

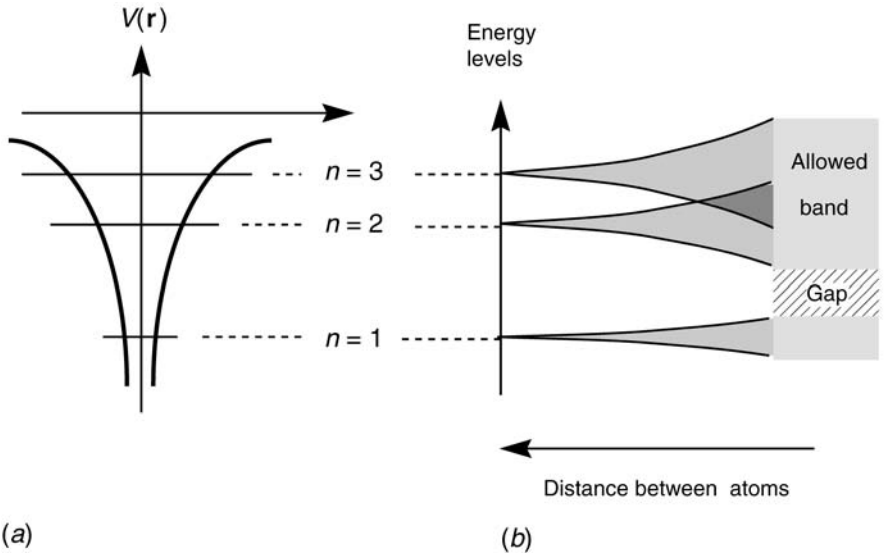


Fig. 5.6. Chemical hybridization and resulting energy bands in solids. As atoms with individual quantized energy levels E_n ($n = 1, 2, \ldots$) are brought closer together (*a*), the levels hybridize leading to the formation of energy bands (*b*). If the broadening of the bands remains inferior to the initial spacing level, a forbidden energy band or *gap* emerges within the crystal.

5.4 Effective mass and density of states

The *tight binding* or *nearly free electron* models both lead to the same result with respect to dispersion $E(\mathbf{k})$ close to the extrema of the valence and conduction bands $\mathbf{k}_{\text{ext}}$ (see Complements 5.A and 5.B). In both cases, we find a quadratic dependence of E versus $\mathbf{k}$ represented by the matrix product:

$$E(\mathbf{k}) = E_{\text{ext}} + \frac{\hbar^2}{2}(\mathbf{k} - \mathbf{k}_{\text{ext}})^t M^{-1}(\mathbf{k} - \mathbf{k}_{\text{ext}}) \tag{5.14}$$

where A^t is the transpose of vector A and M is a real symmetric matrix known as the *effective mass matrix*, for reasons which will be explained shortly. After diagonalizing the M^{-1} matrix, we see that the band structure $E(\mathbf{k})$ may be written in the form:

$$E(\mathbf{k}) = E_{\text{ext}} + \frac{\hbar^2}{2}\left[\frac{(k_1 - k_{\text{ext},1})^2}{m_1} + \frac{(k_2 - k_{\text{ext},2})^2}{m_2} + \frac{(k_3 - k_{\text{ext},3})^2}{m_3}\right] \tag{5.15}$$

where the k_is and $k_{\text{ext},i}$s are components of $\mathbf{k}$ and $\mathbf{k}_{\text{ext}}$ along the eigenvectors of M^{-1}. We recognize the expression relating the energy to the wavevector for free

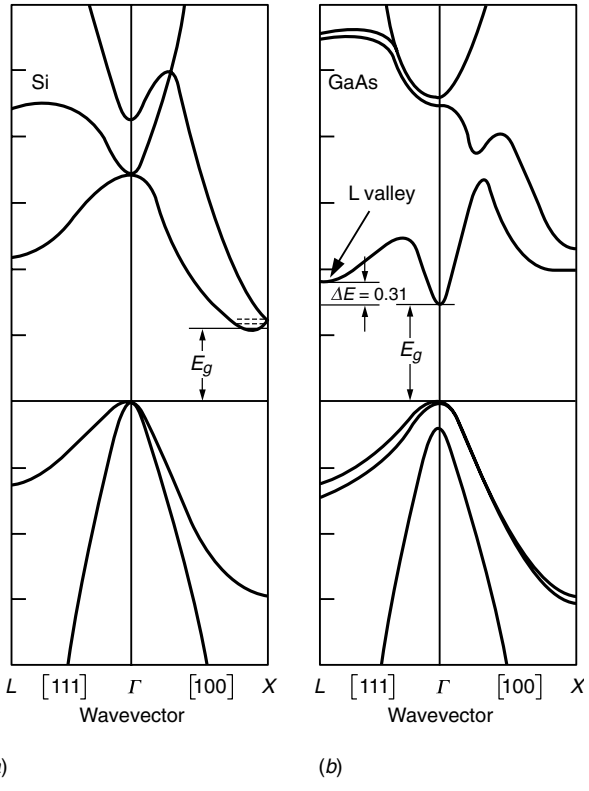


Fig. 5.7. Two examples of forbidden energy bands: (*a*) *indirect* gap for Si and (*b*) *direct* gap for GaAs.

electrons ($E = \hbar^2 k^2/2m$) but with different masses. These are the *effective masses* and result from interaction of the electrons with the periodic potential of the crystal. The effective masses are positive when the energy band curvature is directed upwards and negative when the curvature is downwards. We shall consider the significance of this negative effective mass later on. The constant energy surfaces obtained by setting $E =$ constant form ellipsoids centred about the extrema in the Brillouin zone. Figure 5.8 shows such constant energy ellipsoids in the vicinity of the conduction band extrema in silicon. As indicated in the table on p. xvii, silicon possesses two effective conduction masses – one in the transverse direction m_{et}, and another in the longitudinal direction m_{el}.

For most semiconductors, the valence bands are degenerate at $\mathbf{k} = 0$. The origin of this is the predominantly triplet nature of the sp^3 orbitals which form the valence band. The degeneracy is lifted, however, for $\mathbf{k} \neq 0$, leading to bands with different curvatures. The bands with low curvature are those which lead to heavier effective masses. These bands are thus referred to as *heavy hole* bands (we shall describe later on what we mean by a *hole*). Alternately, bands with a greater degree

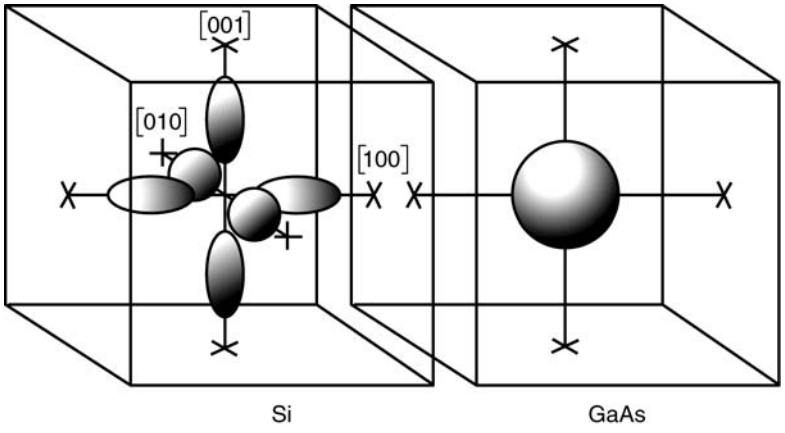


Fig. 5.8. Constant energy ellipsoids in silicon and GaAs.

of curvature are referred to as *light hole* bands. The table on p. xvii gives the different values for effective electron and hole masses in Si and GaAs.

We now wish to calculate the density of states in the various allowed bands. These quantities are of fundamental importance in all calculations involving transitions between bands. To keep from needlessly overburdening the required notation, we will temporarily focus upon the simple case involving a GaAs conduction band, where the effective mass matrix is isotropic. The band structure is then given by:

$$E(\mathbf{k}) = E_c + \frac{\hbar^2}{2m_c}(k_x^2 + k_y^2 + k_z^2) = E_c + \frac{\hbar^2 k^2}{2m_c} \tag{5.16}$$

where k is the norm of the wavevector $\mathbf{k}$ and E_c is the energy at the bottom of the conduction band. We can easily calculate the density of states in $\mathbf{k}$ space. In a volume $\mathrm{d}^3\mathbf{k} = \mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z$, Eq. (5.10) shows that there are $(Na_x)(Na_y)(Na_z)/8\pi^3 = V/8\pi^3$ states assuming cyclic Born–von Karman boundary conditions. If we had used the boundary conditions in (5.9), the density of states would have been V/π^3, but the integration volume (with n_1, n_2, and n_3 all being positive in this case) would have been eight times smaller. The density of states in k space is the infinitesimal number of states $\mathrm{d}^3N = \rho(k)\mathrm{d}^3\mathbf{k}$ situated in the volume element $\mathrm{d}^3\mathbf{k}$:

$$\rho(k) = \frac{V}{(2\pi)^3} \tag{5.17}$$

k space density of states assuming periodic boundary conditions and ignoring electron spin

The number of states $\mathrm{d}N$ situated in the volume contained between two spherical shells of radius k and $k + \mathrm{d}k$ is then given by:

$$\mathrm{d}N = \frac{V}{8\pi^3}4\pi k^2\mathrm{d}k = \frac{V}{2\pi^2}k^2\mathrm{d}k \tag{5.18}$$

This volume $k^2\mathrm{d}k$ in k space corresponds to an equivalent volume in energy space obtained by differentiating (5.16):

$$k^2\mathrm{d}k = 2^{1/2}\left(\frac{m_c}{\hbar^2}\right)^{3/2}(E - E_c)^{1/2}\mathrm{d}E \tag{5.19}$$

The energy density of states $\rho_c(E)$ in the conduction band for GaAs is the number of states situated between energy E and $E + \mathrm{d}E$. To find the number of states, we need only substitute (5.19) into (5.18), without forgetting to take into account the contribution due to electron spin degeneracy (this allows two electrons of opposite spin to occupy the same energy level). Correspondingly, the final expression must be multiplied by a factor of 2, giving:

$$\rho_c(E) = \frac{V}{2\pi^2}\left(\frac{2m_c}{\hbar^2}\right)^{3/2}(E - E_c)^{1/2} \tag{5.20}$$

Density of states for an isotropic band structure

The generalization of this result to arbitrarily shaped ellipsoids draws upon the theory of conic sections and is of little interest to us here. We find, for indirect conduction bands, as in the cases of silicon and germanium:

$$\rho_c(E) = \frac{V}{2\pi^2}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2}(E - E_c)^{1/2} \tag{5.21}$$

where m_c^* is the average effective mass of the conduction band given by:

$$m_c^* = (nm_{\mathrm{el}}^{1/2}m_{\mathrm{et}})^{2/3} \tag{5.22}$$

and where n is the number of equivalent valleys ($n = 6$ for Si and $n = 4$ for Ge). These values are also given in the table on p. xvii. Similarly, the density of states in the valence band is given by:

$$\rho_v(E) = \frac{V}{2\pi^2}\left(\frac{2m_v^*}{\hbar^2}\right)^{3/2}(E_v - E)^{1/2} \tag{5.23}$$

where E_v is the valence band maximum and the effective mass for the valence band density of states is:

$$m_v^* = (m_{\mathrm{hh}}^{3/2} + m_{\mathrm{lh}}^{3/2})^{2/3} \tag{5.24}$$

It is important to remember that the density of states in a three-dimensional crystal increases as $E^{1/2}$ and $m^{3/2}$. Figure 5.9 schematically shows the band structure and density of states obtained using the quadratic approximation in the

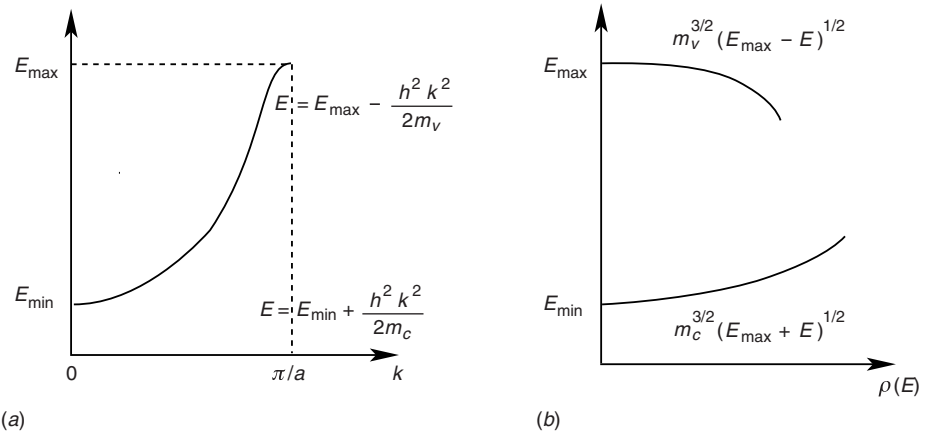


Fig. 5.9. Band structure (*a*) and density of states (*b*) near the allowed energy band extrema.

vicinity of the band extrema. Nothing general can be said using this model about the middle of the band where the quadratic approximation clearly breaks down.

5.5 Dynamic interpretation of effective mass and the concept of holes

Up to this point, we have made reference several times to the notion of a hole. It is in fact a complex notion, which draws upon (in a rigorous description) many-body theory. We content ourselves in this section with presenting a few intuitive elements concerning the theory of holes. We begin by giving a dynamic interpretation of the concept of effective masses.

The electron wavefunctions may be written as a combination of travelling waves:

$$\Psi(\mathbf{r}) = \iiint a(\mathbf{k}) e^{i\left(\mathbf{k}\cdot\mathbf{r} - \frac{E(\mathbf{k})}{\hbar}t\right)} d^3\mathbf{k} \tag{5.25a}$$

or again, in terms of the results of Section 5.2:

$$\Psi(\mathbf{r}) = \sum_{\text{bands } m} \iiint a_m(\mathbf{k}) u_{m,\mathbf{k}}(\mathbf{r}) e^{i\left(\mathbf{k}\cdot\mathbf{r} - \frac{E(\mathbf{k})}{\hbar}t\right)} d^3\mathbf{k} \tag{5.25b}$$

The wavepackets constructed from wavefunctions near a particular $\mathbf{k}$ value have a group velocity given by the dispersion relation $v_g = d\omega/dk$, so that in the case of electronic matter-waves in the band, see Fig. 5.10:

$$\mathbf{v}_g = \frac{1}{\hbar}\nabla_{\mathbf{k}} E \tag{5.26}$$

The work done by an exterior force $\mathbf{F}$ over a time interval dt is:

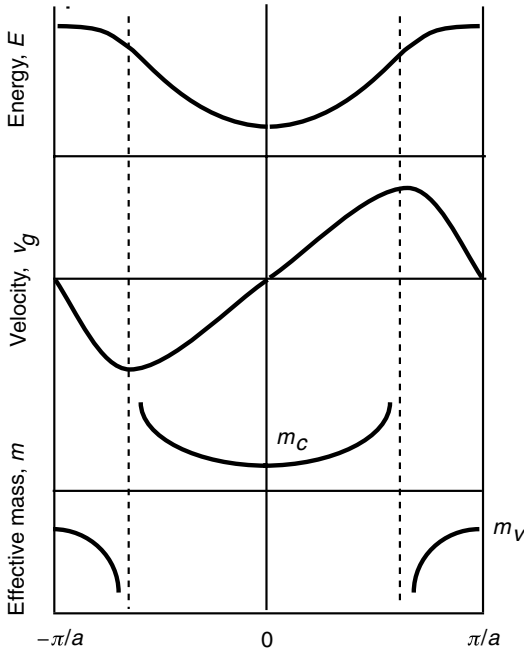


Fig. 5.10. Relationship between band structure, group velocity, and effective mass.

$$\mathrm{d}E = -\mathbf{F}\cdot\mathbf{v}_g\mathrm{d}t \tag{5.27}$$

where the change to wavepacket energy also has as a consequence a shift in the average position of the wavepacket in the band structure given by:

$$\mathrm{d}E = -\nabla_{\mathbf{k}}E\mathrm{d}\mathbf{k} = -\hbar\mathbf{v}_g\mathrm{d}\mathbf{k} \tag{5.28}$$

By inspection, we see therefore that the centre of a wavepacket subjected to a force **F** obeys:

$$\mathbf{F} = \hbar\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} \tag{5.29}$$

We note that if no dissipative mechanisms occur to counteract the effects of this force **F**, the wavepacket will begin to move or circulate through the band structure, see Fig. 5.11. The oscillations which occur as the carriers cycle through the reduced Brillouin zone are referred to as *Bloch oscillations.*

Example

The relaxation time τ between successive collisions for an electron in a solid lies on the order of 0.1 ps. Assuming an applied electric field strength $\mathscr{E}$ of approximately $10^4\,\mathrm{V\,cm^{-1}}$, the wavepacket will take on a $\mathrm{d}\mathbf{k} = q\mathscr{E}\tau/\hbar$ or $\approx 10^6\,\mathrm{cm^{-1}}$. This quantity is much smaller than the typical size of a Brillouin zone ($\pi/a \approx 10^8\,\mathrm{cm^{-1}}$). As a

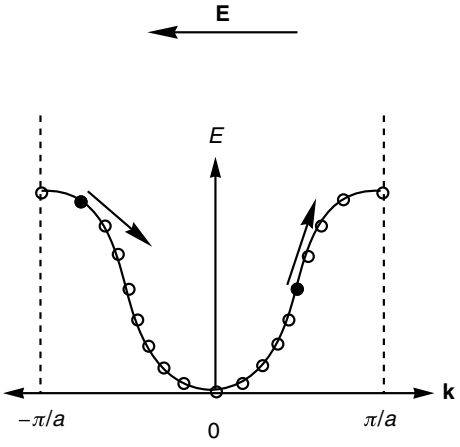


Fig. 5.11. In the absence of collisions, electrons under the influence of an electric field will cycle through the Brillouin zone in a periodic fashion. This gives rise to the phenomenon known as Bloch oscillations.

result, Bloch oscillations have not yet been observed in bulk materials.

In the complete absence of dissipative processes, an electron subjected to a homogeneous force **F** will acquire a uniform motion in **k** space and experience a corresponding acceleration in real space. This acceleration is given by:

$$\frac{\mathrm{d}\mathbf{v}_g}{\mathrm{d}t} = \frac{1}{\hbar}\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathbf{k}}E = \frac{1}{\hbar}[\nabla_{\mathbf{k}}\nabla_{\mathbf{k}}E]\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} \tag{5.30}$$

or, taking (5.29) into account:

$$\frac{\mathrm{d}\mathbf{v}_g}{\mathrm{d}t} = \frac{1}{\$^2}[\nabla_{\mathbf{k}}\nabla_{\mathbf{k}}E]\mathbf{F} \tag{5.31}$$

The matrix $[\nabla_{\mathbf{k}}\nabla_{\mathbf{k}}E]$ is none other than $\hbar^2 M^{-1}$, where M is the effective mass matrix in (5.14). Thus, the response of the wavepacket may then be described in terms of Newtonian dynamics as long as the mass of the particle is replaced by its effective mass m_{eff}.

Under the effect of an electric field $\mathscr{E}$, each element $\mathrm{d}^3\mathbf{k}$ in the band structure contributes a volume of $\mathrm{d}^3\mathbf{k}/8\pi^3$ moving at velocity $\mathbf{v}_g(\mathbf{k})$. An electrical current will then travel through the material as given by:

$$\mathbf{J} = -q\langle\mathbf{v}_g\rangle = -2q \iiint_{\text{Brillouin zone}} \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\frac{1}{\hbar}\nabla_{\mathbf{k}}E \tag{5.32a}$$

or

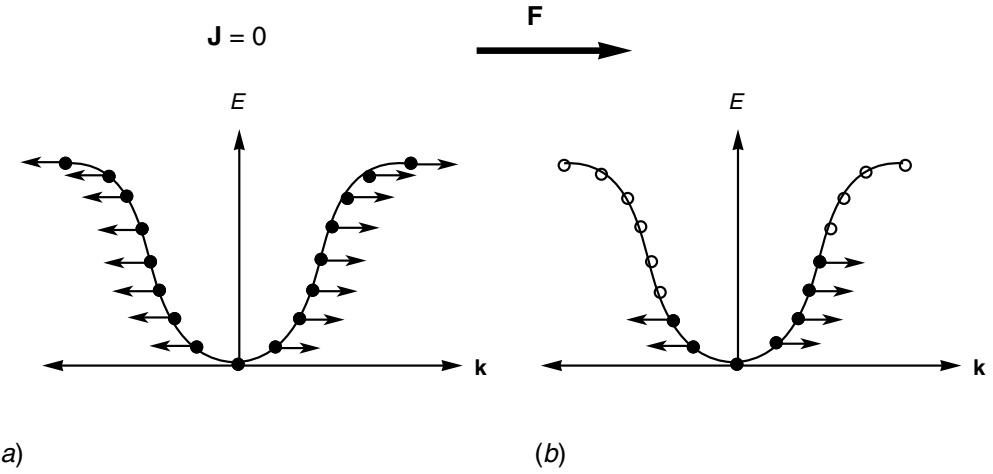


Fig. 5.12. (*a*) In a filled band, the electrical currents resulting from occupied states at **k** and $-\mathbf{k}$ cancel out. In a partially unfilled band (*b*), the electrons acquire a velocity $\mathbf{v}_g$ under the influence of a force **F** given by (5.26–5.29). Electrons are therefore transferred along the symmetry axis, and the sum total of the velocities is non-null. This incomplete band may therefore conduct electricity.

$$\mathbf{J} = -\frac{q}{4\pi^3\hbar} \iiint_{\text{Brillouin zone}} \nabla_{\mathbf{k}} E \mathrm{d}^3\mathbf{k} \tag{5.32b}$$

where we have taken into account the additional factor of 2 resulting from electron spin degeneracy. If the energy band is completely occupied (see Fig. 5.12), the only possible motions in **k** space will involve permutations. As $E(\mathbf{k})$ is symmetric relative to **k**, the volume elements in the band structure with opposing **k** will cancel out, i.e. destructively interfere (Fig. 5.12). Therefore, a completely occupied band will not conduct electricity, i.e:

$$\mathbf{J}(\text{Filled band}) = 0 \tag{5.33}$$

A material in which the last band is completely occupied by electrons, which is separated from the next adjacent empty band by an energy gap, cannot conduct electricity and is referred to as an *insulator*.

If this band, which we refer to naturally as the *conduction band*, is partially filled (to a maximum energy = $E_{\max}$), the electrons will acquire a velocity in **k** space given by (5.31), and move in the direction of the force. The current resulting from these electrons does not cancel out due to the lack of symmetry (Fig. 5.12). These electrons may be contributed by the material itself (as in the case of a metal), or may have been added to an insulator (as in the case of a semiconductor). More specifically, if for example the electric field is applied along the Ox axis, electrons are transferred from states with $k_x < 0$ to the $k_x > 0$ region, and a current will flow

in the crystal, given at $t = 0$ by (see (5.32)):

$$j_x = -\frac{q}{4\pi^3\hbar} \int\limits_{E<E_{\max}} \frac{\mathrm{d}E}{\mathrm{d}k_x}\mathrm{d}k_x \tag{5.34}$$

This current will grow monotonically in time since $E_{\max}$ will gradually slip in the electric field direction. Clearly, within a transitory period lasting a few tenths of a picosecond, collisions will stabilize the current flow. Under stationary state conditions, the situation is described by Boltzmann's equations, which we shall see later in Chapter 6.

Similarly, if we can by some means succeed in emptying electrons from a filled band (the valence band) up to an energy $E_{\min}$, the current will be given by:

$$\begin{aligned} j_x &= -\frac{q}{4\pi^3\hbar} \int\limits_{\text{Occupied states}} \frac{\mathrm{d}E}{\mathrm{d}k_x}\mathrm{d}k_x \\ &= -\frac{q}{4\pi^3\hbar} \int\limits_{\text{Brillouin zone}} \frac{\mathrm{d}E}{\mathrm{d}k_x}\mathrm{d}k_x - \left(-\frac{q}{4\pi^3\hbar} \int\limits_{\text{Unoccupied states}} \frac{\mathrm{d}E}{\mathrm{d}k_x}\mathrm{d}k_x\right) \end{aligned} \tag{5.35}$$

or, given (5.33):

$$j_x = +\frac{q}{4\pi^3\hbar} \int\limits_{E>E_{\min}} \frac{\mathrm{d}E}{\mathrm{d}k_x}\mathrm{d}k_x \tag{5.36}$$

where the integral is evaluated from $E_{\min}$ up to the top of the valance band (Fig. 5.13). We therefore see that the current carried by a nearly filled band, may be interpreted as resulting from the flow of positively charged quasi-particles such as holes. A derivation identical to the one given earlier shows the mass of these particles to be *positive* and given by:

$$M^{-1} = -\frac{1}{\hbar^2}[\nabla_{\mathbf{k}}\nabla_{\mathbf{k}}E] \tag{5.37}$$

since the curvature at the top of the valence band is negative. Figure 5.13 allows us to see how a 'hole' (i.e. an unoccupied electron state in the valence band) may be considered to carry a positive charge. The novice may find some assistance with this concept from simple hydrodynamics. In this case, electrons may be seen to 'fall' to the bottom of the conduction band much like water to the bottom of a glass, whereas holes 'rise' to the top of the valence band in a manner similar to bubbles (*holes* in a fluid) in a glass of champagne.

To complete this section, we should add that the complete wavefunction for a

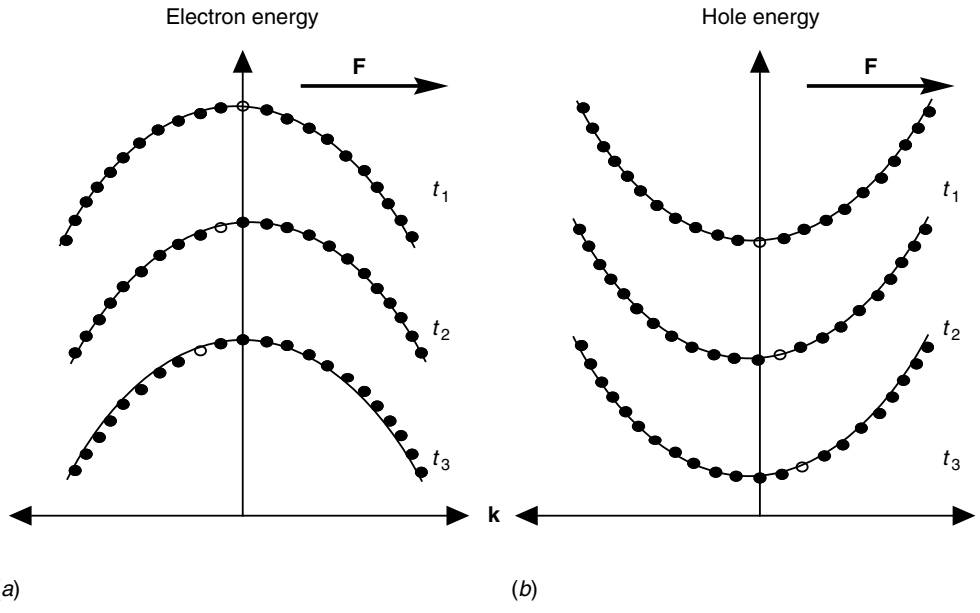


Fig. 5.13. The figure on the left represents electrons in the valence band as having negative charges and effective masses. Under the effect of a positive electric field, the electrons are transferred toward negative k states. Defining a hole at $\mathbf{k}$ as reflecting a missing electron at $-\mathbf{k}$, the hole may be interpreted as a pseudo-particle having a positive charge and mass ($q = e$).

Missing electron at $\mathbf{k}$	$\Leftrightarrow$	Hole present at $-\mathbf{k}$
$E_{\mathrm{el}}(\mathbf{k})$	$\Leftrightarrow$	$-E_h(\mathbf{k})$
$\hbar\dfrac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = -e\mathbf{F}$	$\Leftrightarrow$	$\hbar\dfrac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = e\mathbf{F}$
$\mathbf{v}_{\mathbf{k}}^{\mathrm{el}} = \dfrac{1}{\hbar}\nabla_{\mathbf{k}} E_{\mathrm{el}}(\mathbf{k})$	$\Leftrightarrow$	$\mathbf{v}_{\mathbf{k}}^{h} = \dfrac{1}{\hbar}\nabla_{\mathbf{k}} E_h(\mathbf{k}) = \mathbf{v}_{\mathbf{k}}^{\mathrm{el}}$
$\ddot{M}_{\mathrm{el}}\dfrac{\mathrm{d}\mathbf{v}_{\mathbf{k}}}{\mathrm{d}t} = -e\mathbf{F}$	$\Leftrightarrow$	$\ddot{M}_h\dfrac{\mathrm{d}\mathbf{v}_{\mathbf{k}}}{\mathrm{d}t} = e\mathbf{F}$
$\ddot{M}_{\mathrm{el}} < 0$	$\Leftrightarrow$	$\ddot{M}_h > 0$
charge $-e$	$\Leftrightarrow$	charge e
$\mathbf{j} = \sum_{\substack{\mathbf{k}\ \text{occ.}\\ \text{electrons}}} -e\mathbf{v}_{\mathbf{k}}$	$\Leftrightarrow$	$\mathbf{j} = \sum_{\substack{\mathbf{k}\ \text{occ.}\\ \text{holes}}} e\mathbf{v}_{\mathbf{k}}$

hole is in fact an entanglement of the entire valance band eigenstates (symmetrized into the form of a Slater determinant) from which a single electron has been removed. This sort of function is not particularly practical, but is called upon at times, as for instance in the theory of excitons.

5.6 Carrier statistics in semiconductors

5.6.1 Fermi statistics and the Fermi level

Electrons and holes in semiconductors obey, as must all indistinguishable spin 1/2 particles, Fermi–Dirac statistics. Therefore, the probability at temperature T, of a state with energy E being occupied is given by the *Fermi–Dirac distribution function* $f(E)$. At temperature T, the energy distribution function of such particles is given by:

$$f(E) = \frac{1}{1 + \mathrm{e}^{(E-E_F)/kT}} \tag{5.38}$$

so that the carrier density, for instance in a semiconductor conduction band, is given by:

$$n = \int_{E_c}^{\infty} \rho_c(E) f(E) \mathrm{d}E \tag{5.39a}$$

Indeed, below E_c there are no available states. Given the expression for the density of states in the conduction band (see (5.20)), this last equation may be written in the general form:

$$n = N_c F_{1/2}\left(\frac{E_F - E_c}{kT}\right)$$

$$N_c = \frac{1}{4}\left(\frac{2m_c^* \mathrm{kT}}{\pi\hbar^2}\right)^{3/2} \tag{5.39b}$$

$$F_{1/2}(u) = \frac{1}{\Gamma(3/2)} \int_0^{\infty} \frac{x^{1/2}}{1 + e^{(x-u)}} \mathrm{d}x$$

where $F_{1/2}$ is the *Fermi integral* and N_c is the effective density of states in the conduction band.

Figure 5.14 represents the Fermi–Dirac distribution at zero and non-zero temperatures. At $T = 0$, we must recall that the Fermi energy is the energy of the last occupied state. By definition, the Fermi level is the chemical potential of the particles in the structure. It describes the amount of energy which must be spent to add a particle (initially infinitely far away) to the system. *At thermodynamic equilibrium, this chemical potential is the same for all particles and is identical throughout the structure.* We will return to this very important point during our study of diffusion.

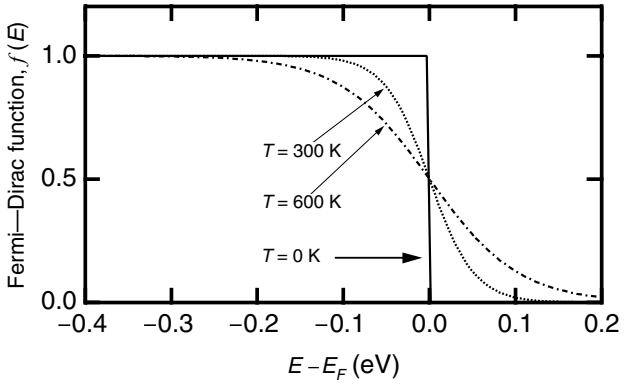


Fig. 5.14. Fermi–Dirac distribution at different temperatures.

In calculating the Fermi level in a semiconductor as a function of the number of electrons, two distinct situations arise depending on whether the Fermi level is situated within a forbidden or an allowed energy band within the sample's band structure.

Occupied Fermi level: degenerate system (Fig. 5.15)
In this case, the Fermi–Dirac function may be approximated by the step function:

$f(E) = 1$, if $E < E_F$

$f(E) = 0$, if $E > E_F$

The volume electron density is then given by the overlap between the Fermi function and the energy density of states per unit volume in the conduction band, i.e:

$$n = \int_0^{E_F} \rho_c(E)\mathrm{d}E = \frac{1}{2\pi^2}\left(\frac{2m_c}{\hbar^2}\right)^{3/2} \int_0^{E_F} (E - E_c)^{1/2}\mathrm{d}E \tag{5.40}$$

or again as:

$$n = \frac{1}{3\pi^2}\left(\frac{2m_c}{\hbar^2}\right)^{3/2} (E_F - E_c)^{3/2} \tag{5.41}$$

We note that the number of carriers is not a function of temperature. This point is characteristic of *degenerate* systems such as metals or highly doped semiconductors.

Example
We seek the number of carriers corresponding to a Fermi level of 30 meV inside the GaAs conduction band.

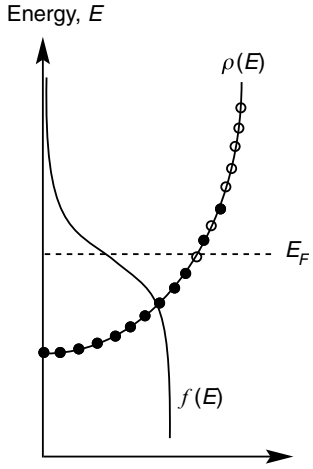


Fig. 5.15. Method for calculating the Fermi level when it is situated in an allowed energy band.

For GaAs, $m_c = 0.067m_e$:

$$n = (2 \times 0.067 \times 0.91 \times 10^{-30}\,\text{kg}/1.1 \times 10^{-68}\,\text{J}^2\,\text{s}^2)^{1.5} \times (0.03\,\text{eV} \times 1.6 \times 10^{-19}\,\text{C})^{1.5}/3\pi^2$$

or,

$$n = 4.4 \times 10^{17}\,\text{cm}^{-3}$$

Thus, for electron concentrations in excess of $10^{17}\,\text{cm}^{-3}$, GaAs behaves as a metal and its Fermi level is occupied even at zero temperature.

Unoccupied Fermi level (Fig. 5.16)

In this case, the Fermi level is situated in the forbidden gap of a semiconductor. The number of electrons is again given by the overlap of the Fermi–Dirac function for the electrons $f_c(E) = 1/(1 - \exp((E - E_F)/kT))$ and the density of states in the conduction band. This time, a different approximation for the Fermi function is required. Here we suppose that the Fermi level is sufficiently deep within the forbidden gap in comparison to kT, so that $E - E_F \gg kT$. That being the case, the Fermi–Dirac function may be approximated by the Boltzmann distribution $\text{e}^{-(E-E_F)/kT}$ (non-quantum regime):

$$n = \int_0^\infty \rho(E)\text{e}^{(E-E_F/)kT}\text{d}E$$

$$= \frac{1}{2\pi^2}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2}\text{e}^{(E_F/kT)}\int_{E_c}^\infty (E - E_c)^{1/2}\text{e}^{-(E/kT)}\text{d}E \qquad (5.42a)$$

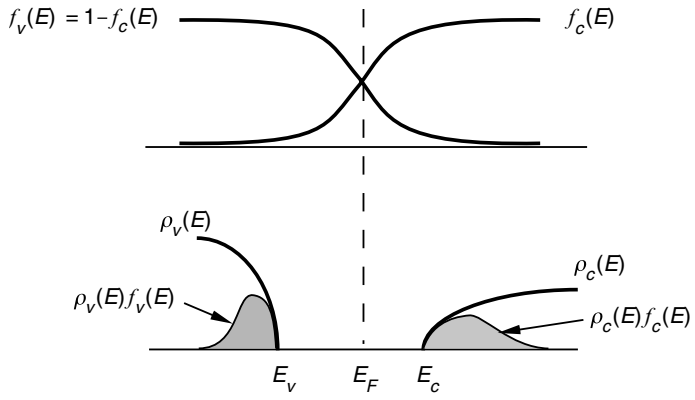


Fig. 5.16. Method for calculating the carrier density in semiconductor bands as a function of the Fermi level position.

i.e.

$$n = N_c e^{-(E_c - E_F)/kT} \tag{5.42b}$$

Electron concentration and Fermi level

where we recall that m_c^* is the effective mass for the density of states in the conduction band given by (5.22). N_c is the *effective density of states* in the conduction band:

$$N_c = \frac{1}{4}\left(\frac{2m_c^* kT}{\pi\hbar^2}\right)^{3/2} \tag{5.43}$$

With regards to holes, the occupation probability $f_v(E)$ of a hole is equal to the probability that this state is unoccupied by an electron, whereby:

$$f_v(E) = 1 - f_c(E) = 1 - \frac{1}{1 + e^{(E-E_F)/kT}} = \frac{1}{1 + e^{-(E-E_F)/kT}} \tag{5.44}$$

Fermi–Dirac statistics for holes

Therefore, in the Boltzmann regime the hole density p in the valence band is related to the Fermi level by:

$$p = \int_0^\infty \rho(E) e^{-(E - E_F/kT)} dE = N_v e^{-(E_F - E_v)/kT} \tag{5.45a}$$

Hole concentration and Fermi level

where N_v is the effective density of states in the valence band given by:

$$N_v = \frac{1}{4}\left(\frac{2m_v^* kT}{\pi\hbar^2}\right)^{3/2} \tag{5.45b}$$

The term 'effective density of states' takes its meaning when one contemplates Eqs. (5.42) and (5.45). The population statistics in semiconductors behave in the same way as a two-level system of energy E_v and E_c, with an energy separation given by the band gap $E_c - E_v = E_g$ and densities of states given by N_v and N_c, respectively (see Fig. 5.18). Introducing values for the relevant physical constants, these effective densities become:

$$N_c = 2.5 \times 10^{19} \left(\frac{m_c^*}{m_e}\right)^{3/2} \left(\frac{T}{300}\right)^{3/2} \text{cm}^{-3}$$

$$N_v = 2.5 \times 10^{19} \left(\frac{m_v^*}{m_e}\right)^{3/2} \left(\frac{T}{300}\right)^{3/2} \text{cm}^{-3} \tag{5.46}$$

Effective density of states for semiconductors

Example
For GaAs, $m_c^* = 0.067$ and $m_v^* = 0.64$, at 300 K Eq. (5.46) leads to:

$N_c = 4.3 \times 10^{17}\,\text{cm}^{-3}$
$N_v = 1.3 \times 10^{19}\,\text{cm}^{-3}$

Equations (5.42) and (5.44) are also useful for calculating the position of the Fermi level once the electron or hole densities are known:

$$E_F = E_c - kT\ln\frac{N_c}{n} = E_v + kT\ln\frac{N_v}{p} \tag{5.47}$$

Fermi level and carrier concentrations

We therefore see that the Fermi level penetrates into the bands, i.e. the semiconductor becomes degenerate once the electron (or hole) density exceeds the effective density of states (see Fig. 5.17).

Equation (5.47) is interesting in that it shows that the Fermi level is nothing other than a change of variables without immediate physical added value. It is a measure of population densities on an energy scale. However, this concept will display all its power in heterogeneous systems found in device physics. In fact, the constancy of the Fermi level throughout the different regions where thermodynamic equilibrium prevails is one of the major tools of device physics.

Finally, if we suppose that the electrons and holes are in thermodynamic equilibrium, they then share the same chemical potential (i.e. the Fermi level is the same everywhere this thermodynamic equilibrium is respected). We will appreciate the power of this concept better when exploring aspects of device physics in Chapter 10. Following (5.42) and (5.44), this has as a consequence that the product np is independent of the Fermi level and is found to be:

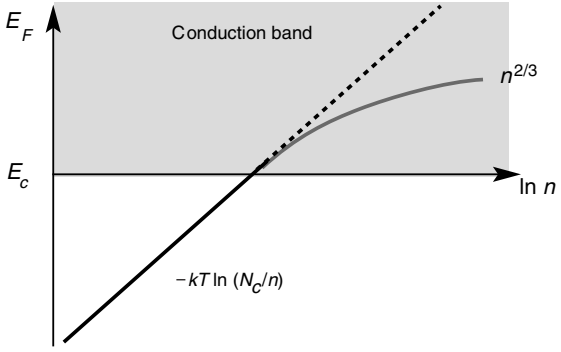


Fig. 5.17. Behaviour of the Fermi level as a function of free carrier density n. For carrier concentrations in excess of the effective density of states N_c the semiconductor becomes degenerate with its Fermi level varying as $n^{2/3}$ as in the case of a metal.

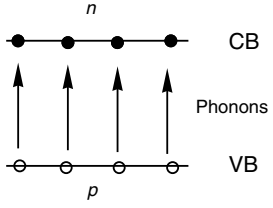


Fig. 5.18. In a lightly doped semiconductor, the bands behave thermodynamically as two discrete levels with concentrations N_c and N_v (CB, conduction band; VB, valence band).

$$np = N_c N_v e^{-E_g/kT} \tag{5.48}$$

Law of mass action in semiconductors

where $E_g = E_c - E_v$ is the width of the bandgap. Expression (5.48) gives the relationship between the semiconductor electron and hole concentrations at equilibrium, and is known as the *law of mass action*.

5.6.2 Intrinsic semiconductors

A semiconductor is said to be *intrinsic* when the origin of the electrons and holes in it is endogenous. In this case, the holes in the valence band result from electrons which have been thermally excited to the conduction band (Fig. 5.18). The electron and hole densities (n and p, respectively) are then equal to the *intrinsic carrier density*:

$$n = p = n_i = \sqrt{N_c N_v} e^{-(E_g/2kT)} \tag{5.49}$$

Intrinsic density of carriers

Example
In GaAs, where the bandgap $E_g = 1.42\,\text{eV}$, the carrier density at $T = 300\,\text{K}$ is $n_i = 3 \times 10^6\,\text{cm}^{-3}$. Clearly, this makes GaAs a poor insulator at room temperature. Residual impurities present within GaAs, however, contribute carrier densities significantly above this value.

For HgCdTe, a small bandgap material (used in infrared detection – see Chapter 11), $E_g = 0.116\,\text{eV}$ and the intrinsic carrier concentration is $n_i = 1.0 \times 10^{16}\,\text{cm}^{-3}$. Such a material will therefore conduct electricity at room temperature without additional doping.

5.6.3 Doped semiconductors

An intrinsic semiconductor on its own is relatively useless. It is both a poor conductor and a poor insulator. Alternately, it becomes quite a fantastic (and useful!) material when *doped*. In fact, when certain chemical impurities are introduced at low levels into semiconductors, they can substitutionally occupy lattice sites within the host crystal. In this case, the chemical impurity wavefunctions hybridize with the atoms in the host lattice and donate an extra electron to the conduction band. This occurs, for example, with pentavalent phosphorous in silicon (Fig. 5.19), or when silicon is introduced as an impurity into GaAs. These impurities are referred to as donors; the doped semiconductor which results from their incorporation is called an *n-type* semiconductor. Other impurity types may capture an electron from the crystalline host (the captured electron becoming bound to the impurity). This occurs, for instance, with boron in silicon (Fig. 5.20). These impurities (or *acceptors*) therefore liberate holes into the semiconductor valence band, turning the crystal into *p-type* material.

It can be shown that such impurity types (donors or acceptors), behave in semiconductors as hydrogenic impurity centres to which the electrons or holes are bound (via the Coulombic interaction) with a characteristic ionization energy of:

$$E_{\text{ion}} = \frac{m^* q^4}{8\varepsilon_R^2 \varepsilon_0^2 h^2} \simeq \frac{m^*}{m_e \varepsilon_R^2} \times 13.6\,\text{eV} \tag{5.50}$$

where m^* is the effective mass and ε_R is the relative permittivity of the semiconductor. This gives rise to ionization energies ranging between 5 and 25 meV. Consequently, these impurities are, for the most part, ionized at room temperature. This, however, is not the case for large gap semiconductors such as GaN, where the ionization energies for such hydrogenic impurities may reach 200 meV.

We suppose in what follows that a density N_D of donors has been introduced into a semiconductor, and that over the expected range of temperatures, all the donors are ionized. If $N_D \gg n_i$, we may easily show that all the carriers in the

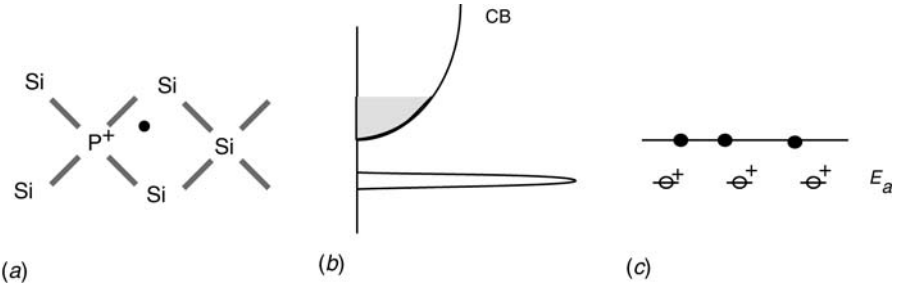


Fig. 5.19. A pentavalent impurity like phosphorous introduced substitutionally into a silicon lattice hybridizes with the tetragonally neighbouring atoms and releases an extra electron into the lattice as a free electron (*a*). This impurity introduces a donor state a few tens of meV below the conduction band (*b*). At room temperature, all these impurity centres are ionized and donate their free electrons to the conduction band of the host material (*c*).

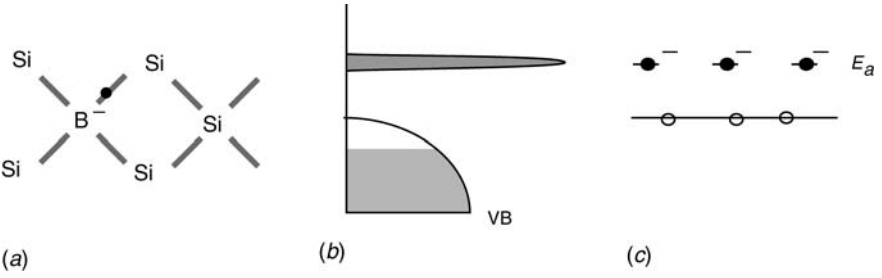


Fig. 5.20. A trivalent impurity like boron introduced substitutionally into a silicon lattice hybridizes with the tetragonally neighbouring atoms and borrows an electron from a nearby covalent bond. This missing electron becomes a hole in the valence band (*a*). This impurity introduces an acceptor state a few tens of meV above the valence band (*b*). At room temperature, all these impurity centre states are occupied by electrons originating from the valence band and a corresponding number of holes are left to populate the valence band of the host material (*c*).

conduction band have been thermally excited from the donor levels, so that:

$$n = N_D \tag{5.51}$$

Consequently, the Fermi level is found to be:

$$E_F = E_c - kT \ln \frac{N_c}{N_D} \tag{5.52}$$

Similarly, if a density of acceptors N_A is introduced in a semiconductor with $N_A \gg n_i$, then all the holes in the valence band result from capture of electrons from the top of the valence band:

$$p = N_A \tag{5.53}$$

and the Fermi level is given by:

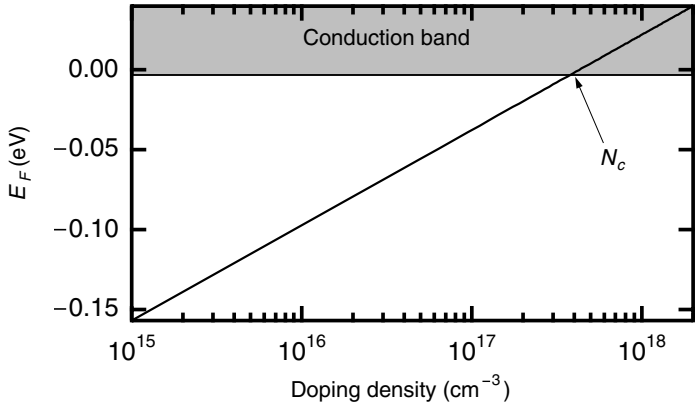


Fig. 5.21. Position of the Fermi level in the GaAs bandgap as a function of doping concentration. The semiconductor becomes degenerate for doping levels in excess of the effective density of states N_c.

$$E_F = E_v + kT \ln \frac{N_v}{N_A} \tag{5.54}$$

Equations (5.52) and (5.54) show that, as long as the doping level is less than the effective density of states in the bands, the semiconductor will remain non-degenerate. For illustrative purposes, Fig. 5.21 shows the position of the Fermi level in the GaAs bandgap as a function of doping.

5.6.4 Quasi-Fermi level in a non-equilibrium system

Up until now, we have considered different cases leading to electron and hole populations in thermodynamic equilibrium. There are, however, numerous instances where the densities of the populations are dominated by non-equilibrium processes. In fact, the operation of all semiconductor components takes place under non-equilibrium conditions either by applying an electric field or by illuminating the structure. In these cases, the electron and hole populations are no longer in thermodynamic equilibrium with each other, nor with the host crystal. We may therefore consider an external process which generates electrons and holes in a semiconductor with a generation rate G_n and G_p (in cm^{-3} s^{-1}) for electrons and holes, respectively. We additionally consider the existence of a recombination process, characterized by time constants τ_n and τ_p and leading to the stabilization of the electron and hole populations to levels where the generation and recombination rates are equal, i.e. $G_n = n/\tau_n$ and $G_p = p/\tau_p$. The stationary concentration of electrons and holes is then given by:

$$n = G_n \tau_n$$

$$p = G_p \tau_p \tag{5.55}$$

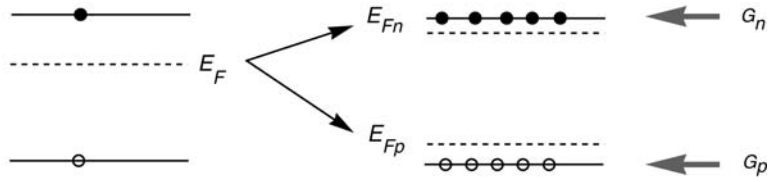


Fig. 5.22. (*a*) At thermal equilibrium, the electrons and holes share the same Fermi level. (*b*) Away from thermodynamic equilibrium (e.g. if the bands are populated by photogenerated or electrically injected carriers), the electron and hole populations are each described by their own Fermi (or quasi Fermi) levels which no longer coincide with (*a*).

In this last formula, we assumed that the created carrier densities are in excess of the thermal generation levels, i.e. $n \gg n_0$ and $p \gg p_0$. A particularly powerful insight given by Shockley was to consider that these populations may continue to be described in terms of a quasi-Fermi level (playfully named *Imref*, which spells Fermi backwards!). If the system is non-degenerate, these quasi-Fermi levels are defined starting from (5.52) and (5.54):

$$E_{Fn} = E_c - kT\ln\left(\frac{N_c}{G_n\tau_n}\right)$$

$$E_{Fp} = E_v + kT\ln\left(\frac{N_v}{G_p\tau_p}\right) \tag{5.56a}$$

Non-degenerate quasi-Fermi level

For the case where the system becomes degenerate, the quasi-Fermi levels are then obtain via (5.41) or:

$$E_{Fn} = E_c + \frac{\hbar^2}{2m_c}(3\pi^2 n)^{2/3}$$

$$E_{Fp} = E_v - \frac{\hbar^2}{2m_v}(3\pi^2 p)^{2/3} \tag{5.56b}$$

Degenerate quasi-Fermi level

Clearly, as $G_n\tau_n \gg n_0$ and $G_p\tau_p \gg p_0$, we see that the quasi-Fermi levels do not coincide (see Fig. 5.22) and, in fact, are separated from the Fermi level by the quantity:

$$(E_{Fn} - E_F) - (E_{Fp} - E_F) = kT\ln\left[\left(\frac{G_n\tau_n}{n_0}\right)\Big/\left(\frac{G_p\tau_p}{p_0}\right)\right] \tag{5.57}$$

This last equation allows one to calculate the displacement of the quasi-Fermi

levels as a function of pump rate and different recombination times. This equation is central to the physics of semiconductor lasers, a topic we shall cover later in Chapter 13.

FURTHER READING

Solid state physics

N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart and Winston, New York (1976).

C. Kittel, *Introduction to Solid State Physics*, 5th Edn, Wiley, New York (1976).

Semiconductors

A. Kireev, *Physique des semiconducteurs*, MIR, Moscou (1975).

F. Lévy, *Physique et technologie des semiconducteurs*, Presses Polytechniques et Universitaires Romandes, Lausanne (1993).

B. Sapoval and C. Hermann, *Physics of Semiconductors*, Springer, Berlin/Heidelberg (1995).

K. Seeger, *Semiconductor Physics*, 3rd Edn, Springer, Berlin (1985).

S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edn, Wiley Interscience, New York (1981).

Complement to Chapter 5

5.A The nearly free electron model

In this complement, we present a first approach to band structure theory founded upon a perturbation of the free electron model, and aptly named the *nearly free electron model*. We take Schrödinger's equation for an electron in a periodic one-dimensional crystal having a lattice spacing of a (keeping in mind that generalization of this discussion to three dimensions merely involves the courageous addition of appropriate indices). The periodicity of the crystal potential then makes this quantity suitable for Fourier analysis:

$$V(x) = \sum_{\substack{\text{Reciprocal} \\ \text{lattice}}} V_G e^{iGx} \tag{5.A.1}$$

where we recall that the vectors $\mathbf{G}$ belong to the reciprocal lattice, i.e. they are of the form $G = \text{integer} \times 2\pi/a$. Schrödinger's equation written in terms of this expression for the potential becomes:

$$H\psi(x) = \left(\frac{p^2}{2m} + \sum_G V_G e^{iGx}\right)\psi(x) = E\psi(x) \tag{5.A.2}$$

The wavefunction $\psi(x)$ may also be written as a Fourier series by drawing upon the Born–von Karman cyclic boundary conditions – $K = n2\pi/(Na)$, where Na refers to the total crystal length and n is a positive (or null) integer – so that:

$$\psi(x) = \sum_K C(K) e^{-iKx} \tag{5.A.3}$$

Substituting (5.A.3) into (5.A.2), we obtain:

$$\sum_K \left[\left(\frac{\hbar^2 K^2}{2m} - E\right) C(K) e^{-iKx} + \left(\sum_G V_G C(K) e^{-i(K-G)x}\right)\right] = 0 \tag{5.A.4}$$

Let us multiply this equation by a wave e^{-ikx} and integrate the whole over the entire crystal volume. The only non-zero terms will be those for which $G - K = k$.

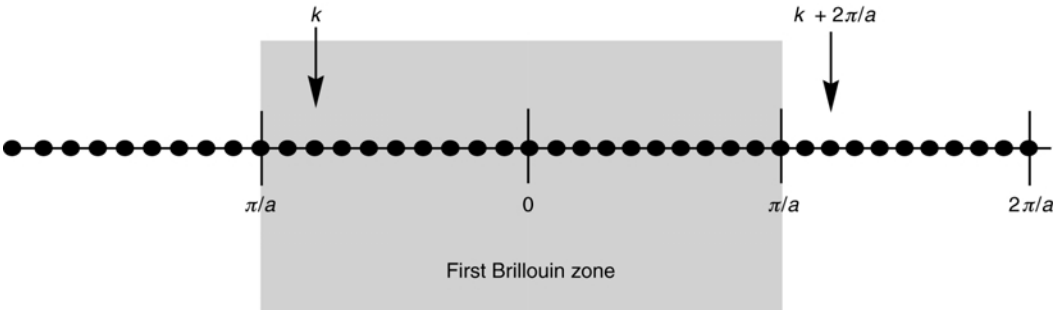


Fig. 5.A.1. The points indicated on this figure represent the allowed wavevectors k, by the Born–von Karman cyclic boundary conditions for a crystal of length $L = 20a$, where a is the lattice constant. If the k term ($= -8$ in this figure) appears in the sum (5.A.4), all the terms $k + 2n\pi/a$ will also be included.

This is to say *that if a term* k *appears in the sum, so will all other vectors which result from the addition of one or more reciprocal lattice vectors* G *to it* (see Fig. 5.A.1). We may therefore index the wavefunctions by a wavevector k chosen *arbitrarily* from the first Brillouin zone, allowing the eigenfunctions to be written as:

$$\psi_k(x) = \sum_G C(k - G)\mathrm{e}^{-\mathrm{i}(k-G)x} \tag{5.A.5}$$

This is nothing else but the *Bloch–Floquet* theorem. Schrödinger's equation (5.A.2) applied to the Bloch functions ψ_k then becomes:

$$\left(E_k - \frac{\hbar^2 k^2}{2m}\right) C(k) + \sum_G V_G C(k - G) = 0 \tag{5.A.6}$$

This is a secular equation in $C(k - G)$, which possesses non-trivial solutions (i.e. non-zero solutions) if and only if the determinant of the linear system is zero. This determinant has a rank equal to the number of points in reciprocal space. It would then appear that we have gained little relative to our original expression for Schrödinger's equation were it only that we had replaced a differential expression with an algebraic one. We now make the hypothesis that the function is *highly periodic*, which is to say that it possesses so few Fourier components that we can consider a single component – as would be the case for a purely sinusoidal potential. A portion of the secular equation in (5.A.6) may then be written as:

$$\begin{vmatrix} E(k) - \dfrac{\hbar^2(k-G)^2}{2m} & V & 0 & 0 \\ V & E(k) - \dfrac{\hbar^2 k^2}{2m} & V & 0 \\ 0 & V & E(k) - \dfrac{\hbar^2(k+G)^2}{2m} & V \\ 0 & 0 & V & \text{etc.} \end{vmatrix} = 0 \tag{5.A.7}$$

For each wavenumber k, the secular equation is a polynomial in $E(k)$ of equal degree to the rank of the matrix (there are thus as many possible solutions as rows in the matrix). *These solutions are continuous in* k *and therefore form energy bands.* A particularly interesting case arises when k is situated near the Brillouin zone boundary, i.e. $k \approx G/2 = \pi/a$. In this case, $G - k$ and k have the same value, and the diagonal elements of the 2×2 submatrix:

$$\begin{bmatrix} E(k) - \dfrac{\hbar^2 G^2}{8m} & V \\ V & E(k) - \dfrac{\hbar^2 G^2}{8m} \end{bmatrix}$$

are identical. Interaction with the nearest neighbour V lifts the degeneracy between the two unperturbed levels $E(k) - \hbar^2 G^2/8m$. This means that the energy bands which otherwise would have the same energy at $k = G/2$ have their degeneracy lifted – we will now look a little deeper into this effect. In proximity to $k = G/2$, the determinant of (5.A.7) is dominated by the 2×2 submatrix near $k = G/2$ and the equation for lifting of the degeneracy may be written as:

$$\begin{vmatrix} E(k) - \dfrac{\hbar^2 k^2}{2m} & V \\ V & E(k) - \dfrac{\hbar^2 (k-G)^2}{2m} \end{vmatrix} = 0 \tag{5.A.8}$$

This equation admits two ensembles of solutions from the secular equation, which after a little effort, may be written in the form:

$$E(q) = \frac{\hbar^2}{2m}\left(q^2 + \frac{G^2}{4}\right) \pm \sqrt{\frac{\hbar^2 G^2}{2m}\frac{\hbar^2 q^2}{2m} + V^2} \tag{5.A.9}$$

where $q = k - G/2$. We introduce ε_G, which is the energy that the free electron would have at $k = G/2$, i.e. $\varepsilon_G = \hbar^2 (G/2)^2/2m$. For small q values, Eq. (5.A.9) becomes:

$$E(q) = \varepsilon_G \pm V + \frac{\hbar^2 q^2}{2m}\left(1 \pm \frac{2\varepsilon_G}{V}\right) \tag{5.A.10}$$

Figure 5.A.2 depicts solutions to Schrödinger's equation in terms of (5.A.7). We see that we may arbitrarily choose to describe the band structure over all the Brillouin zones, or proceed by folding the entire band structure into the first Brillouin zone, with appropriate labelling of the bands. This latter convention is referred to as the *reduced zone scheme* and is most often used. Several essential points should be noted in this figure. First, far away from the boundaries of the Brillouin zone, the band structure takes the form of that for the free electron (this can be shown). Second, a forbidden energy band of approximately $2V$ in width results from the lifting of the degeneracy between the e^{ikr} and the $e^{i(k-G)r}$ waves. We also note that the dispersion relation $E(q)$ is parabolic (in q^2), which allows one to find for the

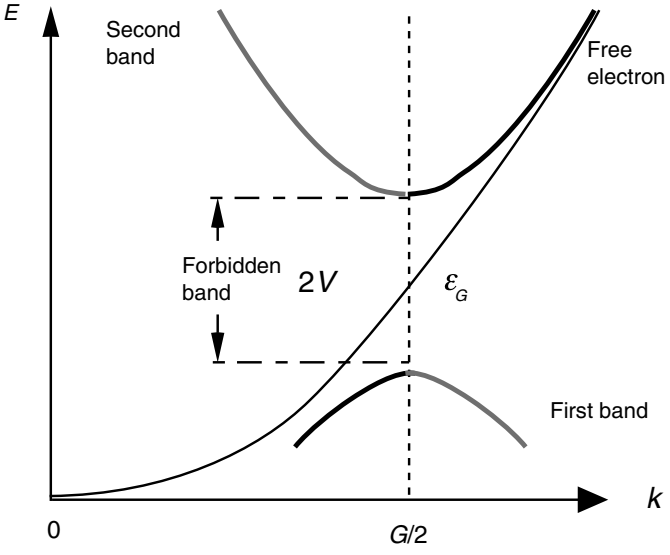


Fig. 5.A.2. Essential features of the nearly free electron model. The periodic perturbation due to the crystal potential lifts the degeneracy near the edges of the Brillouin zone ($G/2 = \pi/a$) giving rise to a forbidden energy band. The grey curves were obtained by folding the energy curves $E(k)$ into the first Brillouin zone – this is referred to as the *reduced zone scheme*.

effective mass in this model:

$$m_{\text{eff}} = \hbar^2 \left(\frac{\mathrm{d}^2 E}{\mathrm{d}q^2} \right)^{-1} = m_e \frac{1}{1 \pm 2(\varepsilon_G / V)} \approx \pm m_e \frac{V}{2\varepsilon_G} \tag{5.A.11}$$

We therefore see that the smaller the width of the forbidden energy band $2V$, the smaller the effective masses. This relationship is also found in other models. Narrow gap semiconductors (very useful in infrared light detection) characteristically possess small effective electron and hole masses.

FURTHER READING

N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart and Winston, New York (1976).

C. Kittel, *Introduction to Solid State Physics*, 5th Edn, Wiley, New York (1976).

5.B Linear combination of atomic orbitals: the tight binding model

In Complement 1.B, we saw that the eigenstates of two identical atoms possessing identical energy levels, hybridize when they are brought together, giving rise to

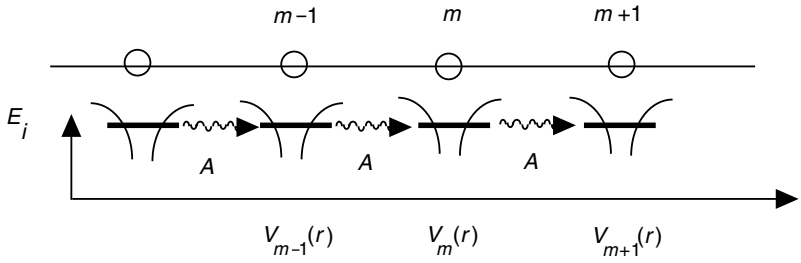


Fig. 5.B.1. One-dimensional model of a crystal potential. Overlap of the orbitals between neighbouring atoms in the crystal allows electrons to travel (tunnel) from one atom to the next.

bonding and anti-bonding levels with distinct eigenenergies. These levels are separated in energy by an amount relating to the degree of overlap (inversely correlated to the physical separation here) between the atoms' eigenstates. We will now demonstrate how to extend this procedure to the description of an entire crystal, and show how one arrives using this model at the notion of allowed and forbidden energy bands. This approach is founded upon a chemical interpretation of the significance of semiconductor band structure and is referred to as the *tight binding model* illustrated in Fig. 5.B.1. This model seeks to give a description of crystal states in terms of linear combinations of constituent atomic orbitals. We consider a one-dimensional, periodic sequence of identical atoms. To begin, we will assume that the atoms are sufficiently separated as to allow one to neglect the influence due to the potential of any given atom on its neighbour.

The Hamiltonian for an electron subject to a single potential due to atom i, situated in the lattice at ia is:

$$H_i = \frac{p^2}{2m} + V(x - ia) \tag{5.B.1}$$

All the atoms therefore have the same energy levels E_n and the same displaced eigenfunctions:

$$|n, i\rangle = \psi_{n,i}(x) = \psi_n(x - ia) \tag{5.B.2}$$

We now bring the atoms close together and seek to find the stationary states and eigenenergies E for the Hamiltonian describing an electron in the ensuing crystal:

$$H = \frac{p^2}{2m} + \sum_i V(x - ia) \tag{5.B.3}$$

As the basis of wavefunctions $\psi_{n,i}$ is complete, we may use it to express the Hamiltonian (5.B.3) as a matrix. Let us have a look at the different terms:

- The diagonal terms $\langle i, n|H|i, n\rangle$ are given by:

$$\langle n,i|H|n,i\rangle = E_n + \sum_{j\neq i}\langle n,i|V(x-x_j)|n,i\rangle \tag{5.B.4}$$

The term in the sum is independent of i and leads to an offset in energy. This term is thus of little physical consequence and will be neglected.

- For the off-diagonal terms, we will only interest ourselves in those corresponding to the nearest neighbour:

$$\begin{aligned}\langle n,i|H|n,i\pm 1\rangle &= E_n\langle n,i|n,i\pm 1\rangle + \sum_{j\neq i}\langle n,i|V(x-x_j)|n,i\pm 1\rangle \\ &\cong \langle n,i|V(x-x_i)|n,i\pm 1\rangle = -A_n\end{aligned} \tag{5.B.5}$$

In (5.B.5), we neglected the overlap between orbitals of type $\langle n,i|n,i\pm 1\rangle$ and kept in the sum only those terms A_n corresponding to tunnelling between nearest neighbours. Making this assumption, the Hamiltonian describing the crystal electrons can be broken down into independent Hamiltonians for each unperturbed energy level E_n:

$$H_n = \begin{vmatrix} E_n & -A_n & 0 & 0 \\ -A_n & E_n & -A_n & 0 \\ 0 & -A_n & E_n & -A_n \\ 0 & 0 & -A_n & \cdots \end{vmatrix} \tag{5.B.6}$$

The stationary wavefunctions are then given by:

$$|\psi_n\rangle = \sum_i C_{n,i}|n,i\rangle \tag{5.B.7}$$

The coefficients $C_{n,i}$ are given by Schrödinger's equation, i.e. solutions to the coupled equations:

$$\begin{cases} -A_nC_{n,i-2} + E_nC_{n,i-1} - A_nC_{n,i} = EC_{n,i-1} \\ -A_nC_{n,i-1} + E_nC_{n,i} - A_nC_{n,i+1} = EC_{n,i} \\ -A_nC_{n,i} + E_nC_{n,i+1} - A_nC_{n,i+2} = EC_{n,i+1} \end{cases} \tag{5.B.8}$$

where E is the energy of the new stationary states. We recognize in Eq. (5.B.8), recurrence relations for the Fibonacci series which admit as solutions:

$$C_{n,i} = \mathrm{e}^{-\mathrm{i}kx_i} \tag{5.B.9}$$

where $x_i = ia$. We may therefore index the new eigenstates of the crystal Hamiltonian as $\psi_{k,n}$. These functions can be readily shown to be equivalent to the Bloch–Floquet functions obtained in (5.B.5) (see Fig. 5.B.2). Substituting (5.B.9) into (5.B.8), we find as energies for the Bloch wavefunctions $\psi_{k,n}$:

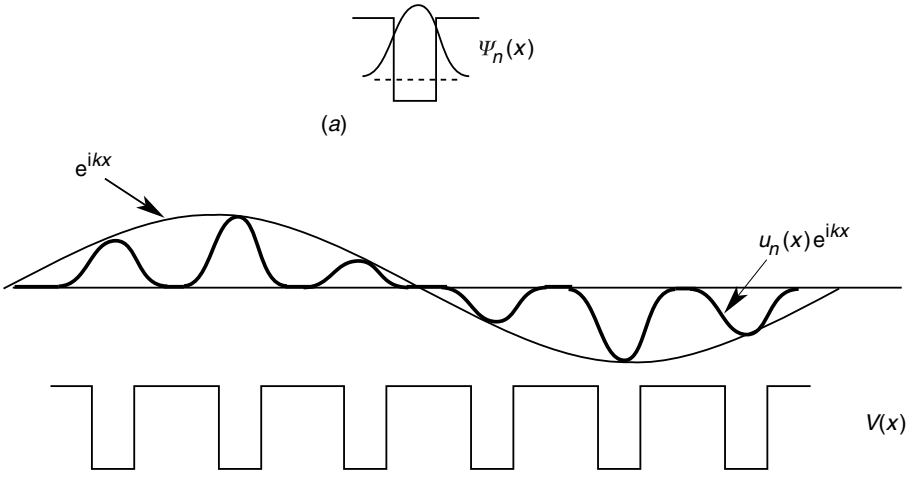


Fig. 5.B.2. Bloch–Floquet functions describing the stationary states of electrons in the periodic potential consist of linear combinations of isolated atomic orbitals (*a*) weighted by coefficients representing a travelling wave in the crystal (*b*).

$$E_n(k) = E_n - 2A_n\cos(ka) \tag{5.B.10}$$

The stationary state energies are thus located on energy bands which depend sinusoidally on k. These bands are depicted schematically in Fig. 5.B.3. This model presents a certain number of advantages over others, namely a greater degree of physical predictability.

1. The bands have a chemical origin and result from the hybridization of levels broadened into bands by tunnelling effects between adjacent atoms. This picture is more satisfying (and closer to reality!) than that employed in the nearly free electron model. We understand more clearly why the conduction band in Si is characterized by the hybridization of the singlet sp^3 states, and why the valence band shares the properties of hybridized triplet states. We also understand better the origin of the degeneracy in the valence band.
2. The forbidden bands have as their origin the empty region left between the unperturbed states E_n after lifting of the degeneracy. If the band broadening is greater than the separation between levels E_n, the bands touch and no energy gap emerges.
3. The deeper the states in energy (small n), the greater the potential (or tunnel) barriers which separate like states in adjacent atoms, and the smaller the values (A_n) which result from the tunnelling integrals. As a result, the deeper lying energy bands (having width $2A_n$) are significantly narrower than the higher lying bands.
4. Near the band extrema, the dispersion relations $E(k)$ are parabolic and lead to

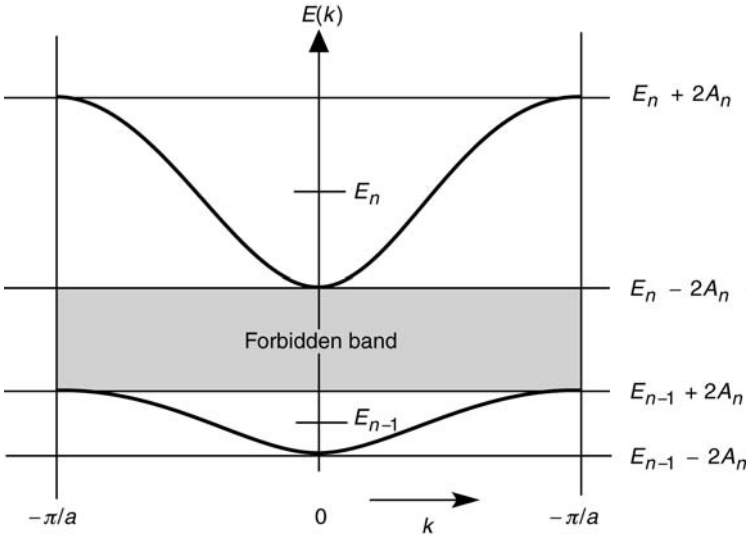


Fig. 5.B.3. Allowed sinusoidal energy bands obtained using the tight binding model. A forbidden energy band results as the allowed bands do not overlap.

an effective mass:

$$m_{\text{eff}} = \frac{\hbar^2}{2Aa^2} \tag{5.B.11}$$

We again find that the effective mass m_{eff} is inversely proportional to the width of the allowed energy band.

FURTHER READING

N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart, and Winston, New York (1976).

5.C Kane's k · p method

The nearly free electron and tight binding models are useful didactically, but both suffer a severe handicap. It is extremely difficult, in practice, to start from experimentally determined parameters (e.g. crystal structure, the electronic structure of constituent atoms, . . .) and obtain predictions such as the size of the bandgap in GaAs, etc. This difficulty emerges as a result of the highly simplistic nature of the approximations used in obtaining the results in Complements 5.A and 5.B. Introduction of the slightest sophistication into these models rapidly leads to intrac-

table calculations. The $\mathbf{k}\cdot\mathbf{p}$ method is a semi-empirical approach attributed to Kane which introduces into the band structure calculations, experimentally accessible data (derived from optical measurements, etc.).

We recall the Schrödinger equation which must be solved:

$$H\psi_{n,\mathbf{k}}(\mathbf{r}) = \left[\frac{\hat{\mathbf{p}}^2}{2m} + V(\mathbf{r})\right]\psi_{n,\mathbf{k}}(\mathbf{r}) = E_{n,\mathbf{k}}\psi_{n,\mathbf{k}}(\mathbf{r}) \tag{5.C.1}$$

where $\psi_{n,\mathbf{k}}$ are the Bloch functions:

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = u_{n,\mathbf{k}}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{r}} \tag{5.C.2}$$

We see that the operator $\hat{\mathbf{p}} = -\mathrm{i}\hbar\nabla$ acts on the Bloch–Floquet wavefunctions in the following manner:

$$\hat{\mathbf{p}}^2[u_{n,\mathbf{k}}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{r}}] = [(\hat{\mathbf{p}} + \hbar\mathbf{k})^2 u_{n,\mathbf{k}}(\mathbf{r})]\mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{r}} \tag{5.C.3}$$

As a result, the periodic portion of the Bloch function $u_{n,\mathbf{k}}$ is a solution to the Schrödinger-type equation:

$$\left[\frac{\hat{\mathbf{p}}^2}{2m} + \frac{\hbar^2\mathbf{k}^2}{2m} + \frac{\hbar}{m}\mathbf{k}\cdot\mathbf{p} + V(\mathbf{r})\right]u_{n,\mathbf{k}}(\mathbf{r}) = E_{n,\mathbf{k}}u_{n,\mathbf{k}}(\mathbf{r}) \tag{5.C.4}$$

For $\mathbf{k} = \mathbf{0}$, we remark that this equation is totally identical to Schrödinger's equation applied to the periodic portion of the Bloch–Floquet functions $u_{n,\mathbf{0}}$:

$$\left[\frac{\hat{\mathbf{p}}^2}{2m} + V(\mathbf{r})\right]u_{n,\mathbf{0}}(\mathbf{r}) = E_{n,\mathbf{0}}u_{n,\mathbf{0}}(\mathbf{r}) \tag{5.C.5}$$

$\mathbf{k}\cdot\mathbf{p}$ theory then uses $u_{n,\mathbf{0}}$ as a complete basis for $u_{n,\mathbf{k}}$:

$$u_{n,\mathbf{k}}(\mathbf{r}) = \sum_m c_m^{(n)}(\mathbf{k})u_{m,\mathbf{0}}(\mathbf{r}) \tag{5.C.6}$$

and Schrödinger's equation in (5.C.4) transforms into:

$$\begin{aligned}\sum_m \left[\frac{p^2}{2m_0} + \frac{\hbar^2k^2}{2m_0} + \frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p} + V(\mathbf{r})\right]c_m^{(n)}(\mathbf{k})u_{m,\mathbf{0}}(\mathbf{r}) \\ = \sum_m E^{(n)}(\mathbf{k})c_m^{(n)}(\mathbf{k})u_{m,\mathbf{0}}(\mathbf{r})\end{aligned} \tag{5.C.7}$$

Thus, taking the projection onto $u_{M,\mathbf{0}}$:

$$\begin{aligned}\left(E_{M,\mathbf{0}} + \frac{\hbar^2k^2}{2m_0}\right)c_M^{(n)}(\mathbf{k}) + \sum_m \frac{\hbar}{m_0}\mathbf{k}\cdot\langle u_{M,\mathbf{0}}|\mathbf{p}|u_{m,\mathbf{0}}\rangle c_m^{(n)}(\mathbf{k}) \\ = E^{(n)}(\mathbf{k})c_M^{(n)}(\mathbf{k})\end{aligned} \tag{5.C.8}$$

This equation represents the usual problem of diagonalizing an effective Hamiltonian matrix. If all the matrix elements $\langle u_{M,\mathbf{0}}|\mathbf{p}|u_{m,\mathbf{0}}\rangle$ are known, we may proceed directly to diagonalizing the matrix. This, however, is not usually the case. Additionally, only the band regions near the gap are of real interest. For $\mathbf{k}$ near $\mathbf{0}$, the terms due to the $\mathbf{k}\cdot\mathbf{p}$ operator can be considered as perturbations. If we define:

$$H^{kp}_{Mm} = \frac{\hbar}{m_0}\mathbf{k}\cdot\langle u_{M,\mathbf{0}}|\mathbf{p}|u_{m,\mathbf{0}}\rangle \tag{5.C.9}$$

(5.C.8) may be written as:

$$\left[E_{M,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} - E^{(n)}(\mathbf{k})\right] c^{(n)}_M(\mathbf{k}) + \sum_m H^{kp}_{Mm} c^{(n)}_m(\mathbf{k}) = 0 \tag{5.C.10}$$

In this last equation, we made use of the superscript (n) to designate one of the solutions (beware of degeneracy!) for which $E^{(n)}(\mathbf{k})$ tends towards $E_{n,\mathbf{0}}$ as $\mathbf{k}$ goes to $\mathbf{0}$. Simplification of the problem is achieved by separating the bands into an ensemble M_1 of interest (which couples strongly to band n) and a second ensemble M_2 of other bands. We then have:

$$\left[E_{M,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} - E^{(n)}(\mathbf{k})\right] c^{(n)}_M(\mathbf{k}) + \sum_{m\in M_1} H^{kp}_{Mm} c^{(n)}_m(\mathbf{k}) + \sum_{m\in M_2} H^{kp}_{Mm} c^{(n)}_m(\mathbf{k}) = 0 \tag{5.C.11}$$

For $n \in M_1$ and $M \in M_2$, the last sum is much smaller than the first. This gives:

$$c^{(n)}_M(\mathbf{k}) \cong \sum_{m\in M_1} \frac{H^{kp}_{Mm} c^{(n)}_m(\mathbf{k})}{E^{(n)}(0) - E_{M,\mathbf{0}}},\ M \in M_2 \tag{5.C.12}$$

For $n \in M_1$ and $M \in M_1$, we insert this result and find:

$$\left[E_{M,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} - E^{(n)}(\mathbf{k})\right] c^{(n)}_M(\mathbf{k}) + \sum_{m\in M_1} H^{kp}_{Mm} c^{(n)}_m(\mathbf{k}) + \sum_{m'\in M_1}\sum_{m\in M_2} \frac{H^{kp}_{Mm} H^{kp}_{mm'} c^{(n)}_{m'}(\mathbf{k})}{E^{(n)}(0) - E_{m,\mathbf{0}}} = 0 \tag{5.C.13}$$

After relabelling $m' \to m, m \to m'$ in the double sum, the equation becomes a diagonalization problem uniquely in the basis of functions belonging to the relevant bands $M \in M_1$:

$$\left[E_{M,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} - E^{(n)}(\mathbf{k})\right] c^{(n)}_M + \sum_{m\in M_1} (H^{kp}_{Mm} + H^{kp''}_{Mm}) c^{(n)}_m = 0$$

$$\text{with } H^{kp''}_{Mm} = \sum_{m'\in M_2} \frac{H^{kp}_{Mm'} H^{kp}_{m'm}}{E^{(n)}(0) - E_{m',\mathbf{0}}} \tag{5.C.14}$$

where we have suppressed the $\mathbf{k}$ argument in the coefficients c_m. Following a perturbative approach, we have therefore rolled the influence of uninteresting

bands into an effective Hamiltonian which acts upon the bands of interest. Using this method (called *Löwdin's* method), we have reduced the size of the problem by restricting the range of validity to the vicinity of $\mathbf{k} = \mathbf{0}$. By focusing on the bands nearest in energy to $E_{M,\mathbf{0}}$, the method becomes equivalent to stationary perturbations – perturbation of a degenerate level if $E_{M,\mathbf{0}}$ is degenerate, and a simple perturbation otherwise.

In III–V semiconductors, cubic zincblende is the crystal structure. In this case, the conduction band is non-degenerate (doubly degenerate if spin is included) and has a minimum at the Brillouin zone centre. Near $\mathbf{k} = \mathbf{0}$ the dispersion in this band is given by (5.C.14) in the simplest version, where the ensemble M_1 contains only the conduction band (c):

$$\left\{\left[E_{c,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} - E^{(c)}(\mathbf{k})\right] + \frac{\hbar^2}{m_0^2}\sum_{m\neq c}\frac{|\langle u_{M,\mathbf{0}}|\mathbf{k}\cdot\mathbf{p}|u_{m,\mathbf{0}}\rangle|^2}{E_{c,\mathbf{0}} - E_{m,\mathbf{0}}}\right\} c_c^{(c)} = 0 \tag{5.C.15}$$

whence:

$$E^{(c)}(\mathbf{k}) = E_{c,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar^2 k^2}{m_0^2}\sum_{m\neq c}\frac{|\langle u_{c,\mathbf{0}}|\hat{\mathbf{k}}\cdot\mathbf{p}|u_{m,\mathbf{0}}\rangle|^2}{E_{c,\mathbf{0}} - E_{m,\mathbf{0}}} \tag{5.C.16}$$

The most important contribution naturally comes from states closest in energy – the states in the valence band. If we neglect the other terms in the sum, we find:

$$E^{(c)}(\mathbf{k}) = E_{c,\mathbf{0}} + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar^2 k^2}{2m_0}\frac{P^2}{E_g} \tag{5.C.17}$$

where:

$$P^2 = \frac{2}{m_0}|\langle u_{c,\mathbf{0}}|p_x|u_{m,\mathbf{0}}\rangle|^2 \tag{5.C.18}$$

is the *Kane matrix element* and E_g is the gap.

From the dispersion $E^{(c)}(\mathbf{k})$, we may deduce the effective mass:

$$\frac{m_c}{m_0} = \left(1 + \frac{P^2}{E_g}\right)^{-1} \tag{5.C.19}$$

It turns out that P^2 can be obtained by optical measurement and that it varies relatively little ($\approx$20–25 eV) from one III–V semiconductor to another. Thus this simple formula predicts that the smaller the bandgap of a semiconductor, the smaller the effective mass in the conduction band. This prediction is well verified experimentally, and the fact that $m_c/m_0 \ll 1$ is due to the strong coupling $P^2/E_g \gg 1$ between the conduction and valence bands.

It is easy to apply the same method to the valence band (assumed to be non-degenerate), in which case we obtain:

$$\frac{m_v}{m_0} = \left(1 - \frac{P^2}{E_g}\right)^{-1} \cong -\frac{m_c}{m_0} \tag{5.C.20}$$

indicating that the hole's effective mass should roughly equal the electron's. Unfortunately, the situation in the valence band is more complicated, as this band is degenerate at maximum.

We cannot enter here into detailed calculations for the valence band, but rather will limit ourselves to offering a qualitative description of the results.

The crystal potential $V(\mathbf{r})$ possesses the symmetry of the zincblende structure. This means that its value will remain unchanged under the application of a symmetry operation on $\mathbf{r}$. Using symmetry arguments, it can be shown that such a potential leads to threefold degenerate eigenstates at the Brillouin zone centre, and that the degeneracy is independent of the magnitude of the potential. The valence band stems from such degenerate eigenstates.

The three basis functions can be labelled $|X\rangle$, $|Y\rangle$, and $|Z\rangle$, where the letters designate the symmetry properties of these wavefunctions. The effect of a symmetry operation on any of these functions is then to map one of the basis functions into one of the other two (or possibly back onto itself). Let us take, for example, mirror reflection in the $(1\,\bar{1}\,1)$ plane (see Fig. 5.1) and which corresponds to the change of co-ordinates $x \to y, y \to x, z \to z$. The basis functions have the properties:

$$\begin{aligned} |X\rangle &= u_X(y,x,z) = u_Y(x,y,z) \\ |Y\rangle &= u_Y(y,x,z) = u_X(x,y,z) \\ |Z\rangle &= u_Z(y,x,z) = u_Z(x,y,z) \end{aligned} \tag{5.C.21}$$

We say that the functions transform among each other as x, y, and z. With these symmetry properties, it can be easily shown that the momentum operator $\hat{\mathbf{p}} = \hbar/\mathrm{i}\nabla$ evaluated between two different functions yield a null matrix element. Let us take as an example $\langle Z|p_z|X\rangle$. A rotation of π about the z axis transforms $|Z\rangle$ to $|Z\rangle$, p_z to p_z, and $|X\rangle$ to $-|X\rangle$. Therefore $\langle Z|p_z|X\rangle = -\langle Z|p_z|X\rangle = 0$. The lifting of the degeneracy at $\mathbf{k} \neq \mathbf{0}$ by $\mathbf{k}\cdot\mathbf{p}$ interaction cannot occur via coupling between the valence band functions, but must necessarily result from interactions with other bands. Given the restricted basis formed by the three valence band functions, the cubic symmetry forces the effective interaction (see (5.C.14)) to take the form:

$\langle i\|\mathbf{k}\cdot\mathbf{p}\|j\rangle$	$\|X\rangle$	$\|Y\rangle$	$\|Z\rangle$
$\langle X\|$	$Lk_x^2 + M(k_y^2 + k_z^2)$	Nk_xk_y	Nk_xk_z
$\langle Y\|$	Nk_yk_x	$Lk_y^2 + M(k_z^2 + k_x^2)$	Nk_yk_z
$\langle Z\|$	Nk_zk_x	Nk_zk_y	$Lk_z^2 + M(k_x^2 + k_y^2)$

(5.C.22)

with three different constants L, M, N. We can now proceed to diagonalize the effective Hamiltonian to obtain the form of the three valence bands for $\mathbf{k} \neq \mathbf{0}$ in terms of these three parameters. Unfortunately, our troubles do not end here, as we must still take into account spin and spin–orbit interaction. If not for this last effect, the effect of spin alone would be to double the number of available states in each valence band with a total degeneracy of 3×2 at $\mathbf{k} = \mathbf{0}$.

Spin–orbit interaction results from a relativistic effect, which may be explained qualitatively as follows: an electron crosses an electrostatic potential at great speed. In the electron's frame of reference, the electric field associated with this potential transforms into a magnetic component which in turn interacts with the electron's spin. Overall, the electron motion and spin become coupled, with the result being that the spin operator no longer commutes with the Hamiltonian. The original 3×2 degenerate levels split into two levels, one being fourfold degenerate and the other being twofold degenerate and situated Δ_0 in energy below the former. The lifting of this degeneracy is in perfect analogy with that of the p_x, p_y, p_z orbitals in an atom under the influence of spin–orbit interaction. As a result, the angular momentum $l = 1$ and spin $s = 1/2$ are no longer good quantum numbers. On the other hand, the new stationary states of the system are eigenvectors of the observables J^2 and J_z, where $\mathbf{J} = \mathbf{L} + \mathbf{S}$. The observable J^2 has as eigenvalues $J^2 = j(j+1)$, with $j = 3/2$ and $1/2$. The state $j = 3/2$ is fourfold degenerate with $j_z = \pm 3/2, \pm 1/2$; and the state $j = 1/2$ is twofold degenerate with $j_z = \pm 1/2$. The new stationary states are now linear combinations of the basis functions $|X\uparrow\rangle$, $|X\downarrow\rangle$, $|Y\uparrow\rangle$, $|Y\downarrow\rangle$, $|Z\uparrow\rangle$, $|Z\downarrow\rangle$, and the new states are designated $|j, j_z\rangle$ as is common practice in atomic physics. This then complicates the form of the $\mathbf{k} \cdot \mathbf{p}$ matrix, but some small compensation is obtained by noting that near $\mathbf{k} = \mathbf{0}$, we may limit ourselves to the $j = 3/2$ level applying Löwdin's method to the ensuing 4×4 matrices. The fact that there are only three parameters is due to the persistence of cubic symmetry. A fairly involved calculation leads to the following form for the effective Hamiltonian in $|3/2, \pm 3/2\rangle$ and $|3/2, \pm 1/2\rangle$ basis:

$$
\begin{array}{c}
\left|\tfrac{3}{2}, +\tfrac{3}{2}\right\rangle \; \left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle \; \left|\tfrac{3}{2}, +\tfrac{1}{2}\right\rangle \; \left|\tfrac{3}{2}, -\tfrac{3}{2}\right\rangle \\
H^{kp''}_{j_z, j'_z} = \begin{vmatrix} H_h & c & b & 0 \\ c^\dagger & H_l & 0 & -b \\ b^\dagger & 0 & H_l & c \\ 0 & -b^\dagger & c^\dagger & H_h \end{vmatrix} \begin{array}{l} \left|\tfrac{3}{2}, +\tfrac{3}{2}\right\rangle \\ \left|\tfrac{3}{2}, -\tfrac{1}{2}\right\rangle \\ \left|\tfrac{3}{2}, +\tfrac{1}{2}\right\rangle \\ \left|\tfrac{3}{2}, -\tfrac{3}{2}\right\rangle \end{array}
\end{array}
\tag{5.C.23}
$$

with the elements:

$$
\begin{aligned}
H_h &= -\frac{\hbar^2}{2m_0}(\gamma_1+\gamma_2)(k_x^2+k_y^2)-\frac{\hbar^2}{2m_0}(\gamma_1-2\gamma_2)k_z^2 \\
H_l &= -\frac{\hbar^2}{2m_0}(\gamma_1-\gamma_2)(k_x^2+k_y^2)-\frac{\hbar^2}{2m_0}(\gamma_1+2\gamma_2)k_z^2 \\
b &= \frac{\hbar^2}{2m_0}2\sqrt{3}\gamma_3(k_x-\mathrm{i}k_y)k_z \\
c &= \frac{\hbar^2}{2m_0}\sqrt{3}[\gamma_2(k_x^2-k_y^2)-2\mathrm{i}\gamma_3 k_x k_y]
\end{aligned}
\tag{5.C.24}
$$

where the three *Luttinger parameters* $\gamma_1, \gamma_2, \gamma_3$, are related to the parameters L, M, and N by $\gamma_1 = 2m_0(L+2M)/3$, $\gamma_2 = 2m_0(L-M)/6$, and $\gamma_3 = 2m_0N/6$. These parameters are given in Table 5.C.1 for GaAs and InP.

Table 5.C.1. *Values of the different Kane–Kohn–Lüttinger parameters used in calculating the band structure in GaAs and InP*

Parameters	GaAs	InP
a(Å)	5.6533	5.8688
E_g(eV)	1.42	1.35
Δ_o(eV)	0.341	0.11
γ_1	7.0	5.04
γ_2	2.25	1.55
γ_3	2.9	2.4

Diagonalization of the effective Hamiltonian (5.C.14) yields two doubly degenerate energies:

$$
\begin{aligned}
E &= E_v+\frac{\hbar^2k^2}{2m_0}+\left[\frac{H_h+H_l}{2}\pm\sqrt{\frac{(H_h-H_l)^2}{4}+b^2+c^2}\right] \\
&= E_v+\frac{\hbar^2}{2m_0}\left[(1-\gamma_1)k^2\pm\sqrt{4\gamma_2^2k^4+12(\gamma_3^2-\gamma_2^2)(k_x^2k_y^2+k_y^2k_z^2+k_z^2k_x^2)}\right]
\end{aligned}
\tag{5.C.25}
$$

If we absorb the contribution of the free electron into the constant γ_1, we see that the $\pm$ before the square root corresponds to a heavy hole mass (+) and a light hole mass (−). For either of these, there is a dependence on the direction of $\mathbf{k}$. In (1 0 0) and equivalent directions, we have:

$$
\frac{m_0}{m_{hh}} = -\gamma_1+2\gamma_2,\ \frac{m_0}{m_{lh}} = -\gamma_1-2\gamma_2 \tag{5.C.26}
$$

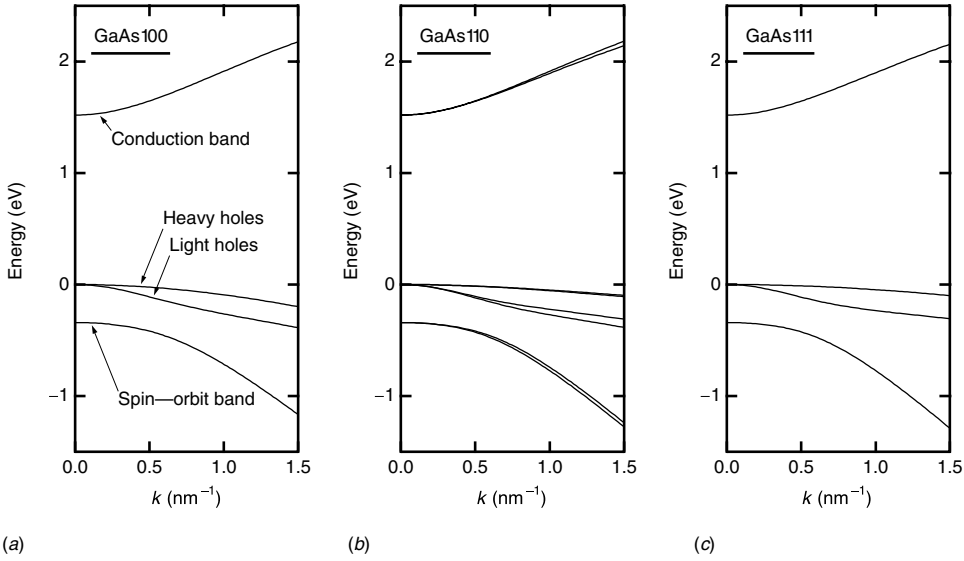


Fig. 5.C.1. Band structure of GaAs calculated using the $\mathbf{k} \cdot \mathbf{p}$ method by including 9×2 bands in (5.C.10). Near $\mathbf{k} = \mathbf{0}$ the heavy hole and light hole bands vary according to (5.C.25). As we move away from the zone centre, however, the effective mass approximation becomes poorer.

and in (1 1 1) directions:

$$\frac{m_0}{m_{hh}} = -\gamma_1 + 2\gamma_3, \frac{m_0}{m_{lh}} = -\gamma_1 - 2\gamma_3 \tag{5.C.27}$$

We may conclude that the equipotential surfaces are spherical only if $\gamma_2 = \gamma_3$.

Finally, we need to determine the Luttinger parameters for a given semiconductor experimentally. In practice, cyclotron resonance measurements are used to extract the shapes of the hole equipotential surfaces to which equation (5.C.25) is then fitted to obtain the 'best fit' Luttinger parameters.

Figure 5.C.1 shows the bandstructure of GaAs calculated by the $\mathbf{k} \cdot \mathbf{p}$ method by including 9×2 bands in (5.C.10). Close to $\mathbf{k} = \mathbf{0}$, we see the dependence on k of the heavy and light hole bands as on (5.C.25), but the more we move away from the centre of the Brillouin zone, the less the effective mass approximation is valid.

FURTHER READING

P. Yu and M. Cardona, *Fundamentals of Semiconductors*, Springer, Berlin (1995).

5.D Deep defects in semiconductors

As we have shown up until now, the existence of forbidden energy bands results from periodicity of the crystal potential. Any deviation from this periodicity will therefore introduce additional, allowed energy levels to the electron in the band-gap. We may use the chemical model as a means of conceptualizing this (refer to Fig. 5.6). In this case, an atom not involved in sp^3 bonding, would leave a level situated right between the bonding and anti-bonding energy levels. A level of this type, necessarily deep in the bandgap relative to the shallow hydrogen-like states of donor and acceptor atoms, is called a *deep defect* and generally contributes undesirable characteristics to its semiconductor host. We will see in Chapters 6, 11, and 13 that such defects often behave as non-radiative recombination centres. We will now give, within the framework of the tight binding model, a brief theoretical description useful in arriving at some degree of understanding of the nature of these defects.

We return to our one-dimensional crystal, as described back in Complement 5.B. In this case, our lattice consisted of a periodic succession of isolated atomic potentials coupled by a tunnel integral A (and identical between all nearest neighbours). We now suppose that a different atomic species, impurity, vacancy, . . ., is located at $m = 0$. This impurity will be assumed to possess a different energy level at E_0'; but for simplicity, the nearest neighbour tunnel integral A, will be left unchanged. The secular equations (5.B.8) then remain unchanged:

$$\begin{cases} -AC_{m-2} + E_0 C_{m-1} - AC_m = EC_{m-1} \\ -AC_{m-1} + E_0 C_m - AC_{m+1} = EC_m \\ -AC_m + E_0 C_{m+1} - AC_{m+2} = EC_{m+1} \end{cases} \tag{5.D.1}$$

except at $m = 0$ where they become:

$$\begin{cases} -AC_{-2} + E_0 C_{-1} - AC_0 = EC_{-1} \\ -AC_{-1} + E_0' C_0 - AC_1 = EC_0 \\ -AC_0 + E_0 C_1 - AC_2 = EC_1 \end{cases} \tag{5.D.2}$$

It can be shown that the allowed bands are not fundamentally changed. We will focus on the fact, however, that there are now spatially localized states at $m = 0$, i.e. of type $e^{m\kappa a}$ for $m < 0$ and $e^{-m\kappa a}$ for $m > 0$, where κ is real.

The functions:

$$C_m = e^{\pm m\kappa a}, \text{ for } m \neq 0 \tag{5.D.3}$$

still remain solutions to (5.D.1) as long as the following relationship holds:

$$E(\kappa) - E_0 = -A(e^{\kappa a} + e^{-\kappa a}) = -2A\cosh(\kappa a) \tag{5.D.4}$$

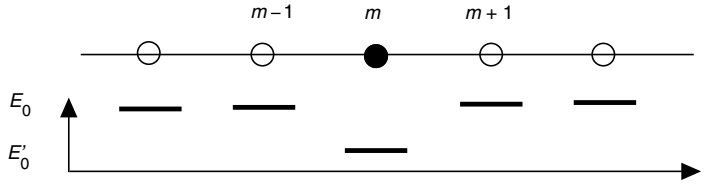


Fig. 5.D.1. Representation of a defect in a one-dimensional periodic potential.

except for $m = 0$. Without the presence of the middle equation in (5.D.2), it is the equality $\kappa = ik$ imposed by the Born–von Karman cyclic boundary conditions which leads to the existence of bands. Alternately, the equation at $m = 0$ (5.D.2) becomes another equation at the limits which (given the above equation) results in:

$$\begin{cases} e^{-2\kappa a} + C_0 = (e^{\kappa a} + e^{-\kappa a})e^{-\kappa a} \\ -Ae^{-\kappa a} + E'_0 C_0 - Ae^{\kappa a} = [E_0 - A(e^{\kappa a} + e^{-\kappa a})]C_0 \\ C_0 + e^{2\kappa a} = (e^{\kappa a} + e^{-\kappa a})e^{\kappa a} \end{cases} \tag{5.D.5}$$

Clearly the upper ($m = -1$) and lower ($m = +1$) equations admit $C_0 = 1$ as a solution, which then leads to a median equation (at $m = 0$) of:

$$E'_0 - E_0 = -2A\sinh(\kappa a) \tag{5.D.6}$$

We note that, as κ is positive, there will only be a solution if $E'_0 < E_0$, i.e. an impurity cannot localize an electronic level unless the energy of its fundamental level is deeper than that of the atoms in the host crystal. We then eliminate κ between both (5.D.4) and (5.D.6) to obtain the energy position of the impurity level $E(\kappa)$, which we label E_{def}:

$$E_{\text{def}} = E_0 - \sqrt{(E'_0 - E_0)^2 + 4A^2} \tag{5.D.7}$$

This level is positioned below the conduction band minimum $E_0 - 2A$ and, therefore, within the forbidden gap of the semiconductor (see Figs. 5.D.1 and 5.D.2). Associated with this deep level is a localized wavefunction, with a decay parameter κ given by:

$$\kappa_{\text{def}} = \frac{1}{a}\operatorname{arg\,cosh}\left(\frac{E_0 - E_{\text{def}}}{2A}\right) \tag{5.D.8}$$

The deeper the level in the bandgap, the more spatially localized the state. The consequence of the presence of impurities in semiconductors is then generally the introduction of allowed energy states within the bandgap which serve to localize or trap carriers. The influence of these levels is undesirable for many reasons. First, these centres trap charge carriers, which decreases the conductivity of the semiconductors. Also, these levels behave both as non-radiative generation and

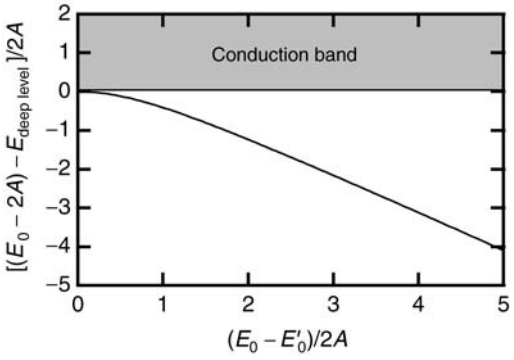


Fig. 5.D.2. Change in the position of the deep level in gap E_{defect} as a function of difference in chemical energy between the starting levels $E_0 - E_0'$ (in units of $2A$).

recombination sites for electrons and holes leading to leakage currents across junctions and non-radiative recombination in lasers. As a result, the development of a novel semiconductor material for device applications, often involves a gigantic metallurgical development stage to find a means of eliminating (or reducing below some critical level) these deep defects from the bulk material.

FURTHER READING

J. S. Blakemore, *Semiconductor Statistics*, Dover, New York (1987).

S. T. Pantelides, Ed., *Deep Centers in Semiconductors*, 2nd Edn, Gordon and Breach, New York (1992).

6 Electronic properties of semiconductors

6.1 Introduction

We saw in Chapter 5 that electrons in a crystalline medium are distributed within allowed energy bands (of conduction or valence type) and that the relative position of the Fermi level with respect to band extrema allows one to explain the isolating, metallic, or semiconducting characteristics of a given material. Using a quantum theory of solids, we also saw that the effect of an externally applied electric field on the electrons in a crystal is to cause them to circulate in periodic fashion through the Brillouin zone (i.e. *Bloch oscillations*) – a prediction that has yet to be observed experimentally in crystals (it has recently been observed in superlattices, though). In this chapter, we will elaborate on a few theoretical elements which will allow us to correlate quantum and classical aspects of the behaviour of electrons in matter. In particular, we will discuss how *scattering mechanisms* between different states within the band structure allow one to account for such macroscopic electronic properties in semiconductors as described by Ohm's or Fick's laws, avalanche, etc. At the end of the chapter, we will be in a position to present an ensemble of equations which will allow us to understand the dominant transport mechanisms in semiconductors.

6.2 Boltzmann's equation

We saw in Chapter 5 that the motion of an electron within a semiconductor band subject to an electric field $\mathbf{F}$ may be said to experience a change of states in the band given by:

$$|a(\mathbf{k}, t)|^2 = \left| a\left(\mathbf{k} - \frac{q\mathbf{F}t}{\hbar}, 0 \right) \right|^2 \tag{6.1}$$

This implies that if an electron is in a state $\mathbf{k}$ in the band at time $t = 0$, then it will be in a state $\mathbf{k} + q\mathbf{F}t/\hbar$ at time t. We recall that this observation forms the basis for the phenomenon of Bloch oscillations. Alternatively, the effect of the field may be described by the equation:

$$\hbar \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = q\mathbf{F} \tag{6.2}$$

This formula is analogous to Newton's law if we consider $\hbar\mathbf{k}$ as a *pseudo-momentum*. In real space, the velocity of an electron in state $\mathbf{k}$ is given by the group velocity:

$$\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) \tag{6.3}$$

and therefore:

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}\mathbf{k}}\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = \frac{q\mathbf{F}}{\hbar^2}\cdot\nabla_{\mathbf{k}}\nabla_{\mathbf{k}} E(\mathbf{k}) = M^{-1}\cdot q\mathbf{F} \tag{6.4}$$

where we may suppose that the inverse matrix M^{-1} for the effective mass does not depend on $\mathbf{k}$ in the vicinity of an energy extremum in reciprocal space. This description is valid for an ideal semiconductor (or any other crystalline solid) with the caveat that the field be sufficiently weak as to prevent band-to-band transitions and more importantly, that scattering by imperfections (phonons, impurities, etc.) resulting in transitions between different $\mathbf{k}$ states may be ignored (this latter condition is generally untenable and severely limits the practicality of this description).

On the other hand, this description forms the basis of the *semiclassical approximation* for electronic transport. In this case, the response of an electron (at time t, situated at a position $\mathbf{r}$ within a semiconductor, and in a state $\mathbf{k}$ within the nth band) to a local field $\mathbf{F}(\mathbf{r})$ is determined according to (6.1) and (6.4). Deviations from the ideal case (impurities, lattice vibrations, defects, . . .) provoke transitions between $\mathbf{k}$ and $\mathbf{k}'$ *without any corresponding effect on the particle's position.* Clearly, this is an approximation, as the simultaneous specification of the position and velocity (to arbitrary precision) of an electron is not possible in quantum mechanics and because the notion of a state $\mathbf{k}$ has no precise sense except in a semiconductor of infinite extent. In spite of these restrictions, this approximation is quite correct as long as the variations in the relevant parameters (e.g. field strength, . . .) are negligible over length scales equal to or greater than the mean distance between scattering events experienced by the electron. Further details concerning justification of the semiclassical approximation are not given here. Clearly, a quantum well will not admit a semiclassical description for the motion of an electron perpendicular to the interfaces. In fact, the main feature of semiclassical approximation is that the wave nature of the electron only comes into play indirectly through the band structure $E(\mathbf{k})$ and *via* scattering processes, which are calculated as transition probabilities per unit time (given by Fermi's golden rule) between wavefunctions corresponding to an initial state $\mathbf{k}$ and a final state $\mathbf{k}'$.

The ensemble of electrons in a band can now be described by a distribution

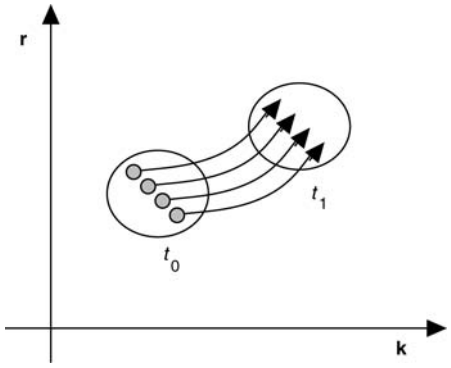


Fig. 6.1. Orbits for a group of particles in phase space between times t_0 and t_1 without collision.

function $f(\mathbf{r}, \mathbf{k}, t)$ which designates either the probability at time t of finding an electron at position $\mathbf{r}$ in state $\mathbf{k}$, or the number of electrons in a phase space volume element of $\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{k}$. The two interpretations are related through the density of states $g_s\mathrm{d}^3\mathbf{r}/(2\pi)^3$, where $g_s = 2$ is the electron spin degeneracy.

Evolution of $f(\mathbf{r}, \mathbf{k}, t)$ in the absence of any collisions between t and $t + \mathrm{d}t$ may be obtained by observing a volume in phase space at time $t + \mathrm{d}t$, Fig. 6.1. The particles now at $(\mathbf{r}, \mathbf{k})$ were initially at $(\mathbf{r} - \mathbf{v}\mathrm{d}t, \mathbf{k} - (\mathrm{d}\mathbf{k}/\mathrm{d}t)\mathrm{d}t)$ at time t, and have therefore drifted:

$$\begin{aligned}\frac{\partial f}{\partial t} &= \frac{f(\mathbf{r}, \mathbf{k}, t + \mathrm{d}t) - f(\mathbf{r}, \mathbf{k}, t)}{\mathrm{d}t} \\ &= \frac{f\left(\mathbf{r} - \mathbf{v}\mathrm{d}t, \mathbf{k} - \dfrac{\mathrm{d}\mathbf{k}}{\mathrm{d}t}\mathrm{d}t, t\right) - f(\mathbf{r}, \mathbf{k}, t)}{\mathrm{d}t} \\ &= -\mathbf{v}(\mathbf{k}) \cdot \nabla_{\mathbf{r}} f(\mathbf{r}, \mathbf{k}, t) - \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} \cdot \nabla_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}, t)\end{aligned} \tag{6.5}$$

To the particles having drifted towards $(\mathbf{r}, \mathbf{k})$, we must add the net number of particles injected into this region of phase space due to collisions from states $\mathbf{k}'$:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \sum_{\mathbf{k}'} S(\mathbf{k}' \to \mathbf{k}) f(\mathbf{r}, \mathbf{k}', t) - S(\mathbf{k} \to \mathbf{k}') f(\mathbf{r}, \mathbf{k}, t) \tag{6.6}$$

where $S(\mathbf{k}' \to \mathbf{k})$ is the probability per unit time for a transition to occur between states $\mathbf{k}'$ and $\mathbf{k}$.

The equation describing the evolution of f under the combined effects of an electric field and collisions is therefore:

$$\frac{\partial f}{\partial t} + \mathbf{v}(\mathbf{k}) \cdot \nabla_{\mathbf{r}} f + \frac{q\mathbf{F}(\mathbf{r})}{\hbar} \cdot \nabla_{\mathbf{k}} f = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \tag{6.7}$$

Boltzmann's equation

This equation, referred to as *Boltzmann's equation*, is of great significance as it relates both the quantum and classical aspects of electrons in solids to each another. Once the solution $f(\mathbf{r}, \mathbf{k}, t)$ to this equation is known, we may calculate such average macroscopic quantities as the carrier density:

$$n(\mathbf{r}, t) = \int \frac{g_s \mathrm{d}^3\mathbf{k}}{(2\pi)^3} f(\mathbf{r}, \mathbf{k}, t) \tag{6.8}$$

the current:

$$\mathbf{j}(\mathbf{r}, t) = q \int \frac{g_s \mathrm{d}^3\mathbf{k}}{(2\pi)^3} \mathbf{v}(\mathbf{k}) f(\mathbf{r}, \mathbf{k}, t) \tag{6.9}$$

the average kinetic energy of the particles:

$$\langle E(\mathbf{r}, t) \rangle = \frac{1}{n(\mathbf{r}, t)} \int \frac{g_s \mathrm{d}^3\mathbf{k}}{(2\pi)^3} E(\mathbf{k}) f(\mathbf{r}, \mathbf{k}, t) \tag{6.10}$$

and the variance of the velocity:

$$\sigma_v^2(\mathbf{r}, t) = \frac{1}{n(\mathbf{r}, t)} \int \frac{g_s \mathrm{d}^3\mathbf{k}}{(2\pi)^3} [v^2(\mathbf{k}) - \langle \mathbf{v} \rangle^2] f(\mathbf{r}, \mathbf{k}, t) \tag{6.11}$$

Equation (6.7), used in conjunction with the collision term (6.6), is of little practical use. We will now seek to obtain a few simple solutions to Boltzmann's equation. First, we shall suppose that the collision integral may be approximated by a single relaxation time (depending only on the energy of state $\mathbf{k}$) which will tend to bring the distribution back to its equilibrium form $f_{\text{eq}}(\mathbf{r}, \mathbf{k}, t)$ (see Fig. 6.2), i.e:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \cong -\frac{f(\mathbf{r}, \mathbf{k}, t) - f_{\text{eq}}(\mathbf{r}, \mathbf{k}, t)}{\tau(E(\mathbf{k}))} \tag{6.12}$$

The approximation (6.12) is of overwhelming physical significance. While all the transport equations introduced up to now are invariant with respect to time reversal, (6.12) introduces a privileged direction of time ('time arrow'). Many of the unsolved bridges between microscopic theory of matter and macroscopic thermodynamics lie hidden in that equation. This last approximation implies that once the perturbing electric field is removed, the function will relax in exponential fashion back to its equilibrium shape, with a relaxation time $\tau(E)$. Seeking a stationary solution close to the equilibrium distribution, we find by iteration the distribution function:

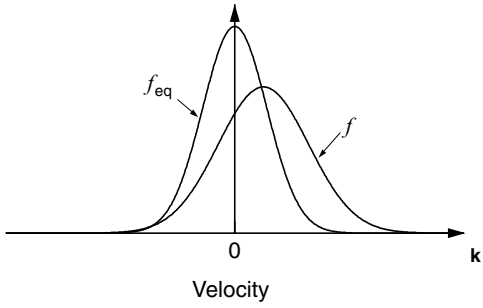


Fig. 6.2. Distribution functions at equilibrium ($f_{eq}(\mathbf{k})$) and under the influence of an electric field ($f(\mathbf{k})$).

$$\begin{aligned} f &\cong f_{eq} - \tau(E(\mathbf{k}))[\mathbf{v}(\mathbf{k})\cdot\nabla_r f_{eq} + q\mathbf{F}\cdot\nabla_k f_{eq}] \\ &= f_{eq} - \tau(E(\mathbf{k}))\left[\mathbf{v}(\mathbf{k})\cdot\nabla_r f_{eq} + q\mathbf{v}(\mathbf{k})\cdot\mathbf{F}\frac{\partial f_{eq}}{\partial E}\right] \end{aligned} \tag{6.13}$$

Using this last expression, we obtain for the average velocity:

$$\langle\mathbf{v}(\mathbf{r})\rangle = \frac{-\int\frac{g_s d^3\mathbf{k}}{(2\pi)^3}\tau(E(\mathbf{k}))\left(\mathbf{vv}\cdot\nabla_r f_{eq} + q\mathbf{vv}\cdot\mathbf{F}\frac{\partial f_{eq}}{\partial E}\right)}{\int\frac{g_s d^3\mathbf{k}}{(2\pi)^3}f_{eq}[\mathbf{r}, E(\mathbf{k})]} \tag{6.14}$$

In the case of a simple non-degenerate semiconductor at temperature T, $f_{eq} \propto \exp(-E/k_B T)$, where k_B is Boltzmann's constant, and for which we may suppose $\mathbf{v}(\mathbf{k}) = \hbar\mathbf{k}/m^*$, Eq. (6.14) may be written:

$$\langle\mathbf{v}(\mathbf{r})\rangle = \mu\mathbf{F} - D\frac{\nabla_r n}{n} \tag{6.15}$$

where μ is the *mobility* defined as:

$$\mu = \frac{q}{3k_B T}\frac{\int v^2\tau\exp\{-[E(\mathbf{k})/k_B T]\}d^3\mathbf{k}}{\int\exp\{-[E(\mathbf{k})/k_B T]\}d^3\mathbf{k}} \tag{6.16}$$

and D is the *diffusion constant*:

$$D = \frac{1}{3}\frac{\int v^2\tau\exp\{-[E(\mathbf{k})/k_B T]\}d^3\mathbf{k}}{\int\exp\{-[E(\mathbf{k})/k_B T]\}d^3\mathbf{k}} \tag{6.17}$$

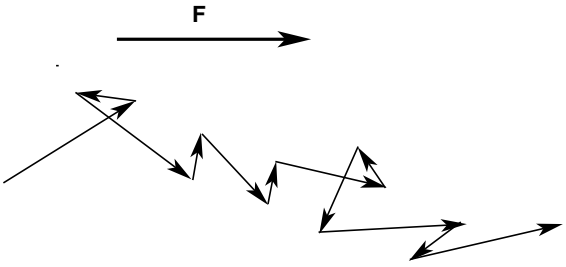


Fig. 6.3. Combined effects of an electric field and scattering in a solid: under the influence of an electric field, carriers in this figure are swept from left to right, while being scattered by collision processes.

Comparing these two last equations, we find the following *Einstein relation* which connects mobility to diffusion:

$$D = \frac{k_B T}{q}\mu \tag{6.18}$$

Einstein relation

Returning to (6.15), we see that the current is given by the sum of two terms:

$$\mathbf{j}(\mathbf{r}) = qn(\mathbf{r})\langle \mathbf{v}(\mathbf{r})\rangle = qn(\mathbf{r})\mu\mathbf{F} - qD\nabla_{\mathbf{r}} n(\mathbf{r}) \tag{6.19a}$$

which may be written (see Fig. 6.3) as:

$$\mathbf{j}(\mathbf{r}) = \mathbf{j}_{\text{cond}}(\mathbf{r}) + \mathbf{j}_{\text{diff}}(\mathbf{r}) \tag{6.19b}$$

The conduction current $\mathbf{j}_{\text{cond}}$ is the current described by *Ohm's law*:

$$\mathbf{j}_{\text{cond}}(\mathbf{r}) = \sigma\mathbf{F} \tag{6.20a}$$

Ohm's law

where σ refers to the electrical conductivity and relates to the mobility μ by:

$$\sigma = nq\mu \tag{6.20b}$$

The diffusion current $\mathbf{j}_{\text{diff}}$ relates to the migration of the carriers to lower density regions. This current is given by Fick's law:

$$\mathbf{j}_{\text{diff}}(\mathbf{r}) = -qD\nabla_{\mathbf{r}} n \tag{6.21}$$

Fick's law

Note: Given that $q = -e$ in our definition of electron mobility, this quantity is negative. A more common practice, however, is to make use of the absolute values for q and μ and to change the sign in front of μ in Eq. (6.15) as necessary.

A particularly important case arises when only the states in the vicinity of an energy minimum are occupied and where the effective mass near the minimum is

isotropic ($E(\mathbf{k}) = \hbar^2 k^2/2m^* + E_c$). The mobility may then be written as a function of a certain average of τ over the kinetic energy:

$$\mu = \frac{q}{3k_B T m^*} \frac{\int_0^\infty E^{3/2}\tau(E)\exp[-(E/k_B T)]\mathrm{d}E}{\int_0^\infty E^{1/2}\exp[-(E/k_B T)]\mathrm{d}E} \tag{6.22}$$

If the relaxation time is independent of energy E, this last expression leads to the simple *Drude model* whereby:

$$\mu = q\frac{\tau}{m^*} \tag{6.23a}$$

Mobility ($\mathrm{cm^2\,V^{-1}\,s^{-1}}$) in the Drude model

More generally, in the case where $\tau(E) = \tau_0(E/E_0)^\alpha$, Eq. (6.22) yields:

$$\mu = \frac{q\tau_0(k_B T)^\alpha}{m^*}\frac{2}{3}\frac{\Gamma[(5/2)+\alpha]}{\Gamma(3/2)} \tag{6.23b}$$

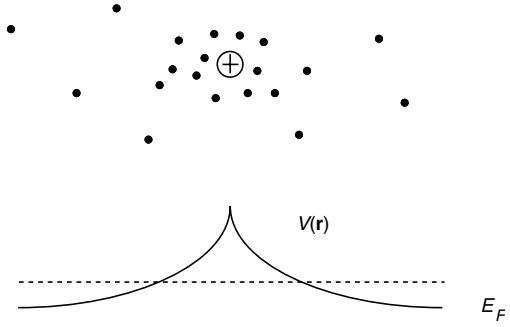
where Γ is the gamma function, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(x+1) = x\Gamma(x)$.

Example
At 300 K, gallium arsenide doped to a carrier concentration of $10^{15}\,\mathrm{cm^{-3}}$ has an electron relaxation time of 0.3 ps. Using the Drude model, the expected electron mobility would be $1.6 \times 10^{-19}\,\mathrm{C} \times 3 \times 10^{-13}\,\mathrm{s}/(0.067 \times 0.9 \times 10^{-30}\,\mathrm{kg})$ or $8000\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$.

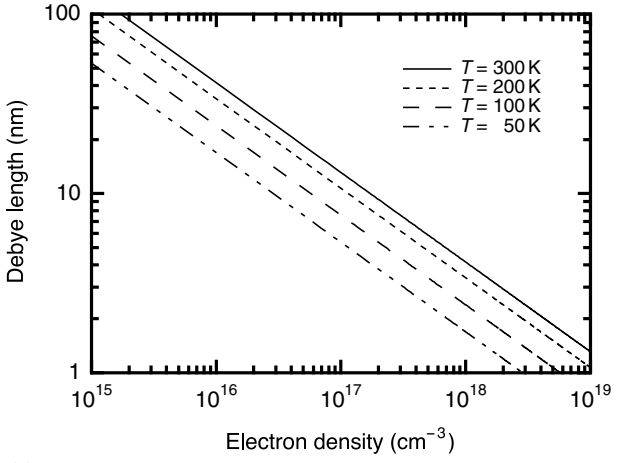
6.3 Scattering mechanisms

Collision mechanisms (also referred to as *scattering mechanisms*) are absolutely essential in describing the electrical properties of materials successfully. Most notably, these processes explain the proportionality between induced current flow and the magnitude of an applied electric field (Ohm's law). Among these mechanisms, the most important contributions generally arise from carrier scattering by charged impurities and phonons. Phonons (lattice vibrations) and electron–phonon interaction will be dealt with in the complements at the end of this chapter. Here we will study carrier scattering by impurities, a process generally dominated by the residual ions used in doping the host semiconductor.

Given the length scale of an electron wavelength in a conduction band, a



(*a*)



(*b*)

Fig. 6.4. (*a*) Conduction electrons accumulate around a positively charged impurity, thereby screening the impurity potential over a length scale given by the Debye wavelength. (*b*) Debye screening length as a function of doping concentration for GaAs ($\varepsilon_R = 12$) at various temperatures.

charged impurity may be treated as a point charge Ze. As the charge is situated in a crystal containing free electrons, we may expect the electron concentration near the impurity to increase thereby screening to a large extent the electric field associated with the defect. The simplest possible treatment (and highly valuable nonetheless) of this phenomenon, which follows, is attributed to Debye. We know that in a non-degenerate semiconductor, the conduction band possesses an average electron density of $n_0 = N_c \exp(-(E_c - E_F)/k_B T)$ (see (5.42) and (5.43)). If we add a slowly varying (with respect to lattice spacing) electrostatic potential $V(\mathbf{r})$, each electron at $\mathbf{r}$ will have, in addition to its lattice energy $E(\mathbf{k})$, a potential energy $-eV(\mathbf{r})$ (see Fig. 6.4). We may visualize the entire conduction band as being displaced vertically in energy by $-eV(\mathbf{r})$. At thermodynamic equilibrium, *the Fermi energy is constant across the material*, and the electron density varies as:

$$n(\mathbf{r}) = N_c \exp\left\{-\frac{[E_c - eV(\mathbf{r}) - E_F]}{k_B T}\right\} = n_0 \exp\left[\frac{eV(\mathbf{r})}{k_B T}\right] \tag{6.24a}$$

The change induced by the potential is then:

$$\begin{aligned}\rho_{\text{ind}}(\mathbf{r}) &= -e(n - n_0) = -en_0\left\{\exp\left[\frac{eV(\mathbf{r})}{k_B T}\right] - 1\right\} \\ &\cong -\frac{e^2 n_0}{k_B T} V(\mathbf{r}) = -\varepsilon q_D^2 V(\mathbf{r})\end{aligned} \tag{6.24b}$$

where the linearization of the charge defines the *Debye wavevector* q_D, or the *Debye length* $\lambda_D = 1/q_D$ as follows:

$$\begin{aligned}q_D^2 &= \frac{e^2 n_0}{\varepsilon k_B T} \\ \lambda_D &= \sqrt{\frac{\varepsilon k_B T}{e^2 n_0}}\end{aligned} \tag{6.25}$$

Debye wavevector and length

This length is important as it *describes the length scale over which an electron gas may screen an external potential*. This point is better illustrated by considering the electrostatic potential due to an external charge density $\rho_{\text{ext}}(\mathbf{r})$. The potential is then determined using Poisson's equation:

$$\nabla^2 V = -\frac{\rho}{\varepsilon} = -\frac{\rho_{\text{ext}} + \rho_{\text{ind}}}{\varepsilon} \cong -\frac{\rho_{\text{ext}}}{\varepsilon} + q_D^2 V \tag{6.26}$$

where ε is the permittivity of the medium ($\varepsilon = \varepsilon_R \varepsilon_0$). This equation may be solved by Fourier transformation:

$$\begin{aligned}\tilde{V}(\mathbf{r}) &= \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} V(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \ldots \\ -k^2 \tilde{V}(\mathbf{k}) &= -\frac{\tilde{\rho}_{\text{ext}}(\mathbf{k})}{\varepsilon} + q_D^2 \tilde{V}(\mathbf{k}) \\ \tilde{V}(\mathbf{k}) &= \frac{\tilde{\rho}_{\text{ext}}(\mathbf{k})}{(k^2 + q_D^2)\varepsilon} = \frac{\tilde{V}_{\text{ext}}(\mathbf{k})}{1 + (q_D^2/k^2)}\end{aligned} \tag{6.27}$$

For a point charge $\rho(\mathbf{r}) = Ze\delta(\mathbf{r})$ such that $\tilde{\rho}(\mathbf{k}) = Ze$, we find (after performing a somewhat lengthy inverse transformation) the total and external potentials:

$$V(\mathbf{r}) = \frac{Ze}{4\pi\varepsilon} \frac{\mathrm{e}^{-q_D r}}{r}, \; V_{\text{ext}}(\mathbf{r}) = \frac{Ze}{4\pi\varepsilon r} \tag{6.28}$$

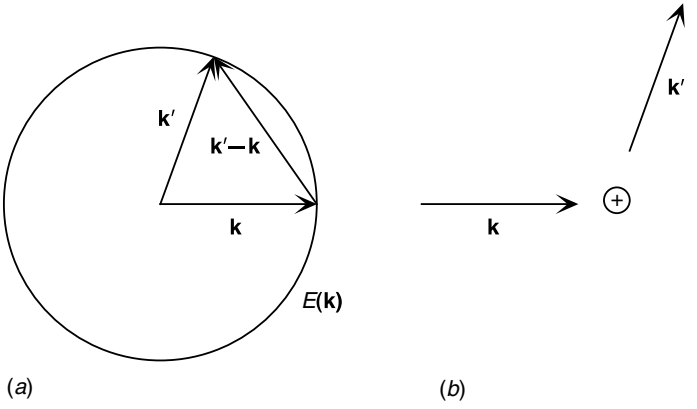


Fig. 6.5. Scattering of an electron by an impurity: (*a*) in reciprocal space and (*b*) in real space.

which indeed gives the screened Coulomb potential over distances on the order of the Debye length.

Example
For GaAs doped with a carrier concentration of $10^{17}\,\mathrm{cm}^{-3}$, we find:

$$\lambda_D = \sqrt{\frac{\varepsilon(k_B T/e)}{e n_0}} = \sqrt{\frac{12 \times 8.8 \times 10^{-12} \times 0.026}{1.6 \times 10^{-19} \times 10^{23}}} = 13\,\mathrm{nm}$$

We will now proceed with calculation of the scattering probability for an electron initially in a momentum state $\mathbf{k}$ within the crystal and having as a wavefunction:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}} u_{\mathbf{k}}(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \tag{6.29}$$

by a distribution of impurities located at $\mathbf{r}_j$:

$$V_{\mathrm{per}}(\mathbf{r}) = \sum_j V_{\mathrm{imp}}(\mathbf{r} - \mathbf{r}_j) \tag{6.30}$$

Returning to time-dependent perturbation theory, developed in Section 1.6.2, along with this stationary potential introduced at $t = 0$, we find following the same arguments the transition rate towards other $\mathbf{k}'$ states using Fermi's golden rule:

$$S(\mathbf{k} \to \mathbf{k}') = \frac{2\pi}{\hbar} |\langle \mathbf{k}' | V_{\mathrm{per}} | \mathbf{k} \rangle|^2 \delta[E(\mathbf{k}') - E(\mathbf{k})] \tag{6.31}$$

which, to within a prefactor, is the same formula given by Eqs. (1.78) and (1.83) for $\omega = 0$. The prefactor here is four times larger as we have not separated out the time dependence for positive and negative frequencies. Energy conservation shows that such a collision is elastic, see Fig. 6.5.

Substituting the wavefunctions for the states $\mathbf{k}$ and $\mathbf{k}'$, we find for the matrix

element:

$$\begin{aligned}\langle \mathbf{k}'|V_{\text{per}}|\mathbf{k}\rangle &= \frac{1}{V}\int u^*_{\mathbf{k}'}(\mathbf{r})u_{\mathbf{k}}(\mathbf{r})\mathrm{e}^{-\mathrm{i}(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}}V_{\text{per}}(\mathbf{r})\mathrm{d}^3\mathbf{r} \\ &= -\frac{1}{V}\frac{Ze^2 I(\mathbf{k}',\mathbf{k})}{\varepsilon[(\mathbf{k}'-\mathbf{k})^2+q_D^2]}\sum_j \mathrm{e}^{-\mathrm{i}(\mathbf{k}'-\mathbf{k})\cdot\mathbf{r}_j}\end{aligned} \tag{6.32}$$

where we have used in addition to the periodicity of the $u_{\mathbf{k}}$ wavefunctions the following approximations: (i) $V_{\text{per}}(\mathbf{r})$ does not vary appreciably over a crystal period, and (ii) the overlap integral:

$$I(\mathbf{k}',\mathbf{k}) = \frac{1}{V_{\text{cell}}}\int_{\text{cell}} u^*_{\mathbf{k}'}(\mathbf{r})u_{\mathbf{k}}(\mathbf{r}) \tag{6.33}$$

is often approximated as being unity for nearby states lying in the same band.

The square modulus of the matrix element will contain double sums of the type:

$$\sum_{j,j'} \mathrm{e}^{-\mathrm{i}(\mathbf{k}'-\mathbf{k})(\mathbf{r}_j-\mathbf{r}_{j'})}$$

A priori, we know nothing about the positions $\mathbf{r}_j$ of the impurities. The most simple approach is to exploit this ignorance by supposing that the positions are random in such a manner as to render the contributions of the different impurities equal to zero, as the phases $(\mathbf{k}'-\mathbf{k})\cdot(\mathbf{r}_j-\mathbf{r}_{j'})$ will be equally random for $j-j'$. As a result, only the sum for $j=j'$ remains, yielding the number of impurities, N_{imp}.

Finally, the transition rate is found to be:

$$S(\mathbf{k}\to\mathbf{k}') = n_{\text{imp}}\frac{2\pi}{\hbar}\frac{|I(\mathbf{k}',\mathbf{k})|^2}{V}\frac{Z^2e^4}{\varepsilon^2}\frac{1}{[(\mathbf{k}'-\mathbf{k})^2+q_D^2]^2}\delta[E(\mathbf{k}')-E(\mathbf{k})] \tag{6.34}$$

where $n_{\text{imp}} = N_{\text{imp}}/V$ is the impurity density and Ze is the charge on a single impurity. We recall that ε is the permittivity of the crystal without, however, taking into account the contribution due to free electrons. Their polarization is taken into account by q_D. We note as well that $S(\mathbf{k}\to\mathbf{k}') = S(\mathbf{k}'\to\mathbf{k})$, as must always be the case for an elastic process.

Knowing $S(\mathbf{k}\to\mathbf{k}')$, we may now calculate the relaxation time $\tau(k)$ for the simple case involving electrons in an isotropic, parabolic band, having an effective mass m^*. Equations (6.6), (6.12), and (6.13) give:

$$\begin{aligned}\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} &= -\frac{f-f_{\text{eq}}}{\tau(k)} = \frac{-e\hbar}{m^*}\mathbf{F}\cdot\mathbf{k}\frac{\partial f_{\text{eq}}}{\partial E} \\ &= \frac{e\hbar}{m^*}\mathbf{F}\cdot\sum_{\mathbf{k}'}(\mathbf{k}'-\mathbf{k})S(\mathbf{k}\to\mathbf{k}')\end{aligned} \tag{6.35}$$

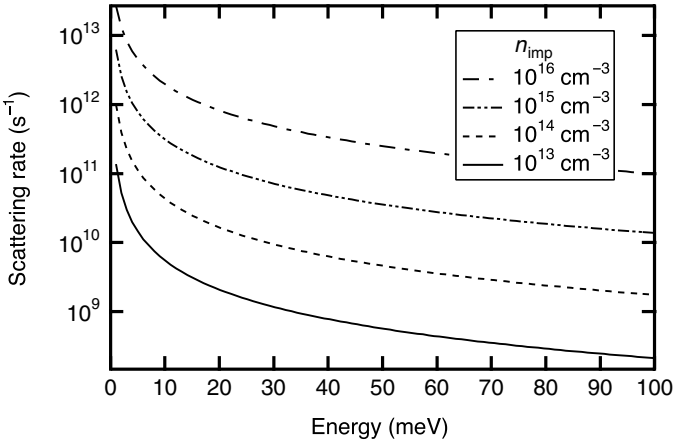


Fig. 6.6. $1/\tau(E)$ for charged impurity scattering at $T = 300$ K in n-type GaAs.

from which we conclude that:

$$\frac{1}{\tau(k)} = \sum_{\mathbf{k}'} \left(1 - \frac{\mathbf{k}' \cdot \mathbf{k}}{k^2}\right) S(\mathbf{k} \to \mathbf{k}') \tag{6.36}$$

Since $S(\mathbf{k} \to \mathbf{k}')$ depends on the angle θ between $\mathbf{k}'$ and $\mathbf{k}$, the pertinent relaxation rate differs from the average collision rate $\Sigma S(\mathbf{k} \to \mathbf{k}')$. As the collisions are weighted by the change in momentum they produce, this quantity refers rather to a *momentum relaxation time*. Approximating $\mathbf{I}(\mathbf{k}', \mathbf{k})$ by 1, we finally obtain:

$$\frac{1}{\tau(k)} = n_{\text{imp}} \frac{2\pi}{\hbar} \frac{Z^2 e^4}{V\varepsilon^2} \int \frac{V \mathrm{d}^3\mathbf{k}'}{(2\pi)^3} \frac{(1 - \cos\theta)\delta[E(k') - E(k)]}{[2k^2(1 - \cos\theta) + q_D^2]^2} \tag{6.37}$$

and the integral can be calculated using the spherical co-ordinates $(k',\ \theta, \phi)$ and the relationship $\delta((\hbar^2/2m^*)(k'^2 - k^2)) = (m^*/\hbar^2 k)\delta(k' - k)$:

$$\frac{1}{\tau(k)} = n_{\text{imp}} \frac{Z^2 e^4}{\varepsilon^2} \frac{m^*}{8\pi\hbar^3 k^3} \left\{ \ln\left[1 + \left(\frac{2k}{q_D}\right)^2\right] - \frac{(2k/q_D)^2}{1 + (2k/q_D)^2} \right\} \tag{6.38}$$

We can convince ourselves that the rate is of order $O(k)$ if $k \ll q_D$ and of order $O(k^3 \ln(k))$ if $k \gg q_D$.

Figure 6.6 shows numerical values for $1/\tau$ for different values of $n_{\text{imp}} = n_0$ at room temperature for electrons in GaAs. In Fig. 6.7, mobility values are obtained by numerically evaluating the integral in (6.16) for τ over the kinetic energy. In fact, we will see that at room temperature, scattering by phonons dominates over impurity scattering with the exception of samples possessing impurity concentrations in excess of $n_{\text{imp}} = 10^{17}\,\text{cm}^{-3}$.

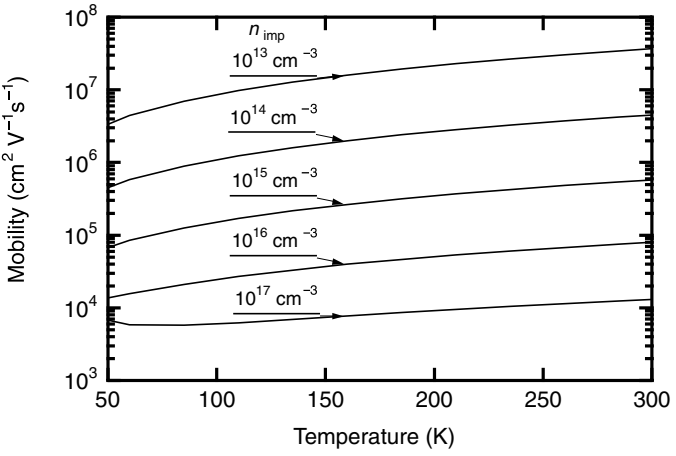


Fig. 6.7. Mobility determined by charged impurity scattering.

6.4 Hot electrons

The solution obtained to Boltzmann's equation (6.13) is only valid provided that both the electric field and the density gradient are weak. In this instance, a linear response is obtained, as in the cases of Ohm's and Fick's laws. In this regime, we can easily show that the kinetic energy of the electrons remains equal to that at equilibrium, $(3/2)k_BTn$: if we replace $\mathbf{k}$ by $-\mathbf{k}$ in (6.13), we see that the deviation from equilibrium of the distribution function simply changes sign and therefore does not contribute to the average energy.

As we increase the field strength $\mathbf{F}$, Boltzmann's equation no longer possesses a general analytic solution. As a result, a large number of numerical methods have been developed to assist with the study of such particular cases. This rather technical area of analysis exceeds the framework of this book, so we shall limit ourselves here to offering here a phenomenological description of some of the more important effects which arise under these conditions.

6.4.1 Warm electrons

When the linear regime is exceeded, the electric field supplies energy to the electron gas. The primary effect of collision is to randomize the velocities of the particles over the time scale set by the momentum relaxation time. As the electrons acquire a mean velocity $\langle v \rangle$ in the direction of the current, the energy contribution of the field to the electrons is given by $\mathbf{j} \cdot e\mathbf{F} = e^2 n \langle v \rangle F$. Initially, this energy heats the electron distribution. Once the system reaches a stationary state, this energy is

dissipated by inelastic processes (most notably electron–phonon interaction, which transfers electronic energy to the vibrational lattice energy, i.e. transfers the heat in the host material). This transfer takes place over a characteristic time, which is referred to as the *energy relaxation time*. This time is usually several orders of magnitude greater than the *momentum relaxation time*. This description implies that the electron gas takes on an average kinetic energy greater than that set by the equilibrium value. The temperature of the electron gas then exceeds that of the crystal lattice. Since electron mobility is temperature dependent, we may expect a corresponding correction to Ohm's law in the form of:

$$\langle v \rangle = (\mu + \alpha F^2)F \tag{6.39}$$

the term in F^2 being excluded by symmetry (we must have $\langle v \rangle \to -\langle v \rangle$ for $F \to -F$), and where α is proportional to $\mathrm{d}\mu/\mathrm{d}T$. We have just seen how impurity scattering leads to increasing mobilities with higher temperatures so that $\alpha > 0$. In Complement 6.B, we will show how scattering by phonons may lead to $\alpha < 0$. Thus, simple determination of the sign of the correction to Ohm's law may give an indication of the nature of the dominant process limiting mobility in a semiconductor.

The mathematical expression which corresponds to this image is an approximation to the distribution function and is called *the displaced Maxwell approximation*:

$$f(\mathbf{k}, \mathbf{k}_0, T_e) = n\left(\frac{2\pi\hbar^2}{m^* k_B T_e}\right)^{3/2} \exp\left[-\frac{\hbar^2(\mathbf{k} - \mathbf{k}_0)^2}{2m^* k_B T_e}\right] \tag{6.40}$$

for which the mean velocity is $\langle \mathbf{v} \rangle = \hbar\mathbf{k}_0/m^*$ and the average kinetic energy is $(3/2)k_B T_e + (1/2)m^*\langle v \rangle^2$ per electron. A method of solving Boltzmann's equation consists of applying this approximation to evaluate the collision integrals and the relaxation times as a function of these two parameters and scattering processes. The parameters are then determined by writing the rate equations of momentum and energy supplied by the field, and the relaxation rates due to collision.

We will not pursue this topic quantitatively any further, but the idea of the electronic temperature being able to differ from the crystalline lattice temperature is a very important point to note. It is of use in distinguishing between two separate relaxation time scales: one reflecting the rapid thermalization of the electron distribution itself, and another reflecting a slower process involving the transfer of electronic energy to the lattice.

6.4.2 Hot electrons: saturation velocity

Under the influence of strong electric fields, the electronic distribution becomes

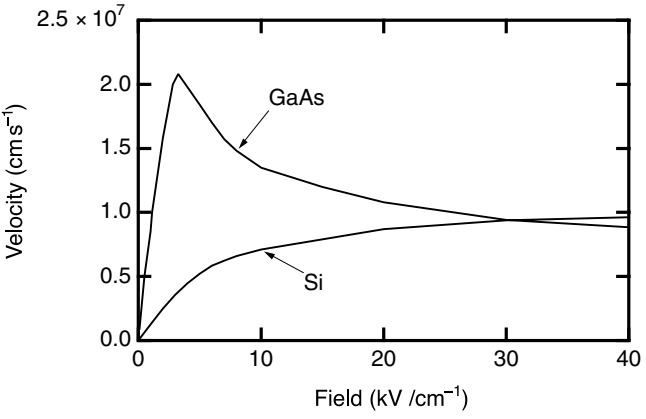


Fig. 6.8. Dependence of velocity on field strength in lightly doped *n*-type Si and GaAs.

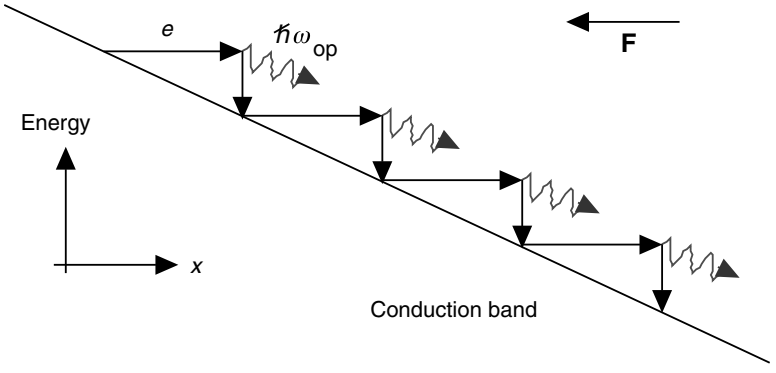


Fig. 6.9. Shockley's model for velocity saturation.

highly skewed. The average velocity usually tends in this case to a saturation value, where the velocity becomes independent of the applied field strength. For semiconductors, this saturation velocity is of the order of 10^7 cm s^{-1}. Figure 6.8 shows the dependence of velocity on field strength in Si and GaAs.

An extremely simplified model put together by Shockley elegantly explains the velocity saturation effect (Fig. 6.9). Let us suppose that the main inelastic scattering mechanism involves the emission of a (optical) phonon having an energy of $\hbar\omega_{op}$, and that this process is so efficient as to guarantee that as soon as an electron accumulates an amount of kinetic energy corresponding to the phonon energy, it will emit a phonon and lose all its energy and speed. The motion of each electron is then completely deterministic: immediately after the emission of a phonon, the electron finds itself at $\mathbf{k} = 0$ and begins to accelerate anew as given by:

$$\mathbf{k} = \frac{-e\mathbf{F}}{\hbar} t \tag{6.41}$$

This acceleration continues up to the threshold set by the phonon energy:

$$\frac{\hbar^2 k^2}{2m^*} = \hbar\omega_{\text{op}},\ v_{\max} = \frac{\hbar k}{m^*} = \sqrt{\frac{2\hbar\omega_{\text{op}}}{m^*}} \tag{6.42}$$

As a result, the velocity distribution becomes an equidistribution between $v = 0$ and $v_{\max}$, in a direction parallel to the applied field, with an average velocity $\langle v \rangle = v_{\max}/2$, which is clearly independent of the electric field **F**.

Example

Inserting the relevant numerical values for GaAs, i.e. $\hbar\omega_{\text{op}} = 36\,\text{meV}$, $m^* = 0.067m_0$, we find:

$$v_{\text{sat}} = \sqrt{\frac{36 \times 1.6 \times 10^{-22}}{2 \times 0.067 \times 9.1 \times 10^{-31}}}\frac{\text{m}}{\text{s}} \cong 2 \times 10^5\,\text{m}\,\text{s}^{-1} \tag{6.43}$$

Which is quite good considering the simplicity of the model.

This model is extreme in several ways. The distribution is completely anisotropic and is very far from being Maxwellian. It would therefore not be meaningful to define an electronic temperature in this case. Also, the assumption that phonon emission proceeds as soon as it is permitted by energy conservation, implies that the average energy $(3/2)\hbar\omega_{\text{op}}$ is itself field independent. Clearly, much room remains for improving this model.

6.4.3 Hot electrons: negative differential velocity

For certain semiconductors, such as GaAs (see Fig. 6.8) and InP, the average velocity as a function of field strength displays a maximum followed by a regime of decreasing velocity. An explanation for this behaviour can be found by considering the band structure for these semiconductors (Fig. 5.7). The conduction band possesses, away from the global minimum at the centre of the Brillouin zone, local minima (the L and X valleys) a few hundred meV above the Γ valley. At thermal equilibrium, states in these valleys are unoccupied. The application of a sufficiently strong electric field, however, will be able to accelerate electrons from the Γ valley to energies approaching that of the L valley to which they may be subsequently scattered. After being transferred to the L valley, the velocities of the electrons are greatly reduced. Even though the electric field will again be able to accelerate these electrons from states in that valley, their velocities will remain significantly smaller than those electrons in the Γ valley, given that the effective mass m_L is greater than m_Γ. At a stationary state, the occupation of states with reduced mobility continues to increase with the field strength, resulting in an overall decrease in the average electron velocity, as depicted in Fig. 6.10.

This *negative differential resistance* (NDR) effect is not simply a matter of

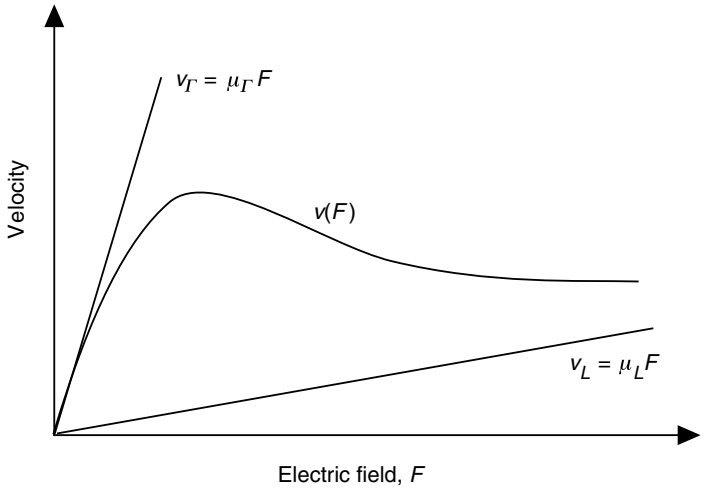


Fig. 6.10. Effect of electron transfer between the Γ and L valleys for increasing field strengths. The low velocity of electrons in the L valley is responsible for the region of negative differential resistance in the average velocity.

curiosity as it can create spatial and/or temporal instabilities. Imagine, for instance, a semiconductor doped to an average density n_0 and polarized by a field strength lying in the NDR regime. Positive fluctuation in the field at location z at time t will reduce the velocity of the carriers at this location so that downstream (i.e. in the direction of the electron flow) the carrier density will dip below n_0, and peak above n_0 upstream. As a result, a charge dipole will be set up whose further effect will be to increase the field (and hence the field inhomogeneity) at z. This process then feeds back on itself, forming a dipolar domain which drifts at a velocity situated somewhere between the saturation velocity and the maximum velocity in the crystal. This dipole structure disappears once it arrives at the anode, with new ones being created near the cathode. In a qualitative sense, we see how the current traversing the circuit will possess an oscillating component having a period of the order of L/v, where L is the length of the active zone of the semiconductor. For $L \approx 1\,\mu\text{m}$ and $v \approx 10^5\,\text{m}\,\text{s}^{-1}$, this corresponds to an oscillation frequency of 100 GHz. Such devices (called *Gunn diodes* after their inventor) allow for the fabrication of compact solid state microwave sources and constitute an important application of NDR.

6.5 Recombination

It is the possibility for electrons and holes to coexist which forms the basis for the application of semiconductors in optoelectronics (and, for a large part, in

electronics as well). We saw in Chapter 5 that, at thermodynamic equilibrium, the density of electrons n_0 and holes p_0, is subject to the relation $n_0 p_0 = n_i^2(T)$, where n_i is the intrinsic density (Eq. (5.49)). The density n_0 (and p_0) can be chosen by doping. This equilibrium can be perturbed, for example, by creating electron–hole pairs, by illuminating a sample (optical pumping) or by injecting carriers on either side of an interface separating alternately doped *n*-type and *p*-type regions (i.e. a *p–n* junction).

Nature is conservative, in that it always provides a number of mechanisms capable of bringing a system back to equilibrium. In the case of electron–hole pairs, these are the generation–recombination mechanisms. *Each process requires an associated interaction.* Interaction with an electromagnetic field allows for radiative recombination in which a photon is emitted, whereas photon absorption corresponds to a generative process for electron–hole pairs. The two processes form the principal subject of this book and will be treated in detail, most particularly, in the following chapter.

Non-radiative generation–recombination mechanisms are often a nuisance in optoelectronics as they proceed in parallel and in competition with radiative processes. It is therefore important to understand them in order to evaluate the performance and limitations of optoelectronic devices. The two dominant processes are:

1. *Shockley–Read–Hall recombination* – in which impurities play an essential role. At first an electron and then a hole become trapped on the same impurity centre, thereby eliminating one electron–hole pair.
2. *Auger recombination* – resulting from electron–electron interaction, whereby an electron recombines with a hole and transfers the resulting energy gained by the recombination to another electron (or hole) in the form of kinetic energy. This is an intrinsic process and does not rely on the participation of defect or impurity centres. The corresponding generation process in this case is *impact ionization*, where an electron with sufficient kinetic energy can create an electron–hole pair by transferring its energy to an electron in the valence band and promoting it to the conduction band.

The Auger effect and impact ionization are addressed in Complement 6.D. Here we will describe Shockley–Read–Hall recombination (Fig. 6.11).

To be specific, let us suppose that a donor impurity population exists at a concentration N_T corresponding to localized states with energy E_T in the gap. We saw in Complement 5.D that an impurity can create a *deep* (*electronic*) *level* (i.e. an allowed state) within the forbidden gap of a semiconductor. The additional electron supplied by the donor can be ionized and released into the conduction band, leaving behind it a positively ionized donor atom as described by:

$$N_T^0 \leftrightarrow N_T^+ + e \tag{6.44}$$

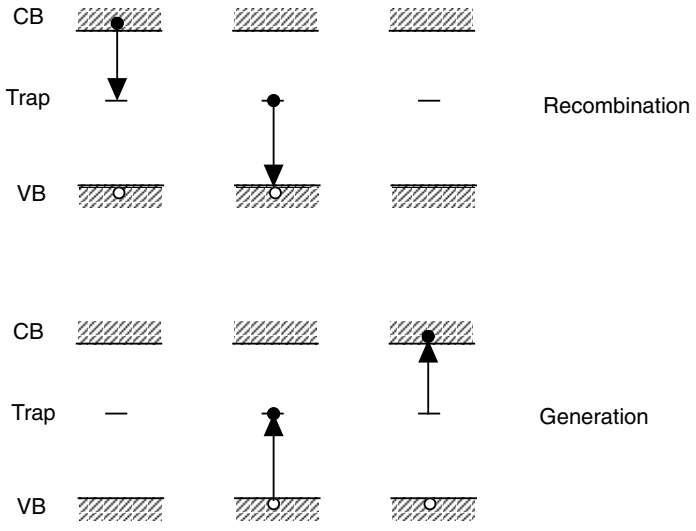


Fig. 6.11. Shockley–Read–Hall recombination and generation processes (CB, conduction band; VB, valence band).

The ionized donor can also capture an electron from the valence band, corresponding to the creation of a hole, according to:

$$N_T^+ \leftrightarrow N_T^0 + h \tag{6.45}$$

In the first reaction, the number of electrons created per second is proportional to the number of neutral traps:

$$G_e = e_e N_T^0 \tag{6.46}$$

where e_e is the emission coefficient (s^{-1}). The inverse process, describing recombination by capture, is proportional to the number of electrons and ionized traps:

$$R_e = c_e N_T^+ n \tag{6.47}$$

with c_e being the capture coefficient ($cm^3\ s^{-1}$).

At equilibrium, which we designate by writing a bar above the relevant quantities, the two rates are equal:

$$\overline{G_e} = \overline{R_e} \Rightarrow \frac{e_e}{c_e} = \frac{\overline{N_T^+} n_0}{\overline{N_T^0}} \tag{6.48}$$

This type of reasoning, in which we extract a *general* expression tying emission and recombination from a *particular* equilibrium situation, is entirely analogous with Einstein's reasoning presented in Complement 3.C (when relating the stimulated and spontaneous emission coefficients B and A). Equation (6.47) is called the

principle of detailed balance. For the second reaction (6.44), we similarly find:

$$\begin{aligned} G_h &= e_h N_T^+ \\ R_h &= c_h N_T^0 p \\ \overline{G_h} &= \overline{R_h} \Rightarrow \frac{c_h}{e_h} = \frac{\overline{N_T^+}}{\overline{N_T^0 p_0}} \end{aligned} \tag{6.49}$$

Away from equilibrium, the net rate of electron generation becomes:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = G_e - R_e = c_e \left(\frac{\overline{N_T^+}}{\overline{N_T^0}} n_0 N_T^0 - n N_T^+ \right) \tag{6.50}$$

and for holes:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = G_h - R_h = c_h \left(\frac{\overline{N_T^0}}{\overline{N_T^+}} p_0 N_T^+ - p N_T^0 \right) \tag{6.51}$$

Given the Fermi distribution, we have:

$$\frac{\overline{N_T^0}}{\overline{N_T^+}} = \frac{\overline{N_T^0}}{N_T - \overline{N_T^0}} = \exp\left(- \frac{E_T - E_F}{k_B T} \right) \tag{6.52}$$

and for a non-degenerate semiconductor:

$$\begin{aligned} n_0 &= N_c exp\left(- \frac{E_c - E_F}{k_B T} \right) \\ p_0 &= N_v exp\left(- \frac{E_F - E_v}{k_B T} \right) \end{aligned} \tag{6.53}$$

so that (6.50) and (6.51) become:

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= c_e \left[N_T^0 N_c \exp\left(- \frac{E_c - E_T}{k_B T} \right) - N_T^+ n \right] \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= c_h \left[N_T^+ N_v \exp\left(- \frac{E_T - E_v}{k_B T} \right) - N_T^0 p \right] \end{aligned} \tag{6.54}$$

In a stationary state (where, for example, a state of non-equilibrium is maintained by continuous optical pumping), we have $\mathrm{d}n/\mathrm{d}t = \mathrm{d}p/\mathrm{d}t$, from which we may express the number of ionized and neutral traps by n and p:

$$\frac{N_T^+}{N_T} = \frac{c_e N_c \exp\left(-\dfrac{E_c - E_T}{k_B T}\right) + c_h p}{c_h\left[N_v \exp\left(-\dfrac{E_T - E_v}{k_B T}\right) + p\right] + c_e\left[N_c \exp\left(-\dfrac{E_c - E_T}{k_B T}\right) + n\right]}$$
$$\frac{N_T^0}{N_T} = \frac{c_h N_v \exp\left(-\dfrac{E_T - E_v}{k_B T}\right) + c_e n}{c_h\left[N_v \exp\left(-\dfrac{E_T - E_v}{k_B T}\right) + p\right] + c_e\left[N_c \exp\left(-\dfrac{E_c - E_T}{k_B T}\right) + n\right]} \tag{6.55}$$

with the recombination rate:

$$-\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{np - n_i^2}{\dfrac{1}{c_e N_T}\left[p + N_v \exp\left(-\dfrac{E_T - E_v}{k_B T}\right)\right] + \dfrac{1}{c_h N_T}\left[n + N_c \exp\left(-\dfrac{E_c - E_T}{k_B T}\right)\right]} \tag{6.56}$$

which may also be expressed in the form of a lifetime $-\mathrm{d}n/\mathrm{d}t = (n - n_0)/\tau_n$:

$$\frac{1}{\tau_n} = \frac{(np - n_i^2)/(n - n_0)}{\tau_{n0}\left[p + N_v \exp\left(-\dfrac{E_T - E_v}{k_B T}\right)\right] + \tau_{p0}\left[n + N_c \exp\left(-\dfrac{E_c - E_T}{k_B T}\right)\right]} \tag{6.57}$$

with $\tau_{n0} = 1/c_e N_T$ and $\tau_{p0} = 1/c_h N_T$.

This expression seems fairly complex. Let us begin by noting that the expressions dependent upon E_T are equal to the electron and hole concentrations we would have at equilibrium assuming the Fermi level was equal to E_T. Next, two important cases give rise to the following simple limits:

1. For a nearly intrinsic semiconductor ($n_0 \approx p_0 \approx n_i$), traps at the centre of the gap, and strong pumping, $\delta n = n - n_0 = \delta p = p - p_0 \gg n_i$:

$$\frac{1}{\tau_n} = \frac{\delta n^2/\delta n}{\tau_{n0}\delta n + \tau_{p0}\delta n} = \frac{1}{\tau_{n0} + \tau_{p0}} = N_T \frac{c_e c_h}{c_e + c_h} \tag{6.58}$$

2. For a p-doped semiconductor ($n_0 = n_i^2/p_0 \ll n_i$), traps at the centre of the gap, and $|\delta n|, |\delta p| \ll p_0$:

$$\frac{1}{\tau_n} = \frac{p_0}{\tau_{n0} p_0} = \frac{1}{\tau_{n0}} = N_T c_e \tag{6.59}$$

In these two situations, the lifetime does not depend on the densities and the notion of a 'lifetime' becomes meaningful.

Example

The capture coefficient c_e is written σv_{th}, where σ corresponds to a *capture cross-section* and v_{th} is the thermal velocity $v_{\mathrm{th}}^2 = \langle v^2 \rangle$. For $\sigma \approx 10^{-15}\,\mathrm{cm}^2$ and

$v_{\text{th}} \approx 10^7\,\text{cm}\,\text{s}^{-1}$, we have $c \approx 10^{-8}\,\text{cm}^3\,\text{s}^{-1}$. The emission coefficient is given by (6.48) and (6.52), where:

$$e_e = c_e N_c \exp\left(-\frac{E_c - E_T}{k_B T}\right)$$

For $N_c = 5 \times 10^{17}\,\text{cm}^{-3}$ and assuming the traps are situated 0.4 eV below the conduction band (O in GaAs), we have $e_e \approx 560\,\text{s}^{-1}$ at room temperature.

6.6 Transport equations in a semiconductor

In principle, the semiclassical description of transport properties in a semiconductor device would be a Boltzmann equation including all the electrons in the conduction and valence bands. In the scattering integral, one would have to specify all the relevant interactions, including those interactions which scatter electrons between bands, corresponding to generation and recombination processes. To this equation, one would also have to add Maxwell's equations for the fields responsible for the forces acting on the electrons.

Such a description would clearly be very difficult to manipulate and the results obtained from it would be far too detailed. We therefore prefer to fall back on equations which are more global in nature and which are also, in part, phenomenological.

The first equations we will derive are the continuity equations, which are obtained from Boltzmann's equation by integrating the distribution function over all **k** states and by separating out the electron and hole contributions:

$$\begin{aligned}\frac{\partial n}{\partial t} - \frac{1}{e}\nabla\cdot\mathbf{j}_n &= G - R \\ \frac{\partial p}{\partial t} + \frac{1}{e}\nabla\cdot\mathbf{j}_p &= G - R\end{aligned} \tag{6.60}$$

Semiconductor equation 1: continuity equation

where G and R are the generation and recombination rates, with the electron current density $\mathbf{j}_n = -ne\langle\mathbf{v}_n\rangle$ and the hole current density $\mathbf{j}_p = pe\langle\mathbf{v}_p\rangle$. For the current densities, we use the equations:

$$\begin{aligned}\mathbf{j}_n &= e\mu_n n\mathbf{E} + eD_n\nabla n \\ \mathbf{j}_p &= e\mu_p p\mathbf{E} - eD_p\nabla p \\ \mathbf{j}_{\text{cond}} &= \mathbf{j}_n + \mathbf{j}_p\end{aligned} \tag{6.61}$$

Semiconductor equation 2: Ohm's and Fick's laws

These equations are inspired from the results obtained in Section 6.1 for weak fields and gradients, whereas the mobility values and diffusion coefficients are parameters obtained either by experiment or by numerical solution of Boltzmann's equation. We also note that the electric field designated by **E** in these equations should not be confused with energy. Depending upon the desired level of sophistication, we may also introduce in the model the dependencies for $\mu_n(E)$, etc., to take into account hot electron effects.

Added to the equations describing the reaction of the particles to the fields, are Maxwell's equations, which determine the dependence of the fields on the material properties:

$$\nabla \cdot \mathbf{D} = \rho \tag{6.62}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{6.63}$$

$$\nabla \times \mathbf{H} = \mathbf{j}_{\text{cond}} + \frac{\partial \mathbf{D}}{\partial t} \tag{6.64}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{6.65}$$

Semiconductor equation 3: Maxwell's equations

where for non-magnetic semiconductors, $\mathbf{B} = \mu_0 \mathbf{H}$ and $\mathbf{D} = \varepsilon \mathbf{E}$, where ε is the permittivity associated with the polarization of localized charges.

In most electronic transport problems in semiconductors, the wavelengths corresponding to the relevant frequencies are much larger than the dimensions of the devices, and the magnetic fields do not play a role. In these cases, the important equations are Poisson's equation (6.62) $\nabla \cdot \mathbf{E} = \rho/\varepsilon$, and the fact that the total current $\mathbf{j} = \mathbf{j}_{\text{cond}} + \partial \mathbf{D}/\partial t$ is conserved, given that its divergence is zero according to (6.64).

The charge density is given by $\rho = e(p - n + N_D^+ - N_A^-)$, where N_D^+ is the density of ionized donors and N_A^- is the density of ionized acceptors.

The equations given here form the basis for the description of classical semiconductor devices, i.e. virtually all electronic devices (e.g. transistors, photodiodes, photoconductors, etc.). Effects which play out over length scales that involve the wave nature of electrons are hidden within parameters in these equations, or in source terms like carrier generation rates by photon absorption, etc.

We now give an important application of these equations. In Section 6.3, we encountered the Debye wavelength (Eq. (6.24)), which represented the typical length scale over which free electrons can screen a weak perturbing potential either by electron accumulation or depletion. In many devices, we apply potentials which are strong enough ($\gg k_B T$) to deplete the electron gas completely in the vicinity of the potential (which then possesses a charge density resulting from the concentration of ionized donors left behind). Poisson's equation evaluated for such a case

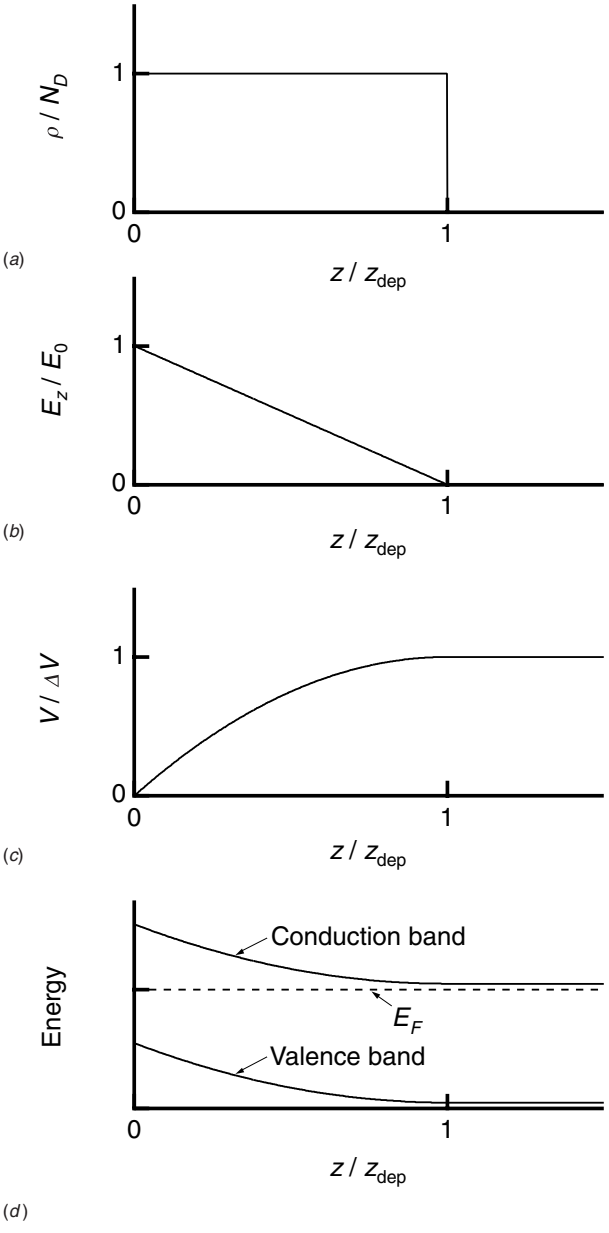


Fig. 6.12. (*a*) Charge distribution (ρ), (*b*) electric field (E_z), (*c*) electrostatic potential (V), and (*d*) band diagram across a depletion region. z_{dep} is given by Eq. (6.68).

leads to another screening length, referred to as the *depletion length.*

Let us imagine a doped semiconductor, with a donor concentration N_D, and further imagine that at $z = 0$, an electric field $-E_0$ pushes the electrons towards $z > 0$. In a zone $0 < z < z_{\text{dep}}$, there will be practically no electrons (see Fig. 6.12), and Poisson's equation may be written:

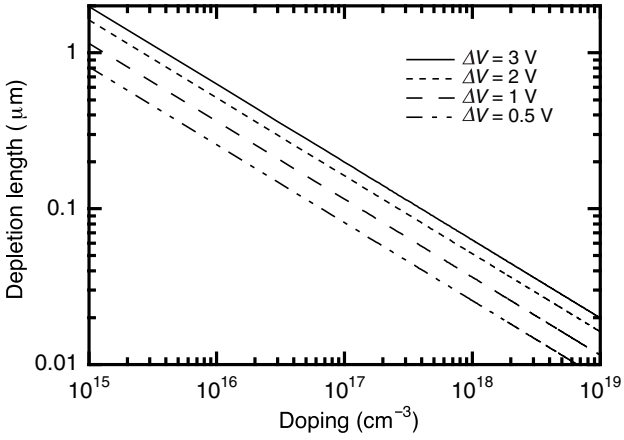


Fig. 6.13. GaAs ($\varepsilon_R = 12$) depletion length for various potential barrier heights ΔV.

$$\frac{\mathrm{d}E_z}{\mathrm{d}z} = \frac{eN_D^+}{\varepsilon}, 0 < z < z_{\mathrm{dep}} \tag{6.66}$$

which may be easily integrated to yield:

$$E_z = -E_0 + \frac{eN_D^+}{\varepsilon} z \tag{6.67}$$

The depletion zone extends all the way up to the point where $E_z = 0$:

$$z_{\mathrm{dep}} = \frac{\varepsilon E_0}{eN_D^+} \tag{6.68}$$

and the potential drop across this region is:

$$V(z_{\mathrm{dep}}) - V(0) = \frac{eN_D^+}{2\varepsilon} z_{\mathrm{dep}}^2 \tag{6.69}$$

so that:

$$z_{\mathrm{dep}} = \sqrt{\frac{2\varepsilon \Delta V}{eN_D^+}} \tag{6.70}$$

Depletion length

This length is important as it gives the scale over which a potential drop extends within a doped semiconductor. We see that it takes on the value of the Debye length if we replace $2\Delta V$ by $k_B T/e$; assuming $e\Delta V \gg k_B T$ implies that $z_{\mathrm{dep}} \gg \lambda_D$. In fact, the transition between the depletion region and the neutral semiconductor is not discontinuous but rather extends over an interval of the order of λ_D, around z_{dep}.

Figure 6.13 shows the thickness of the depletion region for different cases.

Example

Taking $\Delta V = 1\ V$ and $N_D = 10^{17}\,\mathrm{cm}^{-3}$ we have for GaAs:

$$z_{\mathrm{dep}} = \sqrt{\frac{2\varepsilon\Delta V}{eN_D}} = \sqrt{\frac{2 \times 12 \times 8.8 \times 10^{-12} \times 1}{1.6 \times 10^{-19} \times 10^{23}}} m = 0.11\,\mu\mathrm{m}$$

FURTHER READING

A. Kireev, *Physique des semiconducteurs*, MIR, Moscou (1975).
B. Sapoval and C. Hermann, *Physics of Semiconductors*, Springer, Berlin (1995).
K. Seeger, *Semiconductor Physics*, 3rd Edn, Springer, Berlin (1985).
S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edn, Wiley Interscience, New York (1981).

Complement to Chapter 6

6.A The Hall effect

We are interested in studying the influence of an applied magnetic field on a semiconductor layer in which a current flows. Within the context of the semiclassical description, one can show that the effect of the magnetic field **B** on an electron in a semiconductor is simply to add the Lorentz force to Eq. (6.2):

$$\hbar \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t} = q[\mathbf{F} + \mathbf{v}(\mathbf{k}) \times \mathbf{B}] \tag{6.A.1}$$

If, to simplify things, we restrict ourselves to considering electrons residing in those parts of the band structure where the effective mass is isotropic, $\mathbf{v} = \hbar\mathbf{k}/m^*$, the motion of the electrons is classical and remains similar to what would be obtained in a vacuum:

$$m^* \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = q(\mathbf{F} + \mathbf{v} \times \mathbf{B}) \tag{6.A.2}$$

Continuing with our simplification, we take **F** to lie along the Ox axis and **B** to be parallel to the Oz axis. We then model the effect of the scattering by introducing a frictional term $-m^*\mathbf{v}/\tau$ possessing a relaxation time constant. The equations of motion for the electron may then be written as:

$$\begin{aligned} m^* \frac{\mathrm{d}v_x}{\mathrm{d}t} &= q(F + v_y B) - \frac{m^* v_x}{\tau} \\ m^* \frac{\mathrm{d}v_y}{\mathrm{d}t} &= -qv_x B - \frac{m^* v_y}{\tau} \\ m^* \frac{\mathrm{d}v_z}{\mathrm{d}t} &= -\frac{m^* v_z}{\tau} \end{aligned} \tag{6.A.3}$$

and the stationary state solution is:

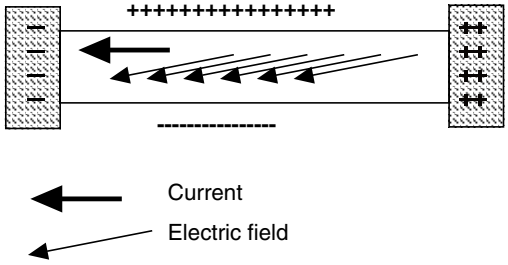


Fig. 6.A.1. Under the effect of a magnetic field (perpendicular to the page) the carriers are deflected from their otherwise rectilinear trajectories between the two contacts. As charges accumulate on both sides of the sample, an additional electric field component is set up in the sample running perpendicular to the current. The resulting field ultimately lies obliquely to the direction of the current flow between the two contacts.

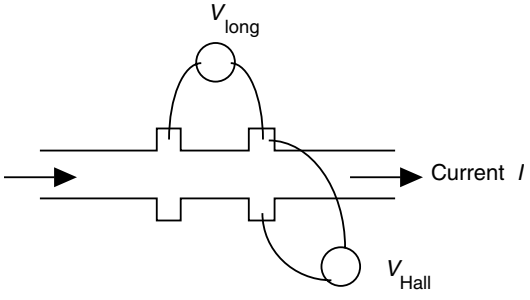


Fig. 6.A.2. Hall effect method for determining the carrier density and type. Measurements of the longitudinal and transverse voltages under the effect of a magnetic field are used to extract the Hall resistivity and angle.

$$\frac{v_x}{F} = \frac{q\tau}{m^*}\frac{1}{1+(\omega_c\tau)^2} = \frac{\mu}{1+(\omega_c\tau)^2}$$

$$\frac{v_y}{F} = -\frac{q\tau}{m^*}\frac{\omega_c\tau}{1+(\omega_c\tau)^2} = -\frac{\mu^2 B}{1+(\omega_c\tau)^2} \tag{6.A.4}$$

$$v_z = 0$$

with the mobility $\mu = q\tau/m^*$ (Eq. (6.20)) and the cyclotron frequency $\omega_c = qB/m^*$. We see that the magnetic field deflects the current from the direction of the electric field, see Fig. 6.A.1. This phenomenon is known as the *Hall effect*. By measuring the current $\mathbf{j} = nq\mathbf{v}$, we may deduce the mobility from $\mu = -j_y/j_x B$, and the carrier density from $n = j/\mu qF$ separately. It is important to note that the sign of mobility is negative for an electron gas and positive for a hole gas. Thus, by measuring the conductivity and the Hall angle between the direction of current flow and the applied electric field, we may obtain the carrier type, density, and mobility.

In practice, we normally fix the direction of the current by fabricating a sample

shaped into a 'Hall bar' (Fig. 6.A.2). The four contacts allow one to measure the longitudinal and transverse components of the field. Such measurements are commonly used in semiconductor characterization.

6.B Optical phonons and the Fröhlich interaction

We will now describe how phonons in a material influence the dynamic behaviour of carriers. Such interactions are extremely important whether in considering transport properties (e.g. electron mobility at room temperature is dominated by interactions with phonons) or relaxation phenomena involving carriers excited optically or by strong electric fields.

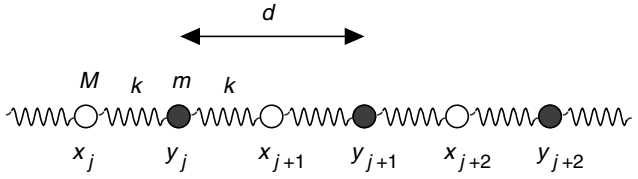


Fig. 6.B.1. Linear chain of atoms with masses M and m. The interatomic forces have been represented by springs so that the force between neighbouring atoms is given by $k(y_j - x_j)$, etc., and where x and y denote the displacement distances of the atoms from their equilibrium positions.

6.B.1 Phonons

Phonons are vibrational waves involving the atoms in a lattice. To begin, we study a simple model consisting of a linear chain of diatomic molecules. We will approximate the interatomic forces as linearly increasing as a function of the displacement distance from equilibrium, and being directed in such a manner as to return the atoms to their equilibrium interatomic positions at $d/2$ (see Fig. 6.B.1). A diatomic model of this sort is particularly well suited to treating compound semiconductors such as GaAs or InP.

The displacement of the jth atom relative to the equilibrium position is taken to be x_j for the heavy atoms (mass M) and y_j for the light atoms (mass m).

The classical Hamiltonian for this system is then:

$$H(x_j, y_j, \dot{x}_j, \dot{y}_j) = \sum_j \frac{1}{2} M\dot{x}_j^2 + \frac{1}{2} m\dot{y}_j^2 + \frac{1}{2} k(y_j - x_j)^2 + \frac{1}{2} k(x_{j+1} - y_j)^2 \tag{6.B.1}$$

with the *conjugate variables*:

$$p_{x_j} = \frac{\partial H}{\partial \dot{x}_j} = M\dot{x}_j, p_{y_j} = \frac{\partial H}{\partial \dot{y}_j} = m\dot{y}_j \tag{6.B.2}$$

The dynamic behaviour of the system is then described by the equations:

$$\begin{aligned}\dot{p}_{x_j} &= M\ddot{x}_j = -\frac{\partial H}{\partial x_j} = -k(x_j - y_j) - k(x_j - y_{j-1}) \\ \dot{p}_{y_j} &= m\ddot{y}_j = -\frac{\partial H}{\partial y_j} = -k(y_j - x_j) - k(y_j - x_{j+1})\end{aligned} \tag{6.B.3}$$

so that:

$$\begin{aligned}M\ddot{x}_j &= k(y_j + y_{j-1} - 2x_j) \\ m\ddot{y}_j &= k(x_{j+1} + x_j - 2y_j)\end{aligned} \tag{6.B.4}$$

These are Newton's equations for the displacement of each atom.

If we seek solutions in the form of waves having frequencies ω (≥ 0):

$$\begin{aligned}x_j &= X\exp(\mathrm{i}qjd - \mathrm{i}\omega t) \\ y_j &= Y\exp(\mathrm{i}qjd - \mathrm{i}\omega t)\end{aligned} \tag{6.B.5}$$

substituting (6.B.5) into (6.B.4) and setting $\omega_M^2 = 2k/M$ and $\omega_m^2 = 2k/m$ we obtain:

$$\begin{aligned}&(\omega_M^2 - \omega^2)X - \omega_M^2 \frac{1 + \mathrm{e}^{-\mathrm{i}qd}}{2} Y = 0 \\ &-\omega_m^2 \frac{\mathrm{e}^{\mathrm{i}qd} + 1}{2} X + (\omega_m^2 - \omega^2) Y = 0\end{aligned} \tag{6.B.6}$$

The condition for the existence of a non-trivial solution gives the mode frequencies:

$$\begin{aligned}&\omega^4 - (\omega_M^2 + \omega_m^2)\omega^2 + \omega_M^2\omega_m^2 \sin^2\left(\frac{qd}{2}\right) = 0 \\ &\omega^2 = \frac{\omega_M^2 + \omega_m^2}{2} \pm \sqrt{\left(\frac{\omega_M^2 + \omega_m^2}{2}\right)^2 - \omega_M^2\omega_m^2 \sin^2\left(\frac{qd}{2}\right)}\end{aligned} \tag{6.B.7}$$

Figure 6.B.2 shows the dispersion curves $\omega(q)$. We note the presence of two branches referred to as the *optical* and *acoustical branches.*

We easily obtain the limits at $q \to 0$:

$$\begin{aligned}\omega_{\mathrm{op}} &\simeq \sqrt{\omega_M^2 + \omega_m^2} = \sqrt{\frac{2k}{m_{\mathrm{red}}}}, \frac{1}{m_{\mathrm{red}}} = \frac{1}{M} + \frac{1}{m} \\ \omega_{\mathrm{ac}} &\simeq \frac{qd}{2} \frac{\omega_M\omega_m}{\sqrt{\omega_M^2 + \omega_m^2}} = \frac{qd}{2}\sqrt{\frac{2k}{M+m}} = v_s q\end{aligned} \tag{6.B.8}$$

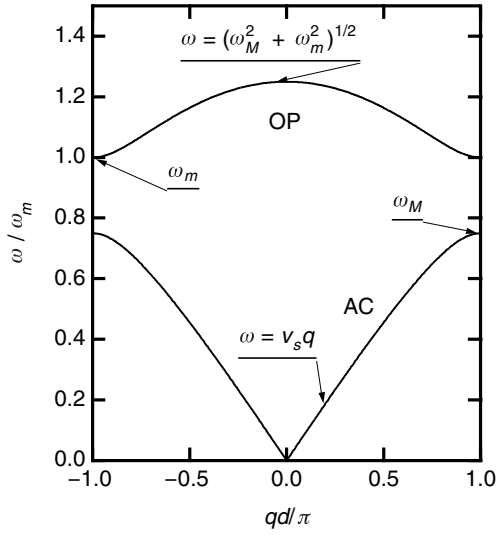


Fig. 6.B.2. Dispersion curves for the vibrational modes of a diatomic chain displaying the acoustical (AC) and optical (OP) branches, along with the characteristic limits.

and at $q = \pi/d$, we find $\omega_{\text{op}} = \omega_m$, and $\omega_{\text{ac}} = \omega_M$.

These solutions lie within the zone given by $-\pi/d < q \leq \pi/d$, because a solution for $q = q_0 + 2n\pi/d$ represents exactly the same atomic displacements as the solution corresponding to q_0. In fact:

$$x_j = X\exp(\mathrm{i}q_0 jd + \mathrm{i}2\pi nj - \mathrm{i}\omega t) = X\exp(\mathrm{i}q_0 jd - \mathrm{i}\omega t) \tag{6.B.9}$$

Furthermore, if the chain is not infinitely long, but has a length $L = Nd$ (see also Eq. (5.10)), application of the periodic boundary conditions $x_{N+j} = x_j$ and $y_{N+j} = y_j$ leads to discrete values for q:

$$qNd = 2\pi n \Rightarrow q_n = \frac{2\pi}{L}n \tag{6.B.10}$$

The *mode density* is then $L/2\pi$ for each branch and the total number of modes is $2N$, i.e. the number of degrees of freedom in a system possessing $2N$ atoms.

Equation (6.B.6) gives for each solution the ratio between the amplitudes $X(q)$ and $Y(q)$, and for the special cases close to the zone centre ($q \approx 0$), we have:

$$Y \simeq -\frac{\omega_m^2}{\omega_M^2}\left(1 + \mathrm{i}\frac{qd}{2}\right)X = -\frac{M}{m}\left(1 + \mathrm{i}\frac{qd}{2}\right)X \tag{6.B.11}$$

for optical phonons and:

$$Y \simeq \left(1 + \mathrm{i}\frac{qd}{2}\right)X \tag{6.B.12}$$

for acoustical phonons.

Optical phonons are therefore characterized by the opposing motion of the two types of atoms, which explains their high frequencies. Acoustical phonons, however, do not produce any displacement in the springs as $q \to 0$ and therefore do not possess elastic energy (explaining why the frequency goes to zero at the centre of the zone).

We will not study the extension of our linear chain model to a three-dimensional crystalline lattice. The ensuing formalism would require an impressive number of indices to keep track of the three-dimensional displacements of each of the atoms and the wavevectors also become three-dimensional vectors in the Brillouin zone. The essence of the model, however, remains essentially unchanged in this case. For a diatomic system like GaAs, the total number of modes becomes $6N$ (where $2N$ is the number of atoms in a volume V). The solutions are then distributed over three acoustical branches having $\omega(\mathbf{q}) \to 0$ for $\mathbf{q} \to 0$ and three optical branches for which $\omega(\mathbf{0}) \neq 0$.

The main difference between the one- and three-dimensional cases is the existence of modes for which the displacements are not parallel to the wavevector $\mathbf{q}$ – these are the *transverse modes*. For a two-dimensional system, Fig. 6.B.3 illustrates the different displacement possibilities for waves propagating towards the right in a symmetric case (in any given lattice, the displacements associated with a mode are generally not perpendicular or parallel to the wavevector).

We now return to the one-dimensional case, which up until now has received a purely classical treatment. To arrive at a more accurate representation of reality, this system must be reconsidered in quantum mechanical terms. We must therefore replace our expression for the Hamiltonian in (6.B.1) with a Hamiltonian operator obtained by replacing $p_{x_j} \to (\hbar/i)\partial/\partial x_j$, etc., with the commutators $[x_j, px'_j] = i\hbar\delta'_{jj}$, etc. It is, however, much more elegant to transform the Hamiltonian function before introducing the operators.

Let us introduce the normal co-ordinates:

$$X_q = \frac{1}{N}\sum_{j=1}^{N} x_j e^{-iqjd}$$
$$Y_q = \frac{1}{N}\sum_{j=1}^{N} y_j e^{-iqjd} \tag{6.B.13}$$

along with the inverse transformations:

$$x_j = \sum_q X_q e^{iqjd}$$
$$y_j = \sum_q Y_q e^{iqjd} \tag{6.B.14}$$

Utilizing the variables X_q and Y_q, the Hamiltonian function becomes:

$$H = N\sum_q \frac{1}{2}M\dot{X}_q\dot{X}_{-q} + \frac{1}{2}m\dot{Y}_q\dot{Y}_{-q} + k\left(X_qX_{-q} + Y_qY_{-q} - X_qY_{-q}\frac{1+\mathrm{e}^{\mathrm{i}qd}}{2} - Y_qX_{-q}\frac{\mathrm{e}^{-\mathrm{i}qd}+1}{2}\right) \tag{6.B.15}$$

The conjugate variables are then:

$$P_q^x = \frac{\partial H}{\partial \dot{X}_q} = MN\dot{X}_{-q},\ P_q^y = \frac{\partial H}{\partial \dot{Y}_q} = MN\dot{Y}_{-q} \tag{6.B.16}$$

and subject to the dynamic equations already put forward in (6.B.3) to (6.B.6):

$$\begin{aligned}\dot{P}_q^x &= NM\ddot{X}_{-q} = -\frac{\partial H}{\partial X_q} = -2kN\left(X_{-q} - Y_{-q}\frac{1+\mathrm{e}^{-\mathrm{i}qd}}{2}\right)\\ \dot{P}_q^y &= Nm\ddot{Y}_{-q} = -\frac{\partial H}{\partial Y_q} = -2kN\left(Y_{-q} - X_{-q}\frac{\mathrm{e}^{\mathrm{i}qd}+1}{2}\right)\end{aligned} \tag{6.B.17}$$

and the eigenmodes, which possess a time dependence given by $\mathrm{e}^{-\mathrm{i}\omega t}$, are given by (6.B.7). For each mode, the amplitudes are related by:

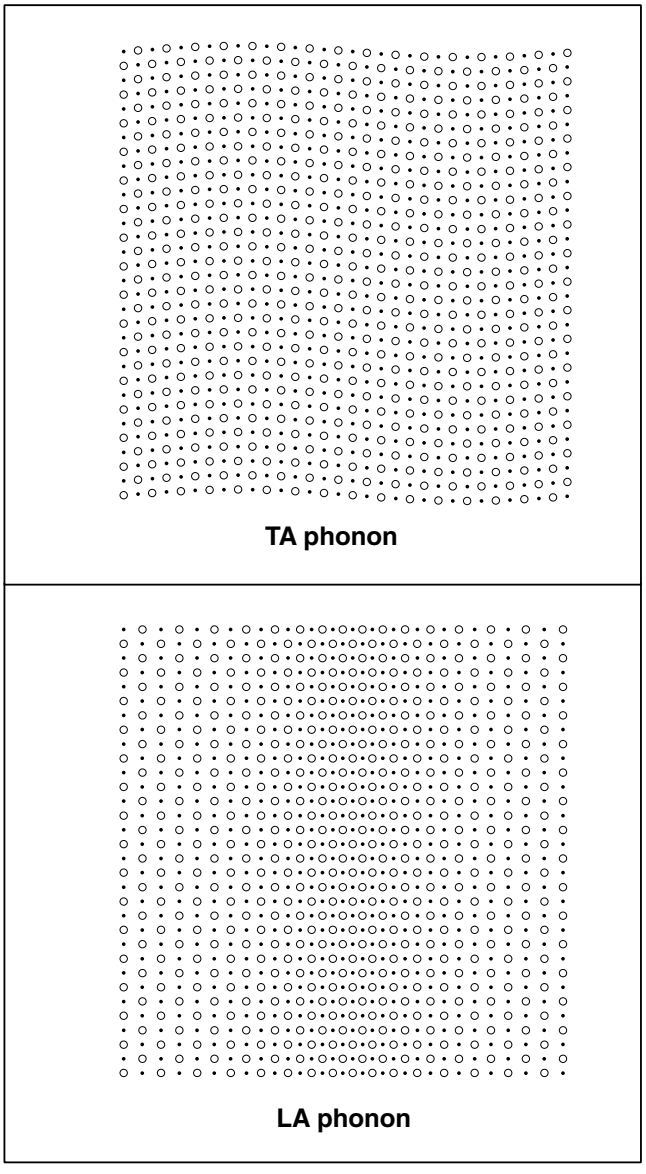
$$\begin{aligned}Y_q^{\pm} &= \left(1 - \frac{\omega_{q,\pm}^2}{\omega_M^2}\right)\frac{2}{1+\mathrm{e}^{-\mathrm{i}qd}}X_q^{\pm}\\ &= \frac{\omega_m^2}{\omega_M^2 - \omega_{q,\pm}^2}\frac{\mathrm{e}^{\mathrm{i}qd}+1}{2}X_q^{\pm}\end{aligned} \tag{6.B.18}$$

which allows us to eliminate Y_q from the Hamiltonian function, but requires us to sum over all the modes:

$$\begin{aligned}H &= N\sum_{q,\pm}\frac{1}{2}M\left(1 + \frac{\omega_M^2 - \omega_{q,\pm}^2}{\omega_m^2 - \omega_{q,\pm}^2}\right)(\dot{X}_q^{\pm}\dot{X}_{-q}^{\pm} + \omega_{q,\pm}^2X_q^{\pm}X_{-q}^{\pm})\\ &= N\sum_{q,\pm}\frac{1}{2}M_{q,\pm}(|\dot{X}_q^{\pm}|^2 + \omega_{q,\pm}^2|X_q^{\pm}|^2)\end{aligned} \tag{6.B.19}$$

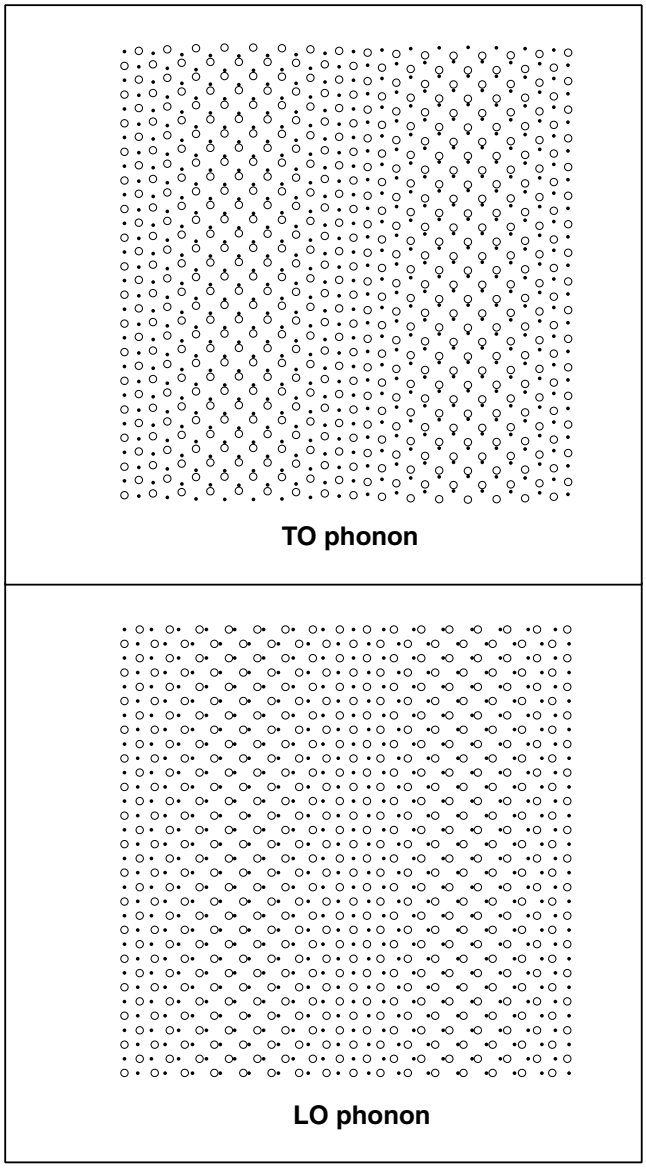
The introduction of the variables X_q lead us to a Hamiltonian function which is the sum of *independent* oscillator Hamiltonian functions possessing eigenfrequencies $\omega_{q,\pm}$. Along with the conjugate variables, we finally obtain:

$$\begin{aligned}P_q^{\pm} &= \frac{\partial H}{\partial \dot{X}_q^{\pm}} = NM_{q,\pm}\dot{X}_{-q}^{\pm}\\ H &= \sum_{q,\pm}\left(\frac{|P_q^{\pm}|^2}{2NM_{q,\pm}} + \frac{1}{2}NM_{q,\pm}\omega_{q,\pm}^2|X_q^{\pm}|^2\right)\end{aligned} \tag{6.B.20}$$



(*a*)

Fig. 6.B.3. (*a*) Atomic displacements in a two-dimensional lattice possessing two types of atoms. Transverse acoustical (TA) and longitudinal acoustical (LA) phonons are depicted in the top and bottom panels, respectively.



(*b*)

Fig. 6.B.3. (*b*) Atomic displacements in a two-dimensional lattice possessing two types of atoms. Transverse optical (TO) and longitudinal optical (LO) phonons are depicted in the top and bottom panels, respectively.

Starting from this Hamiltonian function, the transition to quantum mechanics is now direct: each classical harmonic oscillator becomes a quantum mechanical oscillator (see Complement 1.D) having eigenenergies:

$$E_{q,\pm} = \hbar\omega_{q,\pm}\left(n_{q,\pm} + \frac{1}{2}\right),\ n_{q,\pm} = 0, 1, 2, \ldots \tag{6.B.21}$$

where the 'number of phonons' $n_{q,\pm}$ in the mode is either zero or given by a positive integer.

We note the great similarity between the quantization of atomic motions in a lattice and that of the electromagnetic field. The term *phonon* is meant, therefore, to reflect this close analogy with the photon. In analogy with photons, we may speak of the corpuscular nature of phonons. As such, we may refer to emission/creation and absorption/annihilation of phonons, etc. Clearly, such a language is based upon the creation and annihilation operators of the harmonic oscillator (see Complement 1.D).

As the number of phonons that may exist in mode n is not limited, phonon populations are subject to the same occupation statistics as photons, making them *bosons*. At thermodynamic equilibrium (at temperature T), the average number of phonons in a mode is therefore given by:

$$\langle n_{q,\pm} \rangle = 1/[\exp(\hbar\omega_{q,\pm}/k_B T) - 1] \tag{6.B.22}$$

6.B.2 The Fröhlich interaction

Phonons represent a time-dependent perturbation to the periodic crystal potential that determines the electron states in the bands. This perturbation naturally leads to the possibility of scattering an electron from one state to another according to Fermi's golden rule. In a non-polar semiconductor (such as Si), an interaction potential which couples phonons to electrons results from the fact that the lattice distortions change the band structure proportionately to the relative displacement amplitude of the atoms. For example, local compression due to an acoustic wave increases the local bandgap and raises the energy of the states in the conduction band, see Fig. 6.B.4. Under the adiabatic approximation, an electron is therefore subjected to a potential which is proportional to the deformation amplitude; this potential can then induce transitions, i.e. electron–phonon collisions.

These processes (referred to as *deformation potential scattering*) naturally occur in polar semiconductors (e.g. GaAs) too, but in this case, there are additional interactions due to the fact that phonons create polarization and electrostatic fields that follow the phonon. For acoustical phonons we speak of the *piezoelectric effect*. In the case of optical phonons, the electrostatic field induces electron–phonon interaction, referred to as *Fröhlich interaction*. Fröhlich interaction has a

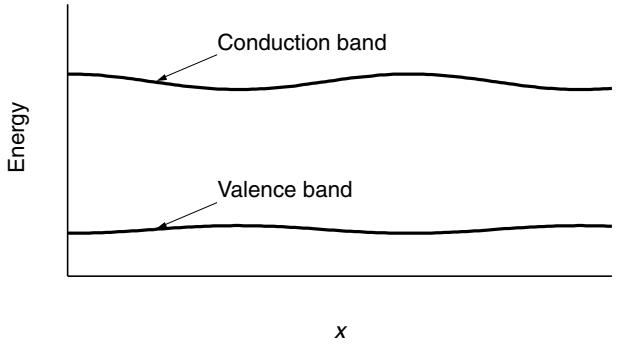


Fig. 6.B.4. A phonon induced variation in the bandgap (referred to as a *deformation potential*) is experienced by electrons as a potential scattering centre.

stronger effect than the deformation potential, and is discussed in this section as an illustration of electron–phonon interaction.

Since the crystal is polar, there is a transfer of charge e^* from the x atoms to the y atoms. The deformation produced by a phonon thereby induces a polarization:

$$\mathbf{P} = e^*(\mathbf{x} - \mathbf{y})\frac{N}{V} \tag{6.B.23}$$

with charge $\rho = -\nabla \cdot \mathbf{P}$. For an optical phonon with a wavevector $\mathbf{q}$, we have:

$$\rho(\mathbf{r}, t) = -e^*\frac{N}{V}\mathrm{i}\mathbf{q} \cdot (\mathbf{X} - \mathbf{Y})\exp(\mathrm{i}\mathbf{q} \cdot \mathbf{r} - \mathrm{i}\omega_{\mathbf{q}}t) \tag{6.B.24}$$

from which we see that only LO phonons can induce a charge. In this last case, Poisson's equation gives us the induced electrostatic potential:

$$V_{\mathrm{ind}}(\mathbf{r}, t) = \frac{e^*}{\varepsilon_\infty \varepsilon_0}\frac{N}{V}\frac{1}{\mathrm{i}q}(X - Y)\exp(\mathrm{i}\mathbf{q} \cdot \mathbf{r} - \mathrm{i}\omega_{\mathbf{q}}t) \tag{6.B.25}$$

This is the potential with which an electron can interact. The matrix element which describes this interaction between an initial electronic state $\mathbf{k}$ and a final state $\mathbf{k}'$ is:

$$\begin{aligned}
&\langle \mathbf{k}| - eV_{\mathbf{q}}(\mathbf{r})|\mathbf{k}'\rangle \\
&= \frac{-ee^*}{\varepsilon_\infty \varepsilon_0}\frac{N}{V}\frac{X - Y}{\mathrm{i}q}\frac{1}{V}\int u^*_{\mathbf{k}'}(\mathbf{r})u_{\mathbf{k}}(\mathbf{r})\mathrm{e}^{\mathrm{i}(\mathbf{q}+\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}\mathrm{d}\mathbf{r} \\
&= \frac{-ee^*}{\varepsilon_\infty \varepsilon_0}\frac{N}{V}\frac{X - Y}{\mathrm{i}q}\Delta_{\mathbf{q}+\mathbf{k}-\mathbf{k}'}\frac{1}{V_c}\int_{\mathrm{cell}} u^*_{\mathbf{k}'}(\mathbf{r})u_{\mathbf{k}}(\mathbf{r})\mathrm{d}\mathbf{r}
\end{aligned} \tag{6.B.26}$$

with the square modulus yielding:

$$|\langle \mathbf{k}| -eV_{\mathbf{q}}(\mathbf{r})|\mathbf{k}'\rangle|^2 = \frac{e^2 e^{*2}}{\varepsilon_\infty^2 \varepsilon_0^2} \frac{N^2}{V^2} \frac{|X-Y|^2}{q^2} \Delta_{\mathbf{q}+\mathbf{k}-\mathbf{k}'} I(\mathbf{k},\mathbf{k}') \tag{6.B.27}$$

Here, Δ is the Kronecker symbol and the overlap integral is the same as introduced in the earlier discussion of impurity scattering (Eq. (6.33)). Before we can make use of Fermi's golden rule, we finally need to obtain e^* and $X - Y$.

For e^*, we use the following argument: as TO phonons have a frequency $\omega_{\mathrm{TO}} = \sqrt{(2k/m_{\mathrm{red}})}$ in our model, the oscillation amplitude near ω_{TO} in response to an oscillating field with frequency ω is given by:

$$-\omega^2(X-Y) = -m_{\mathrm{red}}\omega_{\mathrm{TO}}^2(X-Y) + e^*E \tag{6.B.28}$$

yielding the polarization:

$$P = \frac{e^{*2}N}{m_{\mathrm{red}}V} \frac{E}{(\omega_{\mathrm{TO}}^2 - \omega^2)} \tag{6.B.29}$$

and the permittivity:

$$\varepsilon(\omega) = \varepsilon_0\varepsilon_\infty + \frac{e^{*2}N}{m_{\mathrm{red}}V} \frac{1}{(\omega_{\mathrm{TO}}^2 - \omega^2)} \tag{6.B.30}$$

In particular, we find a 'static' permittivity for $\omega \ll \omega_{\mathrm{TO}}$:

$$\varepsilon_0\varepsilon_{\mathrm{st}} = \varepsilon_0\varepsilon_\infty + \frac{e^{*2}N}{m_{\mathrm{red}}V\omega_{\mathrm{TO}}^2} \tag{6.B.31}$$

In the case of LO phonons, the field increases as a result of the induced polarization field $-e^*(X-Y)(N/V)/\varepsilon_0\varepsilon_\infty$ resulting in the additional restoring force:

$$-\omega^2(X-Y) = -m_{\mathrm{red}}\omega_{\mathrm{TO}}^2(X-Y) + \frac{e^{*2}N}{\varepsilon_0\varepsilon_\infty V}(X-Y) + e^*E \tag{6.B.32}$$

with the response:

$$\frac{P}{E} = \frac{e^{*2}N/m_{\mathrm{red}}V}{[\omega_{\mathrm{TO}}^2 + (e^{*2}N/\varepsilon_0\varepsilon_\infty m_{\mathrm{red}}V) - \omega^2]} = \frac{e^{*2}N/m_{\mathrm{red}}V}{(\omega_{\mathrm{LO}}^2 - \omega^2)} \tag{6.B.33}$$

and a frequency ω_{LO} greater than ω_{TO}. We may therefore determine the effective charge e^* by:

$$\frac{e^{*2}N}{m_{\mathrm{red}}V} = \omega_{\mathrm{TO}}^2\varepsilon_0(\varepsilon_{\mathrm{st}} - \varepsilon_\infty) = \varepsilon_0\varepsilon_\infty(\omega_{\mathrm{LO}}^2 - \omega_{\mathrm{TO}}^2) \tag{6.B.34}$$

and at the same time find the *Lyddane–Sachs–Teller relation*:

$$\frac{\omega_{\mathrm{LO}}^2}{\omega_{\mathrm{TO}}^2} = \frac{\varepsilon_{\mathrm{st}}}{\varepsilon_\infty} \tag{6.B.35}$$

so that:

$$\frac{e^{*2}N}{m_{\mathrm{red}}V} = \omega_{\mathrm{LO}}^2\varepsilon_0\left(1 - \frac{\varepsilon_\infty}{\varepsilon_{\mathrm{st}}}\right) \tag{6.B.36}$$

Finally, the mode amplitude can be determined through the correspondence principle by making the classical energy in the mode equivalent to the number of phonons in the mode (Eqs. (6.B.20) and (6.B.21)):

$$E_{\mathrm{tot}} = \frac{1}{2}m_{\mathrm{red}}\omega_{\mathrm{LO}}^2|X - Y|^2N = n_q\hbar\omega_{\mathrm{LO}} \tag{6.B.37}$$

The near-final result for the transition rate thus becomes:

$$\begin{aligned} S_{\mathrm{abs}}(\mathbf{k}\to\mathbf{k}') &= \frac{2\pi}{\hbar}|\langle\mathbf{k}| - eV_q|\mathbf{k}'\rangle|^2\delta[E(\mathbf{k}') - E(\mathbf{k}) - \hbar\omega_q] \\ &= \frac{\pi}{\hbar}\frac{e^2\hbar\omega_{\mathrm{LO}}}{V|\mathbf{k}' - \mathbf{k}|^2}n_q\left(\frac{1}{\varepsilon_0\varepsilon_\infty} - \frac{1}{\varepsilon_0\varepsilon_{\mathrm{st}}}\right)I(\mathbf{k},\mathbf{k}')\delta[E(\mathbf{k}') - E(\mathbf{k}) - \hbar\omega_q] \end{aligned} \tag{6.B.38}$$

for a collision in which a LO phonon is absorbed and its energy is transferred to an electron. We have cheated somewhat be applying Eq. (6.B.37). In reality, in a correct quantum mechanical treatment of this problem, we would have to write the X–Y operator in terms of creation and annihilation operators corresponding to the Hamiltonian in (6.B.20). Afterwards, the procedure becomes completely analogous to the photon–electron interaction treated in Section 3.5.1. It is therefore not surprising to find that expression (6.B.38) remains valid for phonon absorption. In the case of phonon emission, we must replace n_q by $n_q + 1$, corresponding to stimulated emission plus the spontaneous emission of a phonon:

$$\begin{aligned} S_{\mathrm{em}}(\mathbf{k}\to\mathbf{k}') &= \frac{\pi}{\hbar}\frac{e^2\hbar\omega_{\mathrm{LO}}}{V|\mathbf{k}' - \mathbf{k}|^2}(n_q + 1) \\ &\quad\times\left(\frac{1}{\varepsilon_0\varepsilon_\infty} - \frac{1}{\varepsilon_0\varepsilon_{\mathrm{st}}}\right)I(\mathbf{k},\mathbf{k}')\delta[E(\mathbf{k}') - E(\mathbf{k}) + \hbar\omega_q] \end{aligned} \tag{6.B.39}$$

It is interesting to note that the inclusion of stimulated emission ensures a *detailed balance* at thermodynamic equilibrium:

$$S_{\mathrm{em}}(\mathbf{k}\to\mathbf{k}')f_{\mathrm{eq}}(\mathbf{k})[1 - f_{\mathrm{eq}}(\mathbf{k}')] = S_{\mathrm{abs}}(\mathbf{k}'\to\mathbf{k})f_{\mathrm{eq}}(\mathbf{k}')[1 - f_{\mathrm{eq}}(\mathbf{k})] \tag{6.B.40}$$

when n_q is given by the Bose–Einstein distribution (6.B.22) and f_{eq} is determined by the Fermi–Dirac distribution (5.38). The net flux of particles between any two states is therefore null in that case.

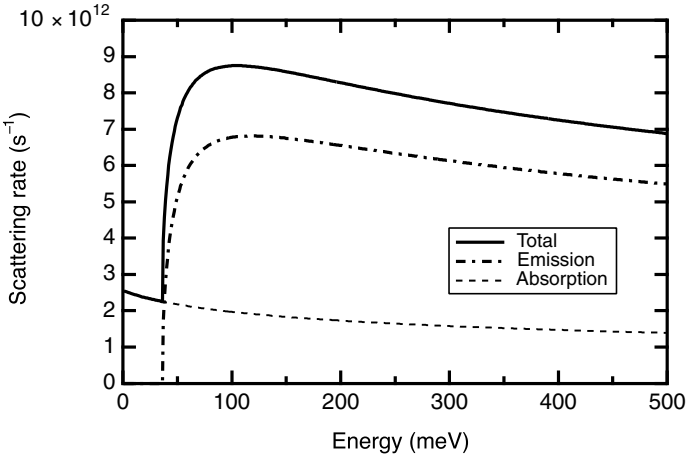


Fig. 6.B.5. Scattering rate due to optical phonons for electrons in GaAs at $T = 300\,\mathrm{K}$. The optical phonon energy is 36 meV, $\varepsilon_{st} = 12.85$, $\varepsilon_{\infty} = 10.88$.

Starting from Eqs. (6.B.38) and (6.B.39) we can calculate the total scattering rate due to optical phonons by Fröhlich interaction:

$$\lambda(k) = \sum_{\mathbf{k}'} S_{\mathrm{abs}}(\mathbf{k} \to \mathbf{k}') + \sum_{\mathbf{k}'} S_{\mathrm{em}}(\mathbf{k} \to \mathbf{k}') \tag{6.B.41}$$

Making the approximation that $I(\mathbf{k}, \mathbf{k}') = 1$, we then have:

$$\begin{aligned} &\sum_{\mathbf{k}'} \frac{1}{|\mathbf{k}' - \mathbf{k}|^2} \delta\left(\frac{\hbar^2 k'^2}{2m} - \frac{\hbar^2 k^2}{2m} \mp \hbar\omega_{\mathrm{op}}\right) \\ &= 2\pi \int_0^{\infty} \frac{V}{(2\pi)^3} k'^2 \mathrm{d}k' \int_{-1}^{1} \mathrm{d}\mu \frac{\delta[(\hbar^2 k'^2/2m) - (\hbar^2 k^2/2m) \mp \hbar\omega_{\mathrm{op}}]}{k^2 + k'^2 - 2kk'\mu} \\ &= \frac{Vm}{4\pi^2 \hbar^2 k} \ln\left|\frac{k + k'}{k - k'}\right| \end{aligned} \tag{6.B.42}$$

with $k'^2 = k^2 \pm 2m\hbar\omega_{\mathrm{op}}/\hbar^2$, which finally yields:

$$\begin{aligned} \lambda[E(k)] = &\frac{e^2 \hbar\omega_{\mathrm{op}} m}{4\pi\varepsilon_0 \varepsilon_{st} \hbar^2 \sqrt{2mE}} \left(\frac{\varepsilon_{st}}{\varepsilon_{\infty}} - 1\right) \\ &\times \left[n_q \ln\left|\frac{\sqrt{E + \hbar\omega_{\mathrm{op}}} + \sqrt{E}}{\sqrt{E + \hbar\omega_{\mathrm{op}}} - \sqrt{E}}\right| + (n_q + 1) \ln\left|\frac{\sqrt{E} + \sqrt{E - \hbar\omega_{\mathrm{op}}}}{\sqrt{E} - \sqrt{E - \hbar\omega_{\mathrm{op}}}}\right| \right] \end{aligned} \tag{6.B.43}$$

Figure (6.B.5) shows λ as a function of E for the case of GaAs. Phonon emission naturally takes place only above a certain energy threshold such that $E > \hbar\omega$, due

to the fact that the electron following phonon emission must possess strictly positive residual energy. We therefore see that electron–phonon interaction is extremely efficient in dissipating excess electron energy (over time scales typically of the order of 100 fs). This interaction therefore plays a fundamental role in optical relaxation mechanisms.

It is also important to note that for scattering (inelastic and anisotropic) occurring via Fröhlich interaction, we cannot in general define a relaxation time. Mobility calculations therefore require a more complicated solution than (6.22) to Boltzmann's equation.

FURTHER READING

B. K. Ridley, *Quantum Processes in Semiconductors*, Clarendon Press, Oxford (1988).

6.C Avalanche breakdown

If the electric field applied to a semiconductor significantly exceeds the saturation velocity field, certain electrons will be able to acquire an additional amount of energy above the conduction band minimum, greater than E_g, the gap energy. In such circumstances, impact ionization processes become possible. The electron interacts with all other electrons in the valence band through electron–electron interaction, and can excite an electron across the forbidden gap, corresponding to generation of an electron–hole pair (a generation process). In this process, energy and pseudo-momentum are conserved. We have, prior to collision, an electron in a state $\mathbf{k}_i$, with energy $E_c(\mathbf{k}_i)$, and an electron in a state $-\mathbf{k}_h$, with energy $E_v(\mathbf{k}_h)$. After collision, there are two electrons in the conduction band in states $\mathbf{k}_1$ and $\mathbf{k}_2$, and the conservation laws may be written (see the inverse process in Fig. 6.D.1):

$$\begin{aligned} &\mathbf{k}_i - \mathbf{k}_h = \mathbf{k}_1 + \mathbf{k}_2 \\ &E_c(\mathbf{k}_i) + E_v(\mathbf{k}_h) = E_c(\mathbf{k}_1) + E_c(\mathbf{k}_2) \end{aligned} \tag{6.C.1}$$

To get an idea of the energy threshold for such a process, we will employ the parabolic band approximation, $E_c(\mathbf{k}_i) = E_g + \hbar^2 k_i^2/2m_e$, and $\mathrm{E}_v(\mathbf{k}_i) = -\hbar^2 k_h^2/2m_h$, along with the effective electron (m_e) and hole (m_h) masses:

$$\begin{aligned} &\mathbf{k}_i = \mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_h \\ &\frac{\hbar^2 k_i^2}{2m_e} = E_g + \frac{\hbar^2 k_1^2}{2m_e} + \frac{\hbar^2 k_2^2}{2m_e} + \frac{\hbar^2 k_h^2}{2m_h} \end{aligned} \tag{6.C.2}$$

We can convince ourselves that the initial energy minimum is obtained when all

the momenta are oriented in the same direction. The minimum is found more elegantly (using Lagrange's method) by minimizing the function:

$$F = E_g + \frac{\hbar^2 k_1^2}{2m_e} + \frac{\hbar^2 k_2^2}{2m_e} + \frac{\hbar^2 k_h^2}{2m_h} + \lambda(k_1 + k_2 + k_h) \tag{6.C.3}$$

where λ is the Lagrange multiplier. Taking the derivative of F with respect to k_1, k_2, and k_h, we find:

$$-\frac{\lambda}{\hbar^2} = \frac{k_1}{m_e} = \frac{k_2}{m_e} = \frac{k_h}{m_h} \tag{6.C.4}$$

from which we obtain a minimum initial kinetic energy of:

$$\min\left(\frac{\hbar^2 k_i^2}{2m_e}\right) = \frac{1 + 2(m_e/m_h)}{1 + (m_e/m_h)} E_g \tag{6.C.5}$$

This leads to impact emission energy thresholds slightly above E_g in the case of heavy holes, but of the order of $3/2E_g$ if the two masses m_e and m_h are close to one another. The threshold is thus lower for the generation of heavy holes.

By symmetry, the analogous process in which the initial particle is a hole, gives the same results as obtained previously, with the roles of m_e and m_h being interchanged. The threshold is generally greater in this case. We must realize, however, that the parabolic band approximation is fairly poor over these energy ranges.

Once a primary electron has generated an electron–hole pair, it will once again be accelerated by the electric field enabling it to participate in subsequent pair generation events. Similarly, after being sufficiently accelerated by the electric field, each one of these secondary particles will in turn be able to play the role of primary particles and participate in pair production, with the result being that the overall free carrier population in the structure increases in a geometric fashion. This situation in a semiconductor is referred to as *avalanche*, with the resulting rise in free carrier concentration producing an increase in conductivity. In practice, we define the ionization coefficients α_n and α_p (which we shall suppose to be equal in what follows) as being the number of secondary pairs produced by a primary electron or hole per centimetre travelled in the medium in the direction of the applied field. These coefficients clearly depend on the local field strength F. While travelling across a region subject to a strong field between $0 < x < w$, a primary electron will therefore generate:

$$N_s = \int_0^w \alpha_n(F)\mathrm{d}x \tag{6.C.6}$$

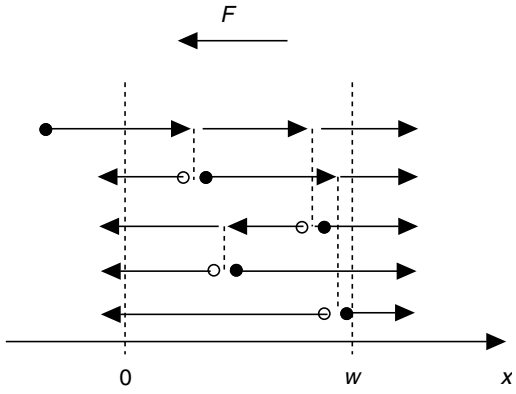


Fig. 6.C.1. Illustration of carrier multiplication by impact ionization. Under the influence of a strong applied field F, a single primary electron enters at $x = 0$ and M_n electrons emerge at $x = w$.

secondary pairs. The secondary and tertiary particles, etc., will generate a total number of electrons in the region given by:

$$N_w = 1 + N_s + N_s^2 + N_s^3 + \cdots = \frac{1}{1 - N_s} \tag{6.C.7}$$

as long as $N_s < 1$. If an electron enters at $x = 0$, a number N_w of electrons will exit at w. We define the *multiplication factor* M_n for electrons as being the ratio of the exit current to the entrance current in the high-field region. In an analogous fashion, we may define M_p for holes as $j_p(0) = M_p j_p(w)$ where holes enter at w and exit at 0. Figure 6.C.1 depicts the avalanche phenomenon.

The simplest model for obtaining ionization coefficients is due to Shockley and goes by the charming name of the *lucky electron model*. Since only those electrons possessing energies in excess of the energy threshold given by (6.C.5) can provoke an avalanche, there are few electrons which will be able to do so, as most will give up their energy in collisions before being accelerated to the threshold energy. We will label as λ_{op} the electron mean free path prior to collision with an optical phonon (and for simplicity's sake assumed to be independent of energy) and λ_i as the mean free path prior to impact with another electron (assuming the energy threshold has been reached). Last, we will assume that each collision causes the electron velocity to drop to zero. To attain the threshold energy E_i, an electron must travel a minimum distance x_i given by $E_i = eFx_i$. The probability that the electron can do so without being scattered by phonons is then:

$$P(E_i) = \exp\left(-\frac{x_i}{\lambda_{\text{op}}}\right) = \exp\left(-\frac{E_i}{eF\lambda_{\text{op}}}\right) \tag{6.C.8}$$

Once the kinetic energy of the electron exceeds the ionization threshold, see Fig.

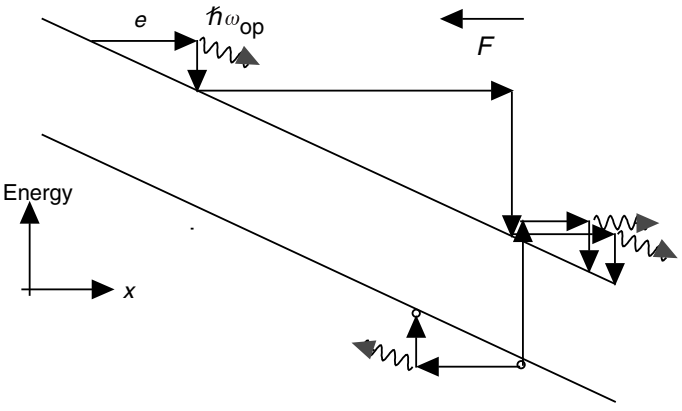


Fig. 6.C.2. The Shockley model for obtaining the ionization coefficient α_n: a primary electron gains sufficient energy to create a secondary electron–hole pair.

6.C.2, the mean free path λ of the electron is given by the composite of both collision mechanisms as $1/\lambda = 1/\lambda_i + 1/\lambda_{op}$. The probability that the first collision event experienced by the electron is an impact ionization event (as opposed to a collision with a lattice phonon) is therefore λ/λ_i. If, however, the first process encountered corresponds to a collision with a phonon, the electron's velocity drops to zero and it must begin to regain its kinetic energy from scratch. From such a 'cold start', the ionization probability is then:

$$P_i = \frac{\lambda}{\lambda_i} \exp\left(-\frac{E_i}{eF\lambda_{op}}\right) \tag{6.C.9}$$

If we define α_{op} as being the number of phonon collisions expected in a travel distance of 1 cm, the balance between the energy supplied by the electric field, and the energy dissipated by ionization and phonon emission processes may be written:

$$eF = \alpha_n E_i + \alpha_{op} E_{op} \tag{6.C.10}$$

where α_n is the ionization coefficient we seek. The number of times an electron must start accelerating from rest (following a collision) over this distance is then $\alpha_n + \alpha_{op}$, which gives for the number of collisions per centimetre:

$$\alpha_n = (\alpha_n + \alpha_{op}) P_i \tag{6.C.11}$$

and by eliminating α_{op} between (6.C.10) and (6.C.11):

$$\alpha_n = \frac{eFP_i}{E_{op} + P_i(E_i - E_{op})} \simeq \frac{eFP_i}{E_{op} + P_i E_i} \tag{6.C.12}$$

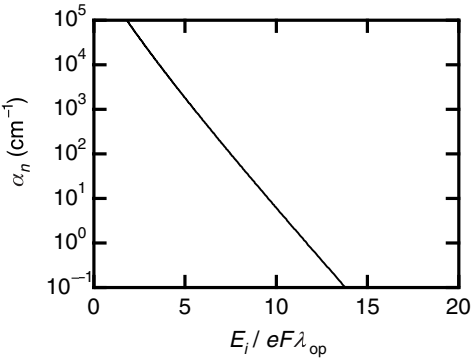


Fig. 6.C.3. Dependence of the ionization coefficient α_n on the applied field strength according to Eq. (6.C.13) for silicon. $E_i = 1.8\,\text{eV}$, $\lambda_{op} = 10\,\text{nm}$, $\lambda_i = 200\,\text{nm}$.

where we have taken advantage of the fact that $E_{op} \ll E_i$ ($\sim$50 meV and $\sim$1 eV, respectively). Reinsertion of (6.C.9) then gives the final result, illustrated in Fig. 6.C.3:

$$\alpha_n = \frac{eF \exp\left[-(E_i/eF\lambda_{op})\right]}{(\lambda_i/\lambda)E_{op} + E_i \exp\left[-(E_i/eF\lambda_{op})\right]} \tag{6.C.13}$$

With the exception of situations involving very high field strengths F, we may neglect the second term in the denominator to write:

$$\alpha_n = \frac{\lambda}{\lambda_1 E_{op}} eF \exp\left(-\frac{E_i}{eF\lambda_{op}}\right) \tag{6.C.14}$$

Impact ionization coefficient

In spite of the extreme simplicity of this model, expression (6.C.14) successfully reproduces the observed experimental dependence of α_n on the applied field strength.

FURTHER READING

J. S. Blakemore, *Semiconductor Statistics*, Dover, New York (1987).

6.D Auger recombination

The inverse process to impact ionization is Auger recombination. In this case, an electron and a hole recombine, and the liberated energy is transferred to another electron or hole. Figure 6.D.1 shows both these processes schematically, assuming

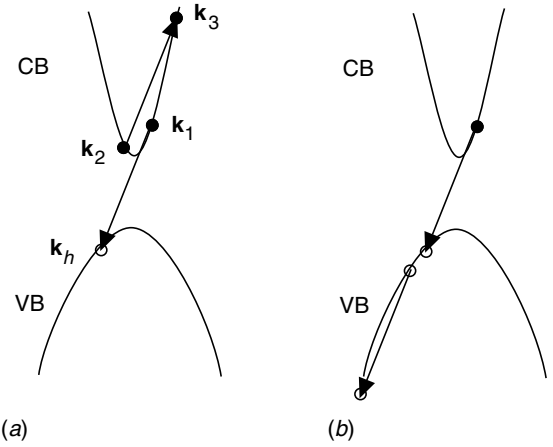


Fig. 6.D.1. Two possible Auger recombination processes: (*a*) an electron with momentum $\mathbf{k}_1$ recombines with a hole $\mathbf{k}_h$ and another electron at $\mathbf{k}_2$ receives the liberated energy and is promoted to $\mathbf{k}_3$ (CHCC); (*b*) a similar process involving two holes and a single electron at the outset (CHHH).

a highly simplified band structure. The effect of a slightly more realistic band structure including heavy and light holes would be to increase the number of possible Auger processes by allowing the second hole to be either light or heavy, and by allowing hole transitions in the same band or between different valence bands.

Figure 6.D.1, equally, shows how the total energy and momentum conservation laws restrict the combinations of possible initial and final states. For example, the recombination between an electron at the bottom of the conduction band and a hole at the top of the valence band is not an allowed Auger process as the secondary particle cannot make a vertical transition.

Qualitatively, it is also evident that since the Auger effect involves three particles, it must be more prevalent under increased electron and hole concentrations. Typically, this is the case in *small gap semiconductors* or in *systems far from thermodynamic equilibrium* (e.g. under high levels of optical pumping or current injection as in *semiconductor lasers*).

The only model which offers any hope of obtaining an analytic description of Auger recombination (and thus some physical insight) is that involving two parabolic bands possessing electrons and holes with respective effective masses m_c and m_v. We will suppose that the electrons can attain equilibrium with one another and that the holes can do the same over a period of time less than the characteristic time scale of the recombination. Also, we will assume that the distribution of particles remains non-degenerate. We therefore have an electron density given by:

$$n = N_c \exp\left(-\frac{E_c - \mu_n}{k_B T}\right) \tag{6.D.1}$$

and a hole density of:

$$p = N_v \exp\left(-\frac{\mu_p - E_v}{k_B T}\right) \tag{6.D.2}$$

where $\mu_n(\mu_p)$ is the electron (hole) quasi-Fermi level. In fact, as we saw in Section 5.6.4, these equations define μ_n and μ_p, and at equilibrium $\mu_n = \mu_p = E_F$. The quasi-equilibrium hypothesis for electrons (holes) implies that the occupation probability of an electronic state $\mathbf{k}$ with energy ε_c:

$$\varepsilon_c(k) = E_c + \frac{\hbar^2 k^2}{2m_c}$$

is given by (bands are parabolic and energy therefore only depends on the wavevector norm k):

$$f_c(\mathbf{k}) = \exp\left[-\frac{\varepsilon_c(k) - \mu_n}{k_B T}\right] \tag{6.D.3}$$

and for a hole:

$$f_v(\mathbf{k}) = \exp\left[-\frac{\mu_p - \varepsilon_v(k)}{k_B T}\right] \tag{6.D.4}$$

with:

$$\varepsilon_v(k) = E_v - \frac{\hbar^2 k^2}{2m_v}$$

The recombination rate, assuming the contribution of a process (see Fig. 6.D.1*a*), with initial states comprising two electrons at $\mathbf{k}_1$ and $\mathbf{k}_2$ and a hole at $\mathbf{k}_h$, and a final state corresponding to a single electron at $\mathbf{k}_3$, will be weighted by the occupation probabilities of the states $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_h$ (and where $\mathbf{k}_3$ may be assumed to be empty with probability 1):

$$\begin{aligned} P(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) &= \exp\left[-\frac{\varepsilon_c(k_1) - \mu_n + \varepsilon_c(k_2) - \mu_n + \mu_p - \varepsilon_v(k_h)}{k_B T}\right] \\ &= \frac{n^2 p}{N_c^2 N_v} \exp\left\{-\frac{\hbar^2[k_1^2 + k_2^2 + (m_c/m_v)k_h^2]}{2m_c k_B T}\right\} \end{aligned} \tag{6.D.5}$$

But again, this process must be permitted by conservation of momentum, i.e:

$$\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_h = \mathbf{k}_3 \tag{6.D.6}$$

and of total energy:

$$\varepsilon_c(k_1) + \varepsilon_c(k_2) - \varepsilon_v(k_h) = \varepsilon_c(k_3) \tag{6.D.7a}$$

so that:

$$\frac{\hbar^2}{2m_c}[k_1 + k_2 + (m_c/m_v)k_h^2 - k_3^2] + E_g = 0 \tag{6.D.7b}$$

Thus, if the interaction depends only slightly upon $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_h$, *the most probable process is obtained by maximizing* $P(k_1, k_2, k_h)$ under the constraints imposed by the conservation laws.

In order to find this maximum value, we will proceed as in Complement 6.C and use the method of Lagrange multipliers to maximize the function:

$$\begin{aligned} F = P(k_1, k_2, k_h) - \gamma \left\{ \frac{\hbar^2}{2m_c}[k_1 + k_2 + (m_c/m_v)k_h^2 - k_3^2] + E_g \right\} \\ - \lambda(k_1 + k_2 - k_h - k_3) \end{aligned} \tag{6.D.8}$$

so that:

$$\begin{aligned} \frac{\partial F}{\partial k_1} &= -\frac{\hbar^2 k_1}{2m_c k_B T}(P - \gamma k_B T) - \lambda = 0 \\ \frac{\partial F}{\partial k_2} &= -\frac{\hbar^2 k_2}{2m_c k_B T}(P - \gamma k_B T) - \lambda = 0 \\ \frac{\partial F}{\partial k_h} &= -\frac{\hbar^2 k_h}{2m_c k_B T}\frac{m_c}{m_v}(P - \gamma k_B T) + \lambda = 0 \end{aligned} \tag{6.D.9}$$

with the optimum value occurring for:

$$k_2 = k_1, k_h = -\frac{m_v}{m_c}k_1 \tag{6.D.10}$$

Momentum conservation gives:

$$k_3 = 2k_1 - k_h = \left(2 + \frac{m_v}{m_c}\right)k_1 \tag{6.D.11}$$

Finally, having found k_2, k_h, and k_3 as a function of k_1, this last quantity is obtained by energy conservation, yielding the result:

$$\frac{\hbar^2 k_1^2}{2m_c} = E_g \frac{\mu^2}{(1 + 2\mu)(1 + \mu)}, \ \mu = \frac{m_c}{m_v} \tag{6.D.12}$$

The maximum probability is therefore:

$$P = \frac{n^2 p}{N_c^2 N_v} \exp\left(-\frac{\mu}{1 + \mu}\frac{E_g}{k_B T}\right) \tag{6.D.13a}$$

or:

$$P = \frac{n}{N_c}\exp\left(-\frac{\mu_n - \mu_p}{k_B T}\right)\exp\left(-\frac{1+2\mu}{1+\mu}\frac{E_g}{k_B T}\right) \tag{6.D.13b}$$

Auger recombination probability (CHCC)

which gives at thermodynamic equilibrium a maximum probability:

$$P_0 = \frac{n_0}{N_c}\exp\left(-\frac{1+2\mu}{1+\mu}\frac{E_g}{k_B T}\right) \tag{6.D.14}$$

The 'mirror' process in which two holes and one electron are involved at the onset may be obtained by interchanging $n \leftrightarrow p, \mu \leftrightarrow 1/\mu, N_c \leftrightarrow N_v$, giving:

$$P^{\mathrm{CHH}} = \frac{p}{N_v}\exp\left(-\frac{\mu_n - \mu_p}{k_B T}\right)\exp\left(-\frac{2+\mu}{1+\mu}\frac{E_g}{k_B T}\right) \tag{6.D.15}$$

Auger recombination probability (CHHH)

and

$$P_0^{\mathrm{CHH}} = \frac{p_0}{N_v}\exp\left(-\frac{2+\mu}{1+\mu}\frac{E_g}{k_B T}\right) \tag{6.D.16}$$

In most semiconductors $\mu = m_c/m_v < 1$, and comparison of the corresponding probabilities indicates that, for comparable electron and hole densities (n and p), the first process will dominate.

The maximum probabilities (6.D.13) to (6.D.16) describe the essentials of the dependence of Auger recombination on temperature and carrier density. Near equilibrium, the probabilities (6.D.14) and (6.D.16) show that the effect is considerable in small gap semiconductors and at elevated temperatures. At a fixed temperature and gap size, the recombination rates R^{CHCC} depend, as we might expect, on the carrier densities n and p as:

$$R^{\mathrm{CHCC}} = C_{\mathrm{CHCC}} n^2 p,\ R^{\mathrm{CHHH}} = C_{\mathrm{CHHH}} n p^2 \tag{6.D.17}$$

where C_{CHHH} is a coefficient tied to the probability P_0 given by (6.D.14). At equilibrium, we find $R_0^{\mathrm{CHCC}} = C_{\mathrm{CHCC}} n_0 n_i^2$, i.e:

$$R^{\mathrm{CHCC}} = R_0^{\mathrm{CHCC}}\frac{n^2 p}{n_0 n_i^2} \tag{6.D.18}$$

For the inverse process, the creation of an electron–hole pair by an electron with sufficient energy is found to be (following the same argument) proportional to n:

$$G^C = G_0^C\frac{n}{n_0} \tag{6.D.19}$$

and, as $G_0^C = R_0^{\mathrm{CHCC}}$, we may write the net recombination as:

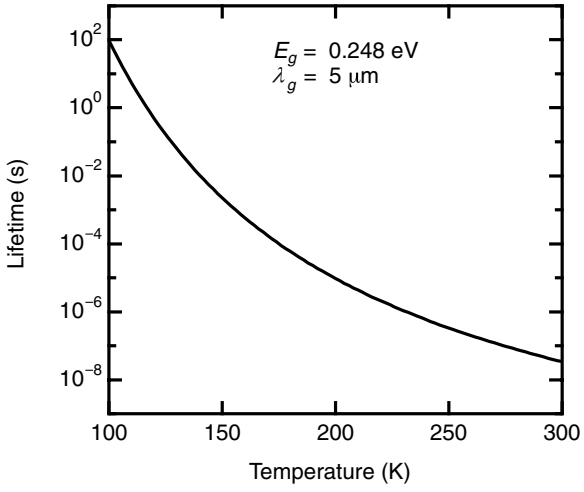


Fig. 6.D.2. Calculated lifetime for an electron in an intrinsic semiconductor with a small gap E_g. The parameters used are those for $InAs_{(1-x)}Sb_x$ with $x = 15\%$.

$$-\left.\frac{dn}{dt}\right|_{GR} = R^{CHCC} - G^C = G_0^C \frac{n}{n_0}\left(\frac{np - n_i^2}{n_i^2}\right) \tag{6.D.20}$$

The evaluation of the prefactors C_{CHCC} and C_{CHHH} is complicated and not particularly informative given that actual band structures are only poorly represented by the two-band parabolic model. We shall, therefore, not pursue the determination of these values here. We present, however, an equation from B. L. Gelmont, *Soviet Physics JETP* **48**, 268 (1978), for the lifetime of an electron $\tau = n_0/R^{CHCC}$ at thermodynamic equilibrium:

$$\tau = \frac{1}{3}\left(\frac{\pi}{2}\right)^{1/2} \frac{(4\pi\varepsilon)^2\hbar^3}{e^4 m_c}\left(\frac{E_g}{k_B T}\right)^{5/2} \exp\left[\frac{E_g}{k_B T}(1 + 2\mu)\right] \tag{6.D.21}$$

Figure 6.D.2 shows that the lifetime of an electron in an intrinsic semiconductor with a small gap depends greatly upon temperature due to the probability P_0 in Eq. (6.D.14).

Away from equilibrium, the limited development in (6.D.20) shows immediately that the rate τ^{-1} associated with the Auger effect possesses a quadratic dependence on the carrier density, i.e. $\tau^{-1} = C_{Auger} n^2$. The proportionality coefficient corresponds to the *Auger coefficient*. Figure 6.D.3 shows experimental values for the Auger coefficients determined at room temperature in two important systems for *infrared detection*: $Hg_xCd_{1-x}Te$ and $InAs_{1-x}Sb_x$. The forbidden gap in these alloys varies as a function of the fractional composition x. Figure 6.D.3 confirms the exponential dependence of the Auger coefficient as a function of the size of the bandgap.

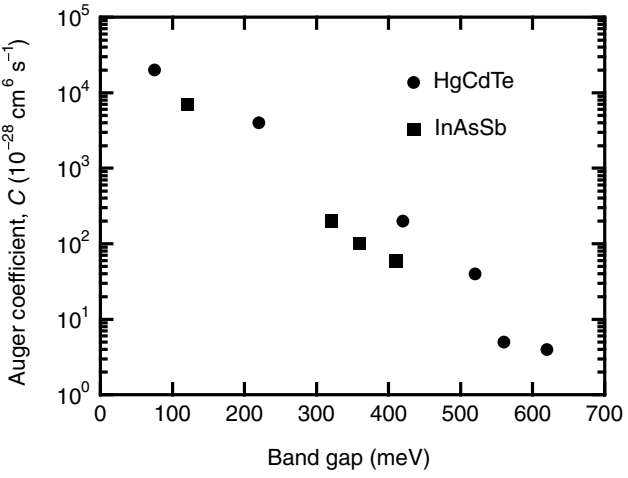


Fig. 6.D.3. Room temperature Auger coefficient for HgCdTe and InAsSb as a function of bandgap (with bandgap variations for a particular semiconductor being achieved by modification of the fractional alloy composition).

FURTHER READING

J. S. Blakemore, *Semiconductor Statistics*, Dover, New York (1987).

P. T. Landsberg, *Recombination in Semiconductors*, Cambridge University Press, Cambridge (1991).

7 Optical properties of semiconductors

7.1 Introduction

In Chapters 1–3, we described the different interaction mechanisms between an electromagnetic wave and a two-level quantum system. We indicated briefly at the time that that these results may be trivially generalized to systems possessing arbitrary numbers of *discrete states*. In Chapter 5, we saw how electrons in a crystalline solid and, more specifically, in a semiconductor (given our interests) are distributed among *energy bands* and can be differentiated according to a continuously varying index: the wavevector $\mathbf{k}$. In this chapter, we will present the mathematical techniques that will allow us to take into account the interactions between light and these continuously distributed energy states. Afterwards, we will be in a position to describe interactions between bulk semiconductor materials and light.

7.2 Dipolar elements in direct gap semiconductors

We consider a volume V of a crystalline semiconductor in which the eigenfunctions $|\Psi_{n\mathbf{k}}\rangle$ with energies $E_{n\mathbf{k}}$ are given by the Bloch–Floquet functions introduced in (5.12):

$$\Psi_{n,\mathbf{k}}(\mathbf{r}) = \frac{u_{n\mathbf{k}}(\mathbf{r})}{\sqrt{V}} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}} = \langle \mathbf{r}|\Psi_{n,\mathbf{k}}\rangle \tag{7.1}$$

We recall that $\mathbf{k}$ is the wavevector of an eigenstate belonging to the first Brillouin zone of a semiconductor, and which designates the position of the state within the nth energy band in the direction $\mathbf{e}_\mathbf{k} = \mathbf{k}/k$ (see Figs. 5.2 and 5.4). The functions $u_{n,\mathbf{k}}(\mathbf{r})$ possess the periodicity of the crystalline lattice (i.e. for every lattice vector $\mathbf{r}_\mathrm{i}$, $u_{n,\mathbf{k}}(\mathbf{r} + \mathbf{r}_\mathrm{i}) = u_{n,\mathbf{k}}(\mathbf{r})$). Furthermore, these functions are normalized by the volume of the primitive cell V_i, i.e. $\int_{V_i} |u_{n\mathbf{k}}(\mathbf{r})|^2 \mathrm{d}^3\mathbf{r} = V_i$. As explained in Complement 1.A, we normalize the stationary functions with respect to a fictitious box of volume $V = NV_i$ so that $\int_V |\Psi_{n\mathbf{k}}|^2 \mathrm{d}^3\mathbf{r} = 1$, resulting in a *pseudo-quantization* of the wavevectors $\mathbf{k}$. We recall that the most common procedure used for pseudo-quantization in crystalline solids consists of applying the *Born–von Karman cyclic*

boundary conditions given in (5.10). Finally, an important point worth remembering (and useful on many occasions) is that spatial variations in the wavefunction, $e^{-i\mathbf{k}\cdot\mathbf{r}}$ are very slow in comparison with those arising from atomic wavefunctions $u_{n,\mathbf{k}}(\mathbf{r})$ (see Fig. 5.3). In other words, $e^{-i\mathbf{k}\cdot\mathbf{r}}$ acts as an envelope to the Bloch functions.

A semiconductor is subjected to an electromagnetic wave for which the perturbation Hamiltonian for the interaction (see (3.13)) may be written:

$$W(\mathbf{r},t) = W\cos(\mathbf{k}_{op}\cdot\mathbf{r} - \omega t) = -q\mathbf{E}\cdot\hat{\mathbf{r}}\cos(\mathbf{k}_{op}\cdot\mathbf{r} - \omega t) \tag{7.2}$$

where $\hat{\mathbf{r}}$ is the position operator, $\mathbf{k}_{op}$ is the wavevector of light, and $\mathbf{E}$ is the electric field. The optical interaction Hamiltonian will then couple the two states $|\Psi_{n,\mathbf{k}}\rangle$ and $|\Psi_{n',\mathbf{k}'}\rangle$. Equations (1.82) and (1.83) allow one to calculate the probability rate (in s^{-1}), for an electron in the nth band with a wavevector $\mathbf{k}$, of being promoted to the n'th band with wavevector $\mathbf{k}'$:

$$P_{n,\mathbf{k},n'\mathbf{k}'} = \frac{\pi}{2\hbar}|\langle\Psi_{n',\mathbf{k}'}|W|\Psi_{n,\mathbf{k}}\rangle|^2\delta(\hbar\omega = E_{n',\mathbf{k}'} - E_{n,\mathbf{k}}) \tag{7.3}$$

We recall that the vectors $\mathbf{k}$ are pseudo-quantized. The energy dependence of the Dirac delta function expresses the requirement that energy be conserved:

$$\hbar\omega = E_{n',\mathbf{k}'} - E_{n,\mathbf{k}} \tag{7.4}$$

We will now calculate the matrix element $W_{n,\mathbf{k},n',\mathbf{k}'}$:

$$\begin{aligned} W_{n,\mathbf{k},n',\mathbf{k}'} &= \langle\Psi_{n',\mathbf{k}'}|W|\Psi_{n,\mathbf{k}}\rangle \\ &= -\frac{q\mathbf{E}}{V}\cdot\iiint_{\text{lattice}} u^*_{n',\mathbf{k}'}(\mathbf{r})e^{i\mathbf{k}'\cdot\mathbf{r}}\mathbf{r}e^{-i\mathbf{k}_{op}\cdot\mathbf{r}}u_{n,\mathbf{k}}(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}d^3\mathbf{r} \end{aligned} \tag{7.5}$$

This last integral may be simplified by noting that the function $e^{i(\mathbf{k}'-\mathbf{k}_{op}-\mathbf{k})\cdot\mathbf{r}}$ varies very slowly in comparison with the product of the two functions $u_{n,\mathbf{k}}(\mathbf{r})$. We may therefore rewrite integral (7.5) in the following manner:

$$W_{n,\mathbf{k},n',\mathbf{k}'} = -\frac{q\mathbf{E}}{V}\cdot\sum_i e^{i(\mathbf{k}'-\mathbf{k}_{op}-\mathbf{k})\mathbf{r}_i}\iiint_{\text{Cell i}} u^*_{n',\mathbf{k}'}(\mathbf{r})\mathbf{r}u_{n,\mathbf{k}}(\mathbf{r})d^3\mathbf{r} \tag{7.6}$$

and the integral $\mathbf{I}$ in (7.6) may be written:

$$\mathbf{I} = \iiint_{\text{Cell } i} u^*_{n',\mathbf{k}'}(\mathbf{R})(\mathbf{R} + \mathbf{r}_\text{i})u_{n,\mathbf{k}}(\mathbf{R})\mathrm{d}^3\mathbf{R}$$
$$= \iiint_{\text{Cell } 0} u^*_{n',\mathbf{k}'}(\mathbf{R})\mathbf{R}u_{n,\mathbf{k}}(\mathbf{R})\mathrm{d}^3\mathbf{R} \tag{7.7}$$

where $\mathbf{R}$ spans the primitive cell about 0. The term in the integral in $\mathbf{r}_\text{i}$ may be eliminated from (7.7) as the Bloch functions are orthonormal for $\mathbf{k} \neq \mathbf{k}'$. As a result, this integral is independent of the ith cell over which it is evaluated. Equation (7.6) then becomes:

$$W_{n,\mathbf{k},n',\mathbf{k}'} = -\frac{q\mathbf{E}}{V_i}\cdot\mathbf{I}\sum_i \mathrm{e}^{\mathrm{i}(\mathbf{k}'-\mathbf{k}_\text{op}-\mathbf{k})\mathbf{r}_\text{i}} = -\frac{q\mathbf{E}}{V_i}\cdot\mathbf{I}\delta(\mathbf{k}' - \mathbf{k}_\text{op} - \mathbf{k}) \tag{7.8}$$

where the Kronecker function in this last term expresses the requirement of momentum conservation in electron–photon interaction:

$$\mathbf{k}' = \mathbf{k} + \mathbf{k}_\text{op} \tag{7.9a}$$

Momentum conservation in optical transitions

The electron and photon wavevectors in the visible and near-visible light regions are quite distinct: $k_\text{op} = 2\pi/\lambda \approx 10^4$–$10^6\,\text{cm}^{-1}$, whereas typical wavevectors in the Brillouin zone are of the order of $k = 2\pi/a$ or $10^8\,\text{cm}^{-1}$ (recalling that a is the interatomic distance). The wavevector of light is therefore negligible in comparison with that of electrons in each of the different energy bands, so that:

$$\mathbf{k}' \approx \mathbf{k} \tag{7.9b}$$

The optical transitions are 'vertical'

Figure 7.1 illustrates that optical transitions within the band structure must take place at constant $\mathbf{k}$. It is in this sense that optical transitions are referred to as being *vertical*. Clearly, this transition rule can only be respected if the forbidden gap is direct, i.e. if the valence and conduction band extrema are aligned vertically in the Brillouin zone. The case involving transitions between indirect bands is more complicated and is not treated in this book. In future calculations, we will therefore be able to eliminate the index $\mathbf{k}'$ from (7.8).

From this point onwards, we will limit our investigations to optical transitions between the valence and conduction bands. We will denote as $u_{c\mathbf{k}}$ and $u_{v\mathbf{k}}$ the periodic portions of the Bloch functions in the conduction and valence bands, respectively. The dipolar matrix element $W_{vc}(\mathbf{k})$ between the valence and conduction bands is then:

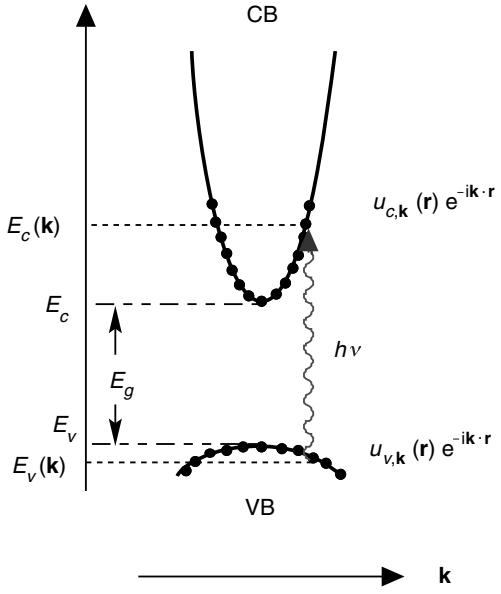


Fig. 7.1. Optical transitions between states in the conduction and valence bands take place at constant **k**. These transitions are referred to as being 'vertical'.

$$W_{v,c}(\mathbf{k}) = -q\mathbf{E}\cdot\mathbf{r}_{vc}\delta(\mathbf{k}' - \mathbf{k}_{\mathrm{op}} - \mathbf{k}) \tag{7.10}$$

$$\mathbf{r}_{vc} = \iiint_{\mathrm{cell}} u^*_{c,\mathbf{k}}(\mathbf{R})\mathbf{R}u_{v,\mathbf{k}}(\mathbf{R})\mathrm{d}^3\mathbf{R}$$

Dipole matrix element for the transition valence band → conduction band

We can now utilize Eq. (3.D.6) to relate the matrix elements of the optical interaction Hamiltonians $\mathbf{A}\cdot\mathbf{p}$ and $\mathbf{D}\cdot\mathbf{E}$:

$$\mathbf{r}_{vc} = -\frac{i}{m_0\omega_k}\mathbf{p}_{vc}$$
$$\mathbf{r}_{vc} = \langle u_{c,\mathbf{k}}|\mathbf{r}|u_{v,\mathbf{k}}\rangle;\ \mathbf{p}_{vc} = \langle u_{c,\mathbf{k}}|\mathbf{p}|u_{v,\mathbf{k}}\rangle \tag{7.11a}$$

where m_0 is the vacuum electron mass, and $\hbar\omega_k$ is the energy separating the two **k** states in the conduction and valence bands. We saw in Complement 5.C that the element $\mathbf{p}_{vc}$ may be easily obtained starting from the Kane parameter, P. The formula is extremely useful as P is largely constant across all III–V semiconductors (see Table 7.1), making (7.10) quite predictive. Furthermore, by using Kane's theory, we can easily show that *to first order in k*, the matrix elements (7.11a) are *constant*. The equation which couples $\mathbf{p}_{vc}$ to the Kane element P, can be found across the literature in as many different forms as there are definitions for the Kane

Table 7.1. *Values of* E_g (*bandgap*), m_c/m_0 (*conduction band effective mass*), E_P (*Kane energy*), *and the optical transition dipolar matrix element* r_{vc} *for various semiconductors* (*from G. Bastard*, Wave Mechanics Applied to Semiconductor Heterostructures, *Wiley*, *New York* (1991))

	GaAs	InP	GaSb	InAs	InSb
E_g (eV)	1.5192	1.4236	0.811	0.418	0.2352
m_c/m_0	0.0665	0.079	0.0405	0.023	0.0139
E_P (eV)	22.71	17	22.88	21.11	22.49
r_{vc} (Å)	6.14	5.67	11.5	21.5	39.5

parameter. In (5.C.18), we have utilized the convention that 'P^2 is homogeneous with energy'. Other authors use the convention that 'P is homogeneous with velocity' and others 'with momentum'. We prefer to use the Kane energy E_P ($=P^2$ in our convention), which is more unanimously accepted (20–25 eV), leading to the result:

$$|\mathbf{r}_{vc}| = \frac{\hbar}{E_g}\sqrt{\frac{E_P}{2m_0}} \tag{7.11b}$$

Dipole matrix element and Kane energy

where E_g is the semiconductor bandgap. An equally popular notion in the literature is that of the *matrix element* x_{vc}. This term is introduced to take into account the fact that only the heavy and light hole bands participate in the optical transitions and not the hole spin–orbit band, which is too far away in energy. We could use the transition elements introduced in (5.C.22) but the idea behind this calculation is quite simple: only two-thirds of the oscillator strength (i.e. two bands out of three) is involved, leading to:

$$x_{vc}^2 = \frac{1}{3}\frac{\hbar^2}{E_g^2}\frac{E_P}{m_0} \tag{7.11c}$$

Table 7.1 shows the relevant physical parameters E_g, E_P, m_c/m_0, and r_{vc} for the principal semiconductors used in optoelectronics applications. We note that for small gap semiconductors, the dipolar matrix element becomes very large in comparison to the interatomic distance; for optical properties, the electronic wavefunctions become more and more delocalized as the bandgap decreases. It is therefore customary to speak of semiconductors becoming 'more quantum in nature as their bandgaps shrink'.

Example
The Kane energy E_P in GaAs is 22.7 eV and corresponds to a Kane velocity of:

$$v_{\text{Kane}} = \sqrt{(E_P/2m_0)} = \sqrt{[(22.7\,\text{V} \times 1.6 \times 10^{-19}\,\text{C})/(2 \times 0.9 \times 10^{-30}\,\text{kg})]} = 1.42 \times 10^{6}\,\text{m s}^{-1}$$

The matrix element r_{vc} given by (7.11b) is then:

$$r_{vc} = 1.05 \times 10^{-34}\,\text{J s} \times 1.42 \times 10^{6}\,\text{m s}^{-1}/(1.5\,\text{V} \times 1.6 \times 10^{-19}\,\text{C}) = 6.14\,\text{Å}$$

The element x_{vc} enters into the absorption calculation and is equal to $r_{vc} \times \sqrt{(2/3)}$ or 5.0 Å.

7.3 Optical susceptibility of a semiconductor

We are now in a position to calculate the optical susceptibility associated with transitions from the valence to the conduction bands. We will consider a semiconductor crystal with volume V subjected to a pseudo-quantization condition. The structure of the valence and conduction bands will be assumed to be parabolic, and well described by their corresponding effective masses m_v and m_c. Given that optical transitions are vertical, we will interest ourselves in the optical susceptibility $\chi_{\mathbf{k}}(\omega)$ resulting from constant $\mathbf{k}$ transitions between the valence and conduction bands. Equations (7.4) and (7.9) give the relationship between the conduction $E_c(\mathbf{k})$ and valence bands $E_v(\mathbf{k})$ coupled by the optical transition (see Fig. 7.1):

$$E_c(\mathbf{k}) - E_v(\mathbf{k}) = \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_c} + \frac{1}{m_v}\right) + E_g \tag{7.12}$$

where k is the norm of the vector $\mathbf{k}$. It is useful to introduce the *reduced effective mass* as it simplifies (7.12):

$$\frac{1}{m_r} = \frac{1}{m_c} + \frac{1}{m_v} \tag{7.13}$$

Reduced effective mass

The (dimensionless) optical susceptibility associated with a transition between quasi-discrete levels $E_v(\mathbf{k})$ and $E_c(\mathbf{k})$ is given by (3.26) to be:

$$\chi_{\mathbf{k}}(\omega) = \frac{q^2 x_{vc}(\mathbf{k})^2 T_2}{\varepsilon_0 \hbar} \frac{[\omega - \omega_{vc}(\mathbf{k})]T_2 - i}{[\omega - \omega_{vc}(\mathbf{k})]^2 T_2^2 + 1}[N_v(\mathbf{k}) - N_c(\mathbf{k})] \tag{7.14a}$$

where $x_{vc}(\mathbf{k})$ ($= x_{vc}$ as we have supposed the dipolar matrix elements are independent of $\mathbf{k}$) is the transition element in the Ox direction (arbitrarily selected to

simplify notation), $\omega_{vc}(\mathbf{k}) = (E_c(\mathbf{k}) - E_v(\mathbf{k}))/\hbar$, and T_2 is the relaxation time for electrons in the bands. $N_c(\mathbf{k})$ and $N_v(\mathbf{k})$ represent the carrier densities (cm^{-3}) in the levels E_c and E_v. The total susceptibility $\chi(\omega)$ is then found by summing the susceptibility terms (7.14a) over all wavevectors $\mathbf{k}$ in the band structure, i.e:

$$\chi(\omega) = 2\sum_{\mathbf{k}} \frac{q^2 x_{vc}(\mathbf{k})^2 T_2}{\varepsilon_0 \hbar} \frac{[\omega - \omega_{vc}(\mathbf{k})]T_2 - i}{[\omega - \omega_{vc}(\mathbf{k})]^2 T_2^2 + 1} [N_v(\mathbf{k}) - N_c(\mathbf{k})] \tag{7.14b}$$

where a factor of 2 results from the fact that there are two spin states available for every $\mathbf{k}$ vector. We will then suppose that the volume of the fictitious box is large enough to allow us to replace the summation over $\mathbf{k}$ by an integral:

$$2 \sum_{\mathbf{k}_n \in BZ} \leftrightarrow \int_{\mathbf{k}} \rho(\mathbf{k}) \cdot \mathrm{d}^3\mathbf{k} \leftrightarrow \int_{E} \rho(E) \cdot \mathrm{d}E \tag{7.15}$$

Equivalence rule between the summation over $\mathbf{k}$, and the integrals over $\mathbf{k}$ and over energy E

where the density $\rho(\mathbf{k})$ is in units of $1/(\mathrm{cm}^{-3})$, $\rho(\mathrm{E})$ is in units of 1/J, and a factor of 2 is included to take into account *spin degeneracy*. In an isotropic medium, the microscopic densities $N_c(\mathbf{k})$ and $N_v(\mathbf{k})$ in (7.14) are given by $\rho_c(\mathbf{k})\mathrm{d}^3\mathbf{k}$ and $\rho_v(\mathbf{k})\mathrm{d}^3\mathbf{k}$, where $\rho_c(\mathbf{k})$ and $\rho_v(\mathbf{k})$ are equal and given by (5.17) with $\rho_c(\mathbf{k}) = \rho_v(\mathbf{k}) = 2V/8\pi^3 = V/4\pi^3$ (we have multiplied by 2 to account for spin degeneracy and recall that V is the crystal volume). To these terms we must add an expression corresponding to thermal quasi-equilibrium: i.e. introduce the Fermi–Dirac functions $f_c(E_c(\mathbf{k}))$ and $f_v(E_v(\mathbf{k}))$ which describe the occupation probabilities for the levels $E_c(\mathbf{k})$ and $E_v(\mathbf{k})$ given in (5.38) and in Section 5.6:

$$f_c(E) = \frac{1}{1 + \mathrm{e}^{(E - E_{Fc})/kT}}$$
$$f_v(E) = \frac{1}{1 + \mathrm{e}^{(E - E_{Fv})/kT}} \tag{7.16a}$$

E_{Fc} and E_{Fv} are the quasi-Fermi levels for the electrons and holes. (Note that f_v is different from (5.44) since here we use the probability that a valence band state is occupied by an electron). The infinitesimal density $N_v - N_c$ in (7.14) for the element $\mathrm{d}^3\mathbf{k}$ is then replaced by:

$$\begin{aligned} N_v - N_c &\leftrightarrow \rho_c \mathrm{d}^3\mathbf{k}[\![f_v[E_v(\mathbf{k})]\{1 - f_c[E_c(\mathbf{k})]\} - f_c[E_c(\mathbf{k})]\{1 - f_v[E_v(\mathbf{k})]\}]\!] \\ &= \rho_c \mathrm{d}^3\mathbf{k}\{f_v[E_v(\mathbf{k})] - f_c[E_c(\mathbf{k})]\} \end{aligned} \tag{7.16b}$$

The optical susceptibility is then obtained by integrating (7.14) over the first Brillouin zone or:

$$\chi(\omega) = \frac{q^2 x_{vc}^2 T_2}{\varepsilon_0 \hbar} \int \rho_c \mathrm{d}^3\mathbf{k} \{ f_v[E_v(\mathbf{k})] - f_c[E_c(\mathbf{k})] \} \frac{[\omega - \omega_{vc}(\mathbf{k})]T_2 - i}{[\omega - \omega_{vc}(\mathbf{k})]^2 T_2^2 + 1} \tag{7.17}$$

We emphasize again that within the framework of Kane's theory (seen in Complement 5.C), we suppose that the matrix element x_{vc} is independent of $\mathbf{k}$. The energy conservation requirement can then be written in terms of (7.4), (7.12), and (7.13) as:

$$\hbar\omega_{vc} = E_c(\mathbf{k}) - E_v(\mathbf{k}) = \frac{\hbar^2 k^2}{2m_r} + E_g$$

$$E_c(\mathbf{k}) = E_g + \frac{\hbar^2 k^2}{2m_c} \tag{7.18}$$

$$E_v(\mathbf{k}) = -\frac{\hbar^2 k^2}{2m_v}$$

The wavevectors $\mathbf{k}$ and the photon energies $\hbar\omega_{vc}$ being thus related, we may integrate (7.17) over the frequency range of the incident photons with the help of expressions (5.19) and (5.20):

$$\rho_c \mathrm{d}k = \rho_j(\omega_{vc})\mathrm{d}\omega_{vc} = \frac{1}{2\pi^2}\left(\frac{2m_r}{\hbar}\right)^{3/2} (\omega_{vc} - E_g/\hbar)^{1/2} \mathrm{d}\omega_{vc} \tag{7.19a}$$

or again as:

$$\rho_j(\omega) = \frac{1}{2\pi^2}\left(\frac{2m_r}{\hbar}\right)^{3/2} (\omega - E_g/\hbar)^{1/2} \tag{7.19b}$$

$$\rho_j(E) = \frac{1}{2\pi^2}\left(\frac{2m_r}{\hbar^2}\right)^{3/2} (E - E_g)^{1/2} \tag{7.19c}$$

Joint density of states (J^{-1} cm^{-3})

Expression (7.19) describes the density of states *joined by a photon of energy* $\hbar\omega$. Substituting the last equation into the integral in (7.17), we obtain the principal finding of this section:

$$\chi(\omega) = \frac{q^2 x_{vc}^2 T_2}{\varepsilon_0 \hbar} \int_{E_g/\hbar}^{\infty} \rho_j(\omega_{vc})\mathrm{d}\omega_{vc} [f_v(E_v) - f_c(E_c)] \frac{(\omega - \omega_{vc})T_2 - i}{(\omega - \omega_{vc})^2 T_2^2 + 1} \tag{7.20}$$

Optical susceptibility associated with an interband transition in a semiconductor

In this last expression, the joint energies E_v and E_c are functions of $\hbar\omega$ by intermediary of the relations put forth in (7.18). The absorption coefficient is related to the imaginary part of the optical susceptibility by (3.36), i.e:

$$\alpha(\omega) = \frac{\omega}{cn_{\text{op}}}\chi_{\text{Im}} \tag{7.21}$$

or

$$\alpha(\omega) = \frac{\pi q^2 x_{vc}^2 \omega}{\varepsilon_0 \hbar n_{\text{op}} c} \int_{\omega_g}^{\infty} \rho_j(\omega_{vc}) \mathrm{d}\omega_{vc} [f_v(E_v) - f_c(E_c)] \frac{1/\pi T_2}{(\omega - \omega_{vc})^2 + 1/T_2^2} \tag{7.22}$$

This last formula may be considerably simplified by noticing that the Lorentzian distribution behaves as a Dirac function in comparison with the more slowly varying functions in the integral. We then obtain for the absorption coefficient α (in cm^{-1})

$$\alpha(\omega) = \frac{\pi q^2 x_{vc}^2 \omega}{\varepsilon_0 \hbar n_{\text{op}} c} \rho_j(\omega) [f_v(\hbar\omega) - f_c(\hbar\omega)] \tag{7.23}$$

or again:

$$\alpha(\omega) = -\gamma(\omega) = \alpha_0(\omega) [f_v(\hbar\omega) - f_c(\hbar\omega)] \tag{7.24}$$

Absorption and gain in a semiconductor (cm^{-1})

where $\gamma(\omega)$ is the gain of the semiconductor medium and $\alpha_0(\omega)$ is the empty conduction band absorption with:

$$\alpha_0(\omega) = \frac{q^2 x_{vc}^2}{\lambda_0 \varepsilon_0 \hbar n_{\text{op}}} \left(\frac{2m_r}{\hbar}\right)^{3/2} \sqrt{(\omega - E_g/\hbar)} \tag{7.25a}$$

Empty conduction band absorption (cm^{-1})

The functions $f_c(\hbar\omega)$ and $f_v(\hbar\omega)$ are the Fermi–Dirac functions:

$$f_c(\hbar\omega) = \frac{1}{1 + \exp\{[E_c(\hbar\omega) - E_{Fc}]/kT\}}$$
$$E_c(\hbar\omega) = E_g + \frac{m_r}{m_c}(\hbar\omega - E_g) \tag{7.25b}$$

and

$$f_v(\hbar\omega) = \frac{1}{1 + \exp\{[E_v(\hbar\omega) - E_{Fv}]/kT\}}$$
$$E_v(\hbar\omega) = -\frac{m_r}{m_v}(\hbar\omega - E_g) \tag{7.25c}$$

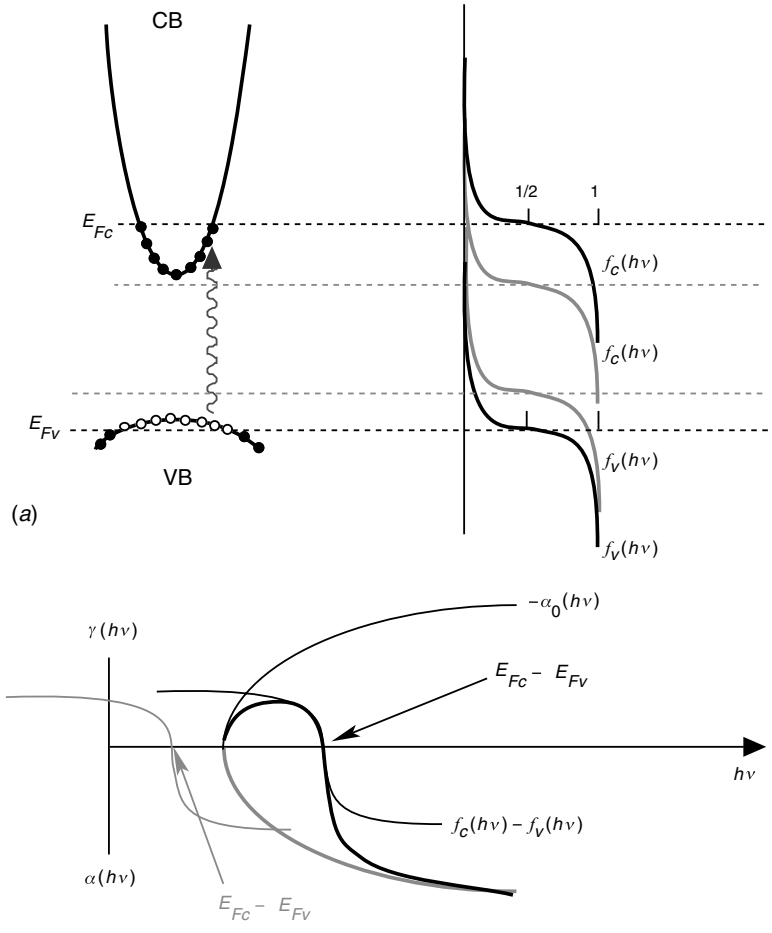


Fig. 7.2. These figures show the evolution of the absorption and gain curves as a function of the position of the quasi-Fermi level. The grey (dark) curves correspond to a small (large) displacement from equilibrium. In this case, the medium absorbs all photons having energies in excess of the bandgap. Once the energy separation between the two quasi-Fermi levels exceeds the bandgap, all photons possessing energies between E_g and $E_F - E_{Fv}$ are amplified (the Bernard–Durrafourg condition).

Equation (7.24) indicates that the absorption in direct gap semiconductors, for photons with energies greater than the bandgap, grows as $\sqrt{(\hbar\omega - E_g)}$ (see Fig. 7.2). Equations (7.23)–(7.25) allow one to describe the optical behaviour of a semiconductor away from thermodynamic equilibrium.

The condition for optical gain (i.e. $\alpha(\omega) < 0$) is given by:

$$f_c(\hbar\omega) - f_v(\hbar\omega) > 0 \tag{7.26a}$$

or again, given (7.25b) and (7.25c):

$$E_{Fc} - E_{Fv} > \hbar\omega \tag{7.26b}$$

Bernard–Durrafourg condition

This last equation stipulates that only those photons with energy less than the energy separation between the two quasi-Fermi levels will be amplified. These quasi-Fermi levels are related to the non-equilibrium population N by the condition put forward in Section 5.6:

$$N = \int_0^\infty \rho_c(E) \frac{1}{1 + \exp[(E - E_{Fc})/kT]} \mathrm{d}E \tag{7.27}$$

for the electron quasi-Fermi level E_{Fc} (and a similar formula for the holes). Under positive gain conditions, the curve described by (7.23) is referred to as the semiconductor *gain curve*. Figure 7.2 depicts the evolution of the gain curve as a function of increasing deviation from thermodynamic equilibrium.

7.4 Absorption and spontaneous emission

We know from Chapter 3 that the approach founded on optical susceptibility does not allow us to take into account spontaneous emission. A natural method for dealing with this effect is to make use of the approach employed in Einstein's rate equations. We presented this theory in Section 3.6 for a *broad spectrum electromagnetic wave*. We will rederive this approach here for the following reasons: (i) it is a very subtle calculation (keep an eye on the dimensions!) of great practical importance, (ii) the notations used for semiconductors vary for historical reasons, and finally (iii) this context is fairly different: we are now dealing with a *broadened energy spectrum* (we have, however, already mentioned in Complement 3.C the equivalence between a *broad spectrum wave* and a *broadened transition energy spectrum*).

We consider a semiconductor of volume V, a level $E_v(\mathbf{k})$ in the valence band, and a level $E_c(\mathbf{k})$ in the conduction band for a specific $\mathbf{k}$ value (as we prepare the way for pseudo-quantization . . .). As we saw in Section 3.6 for an electromagnetic wave with a *wide spectral distribution* $\rho_{\mathrm{ph}}(\hbar\omega)$, the optical transition rate g_{vc} (in s^{-1}), *over the entire volume* V and between the levels E_v and E_c, is proportional to the probability that the state E_v is occupied and that the state E_c is empty.

$$g_{vc}(\mathbf{k}) = B_{vc} f_v(E_v)[1 - f_c(E_c)]\rho_{\mathrm{ph}}(h\nu) \tag{7.28}$$

where ρ_{ph} (in J^{-1}) is the photon density per unit energy in the volume V, and B_{vc} (in $\mathrm{J\,s}^{-1}$) is the transition rate *per photon within the cavity*, given by Fermi's golden rule (see (3.76)–(3.77)) in the form:

$$B_{vc} = \frac{\pi}{2\hbar} |\langle \Psi_{c,\mathbf{k}} | \mathrm{W} | \Psi_{v,\mathbf{k}} \rangle|^2 = \frac{\pi}{2\hbar} q^2 E_0^2 x_{vc}^2 \tag{7.29}$$

where E_0 is the electric field *associated with one photon in volume V* (B_{vc} is given for a single photon in the cavity!) given by (2.77):

$$\hbar\omega = \frac{1}{2} n_{\mathrm{op}}^2 \varepsilon_0 E_0^2 V \tag{7.30a}$$

or

$$E_0^2 = \frac{2\hbar\omega}{\varepsilon_0 n_{\mathrm{op}}^2 V} \tag{7.30b}$$

from which:

$$B_{vc} = \frac{\pi q^2 x_{vc}^2 \omega}{\varepsilon_0 n_{\mathrm{op}}^2 V} = \frac{\pi q^2 E_P}{3\varepsilon_0 n_{\mathrm{op}}^2 m_e \omega V} \tag{7.31}$$

where we have made use of (7.11b). Note the presence of the term V in the denominator: this results from the delocalized nature of the Bloch wavefunctions, which requires that we take as an interaction region the entire volume V of the crystal (see Complement 1.A). We recall that energy conservation requires that $\hbar\omega = \hbar\omega_{vc} = E_{vc}(k) = E_c(\mathbf{k}) - E_v(\mathbf{k})$. Two differences exist from the conditions assumed in Section 3.6. First, the functions f_c and f_v given in (7.25a) and (7.25b) allow the description of non-equilibrium situations in terms of quasi-Fermi levels, and second, the Einstein coefficient B_{vc} is this time expressed in J s^{-1} (we propose as an exercise that the reader verify the dimensions of B_{vc} in (7.31)).

In a similar way, the stimulated emission rate is given by:

$$g_{cv}(\mathbf{k}) = B_{cv} f_c(E_c)[1 - f_v(E_v)]\rho_{\mathrm{ph}}(h\nu) \tag{7.32}$$

To this, we must clearly add the spontaneous emission rate g_{cv}^{spon} (in s^{-1}), which is independent of the photon density in the semiconductor and given by:

$$g_{cv}^{\mathrm{spon}}(\mathbf{k}) = A_{cv} f_c(E_c)[1 - f_v(E_v)] \tag{7.33}$$

A_{cv} is the spontaneous transition rate in the volume V. To begin with, we will concern ourselves with the case of thermodynamic equilibrium. The photon density per unit energy in the volume V is given by the blackbody emission spectrum in J s^{-1} m^{-3} (Eq. (2.91)) divided by $h\nu$ to obtain a numerical density (and not an energy density) and again by h to bring it into units of energy and not frequency, i.e:

$$\rho_{\mathrm{ph}}(h\nu) = \frac{8\pi n_{\mathrm{op}}^3 (h\nu)^2}{h^3 c^3} \frac{1}{\exp(h\nu/kT) - 1} V \tag{7.34}$$

Similarly, at equilibrium, the Fermi–Dirac distributions are identical between the two bands ($f_c = f_v = f$ and $E_{Fc} = E_{Fv} = E_F$).

The stationary condition at thermodynamic equilibrium can be written in the same fashion as in (3.C.4):

$$\begin{aligned} &B_{vc} f(E_v)[1 - f(E_c)]\rho_{\text{ph}}(E_{vc}) \\ &= B_{cv} f(E_c)[1 - f(E_v)]\rho_{\text{ph}}(E_{vc}) + A_{cv} f(E_c)[1 - f(E_v)] \end{aligned} \tag{7.35a}$$

or again:

$$\frac{8\pi n_{\text{op}}^3 E_{vc}^2 V}{h^3 c^3 [\exp(E_{vc}/kT) - 1]} = \frac{A_{cv}}{B_{vc}\exp(E_{vc}/kT) - B_{cv}} \tag{7.35b}$$

This can be true at all temperatures only if the Einstein coefficients B_{cv}, B_{vc}, and A_{cv} are related by:

$$B_{vc} = B_{cv} \tag{7.36a}$$

$$A_{cv} = \frac{8\pi n_{\text{op}}^3 E_{vc}^2 V}{h^3 c^3} B_{cv} \tag{7.36b}$$

These equations are similar to those expressed in (3.77) and (3.78) with the exception that B_{cv} and B_{12} do not share the same units. We wish to emphasize the concept that underlies Einstein's approach: that although the relationships between the coefficients A and B are established under the umbrella of thermodynamic equilibrium, being intrinsic to the system, *they remain valid in all instances*, even in situations which place the system far from thermodynamic equilibrium.

We can now replace B_{vc} in (7.36) by the expression given for it in (7.31), to find:

$$A_{cv} = \frac{1}{\tau_R} \tag{7.37a}$$

where τ_R is the *radiative lifetime* given by:

$$\frac{1}{\tau_R} = \frac{q^2 x_{vc}^2 n_{\text{op}} \omega_{vc}^3}{\pi c^3 \hbar \varepsilon_0} = \frac{q^2 n_{\text{op}} E_g E_P}{3\pi c^3 \hbar^2 \varepsilon_0 m_e} \tag{7.37b}$$

Radiative lifetime in a semiconductor

We note that the same tendency for radiative lifetimes exists in both semiconductors and atoms. The greater the magnitude of the forbidden gap, the shorter the radiative lifetime, and the more difficult it will be to invert the carrier population in the system. This behaviour is demonstrated, for instance, when moving from a semiconductor possessing a near-infrared bandgap, such as GaAs, to a wide gap semiconductor, such as GaN.

We can now go ahead and calculate the absorption rate $r_{\text{abs}}(\mathbf{k})$ (in s^{-1}) for an arbitrary wavevector $\mathbf{k}$, resulting from competition between the transitions $c \to v$ and $v \to c$:

$$r_{\text{abs}}(\mathbf{k}) = B_{vc} f_v(E_v)[1 - f_c(E_c)]\rho_{\text{ph}}(h\nu) - B_{cv} f_c(E_c)[1 - f_v(E_v)]\rho_{\text{ph}}(h\nu) \tag{7.38}$$

or

$$r_{\text{abs}}(\mathbf{k}) = B_{vc}[f_v(E_v) - f_c(E_c)]\rho_{\text{ph}}(h\nu) \tag{7.39}$$

where E_v and E_c are the states joined by the photon $h\nu$ given in (7.18) and (7.25).

We will now consider a radically different situation. We seek the absorption for a *monochromatic wave* with frequency ν incident upon a crystal (i.e. ρ_{ph} is a Dirac delta function $\delta(E - h\nu)$ – there is a single photon in the cavity volume V). The absorption $R_{\text{abs}}(h\nu)$ in the volume V is due to transitions over the entire Brillouin zone and (see (7.15)) is given by:

$$R_{\text{abs}}(h\nu) = 2\sum_{\mathbf{k}_n} r_{\text{abs}}(\mathbf{k}_n) = \int_k \rho_c r_{\text{abs}}(\mathbf{k}) \mathrm{d}^3\mathbf{k} \tag{7.40}$$

Given the requirement of energy conservation (7.18), and the transformation of variables in (7.19), this leads to an absorption rate of:

$$\begin{aligned} R_{\text{abs}}(h\nu) &= \int_E \rho_j(E) V B_{vc}[f_v(E_v) - f_c(E_c)]\delta(E - h\nu)\mathrm{d}E \\ &= \rho_j(h\nu) V B_{vc}[f_v(E_v) - f_c(E_c)] \end{aligned} \tag{7.41}$$

(Note that the volume term V will be eliminated by the $1/V$ term in B_{vc}, which is the reward for pseudo-quantization).

The absorption coefficient $\alpha(h\nu)$ is obtained by examining the experiment described in Fig. 7.3. In this case, a photon of energy $\hbar\omega$ impinges upon a semiconductor slab with surface S and thickness Δz. The energy absorbed during time Δt is $R_{\text{abs}}\hbar\omega\Delta t$ (with volume $V = S\Delta z$) and the energy crossing the surface is $\hbar\omega c/n_{\text{op}}\Delta t$. The ratio between these two quantities is simply $\alpha(h\nu)\Delta z$ and may be written:

$$\begin{aligned} \alpha(\hbar\omega) &= \frac{\text{power absorbed per unit volume}}{\text{power incident per unit surface}} \\ &= \frac{R_{\text{abs}}\hbar\omega}{\hbar\omega c/n_{\text{op}}} \end{aligned} \tag{7.42}$$

or given (7.41) and (7.31) as:

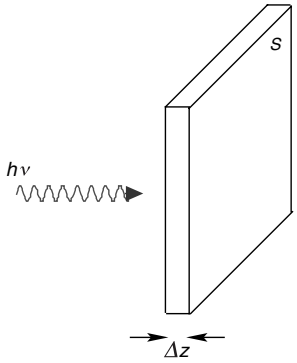


Fig. 7.3. Experimental geometry for the determination of absorption coefficient α.

$$\alpha(\omega) = \frac{\pi q^2 x_{vc}^2 \omega}{\varepsilon_0 n_{\text{op}} c} \rho_j(h\nu)[f_v(h\nu) - f_c(h\nu)] \tag{7.43}$$

which is exactly the equation obtained using density matrix formalism in (7.22) recalling that the densities per unit frequency and energy are related by $\rho_j(h\nu) = \rho_j(\omega)/\hbar$.

Clearly, we have not invested this effort solely to verify the correspondence between density matrix formalism and Einstein's rate equations. In addition, using this approach we have gained a means of calculating the spontaneous emission rate $R_{\text{spon}}(h\nu)$ due to the distribution of carriers in quasi-thermodynamic equilibrium in a semiconductor. This rate is given by summation of (7.33) over the Brillouin zone:

$$R_{\text{spon}}(h\nu) = 2\sum_{\mathbf{k}} r_{\text{spon}}(\mathbf{k}) = 2\sum_{\mathbf{k}} \frac{1}{\tau_R(\mathbf{k})} f_c(\mathbf{k})[1 - f_v(\mathbf{k})]\delta(E_c - E_v = h\nu) \tag{7.44}$$

with this sum being carried out for all wavevectors $\mathbf{k}$ that verify the energy conservation condition (expressed by the Dirac delta function δ), i.e:

$$E_c(\mathbf{k}) - E_v(\mathbf{k}) = h\nu = \frac{\hbar^2 k^2}{2m_r} + E_g \tag{7.45}$$

Using the equivalence (7.15) and the relation in (7.19) we find:

$$R_{\text{spon}}(h\nu) = \int_0^\infty r_{\text{spon}}(E)\rho_j(E)\delta(E = h\nu)\mathrm{d}E = r_{\text{spon}}(h\nu)\rho_j(h\nu) \tag{7.46}$$

from which:

$$R_{\text{spon}}(h\nu) = \frac{1}{\tau_R}\rho_j(h\nu) f_c[\mathrm{E}_c(h\nu)]\{1 - f_v[E_v(h\nu)]\} \tag{7.47}$$

Spectral distribution of the radiative recombination rate in a semiconductor (s^{-1} cm^{-3} J^{-1})

where we recognize an expression of type n/τ_R obtained for atoms (see Chapter 4); or again, by using (7.37b):

$$R_{\text{spon}}(\hbar\omega) = \frac{q^2 x_{vc}^2 n_{\text{op}} \omega^3}{\pi c^3 \hbar \varepsilon_0}\rho_j(\hbar\omega) f_c[E_c(\hbar\omega)]\{1 - f_v[E_v(\hbar\omega)]\} \tag{7.48}$$

Dividing expression (7.48) by the absorption formula (7.43), we obtain a relationship between absorption and spontaneous emission which remains valid for all quasi-thermodynamic equilibrium situations:

$$R_{\text{spon}}(h\nu) = \alpha(h\nu)\frac{8\pi n_{\text{op}}^2 (h\nu)^2}{h^3 c^2}\frac{1}{\mathrm{e}^{h\nu/kT} - 1} \tag{7.49}$$

Van Roosbroeck–Shockley equation

Equation (7.49) allows one to calculate the spontaneous emission spectrum $R_{\text{spon}}(h\nu)$ starting from the absorption spectrum, and expresses nothing more than the state of micro-equilibrium that exists for each photon of energy $h\nu$ between spontaneous emission and absorption due to a blackbody spectrum.

We now seek to determine the spectral distribution $R_{\text{spon}}(h\nu)$ for spontaneous emission of a semiconductor close to thermodynamic equilibrium. The Fermi–Dirac functions in (7.41) can be approximated by the Boltzmann functions:

$$\begin{aligned} f_c(E_c) &\approx \exp\left(-\frac{E_c - E_{Fc}}{kT}\right) \\ 1 - f_v(E_v) &\approx \exp\left(-\frac{E_{Fv} - E_v}{kT}\right) \end{aligned} \tag{7.50}$$

whence:

$$f_c[E_c(h\nu)]\{1 - f_v[E_v(h\nu)]\} \approx \exp\left(-\frac{h\nu}{kT} + \frac{\Delta E_F}{kT}\right) \tag{7.51}$$

and where ΔE_F is the difference between the quasi-Fermi levels $\Delta E_F = E_{Fc} - E_{Fv}$. We thus arrive at the spectral distribution $R_{\text{spon}}(h\nu)$:

$$R_{\text{spon}}(h\nu) = K_{\text{spon}}(h\nu - E_g)^{1/2}\exp\left(-\frac{h\nu - E_g}{kT}\right) \tag{7.52a}$$

Spectral distribution of the spontaneous emission rate (s^{-1} cm^{-3} J^{-1})

and the constant K_{spon} is given by (7.47) to be:

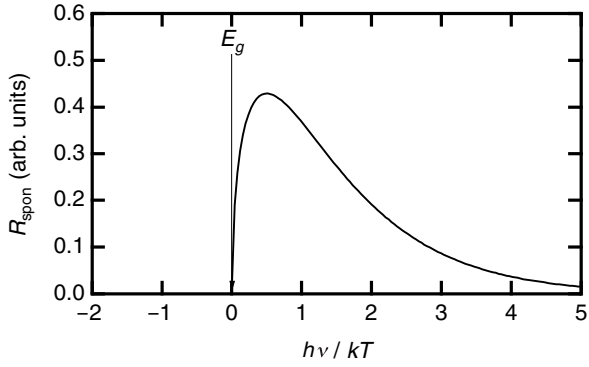


Fig. 7.4. Spectral distribution of the spontaneous emission rate. The width of the spectrum is of the order of $2\,kT$.

$$K_{\text{spon}} = \frac{(2m_r)^{3/2}}{\pi\hbar^2\tau_R}\exp\left(\frac{\Delta E_F - E_g}{kT}\right) \tag{7.52b}$$

Figure 7.4 depicts the behaviour of the spontaneous emission rate as a function of energy of the emitted photons. We note that practically the entire emitted optical power is distributed over an interval of $1.8kT$ above the absorption threshold.

Equation (7.52b) shows that the emitted optical power increases as $\exp(\Delta E_F)$, i.e. the output power increases the further the semiconductor is driven from thermal equilibrium.

Example

We will calculate the absorption coefficient $\alpha(E)$ and the spontaneous emission rate τ_R for GaAs:

$E_g = 1.45\,\text{eV}$
$m_v = 0.46m_e$
$m_c = 0.067m_e$
$x_{vc} = 3.2\,\text{Å}$
$n_{\text{op}} = 3.6$

The absorption coefficient is then found by (7.25a) to be:

$$\alpha(h\nu) = K_{\text{abs}}\left(\frac{h\nu - E_g}{q}\right)^{1/2}$$
$$\text{with } K_{\text{abs}} = \frac{q^{5/2}x_{vc}^2(2m_r)^{3/2}}{\lambda_0\varepsilon_0 n_{\text{op}}\hbar^3} \tag{7.53}$$

so that $K_{\text{abs}} = (1.6 \times 10^{-19}\,\text{C})^{5/2} \times (3.2 \times 10^{-10}\,\text{m})^2 \times (2 \times 0.058 \times 0.9 \times 10^{-30})^{3/2}/(0.8 \times 10^{-6}\,\text{m} \times 8.85 \times 10^{-12}\,\text{Fd m}^{-1} \times (1.05 \times 10^{-34})^3 \times 3.6)$ or $K_{\text{abs}} = 12\,000\,\text{cm}^{-1}\,\text{eV}^{-1/2}$. Thus, for a photon with an energy which exceeds the GaAs

bandgap by 0.01 eV, the absorption coefficient will be $1200\,\mathrm{cm}^{-1}$.

The radiative lifetime is then given by:

$$\frac{1}{\tau_R} = \frac{q^5 x_{vc}^2 n_{\mathrm{op}} (\hbar\omega_{vc}/q)^3}{\pi c^3 \hbar^4 \varepsilon_0} \tag{7.54}$$

so

$1/\tau_R = (1.5\,\mathrm{V})^3 \times (1.6 \times 10^{-19}\,\mathrm{C})^5 \times (3.2 \times 10^{-10}\,\mathrm{m})^2 \times 3.6/(3.14 \times (3 \times 10^8\,\mathrm{m\,s^{-1}})^3 \times (1.05 \times 10^{-34}\,\mathrm{J\,s})^4 \times 8.85 \times 10^{-12}\,\mathrm{Fd\,m^{-1}})$ or $\tau_R = 0.7\,\mathrm{ns}$.

Note that the x_{vc} value is not the one derived from Kane's theory (Table 7.1). The value has been chosen in order to fit the experimental absorption data. This reveals that the parabolic two-band Kane theory is oversimplified since it does not take into account non-parabolicity, valence band anisotropy, etc.

7.5 Bimolecular recombination coefficient

Equation (7.44), which was described in the last section, expresses the distribution of the optical transition rate (in $\mathrm{cm^{-3}\,s^{-1}\,J^{-1}}$) as a function of photon energy for two non-equilibrium carrier distributions as described by their respective quasi-Fermi levels. We will now seek the total radiative recombination rate for photons of all energy, i.e. involving the integral of the distribution (7.44) taken over hv:

$$R_{\mathrm{spon}} = \frac{1}{\tau_R} \int_{E_g}^{\infty} \rho_j(hv) f_c[E_c(hv)]\{1 - f_v[E_v(hv)]\}\mathrm{d}hv \tag{7.55}$$

where the product of the two Fermi–Dirac distributions is given by (7.51), or:

$$R_{\mathrm{spon}} = \frac{\mathrm{e}^{(E_{Fc}-E_{Fv})/kT}}{\tau_R} \int_{E_g}^{\infty} \rho_j(hv) \mathrm{e}^{-hv/kT}\mathrm{d}hv \tag{7.56}$$

We will now utilize the expression for the joint density of states (7.19c) so that:

$$\begin{aligned} R_{\mathrm{spon}} &= \frac{\mathrm{e}^{(E_{Fc}-E_{Fv})/kT}}{\tau_R} \frac{1}{2\pi^2}\left(\frac{2m_r}{\hbar^2}\right)^{3/2} \int_{E_g}^{\infty} (hv - E_g)^{1/2} \mathrm{e}^{-hv/kT}\mathrm{d}hv \\ &= \frac{\mathrm{e}^{[(E_{Fc}-E_{Fv})-E_g]/kT}}{\tau_R} \frac{1}{2\pi^2}\left(\frac{2m_r}{\hbar^2}\right)^{3/2} (kT)^{3/2} \int_0^{\infty} u^{1/2}\mathrm{e}^{-u}\mathrm{d}u \end{aligned} \tag{7.57}$$

This equation can be greatly simplified once we recognize in the above expression (7.57) the effective density of states given in (5.42) and (5.43):

$$N_j = \frac{1}{4}\left(\frac{2m_r kT}{\pi\hbar^2}\right)^{3/2} \tag{7.58}$$

where we have replaced the effective mass in the conduction band by the reduced effective mass m_r so that:

$$R_{\text{spon}} = \frac{1}{\tau_R} N_j e^{(E_{Fc}-E_{Fv}-E_g)/kT} \tag{7.59}$$

Taking into account expressions (5.42)–(5.45) for the effective state densities in the two bands and the carrier density as a function of Fermi energy, Eq. (7.59) becomes:

$$R_{\text{spon}} = \frac{1}{\tau_R}\frac{N_j}{N_c N_v} np \tag{7.60}$$

This last equation is better known under the form:

$$R_{\text{spon}} = Bnp \tag{7.61}$$

$$B = \frac{1}{\tau_R}\frac{N_j}{N_c N_v} = \frac{1}{\tau_R N_c}\left(\frac{m_r}{m_v}\right)^{3/2} \tag{7.62}$$

Radiative recombination rate ($cm^{-3}\,s^{-1}$) and the bimolecular radiative recombination coefficient ($cm^3\,s^{-1}$)

where B is the *bimolecular recombination coefficient*. Table 7.2 shows experimentally determined values for B coefficients in some of the more important III–V semiconductors.

Table 7.2. *Bimolecular recombination coefficients for various direct gap semiconductors* (*after V. P. Varshni*, Phys. Stat. Sol. **19**, 459 (1967)).

Material	$B\,(cm^3\,s^{-1})$
GaAs	7.2×10^{-10}
GaSb	2.4×10^{-10}
InP	1.3×10^{-9}
InAs	8.5×10^{-11}
InSb	4.6×10^{-11}

Example

For GaAs, the effective density of states in the conduction band is $N_c = 4.3 \times 10^{17}\,cm^{-3}$ (Section 5.6.1) and the lifetime τ_R is 7×10^{-10} s. The reduced mass is $1/(1/0.067) + (1/0.50) = 0.059$. The bimolecular recombination coefficient B is then $1/(7 \times 10^{-10}\,s \times 4.3 \times 10^{17}\,cm^{-3}) \times (0.059/0.50)^{3/2}$ or $1.34 \times 10^{-10}\,cm^3\,s^{-1}$,

which is fairly close to the experimental result of $\sim 10^{-10}\,\mathrm{cm^3\,s^{-1}}$.

More generally, the bimolecular recombination coefficient may be calculated using the expression for B in (7.62) and completed using the expressions for effective density of states N_c given in (5.43) and effective masses m_r and m_v given in (5.C.19) and (5.C.27) within the context of Kane's theory. Taking the Kane energy E_P as being constant for III–Vs, and equal to $\sim 20\,\mathrm{eV}$, we can thus derive the value of the B coefficient as a function of the III–V semiconductor bandgap. Figure 7.5 compares the results of this simple theory with the experimental results given in Table 7.2. We conclude that the agreement is satisfactory. This theory therefore affords a considerable degree of predictive capacity and can be further improved by taking into account additional contributions from the spin-orbit band, etc.

Expression (7.61) constitutes a *law of mass action*: the electrons and holes recombine individually with one another, and so the recombination rate of a particular carrier type is clearly proportional to the density of the other type. We will suppose that the commonly occurring situation consists of a small displacement of the system from equilibrium. The equilibrium carrier densities are given by n_0 and p_0 and are related to each other by the law of mass action $n_0 p_0 = n_i^2$, where n_i is the intrinsic carrier density (5.49). The non-equilibrium carrier densities are then given by:

$$\begin{aligned} n &= n_0 + \Delta n \\ p &= p_0 + \Delta p \end{aligned} \tag{7.63}$$

where Δn and Δp are assumed to be small in comparison with the majority carrier density and (to maintain charge neutrality) $\Delta n = \Delta p$. An additional recombination rate is then given by:

$$\Delta_{\mathrm{spon}} = B(np - n_0 p_0) \approx B(n_0 + p_0)\Delta n \tag{7.64}$$

and the *radiative recombination rate* for non-equilibrium carriers is then:

$$\begin{aligned} R_{\mathrm{rad}} &= \frac{\Delta n}{t_{\mathrm{rad}}} \\ t_{\mathrm{rad}} &= \frac{1}{B(n_0 + p_0)} = \frac{1}{BN_{\mathrm{dop}}} \end{aligned} \tag{7.65}$$

Radiative recombination time in a semiconductor

where N_{dop} is the doping level in the semiconductor. The radiative recombination time t_{rad} may also be written:

$$t_{\mathrm{rad}} = \tau_R \frac{N_c}{N_{\mathrm{dop}}} \left(\frac{m_v}{m_r} \right)^{3/2} \tag{7.66}$$

The *radiative recombination time* t_{rad} is therefore generally rather close to the

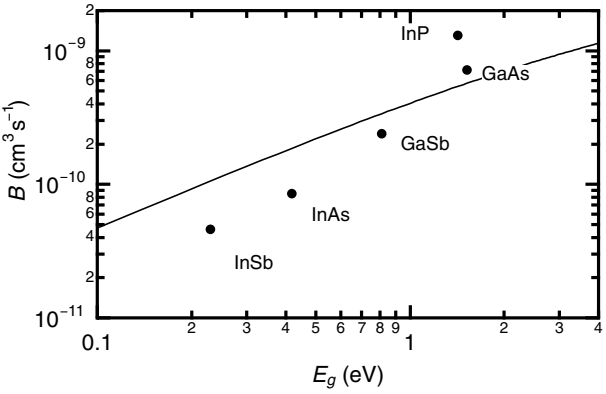


Fig. 7.5. Bimolecular recombination coefficients for different III–V semiconductors. The continuous line represents results generated using (7.62) assuming a constant Kane energy E_P of 20 eV for all semiconductors. The plotted points are experimentally determined values.

spontaneous lifetime τ_R, i.e. of the order of nanoseconds. These two concepts are often confused with one another.

Example
For a $10^{17}\,\text{cm}^{-3}$ GaAs doped crystal, the effective density of states in the conduction band is $N_c = 4.3 \times 10^{17}\,\text{cm}^{-3}$ (Section 5.6.1), the spontaneous lifetime τ_R is 7×10^{-10} s, and the reduced effective mass is $1/(1/0.067) + (1/0.50) = 0.059$. The radiative lifetime is therefore:

$$t_{\text{rad}} = 7 \times 10^{-10}\,\text{s} \times (4.3 \times 10^{17}\,\text{cm}^{-3}/10^{17}\,\text{cm}^{-3}) \times (0.5/0.059)^{3/2} \text{ or } t_{\text{rad}} = 5\,\text{ns}.$$

7.6 Conditions for optical amplification in semiconductors

The analysis in Section 7.3 has shown us that a semiconductor driven from thermodynamic equilibrium can provide optical gain to photons with energies satisfying the Bernard–Durrafourg condition (7.26b). It is therefore important to uncover the experimental circumstances which will allow this condition to be achieved.

Assume that a significant number of carriers (Δn and Δp) are introduced under non-equilibrium conditions so that the quasi-Fermi levels are given by:

$$E_{Fc} = E_c + kT\tilde{F}_{1/2}\left(\frac{n}{N_c}\right)$$
$$E_{Fv} = E_v - kT\tilde{F}_{1/2}\left(\frac{p}{N_v}\right) \tag{7.67}$$

where $\tilde{F}_{1/2}$ is the inverse function of $F_{1/2}$, the integral of the Fermi–Dirac function of order 1/2 is defined (see Section 5.6) as:

$$F_{1/2}(u) = \frac{1}{\Gamma(3/2)} \int_0^\infty \frac{x^{1/2}}{1 + e^{(x-u)}} dx \tag{7.68}$$

We will suppose the semiconductor is lightly doped, and that $n = p \gg n_0$ and p_0. The material will then no longer absorb photons with energy $h\nu = E_g$ once the non-equilibrium carrier density n_{transp} satisfies the following *transparency condition* (see (7.26b)):

$$\tilde{F}_{1/2}\left(\frac{n_{\text{transp}}}{N_c}\right) + \tilde{F}_{1/2}\left(\frac{n_{\text{transp}}}{N_v}\right) > 0 \tag{7.69}$$

Transparency condition for a bulk semiconductor

Figure 7.6 shows the difference $E_{Fc} - E_{Fv}$ as a function of non-equilibrium carrier density n. The transparency density is achieved when $E_{Fc} - E_{Fv} = E_g$ (equivalent to the transparency condition in (7.69)).

Example

The MATHEMATICA program below gives an example of the behaviour of quasi-Fermi levels in GaAs as a function of non-equilibrium carrier density. The transparency condition is attained for $n_{\text{transp}} = 1.2 \times 10^{18}\,\text{cm}^{-3}$.

```
gamm32=Sqrt[N[Pi]]/2
f[x__,u__]=Sqrt[x]/(1+Exp[x−u])
tifermi=Table[{NIntegrate[f[x,u],{x,0,Infinity}]/gamm32,u},{u,-5,5,0.2}]
ifermi=Interpolation[tifermi]
Nc=4.7*10^17 (*in cm-3*)
Nv=7 10^18 (*in cm-3*)
efc[n__]=ifermi[n/Nc]
efv[p__]=ifermi[p/Nv]
deltaef[x__]= 25.9*(efc[x]+efv[x])
Plot[deltaef[x],{x,10^16,2 10^18}]
```

Above this transparency threshold, the medium begins to amplify those photons possessing energies which satisfy the Bernard–Durrafourg condition. The gain spectrum is then given by:

$$\gamma(h\nu) = K_{\text{abs}}(h\nu - E_g)^{1/2}[f_c(h\nu) - f_v(h\nu)] \tag{7.70}$$

where K_{abs} is given by (7.52b), and the Fermi–Dirac functions $f_c(h\nu)$ and $f_v(h\nu)$ are the occupation rates for the levels in the conduction and valence bands satisfying $E_c - E_v = h\nu$, and given by (7.25a) and (7.25b). Figure 7.7 shows how the gain curve evolves as a function of non-equilibrium carrier density in GaAs. The greater

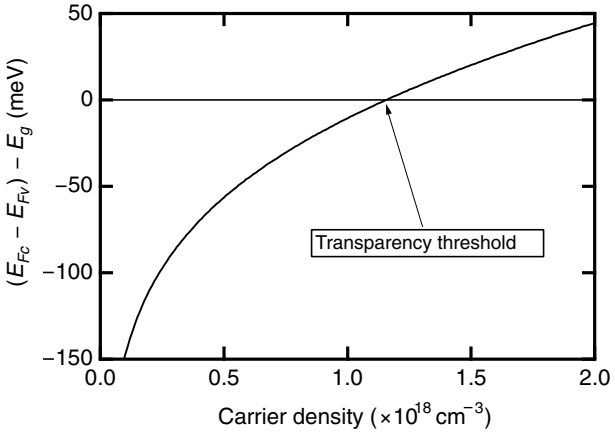


Fig. 7.6. Evolution of the quasi-Fermi level difference as a function of non-equilibrium carrier density in GaAs. The transparency threshold is attained once the Bernard–Durrafourg condition, with $h\nu = E_g$, has been satisfied.

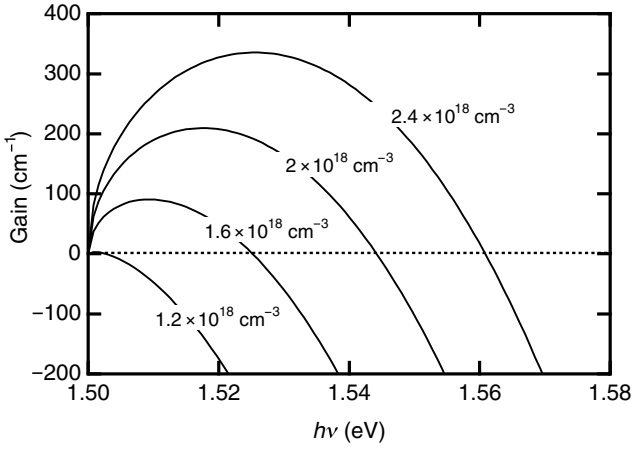


Fig. 7.7. Evolution of the gain curve in GaAs as a function of the density of electron–hole pairs. Above a concentration of 1.2×10^{18} cm^{-3}, the medium provides optical gain over the range of photon energies satisfying the Bernard–Durrafourg condition.

the amount by which this density exceeds the transparency threshold, the broader the corresponding gain curve becomes. Figure 7.8 shows how the maximum gain varies as a function of non-equilibrium carrier density. Once transparency has been achieved, the maximum gain increases linearly with charge density above threshold. This purely phenomenological relationship is very useful in modelling the behaviour of semiconductor lasers (see Chapter 13).

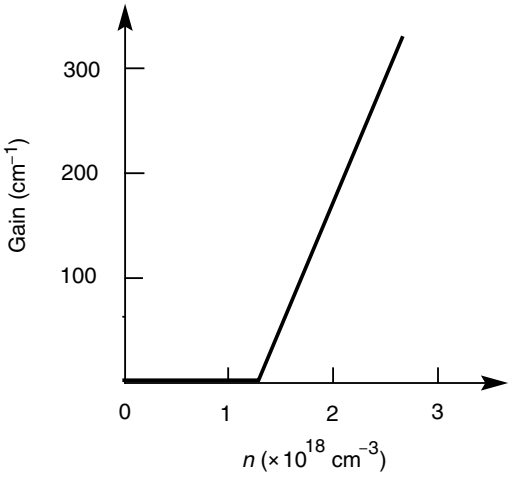


Fig. 7.8. Evolution of the gain curve maximum in GaAs as a function of electron–hole pair density.

Example

Below, we give a MATHEMATICA program which describes the behaviour of amplification in a non-equilibrium semiconductor. We encourage the reader to try out this program using numerical constants for the different semiconductors listed in the table on p. xvii.

```
gamm32=Sqrt[N[Pi]]/2;
f[x_,u_]=Sqrt[x]/(1+Exp[x-u]);
tifermi=Table[{NIntegrate[f[x,u],{x,0,Infinity}]/gamm32,u},{u,-5,5,0.2}]
ifermi=Interpolation[tifermi];
kb=8.63 10^-5;T=300;
Nc=4.7*10^17 (*in cm-3*);Nv=7. 10^18 (*in cm-3*);
Eg=1.5;mc=0.067;mv=0.46;Kabs=12000;
mr=1/(1/mc+1/mv);
n=1.2 10^18;
nc=n/Nc;efc= 25.9 10^-3 *ifermi[nc]
nv=n/Nv;efv=25.9 10^-3 *ifermi[nv]
Ec=mr/mc*(hnu-Eg);
Ev=-mr/mv*(hnu-Eg);
fc=1/(Exp[(Ec-efc)/(kb*T)]+1);
fv=1/(Exp[(Ev+efv)/(kb*T)]+1);
gain=Kabs*(hnu-Eg)(1/2)*(fc-fv);
plot12=Plot[gain,{hnu,1.5,1.6}]
```

FURTHER READING

A. Kireev, *Physique des Semiconductors*, MIR, Moscow (1975).
J. I. Pankove, *Optical Processes in Semiconductors*, Dover Publications, New York (1971).
B. K. Ridley, *Quantum Processes in Semiconductors*, Clarendon Press, Oxford (1988).
B. Sapoval and C. Hermann, *Physics of Semiconductors*, Springer, Berlin (1995).
K. Seeger, *Semiconductor Physics*, 3rd Edn, Springer, Berlin (1985).
P. Yu and M. Cardona, *Fundamentals of Semiconductors*, Springer, Berlin (1995).

Complement to Chapter 7

7.A The Franz–Keldysh-effect electromodulator

When we apply a static electric field to a semiconductor, the Bloch wavefunctions do not represent the stationary states of the crystal as the potential no longer possesses the periodicity of the lattice spacing. This corresponding change in the wavefunctions has a profound effect on optical absorption. The semiconductor can now absorb photons possessing an energy $\hbar\omega$ inferior to bandgap E_g. Furthermore, above the gap, the absorption spectrum displays oscillations superimposed upon the usual spectral features in $(\hbar\omega - E_g)^{1/2}$ obtained for $F = 0$. This is the *Franz–Keldysh effect*. As we shall see later, this effect is successfully harnessed in electromodulators. We will begin by presenting the formalism which will allow us to describe this effect.

We will arbitrarily assume that the electric field is oriented in the z direction. Schrödinger's equation for an electron is then:

$$\left[\frac{p^2}{2m_0} + V_c(\mathbf{r}) - eFz\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{7.A.1}$$

We may expand the wavefunction ψ in terms of Bloch basis functions $|n\mathbf{k}\rangle = u_{n,\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$ (solutions to the electric field free equation) as:

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{V}}\sum_{n\mathbf{k}} a_n(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r}) \tag{7.A.2}$$

and then project the resulting equation upon the basis functions $|N\mathbf{K}\rangle$, yielding:

$$\varepsilon_N(\mathbf{k})a_N(\mathbf{k}) - eF\sum_{n\mathbf{k}}\langle N\mathbf{K}|z|n\mathbf{k}\rangle a_n(\mathbf{k}) = Ea_N(\mathbf{k}) \tag{7.A.3}$$

the second term in this last equation (due to the field) may be transformed in the following manner:

$$\sum_{n\mathbf{k}}\langle N\mathbf{K}|z|n\mathbf{k}\rangle a_n(\mathbf{k}) = \frac{1}{V}\sum_{n\mathbf{k}}\int \mathrm{d}\mathbf{r}e^{-i\mathbf{K}\cdot\mathbf{r}}u^*_{N\mathbf{K}}(\mathbf{r})za_n(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})$$

$$
\begin{aligned}
&= \sum_n \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^3} \int \mathrm{d}\mathbf{r} \mathrm{e}^{-\mathrm{i}\mathbf{K}\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) \frac{1}{\mathrm{i}} \frac{\partial}{\partial k_z} [a_n(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r})] \\
&\quad - \sum_n \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^3} \int \mathrm{d}\mathbf{r} \mathrm{e}^{-\mathrm{i}\mathbf{K}\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \frac{1}{\mathrm{i}} \frac{\partial a_n(\mathbf{k})}{\partial k_z} \\
&\quad - \sum_n \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^3} \int \mathrm{d}\mathbf{r} \mathrm{e}^{-\mathrm{i}\mathbf{K}\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} a_n(\mathbf{k}) \frac{1}{\mathrm{i}} \frac{\partial u_{n\mathbf{k}}(\mathbf{k})}{\partial k_z}
\end{aligned}
\tag{7.A.4}
$$

The first term cancels out as the integral over k_z is a periodic function in k_z having a period of $2\pi/a_z$, where a_z is the primitive cell spacing. The other terms give, using the periodicity of $u_{n\mathbf{k}}(\mathbf{r})$ in $\mathbf{r}$ and the orthogonality of the basis:

$$
\begin{aligned}
&\sum_{n,\mathbf{k}} \langle N\mathbf{K}|z|n\mathbf{k}\rangle a_n(\mathbf{k}) \\
&= \mathrm{i} \frac{\partial a_N(\mathbf{K})}{\partial k_z} + \sum_{n\mathbf{k}} \mathrm{i} a_n(\mathbf{k}) \Delta_{\mathbf{K}\mathbf{k}} \int \mathrm{d}\mathbf{r} u^*_{N\mathbf{K}}(\mathbf{r}) \frac{\partial u_{n\mathbf{k}}(\mathbf{r})}{\partial k_z} \\
&= \mathrm{i} \frac{\partial a_N(\mathbf{K})}{\partial k_z} + \sum_n a_n(\mathbf{K}) z_{Nn}(\mathbf{K})
\end{aligned}
\tag{7.A.5}
$$

with

$$
z_{Nn}(\mathbf{K}) = \frac{\mathrm{i}}{V} \int \mathrm{d}\mathbf{r} u^*_{N\mathbf{K}}(\mathbf{r}) \frac{\partial u_{n\mathbf{K}}(\mathbf{r})}{\partial K_z} = \frac{\mathrm{i}}{V_{\text{cell}}} \int_{\text{cell}} \mathrm{d}\mathbf{r} u^*_{N\mathbf{K}}(\mathbf{r}) z u_{N\mathbf{K}}(\mathbf{r})
$$

Schrödinger's equation for the coefficients $a_n(\mathbf{k})$ then becomes:

$$
-\mathrm{i}eF \frac{\partial a_N(\mathbf{K})}{\partial K_z} - eF \sum_n z_{Nn}(\mathbf{K}) a_n(\mathbf{K}) = [E - \varepsilon_N(\mathbf{K})] a_N(\mathbf{K}) \tag{7.A.6}
$$

Schrödinger's equation for a crystal in the presence of an electric field

This last equation turns up in all problems involving the application of an electrostatic field superimposed upon a crystal potential, as in a tunnel effect across a gap (or *Zener effect*).

Following a perturbative approach, we will seek a solution close to $\varepsilon_N(\mathbf{0})$; we may first neglect the influence of the terms contributed by the bands $n \neq N$ (the bands are then decoupled). The term in z_{NN} remaining in the sum could be moved over to the right-hand side, but we will elect to neglect it, and hope that $a_N(\mathbf{K})$ will only be important for $\mathbf{K} \cong \mathbf{0}$ and $z_{NN}(\mathbf{0}) = 0$ for an extremum in band N.

For $E \cong \varepsilon_N(0)$, the solution to (7.A.6) is:

$$
\begin{aligned}
a_N(\mathbf{K}) &= a_N(\mathbf{K}_{\|}, K_z) \\
&= a_N(\mathbf{K}_{\|}, 0)\exp\left\{\frac{\mathrm{i}}{eF}\int_0^{K_z}[E - \varepsilon_N(\mathbf{K}')]\mathrm{d}K_z'\right\} \\
&= a_N(\mathbf{K}_{\|}, 0)\exp\left[\frac{\mathrm{i}}{eF}\left\{\left[E - \varepsilon_N(\mathbf{0}) - \frac{\hbar^2 K_{\|}^2}{2m_N}\right]K_z + \frac{\hbar^2 K_z^3}{6m_N}\right\}\right]
\end{aligned}
\tag{7.A.7}
$$

where we have introduced the effective mass m_N for the Nth band and $\mathbf{K}_{\|} = (K_x, K_y)$. In fact, these approximations are those of the *effective mass approximation*, which we will study in a more general manner in Chapter 8. We are now in a position to calculate the following function:

$$
\zeta_{N\mathbf{K}\|}(z) = \frac{1}{L_z}\sum_{K_z} a_N(K_z)\mathrm{e}^{\mathrm{i}K_z z} = \int \frac{\mathrm{d}K_z}{2\pi} a_N(K_z)\mathrm{e}^{\mathrm{i}K_z z} \tag{7.A.8}
$$

which is none other than the envelope function describing the variation in amplitude of the wavefunction for a stationary state along the direction of the electric field. For this, we will use the following formula (see Eq. (10.4.32) of Abramowitz and Stegun):

$$
(3a)^{-1/3}\pi\mathrm{Ai}[(3a)^{-1/3}x] = \int_0^{\infty}\cos(at^3 + xt)\mathrm{d}t \tag{7.A.9}
$$

We thus obtain the following expression for the envelope function:

$$
\zeta_{N\mathbf{K}}(z) = a_N(\mathbf{K}_{\|}, 0)\mathrm{Ai}\left\{\left(\frac{2m_N}{\hbar^2 e^2 F^2}\right)^{1/3}\left[E - \varepsilon_N(\mathbf{0}) - \frac{\hbar^2 K_{\|}^2}{2m_N} - eFz\right]\right\} \tag{7.A.10}
$$

Envelope function for an electron in a band *N*

In this expression, Ai(z) is the *Airy function*, which by definition is the solution to the equation:

$$
\frac{\mathrm{d}^2\mathrm{Ai}}{\mathrm{d}z^2} - z\mathrm{Ai}(z) = 0 \tag{7.A.11}
$$

and which tends towards zero as $z \to \infty$. Figure 7.A.1 shows the general form of this function. We see that as $z \to \infty$, the function tends towards zero in an essentially exponential fashion. In fact,

$$
\mathrm{Ai}(z) \approx \frac{1}{2\sqrt{\pi}z^{1/4}}\exp\left(-\frac{2}{3}z^{3/2}\right), z \to \infty \tag{7.A.12}
$$

For $z < 0$, the function oscillates with a decreasing period:

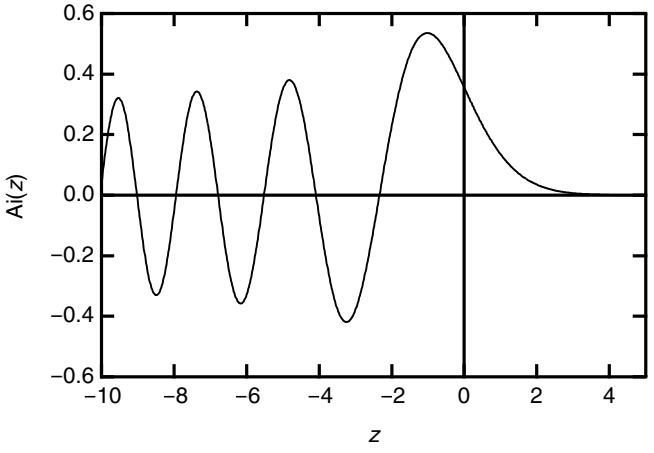


Fig. 7.A.1 The Airy function.

$$\mathrm{Ai}(-z) \approx \frac{1}{2\sqrt{\pi} z^{1/4}} \sin\left(-\frac{2}{3} z^{3/2} + \frac{\pi}{4}\right), z \to \infty \tag{7.A.13}$$

The solution (7.A.10) may seem somewhat complicated at first. We will see, however, that its behaviour can be easily understood. We note at first that the argument of the Airy function in (7.A.10) is 0 for $z_0 = (E - \varepsilon_N(0) - \hbar^2 K_{\|}^2/2m_N)/eF$. For $z < z_0$, the function decreases exponentially. Figure 7.A.2 shows that this effect is similar to a tunnel effect wherein the wavefunction penetrates into the semiconductor gap. For $z > z_0$, the function oscillates more and more rapidly, corresponding to a situation of increasing kinetic energy (see Fig. 7.A.1). Furthermore, the solutions for other energies are the same, except that they are translated by $\Delta z_0 = \Delta E/eF$. One consequence of this property is that the normalization constant $a_N(0)$ does not depend on $K_{\|}$ or E.

We are now able to understand the quantum effects associated with absorption occurring in semiconductors under an applied electric field. Figure 7.A.2 shows two electron wavefunctions in the conduction band, as well as two hole wavefunctions in the valence band. The solutions for the holes in the valence band are equivalent to those for the electrons: we need only replace m_e by $-m_h$, where m_h is the (positive) mass of the holes, which has as a consequence that the direction of motion of the particles in the field changes. Also, we see how the electrons (and holes) can travel across the forbidden gap by tunnelling. The probability of the occurrence of this (*Zener*) effect is tied to the overlap integral between the exponential wavefunction and the valence band facing it.

Similarly, we may guess that it is possible to promote an electron from the valence band to the conduction band by absorbing a photon, even if its energy $\hbar\omega$ is less than E_g. All we need for this to happen is for the two wavefunctions to have a

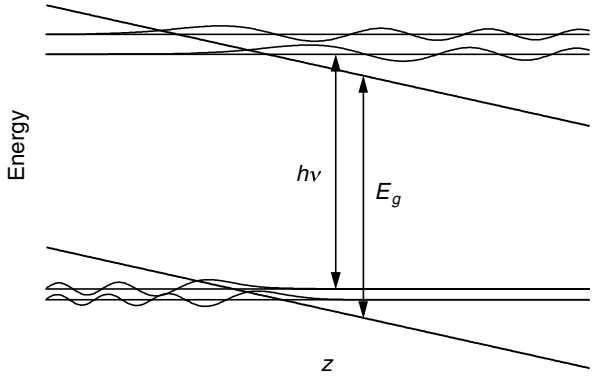


Fig. 7.A.2. In the presence of an electric field, the wavefunctions ζ for the electrons and holes overlap even if they are separated in energy by an amount smaller than the size of the bandgap E_g. This permits optical absorption of photons with energy less than the forbidden gap.

non-zero overlap. For a fixed value of $\hbar\omega$, this overlap decreases exponentially as $F \to 0$.

More precisely, the transition rate between a state $|\mathbf{K}_{\|}, E_1\rangle$ in the valence band and a state $|\mathbf{K}_{\|}, E_2\rangle$ in the conduction band is given by Fermi's golden rule to be:

$$S(1 \to 2) = \frac{\pi}{\hbar}|\langle 2|e\mathbf{E}_\omega \cdot \mathbf{r}|1\rangle|^2 \delta(E_2 - E_1 - \hbar\omega) \tag{7.A.14}$$

If we substitute into this last expression the wavefunctions obtained in (7.A.10), we find in an analogous fashion to (7.10) the optical matrix element:

$$\langle 2|e\mathbf{E}_\omega \cdot \mathbf{r}|1\rangle = \Delta_{\mathbf{K}\mathbf{K}'}|\langle u_c|e\mathbf{E}_\omega \cdot \mathbf{r}|u_v\rangle|^2 \sum_{K_z} a_c^*(K_z)a_v(K_z) \tag{7.A.15}$$

where the part which has changed with respect to the $F = 0$ case is the sum:

$$\begin{aligned} I_z &= \sum_{K_z} a_c^*(K_z)a_v(K_z) \\ &= \int \frac{\mathrm{d}K_z}{2\pi} a_c^*(0)a_v(0)\exp\left\{\frac{-\mathrm{i}}{eF}\left[\left(E_2 - E_1 - E_g - \frac{\hbar^2 K_{\|}^2}{2m_r}\right)K_z + \frac{\hbar^2 K_z^3}{6m_r}\right]\right\} \end{aligned} \tag{7.A.16}$$

where m_r is the reduced mass. From (7.A.9) we obtain once again an Airy function:

$$I_z = a_c^*(0)a_v(0)\left(\frac{2m_r eF}{\hbar^2}\right)^{1/3} \mathrm{Ai}\left[-\left(\frac{2m_r}{\hbar^2 e^2 F^2}\right)^{1/3}\left(E_2 - E_1 - E_g - \frac{\hbar^2 K_{\|}^2}{2m_r}\right)\right] \tag{7.A.17}$$

The essential point is that I_z is different from 0 even when $E_2 - E_1 = \hbar\omega$ is smaller than E_g. This is the term that is responsible for the Franz–Keldysh effect. To calculate the total transition rate at a given frequency, we need to sum over all

initial states, i.e. over $\mathbf{K}_{\|}$ and E_1, the final energy E_2 being equal to $E_1 + \hbar\omega$. To do so, we will temporarily drop the constants and pick them up again later by forcing a correspondence with the result for $F = 0$. Introducing the Franz–Keldysh *characteristic energy*:

$$\beta = \left(\frac{2m_r}{\hbar^2 e^2 F^2}\right)^{-1/3} \tag{7.A.18a}$$

Franz–Keldysh characteristic energy

and the dimensionless variable:

$$\xi = \frac{\hbar\omega - E_g}{\beta} \tag{7.A.18b}$$

the sum over $\mathbf{K}_{\|}$ yields:

$$\begin{aligned}\sum_{\mathbf{K}\|} |I_z|^2 &= \int \frac{\mathrm{d}\mathbf{K}_{\|}}{(2\pi)^2} |I_z|^2 \propto \int_0^\infty \mathrm{Ai}^2\left(\frac{E_{\|}}{\beta} - \xi\right) \mathrm{d}E_{\|} = \beta \int_{-\xi}^\infty \mathrm{Ai}^2(z)\mathrm{d}z \\ &= \beta[\mathrm{Ai}'^2(-\xi) + \xi \mathrm{Ai}^2(-\xi)]\end{aligned}$$

This last integral may be verified by taking its derivative and by using the equation for the definition of Ai(z): $\mathrm{Ai}''(z) = z\mathrm{Ai}(z)$.

For $F \to 0$ the absorption as a function of $\hbar\omega$ must agree with the previously obtained result (Eq. (7.24)):

$$\alpha(\hbar\omega, F = 0) = \alpha_b \sqrt{\hbar\omega - E_g} \tag{7.A.19}$$

Taking $F \to 0$ and $\hbar\omega < E_g, \xi \to -\infty$, (7.A.18) tends towards zero, as it must. For $F \to 0, \hbar\omega > E_g, \xi \to \infty$ and we must use the asymptotic form (Eq. (7.A.13)), which leads to:

$$\alpha(\hbar\omega, F = 0) = A\frac{\beta}{\pi}\xi^{1/2} = A\frac{\beta}{\pi}\sqrt{\frac{\hbar\omega - E_g}{\beta}}, F \to 0 \tag{7.A.20}$$

which behaves essentially as Eq. (7.A.19). The comparison between (7.A.19) and (7.A.20) therefore determines A, with the final result being:

$$\alpha(\hbar\omega, F) = \alpha_b \beta^{1/2} \pi[\mathrm{Ai}'^2(-\xi) + \xi \mathrm{Ai}^2(-\xi)] \tag{7.A.21}$$

Absorption in the presence of an electric field

Figure 7.A.3 shows the absorption in a semiconductor under an applied field. A non-zero absorption coefficient is observed for photons possessing energies below the bandgap, while oscillations in the absorption as a function of photon energy are apparent for $\hbar\omega > E_g$.

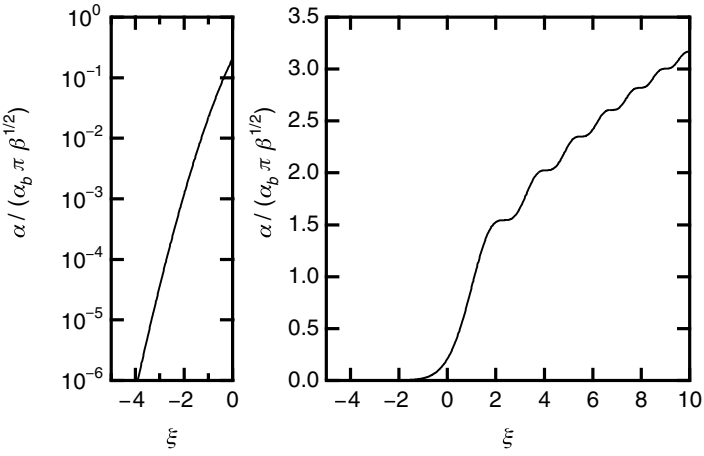


Fig. 7.A.3. Absorption under the influence of an electric field as a function of photon energy separation from the bandgap $\xi = (\hbar\omega - E_g)/\beta$, where $\beta = (\hbar^2 e^2 F^2 / 2m_r)^{1/3}$. The response below the gap (left panel) is plotted on a logarithmic scale.

Example

Under an applied field strength of $F = 10\,\mathrm{kV\,cm^{-1}}$ and $m_r = m_e m_h/(m_e + m_h) = 0.059 m_0$ (for GaAs), we obtain a characteristic Franz–Keldysh energy of:

$$\beta = \left(\frac{\hbar^2 e^2 F^2}{2m_r}\right)^{1/3} = \left[\frac{(1.05 \times 10^{-34})^2 (10^6)^2}{2 \times 0.059 \times 9.1 \times 10^{-31} \times 1.6 \times 10^{-19}}\right]^{1/3} \mathrm{eV} = 8.6\,\mathrm{meV}$$

The absorption at $\hbar\omega - E_g = -8.6\,\mathrm{meV}$ is then (see Fig. 7.A.3) $f(-1) = 0.022$ or 2.2% of what it would be for $\hbar\omega - E_g = +8.6\,\mathrm{meV}$ in the absence of an applied field.

This effect is harnessed in *Franz–Keldysh-effect modulators*. The governing principle is quite simple (see Fig. 7.A.4). On a semiconductor waveguide (with bandgap E_g), we deposit a metallic electrode to form a Schottky barrier with the semiconductor (see Chapter 10). By applying a voltage V between this top electrode and the ohmic contact on the back of the semiconductor, we introduce an electric field $F = V/d$, where d is the separation between the upper and lower contacts (the semiconductor doping level is kept sufficiently low to minimize any potential drop due to space charge effects). Applying a 5 V potential across a 1 μm contact separation leads to an electric field strength of $50\,\mathrm{kV\,cm^{-1}}$, or a characteristic energy β of 25 meV. As a result, waveguided photons with energy $E_g - 25\,\mathrm{meV}$ will be absorbed by a factor of 100 times more than under zero-field conditions. We may then easily see how we might use such a device to transform an electrical (time-varying) signal into an optical one. Such a device is referred to as a

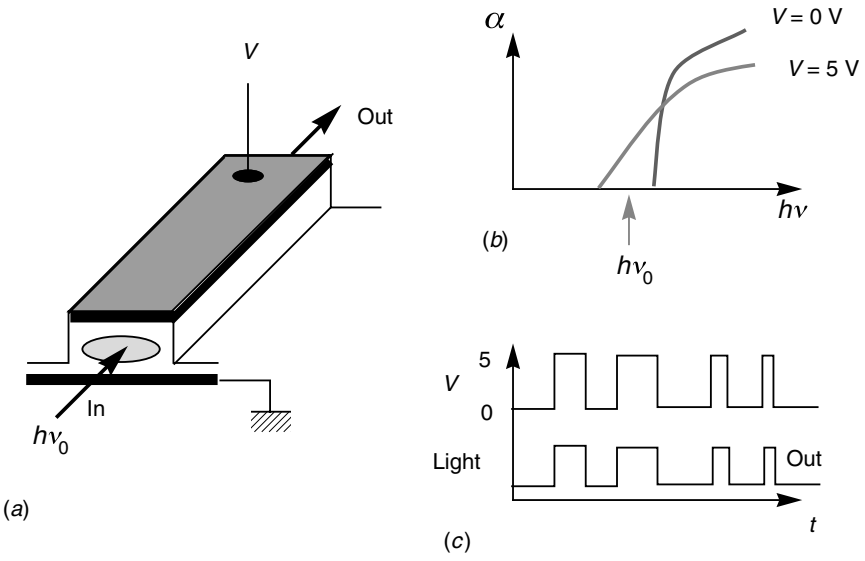


Fig. 7.A.4. Operation of an electro-optic modulator based upon the Franz–Keldysh effect.

Franz–Keldysh electromodulator. This type of modulator is extremely rapid (with a maximum operational frequency in the tens of GHz range) as it does not require the bulk displacement of electrical charges to operate. This is not, for instance, the case with SEED (self-electro-optic effect devices) modulators (see Complement 8.C).

FURTHER READING

M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York, Dover (1970).

H. Haug and S. W. Koch, *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, Singapore, World Scientific (1993).

J. I. Pankove, *Optical Processes in Semiconductors*, Dover Publications, New York (1971).

7.B Optical index of semiconductors

Knowledge of the optical index of semiconductors turns out to be of fundamental importance in the implementation of optoelectronic components. This is a fairly complex subject as the models differ significantly depending upon the wavelength ranges being considered. In particular, we distinguish on the basis of the behaviour of two optical regions.

7.B.1 Mid- and far-infrared regions

The interaction between a semiconductor and an electromagnetic wave in the mid- and far-infrared regions brings into play the material's optical phonons (see Complement 6.B). The far-infrared regime is called the *Reststrahlen region* where the material becomes overwhelmingly absorbent and dispersive. In GaAs, for example, this region is situated around 30 μm.

In this range of wavelengths, the absorption and the optical index are well described by a model employing harmonic oscillators with eigenfrequencies equalling those of transverse optical phonons (ν_{TO}). This model requires four parameters: the static dielectric constant ε_s, the dielectric constant at infinite frequency ε_∞, the resonant frequency ν_{TO}, and the damping coefficient γ. A trivial extension of the *Lorentz model* then gives for the dielectric constant:

$$\varepsilon = \varepsilon_1 - i\varepsilon_2 = (n_{op} - ik_{op})^2 = \varepsilon_\infty + (\varepsilon_0 - \varepsilon_\infty)\frac{\nu_{TO}^2}{\nu_{TO}^2 - \nu^2 + i\gamma\nu} \tag{7.B.1}$$

In GaAs, the optical phonon energy is $h\nu_{TO} = 33.25\,\text{meV}$, the relevant dielectric constants are $\varepsilon_\infty = 10.88$, $\varepsilon_s - \varepsilon_\infty = 1.97$, and the damping coefficient is $\gamma \approx \nu_{TO}/133$. Far from resonance (for photon energies in excess of 33 meV, relation (7.B.1) may be simplified easily to yield:

$$n_{op} \approx \sqrt{\varepsilon_1} \approx \left\{\varepsilon_\infty - (\varepsilon_0 - \varepsilon_\infty)\frac{1}{[(\nu/\nu_{TO})^2 - 1]}\right\}^{1/2} \tag{7.B.2a}$$

Sellmeier's equation

Table 7.B.1 gives the various parameters required to perform these calculations in the GaAs/AlGaAs system.

Table 7.B.1. *Dielectric constants and transverse optical phonon energies in* $Al_xGa_{1-x}As$ *as a function of compositional fraction* x

Static dielectric constant, ε_s	$13.18–3.12x$
HF dielectric constant, ε_∞	$10.89–2.73x$
Optical phonon energy, $h\nu_{TO}$	$33.29–0.64x–1.16x^2$

In fact, many phenomenological corrections to Sellmeier's formula exist, with the most precise for $Al_xGa_{1-x}As$ being (with $h\nu$ in eV):

$$n_{op} \approx \left\{7.10 + \frac{3.78}{1 - 0.180(h\nu)^2} - \frac{1.97}{[(30.08h\nu)^2 - 1]}\right\}^{1/2} - 0.528x \tag{7.B.2b}$$

Figure 7.B.1 shows the variation in optical index for GaAs between 5 and 12 μm.

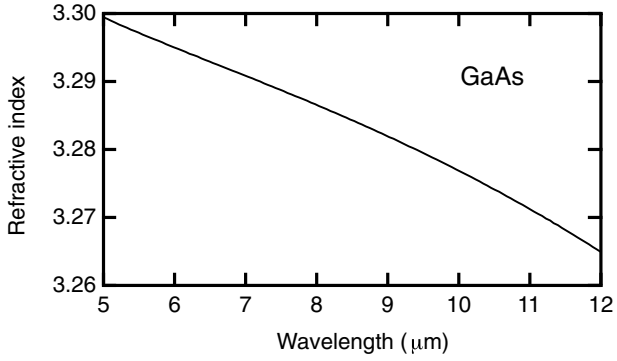


Fig. 7.B.1. Optical index in the mid-infrared region for GaAs obtained using Sellmeier's extended equation (7.B.2b).

7.B.2 Near gap regime

Clearly, the optical properties of a material in this wavelength range are dominated by the absorption band of the semiconductor. The simplest approach which takes this effect into account is the *effective harmonic oscillator model* developed by Wemple and DiDomenico. This approach consists of combining the ensemble of transitions between the valence and conduction bands into a single effective oscillator transition with a resonant frequency v_{eo} and oscillator strength $\pi E_d/2$ (for historical reasons E_d is called the *dispersion energy*). Sellmeier's equation may then be written:

$$n_{op}^2 = 1 + \frac{E_{eo}E_d}{E_{eo}^2 - (hv)^2} \tag{7.B.3}$$

This relationship holds surprisingly well for a large number of important semiconducting and insulating materials. Nonetheless, agreement between theory and experiment is insufficient for optoelectronics applications where precise control over the optical indices is of crucial importance. We therefore return to a more phenomenological approach.

To do so, we recall that the real ε_1 and imaginary ε_2 parts of the dielectric constant are related through the *Kramers–Kronig relation*:

$$\varepsilon_1(E) = 1 + \frac{2}{\pi}\mathrm{PV}\int_0^\infty \frac{E'\varepsilon_2(E')}{E'^2 - E^2}\,\mathrm{d}E' \tag{7.B.4}$$

The Kramers–Kronig relation

where PV signifies the 'principal value'. We note that the integral in (7.B.4) shows that although absorption only brings into play energies which couple to the photon energy, the real portion of the dielectric constant ε_1 is influenced by the

entire band structure, rendering its calculation more difficult. Equation (7.B.4) can be expanded as a limited series:

$$\varepsilon_1(E) = 1 + \frac{2}{\pi}\int_{E_g}^{\infty} \varepsilon_2(E')\left(\frac{1}{E'} + \frac{E^2}{E'^3} + \frac{E^4}{E'^5} + \cdots\right)\mathrm{d}E' \tag{7.B.5}$$

or again:

$$\varepsilon_1(E) = 1 + \mathrm{M}_{-1} + \mathrm{M}_{-3}E^2 + \mathrm{M}_{-5}E^4 + \cdots \tag{7.B.6a}$$

where the moments M_i are given by:

$$M_i = \frac{2}{\pi}\int_{E_g}^{\infty} \varepsilon_2(E)E^i\mathrm{d}E \tag{7.B.6b}$$

We therefore seek an expression for $\varepsilon_2(E)$ (which is nothing else but the absorption in the material). Different models exist which all suffer a fair degree of arbitrariness as they incorporate phenomenological adjustments to the semiconductor absorption curve (usually quite complex). We will content ourselves here with a presentation of the *Afromowitz model* as it is the one best suited to applications.

This model involves approximating $\varepsilon_2(E)$ by the function:

$$\varepsilon_2(E) = \begin{cases} \eta E^4, E_g < E < E_f \\ 0, \text{elsewhere} \end{cases} \tag{7.B.7}$$

We will see shortly how the parameters relate to E_{eo} and E_d. Figure 7.B.2 shows the level of agreement with GaAs.

This last equation allows one to calculate the moments M_{-1} and M_{-3} in (7.B.6b):

$$\begin{aligned} M_{-1} &= \frac{\eta}{2\pi}(E_f^4 - E_g^4) \\ M_{-3} &= \frac{\eta}{\pi}(E_f^2 - E_g^2) \end{aligned} \tag{7.B.8}$$

Comparing (7.B.3) with (7.B.6), we see that the parameters η and E_f are tied to the dispersion energy E_d and to the effective oscillator energy E_{eo} by:

$$\begin{aligned} E_f &= (2E_{\text{eo}}^2 - E_g^2)^{1/2} \\ \eta &= \frac{\pi}{2}\frac{E_d}{E_{\text{eo}}^3(E_{\text{eo}}^2 - E_g^2)} \end{aligned} \tag{7.B.9}$$

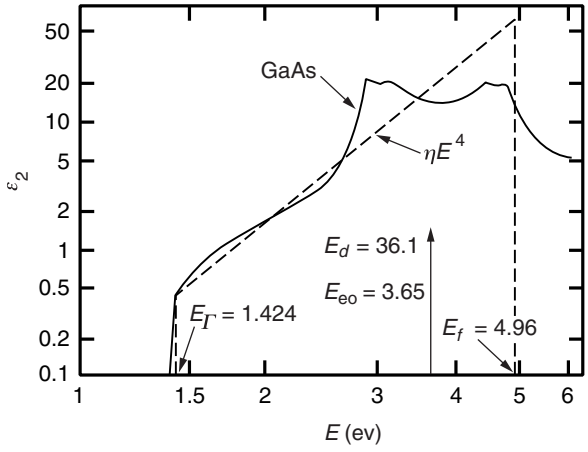


Fig. 7.B.2. ε_2 as a function of photon energy in GaAs and its approximation by (7.B.7) (reprinted *Solid State Comm.* **15**, M. A. Afromowitz. Refractive index of $Ga_{1-x}Al_xAs$, p. 59 (1974), with permission from Elsevier Science).

In fact, substituting the formula for ε_2 (7.B.7) into the Kramers–Kronig relation, we finally obtain:

$$\varepsilon_1(E) = 1 + M_{-1} + M_{-3}E^2 + \frac{\eta}{\pi}\ln\left[\frac{E_f^2 - E^2}{E_g^2 - E^2}\right]E^4 \qquad (7.B.10)$$

The Afromowitz relation

The ensemble of Eqs. (7.B.8), (7.B.9), and (7.B.10) allows the calculation of refraction indices for different semiconductors (comparing well with experimental values) once the values for E_g, E_{eo}, and E_d have been determined. Table 7.B.2 gives these values for three ternary semiconductor material systems: AlGaAs, GaAsP, and GaInP. Figure 7.B.3 shows the calculated results for the optical indices across a range of AlGaAs alloy compositions.

Table 7.B.2. *Effective oscillator energies* E_{eo}, *dispersion energies* E_d, *and forbidden gaps* E_g *for different semiconductor alloys*

	$Ga_{1-x}Al_xAs$	$GaAs_{1-x}P_x$	$Ga_xIn_{1-x}P$
E_{eo} (eV)	$3.65 + 0.871x + 0.179x^2$	$3.65 + 0.721x + 0.139x^2$	$3.391 + 0.524x + 0.595x^2$
E_d (eV)	$36.1 - 2.45x$	$36.1 + 0.35x$	$28.91 + 7.54x$
E_g (eV)	$1.424 + 1.266x + 0.26x^2$	$1.441 + 1.091x + 0.21x^2$	$1.34 + 0.668x + 0.758x^2$

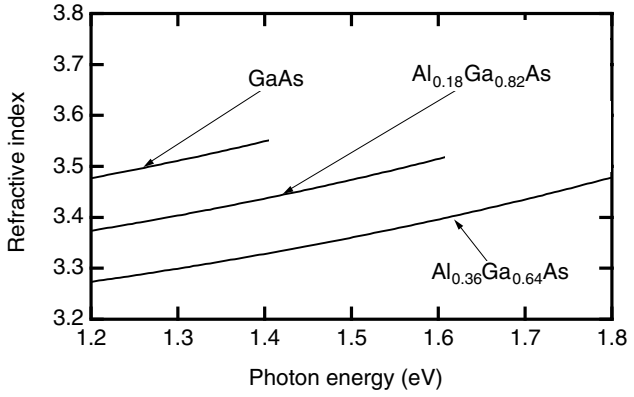


Fig. 7.B.3. Optical indices as a function of photon energy for different AlGaAs compositions obtained using Afromowitz' formula (7.B.10).

FURTHER READING

M. A. Afromowitz, *Solid State Comm.* **15**, 59 (1974).
J. S. Blakemore, *J. Appl. Phys.* **53**, R123 (1982).
S. H. Wemple and M. DiDomenico, *Phys. Rev.* **B3**, 1338 (1971).

7.C Free-carrier absorption

One of the most astounding aspects of semiconductors is the capacity of their constituent electrons to either absorb or emit light by shuttling between the conduction and valence bands in what are referred to as *interband transitions*. There exists yet another type of optical transition involving states within a single band called *intraband transitions*. We will begin by giving a phenomenological description before moving on to discuss the microscopic (and therefore quantum) origin of these transitions.

We consider a semiconductor with an index of refraction $n_{\text{op}}(\varepsilon_0 n_{\text{op}}^2 = \varepsilon)$. The absorption due to free carriers is taken into account by an electrical conductivity σ. The semiconductor is then subjected to an electromagnetic wave $\mathbf{E}$ with frequency ω, propagating in the z direction and incident upon its surface at $z = 0$. This causes an electric current $\mathbf{j} = \sigma\mathbf{E}$ to propagate within the bulk. We will suppose in this first part that the electrical conductivity is independent of frequency ω.

In the semiconductor, the electromagnetic wave is a solution to Maxwell's equations (see (2.1)):

$$\begin{aligned}\nabla \times \mathbf{B} &= \mu_0 \mathbf{j} + \mu_0 \frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t} \\ \nabla \times \mathbf{E} &= -\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}\end{aligned} \tag{7.C.1}$$

where $\mathbf{D}$ is the displacement vector related to the electric field $\mathbf{E}$ and the polarization $\mathbf{P}$ by:

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} = \varepsilon_0 \mathbf{E} + \varepsilon_0 \chi^{(1)} \mathbf{E} = \varepsilon \mathbf{E} \tag{7.C.2}$$

As seen in Chapter 2, the system of Eqs. (7.C.1) is enriched by $\nabla \cdot \mathbf{B} = 0$ and, assuming the medium to be neutral ($\rho = 0$), by Poisson's law $\nabla \cdot \mathrm{E} = 0$. Making use of (7.C.1), (7.C.2), and these two null divergences, Maxwell's equations lead to the simple wave equation:

$$\nabla^2 \mathbf{E} = \mu_0 \sigma \frac{\mathrm{d}\mathbf{E}}{\mathrm{d}t} + \mu_0 \varepsilon \frac{\mathrm{d}^2 \mathbf{E}}{\mathrm{d}t^2} \tag{7.C.3}$$

The dispersion relation $\omega(k)$ relating the wavevector and the frequency may be obtained by substituting the wave equation $E = E_0 \mathrm{e}^{\mathrm{i}(\omega t - kz)}$ into (7.C.3) giving:

$$k^2 = \mu_0 \varepsilon \omega^2 - \mathrm{i}\omega \mu_0 \sigma \tag{7.C.4}$$

We therefore find two distinct regimes dependent upon the relative magnitudes of the real and imaginary parts of (7.C.4).

Strong conductivity ($\sigma \gg \varepsilon\omega$)
In this case, the imaginary portion dominates (7.C.4) and the wavevector is given by:

$$k^2 \approx -\mathrm{i}\omega \mu_0 \sigma \tag{7.C.5}$$

We also see that the wave amplitude decreases exponentially with an absorption coefficient given by twice the square root of (7.C.5). The inverse of this coefficient leads to a characteristic distance referred to as the *skin depth* δ and essentially reflects the distance over which the wave is absorbed by free carriers.

$$\delta = \frac{1}{\sqrt{2\mu_0 \omega \sigma}} \tag{7.C.6}$$

Skin depth

Example
We want to construct a plane radar wave reflector in silicon ($\varepsilon_{\mathrm{Si}} = 13$) operating in the X band, i.e. around 10 GHz (see Fig. 7.C.1). In the dark, the silicon layer is highly resistive ($\rho > 1000\,\Omega\,\mathrm{cm}$). Under illumination its conductivity

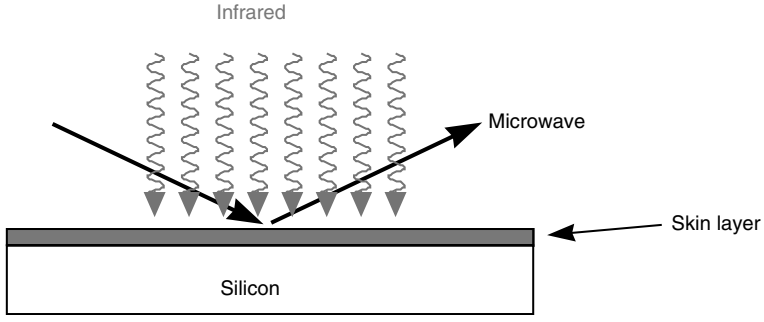


Fig. 7.C.1. An illuminated slab of silicon becomes highly conducting thereby screening an incident microwave over a distance corresponding to the skin depth. A portion of the incident microwave is absorbed over this region, with the remainder of the signal being reflected.

increases. The inequality $\sigma \gg \varepsilon\omega$ is satisfied as soon as the resistivity ρ gets smaller than $1/(2\pi \times 10^{10}\,\mathrm{s}^{-1} \times 13 \times 8.85 \times 10^{-14}\,\mathrm{Fd\,cm}^{-1})$ or $14\,\Omega\,\mathrm{cm}$. This value is easy to reach by illumination. To be specific, let us consider a silicon layer (mobility $\mu = 1000\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$, carrier lifetime $\tau = 1\,\mathrm{ms}$) of $d = 1\,\mu\mathrm{m}$ thickness used as a reflecting layer at 10 GHz. First, in order to reflect the radar wave effectively, the skin depth given by (7.C.6) must be smaller than the silicon layer thickness, or $\sigma > 1/(2 \times 1.2 \times 10^{-6}\,\mathrm{H\,m}^{-1} \times 2\pi \times 10^{10}\,\mathrm{s}^{-1} \times 10^{-12}\,\mathrm{m}^2)$, i.e. $6.6 \times 10^6\,\Omega^{-1}\,\mathrm{m}^{-1}$. This resistivity corresponds to a photocreated carrier density of $n = \sigma/q\mu$ or 4×10^{20} electrons cm^{-3}. This requires an incident illumination power $P(=ndhv/\tau)$ of $6.4\,\mathrm{W\,cm}^{-2}$ for photons of 1 eV energy.

Weak conductivity ($\sigma \ll \varepsilon\omega$)
In this case, (7.C.4) takes the form:

$$k = \sqrt{\mu_0 \varepsilon}\omega \left(1 - \mathrm{i}\frac{\sigma}{\varepsilon\omega}\right)^{1/2} \approx \sqrt{\mu_0 \varepsilon}\omega \left(1 - \mathrm{i}\frac{\sigma}{2\varepsilon\omega}\right) \tag{7.C.7}$$

We see that the wave is absorbed ($E_0 \mathrm{e}^{-\kappa z}$) and that the intensity I decreases with an attenuation constant α ($I_0 \mathrm{e}^{-2\kappa z} = I_0 \mathrm{e}^{-\alpha z}$):

$$\alpha = \frac{Z_0}{n_{\mathrm{op}}}\sigma \tag{7.C.8}$$

where Z_0 is the vacuum impedance ($Z_0 = (\mu_0/\varepsilon_0)^{1/2} = 377\,\Omega$). In fact, this last expression is of little use as it ignores the dynamic behaviour of electrons, i.e. it supposes that electrical conductivity is a function of the instantaneous value of the electric field strength, which is generally not the case. We thus present a more pertinent model, one which takes into account the response time of the electron gas, known as the *Drude model*.

We consider a free electron gas of density n consisting of electrons with an

effective mass m_{eff} (see Chapter 5). We saw in Chapter 6 how the entire band structure of a material can be taken into account by means of the simple notion of an effective mass, and that the dynamic behaviour of the electron can be described by the equation:

$$m_{\text{eff}}\dot{v} + \frac{m_{\text{eff}}v}{\tau} = qE(t) \tag{7.C.9}$$

where v is the velocity of the electron and τ is the mean time between collisions. If the applied field is constant $E(t) = E_0$, the velocity will also be a constant given by $v_0 = q\tau E_0/m_{\text{eff}}$ with the electrical current resulting from the general motion of the gas being $j_0 = nqv_0 = \sigma_0 E_0$. The electrical conductivity is therefore $\sigma_0 = nq^2\tau/m_{\text{eff}}$, which is what we found earlier in Section 6.2. If the field E is an electromagnetic field $Ee^{i\omega t}$, (7.C.9) then shows that the conductivity $\sigma(\omega)$ depends upon the frequency:

$$\sigma(\omega) = \frac{\sigma_0}{1 + i\omega\tau} \tag{7.C.10}$$

The dispersion relation (7.C.4) describing the relationship between the wavevector $\mathbf{k}$ and the frequency ω remains valid by replacing σ by the above expression for $\sigma(\omega)$. We therefore immediately obtain a new expression for absorption (one more exact than that given in (7.C.8)):

$$\alpha = \frac{1}{n_{\text{op}}} \frac{Z_0\sigma_0}{1 + (\omega\tau)^2} \tag{7.C.11}$$

The product $Z_0\sigma_0$ is indeed homologous to an inverse length. In crystalline semiconductors, the collision time τ is generally in the range 0.1–10 ps and the product $\omega\tau$ is therefore very large for optical frequencies. Expression (7.C.11) may therefore be simplified to:

$$\alpha = \frac{Z_0}{4\pi c^2 n_{\text{op}}} \frac{nq^2\lambda_0^2}{m_{\text{eff}}\tau} \tag{7.C.12}$$

Drude model for free carrier absorption

Example

A slab of silicon ($n_{\text{op}} = 3.6$, $m_{\text{eff}} = 0.321m_0$) is doped to a concentration of $10^{20}\,\text{cm}^{-3}$, or $10^{26}\,\text{m}^{-3}$. At this impurity-concentration level, the mean collision time τ is 10^{-14} s. This results in a conductivity of:

$$\sigma_0 = \frac{nq^2\tau}{m_{\text{eff}}} = \frac{10^{26}\,\text{m}^{-3}(1.6 \times 10^{-19}\text{C})^2 10^{-14}\,\text{s}}{0.321 \times 0.9 \times 10^{-30}\,\text{kg}} = 8.8 \times 10^4\,\Omega^{-1}\,\text{m}^{-1}$$

or a resistivity of the order of $1\,\text{m}\Omega\,\text{cm}^{-1}$. The absorption coefficient for a 10 μm ($2 \times 10^{14}\,\text{s}^{-1}$) wave is then given by (7.C.11) to be:

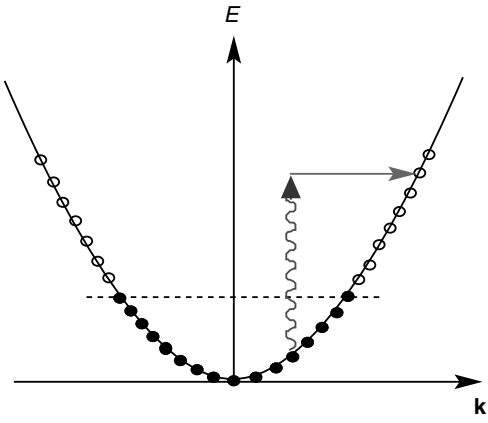


Fig. 7.C.2. The absorption of a photon by a single band is only possible with the assistance of a scattering mechanism which can provide the required momentum to make up the difference between the initial and final states.

$$\alpha = \frac{Z_0\sigma_0}{n_{\text{op}}(\omega\tau)^2} = \frac{377\,\Omega \times 8.8 \times 10^4\,\Omega^{-1}\,\text{m}^{-1}}{3.6 \times 4} = 2.3 \times 10^6\,\text{m}^{-1}$$

Infrared light is therefore absorbed over a distance of approximately 0.4 μm.

This very simple model has a number of drawbacks:

- It predicts a λ^2 dependence of the absorption coefficient which is not borne out by experiment. Generally, the absorption coefficient α is proportional to a power of wavelength λ other than 2 (α = constant $\times$ λ^p, where p varies anywhere between 1.5 and 3).
- It says nothing concerning the microscopic origin of the free carrier absorption process (interactions with impurities, acoustical phonons, . . .).

In order to resolve these problems, it is necessary to resort to a quantum mechanical treatment. We must nonetheless recognize that a quantum mechanical analysis will have to deal with a certain number of difficulties before it can offer a satisfactory description of absorption by a free electron gas. Figure 7.C.2 explains one such consideration.

As we saw at the beginning of this chapter, the matrix element describing the coupling between a band electron and an electromagnetic wave is non-zero only for nearly vertical transitions which conserve wavevector **k**. As the quasi-free electrons are distributed across the parabolic conduction band of the semiconductor, there are no available states 'above' this band which may receive an excited electron. In other words, we cannot satisfy both the momentum ($\mathbf{k} = \mathbf{k}'$) and energy conditions ($\hbar^2k^2/2m_{\text{eff}} + h\nu = \hbar^2k^2/2m_{\text{eff}}$). We therefore require another

interaction to supply the additional momentum needed to bring the excited electron back to the conduction band.

To do this, we have at our disposal all the mechanisms described in Chapter 6 (impurities, optical phonons, . . .). We will take as an example interaction with an optical phonon described by the interaction Hamiltonian $1/2(H_{\text{ph}}e^{i\omega t} + \text{c.c.})$. Since this transition requires the participation of two simultaneous perturbations, we must appeal to second-order perturbation theory, as touched upon in Complement 3.B. In that case, Eq. (3.B.26) gave the transition rate per second between an initial state $|i\rangle$ and a final state $|f\rangle$ to be:

$$S(i \to f) = \frac{\pi}{\hbar} \left| \sum_n \frac{\langle f|H_{\text{per}}|n\rangle\langle n|H_{\text{per}}|i\rangle}{E_i - E_n} \right|^2 \delta(E_{\text{tot}}) \tag{7.C.13}$$

This formula was applied in the context of two-photon absorption. Here we have $H_{\text{per}} = H_{\text{op}} + H_{\text{ph}}$, and the relevant terms are:

$$S(i \to f) = \frac{\pi}{\hbar} \left| \sum_n \frac{\langle f|H_{\text{op}}|n\rangle\langle n|H_{\text{ph}}|i\rangle}{E_i - E_n} + \frac{\langle f|H_{\text{op}}|n\rangle\langle n|H_{\text{ph}}|i\rangle}{E_i - E_n} \right|^2 \delta(E_{\text{tot}}) \tag{7.C.14}$$

The two other terms in the sum correspond to two-photon absorption (for terms of type $\langle i|H_{\text{op}}|n\rangle\langle n|H_{\text{op}}|f\rangle$) and to two-phonon scattering (for terms of type $\langle i|H_{\text{ph}}|n\rangle\langle n|H_{\text{ph}}|f\rangle$). Such transitions clearly do not preserve energy and momentum simultaneously. Total energy conservation during the various allowed second-order transitions may be expressed as:

$$\delta(E_{\text{tot}}) = \delta(E_f - E_i - \hbar\omega_{\text{op}} \pm \hbar\omega_{\text{ph}}) \tag{7.C.15}$$

depending upon whether phonon emission or absorption is involved. While the principle behind this calculation is clear, the summation over intermediate states $|n\rangle$ is rather involved in practice. This calculation can be examined in the work by B. K. Ridley (1988).

Table 7.C.1. *Coefficients required for* (7.C.16) *to calculate free electron absorption in different semiconductors* (*H. Y. Fan*, Semiconductors and Semimetals, *vol.* 3, *R. K. Willardson and A. C. Beer*, *eds*, *Academic Press*, *New York* (1967), *Ch.* 9)

Material	$N\,(\times 10^{17}\,\text{cm}^{-3})$	K	p
GaAs	1–5	3	3
InP	0.4–4	4	2.5
GaSb	0.5	6	3.5
InAs	0.3–8	4.7	3
InSb	1–3	2.3	2
AlSb	0.4–4	15	2

Given the complexity of these quantum models, use is often made of phenomenological formulas drawn from experiments. Free carrier absorption can be expressed as:

$$\alpha = K_\alpha \left[\frac{N(\text{cm}^{-3})}{10^{17}}\right]\left[\frac{\lambda(\mu\text{m})}{9}\right]^p \tag{7.C.16}$$

where K_α is the material-dependent parameter given in Table 7.C.1 for different semiconductors.

Figure 7.C.3 shows the absorption coefficients in several important semiconductors for optoelectronic applications (GaAs, InP, and GaSb) as a function of free electron concentration at two commonly employed wavelengths: 1.55 μm for telecommunications and 10 μm for optical defence applications. We see that parasitic absorption can be significant and must be taken into account in semiconductor laser threshold calculations.

Free carriers cause a change $\Delta\varepsilon$ in the optical index of semiconductor materials. To calculate the size of the effect, we need only introduce expression (7.C.10) into the dispersion relation (7.C.4) and sum up the real portions of k^2. We then find:

$$\Delta\varepsilon = -\frac{\omega_p^2}{\omega^2} \tag{7.C.17}$$

where ω_p is the *plasma frequency* of the electron gas given by:

$$\omega_p^2 = \frac{nq^2}{\varepsilon_0 m_{\text{eff}}} \tag{7.C.18}$$

Plasma frequency

This variation in the optical index is harnessed in *electro-optic modulators*. As shown in Fig. 7.C.4, the electromagnetic wave is guided (see Chapter 11) by a space charge region (see Chapter 10) placed between two mirrors deposited on either extremity of the guide. This guide of length L behaves as a Fabry–Pérot cavity, with a transmittance given by (9.D.21). Intuitively, the last formula says that the guide lets through light of wavenumber k when $kL = m\pi$ (ON state) and blocks light when $kL = (2m + 1)\pi/2$ (OFF state). Under the influence of an externally applied voltage, electrons are injected into the space charge region and act to modulate the effective guide index ε according to (7.C.17) and allow the guide to move from an ON state to an OFF state. The voltage required to switch the modulator state is given by:

$$k(V_{\text{OFF}}) - k(V_{\text{ON}}) = \frac{\pi}{2L} \tag{7.C.19}$$

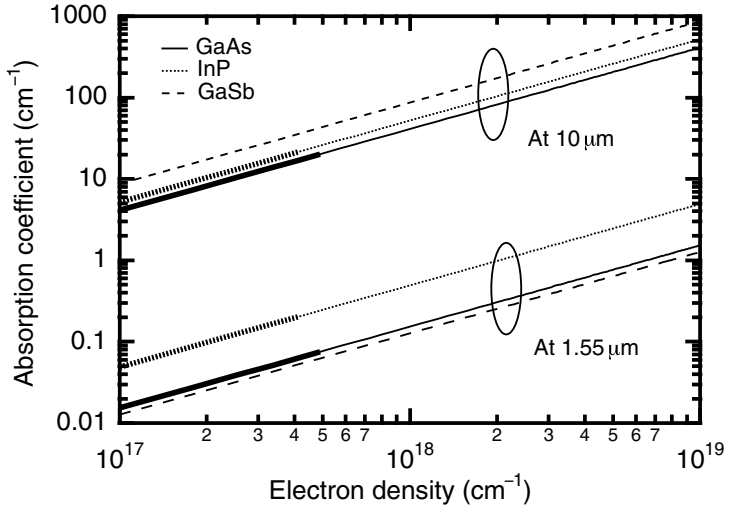


Fig. 7.C.3. Free carrier absorption coefficients as a function of electron density (calculated from (7.C.16) using the values in Table 7.C.1), for GaAs, InP, and GaSb at 1.55 and 10 μm.

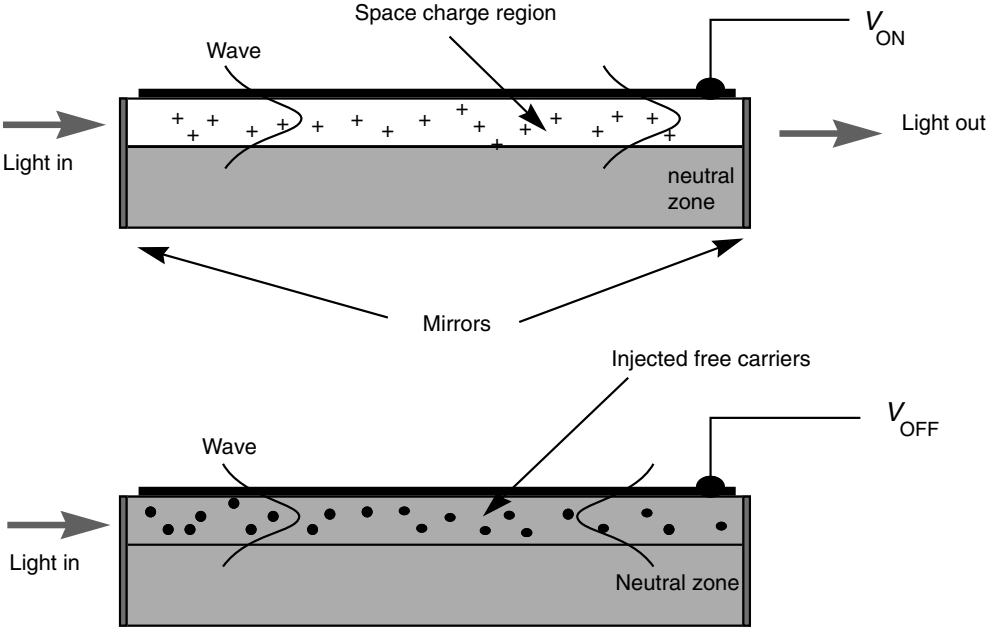


Fig. 7.C.4. Operation of a free carrier electromodulator.

or, given the dispersion relation $k = \omega\varepsilon^{1/2}/c = 2\pi n_{op}/\lambda_0$, where λ_0 is the vacuum wavelength:

$$\Delta\varepsilon = \frac{n_{op}}{2}\frac{\lambda_0}{L} \tag{7.C.20}$$

Example

We consider an InP waveguide ($n_{op} = 3.5$, $m_{eff} = 0.077$). We wish to know the guide length necessary to switch the modulator at $\lambda_0 = 1.55\,\mu m$ by injecting a space charge region with 10^{17} carriers cm^{-3}. The plasma frequency ω_p of the electron gas is given by (7.C.18) to be:

$$\omega_p = \sqrt{\frac{10^{23}\,m^{-3}}{8.86 \times 10^{-12}\,Fd\,m^{-1} \times 0.077 \times 0.9 \times 10^{-30}\,kg}}\,1.6 \times 10^{-19}\,C$$

$$= 6.4 \times 10^{13}\,s^{-1}$$

At 1.55 μm, the frequency ω is $1.2 \times 10^{15}\,s^{-1}$, which leads to a variation $\Delta\varepsilon$ of -3×10^{-3}. The length of the device given by (7.C.20) is then:

$$L = 3.5 \times 1.55 \times 10^{-3}\,mm/(2 \times 3 \times 10^{-3}) = 0.9\,mm$$

FURTHER READING

N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart, and Winston, New York (1976).

J. I. Pankove, *Optical Processes in Semiconductors*, Dover Publications, New York (1971).

B. K. Ridley, *Quantum Processes in Semiconductors*, Clarendon Press, Oxford (1988).

8 Semiconductor heterostructures and quantum wells

8.1 Introduction

In Chapter 5 we saw that the principal characteristic of a semiconducting material is the existence of forbidden energy bands, or *gaps*, acting to separate the electron-rich valence band from the electron-poor conduction band. Both the bandgap and the energy bands are determined by the bulk potential of the crystalline material.

At the basis of bandstructure engineering is the *heterojunction*, which is obtained by growing one semiconductor layer onto another. For certain carefully selected semiconductors possessing compatible crystal structures and lattice spacings, it is possible to achieve epitaxial growth of one material onto another. In this case, the atomic positions of the second material form a virtually perfect continuation of the underlying substrate lattice. Under carefully controlled conditions, the compositional transition between the two materials can be made almost perfectly abrupt (i.e. with heterointerfaces in many instances being defined on a monolayer scale).

Away from the heterojunction, and deep within the bulk of the two materials, the electrons are subject to volumetric potentials (bandgaps and band structures) characteristic of each of the constituent bulk semiconductors. In the vicinity of the heterojunction, the crystal potential changes abruptly from one material to the other. A quantitative description of this change in potential requires that calculations be performed at the atomic level. These calculations (performed numerically on computers) are extremely involved and lie outside the scope of this book. Such calculations indicate, however, that over the scale of *a few atomic layers* near the interface, there is a transfer of electrical charge. This transfer creates an *interface charge dipole* and is responsible for an abrupt jump in the electrostatic potential between both interfaces and is superimposed on the crystalline potential. On either side of this transition region (extending over distances of the order of monolayers), the electrons take on the characteristics of the constituent semiconductors (gaps, effective masses, . . .). The relative positions of the valence bands, however, are determined by this charge transfer process. From a macroscopic viewpoint, we can consider the heterojunction as possessing a valence band discontinuity or *offset* ΔE_v. For our purposes, we will take ΔE_v to be a parameter (determined either through experiment or predicted by more comprehensive theories) that depends uniquely on the pair of materials forming the heterojunction.

Once the valence band offset has been determined, the conduction band offset ΔE_c may be trivially deduced from our knowledge of the bulk semiconductor bandgaps. It is important to appreciate that given the chemical origin of the interfacial charge dipole, its spatial influence is limited to the relatively small number of chemical bonds that span the heterojunction interface.

Three types of structures are possible (see Fig. 8.1): in one case the entire gap of the small bandgap semiconductor resides within the conduction and valence bands of the large gap material (i.e. a *type I* heterostructure); alternatively, one of the band offsets may be larger than the difference between the semiconductor bandgaps, but smaller than the magnitude of the larger semiconductor bandgap (*type II*); and, finally, one of the offsets may be larger than the gap of the largest bandgap semiconductor (*type III*).

For an electron in the conduction band of a heterostructure, the offset represents in essence a discontinuity in the potential energy across the heterojunction. For a given total energy E_{tot}, the electron will have a kinetic energy $E_{\text{tot}} - E_{c1}$ in semiconductor 1, and $E_{\text{tot}} - E_{c2}$ in semiconductor 2, where E_{ci} is the energy at the bottom of the conduction band of semiconductor i. As a result, by depositing sequences of semiconductor layers (each of appropriate thickness and composition) it is possible to create any potential profile along the crystal growth direction *on demand*. The possibilities are principally limited by the number of available materials and by their capacity to be grown epitaxially. It is clear that the crystal structures of the materials must have nearly identical lattice constants. If not, the semiconductor deposited over the first will acquire a very large elastic deformation energy which the crystal will only be able to accommodate as long as the thickness of the deposited layer remains small. Beyond this *critical thickness* the strain in the epilayers will relax, disrupting the near perfect crystallinity of the structure and degrading the optoelectronic properties otherwise afforded by these materials. In the case of InAs, for instance, only a little less than two monolayers can be epitaxially deposited on GaAs before the film relaxes by forming three-dimensional islands.

Figure 8.2 shows the bandgaps and lattice constants for various bulk semiconductors. The requirement of closely matching lattice constants separates these semiconductors into five distinct families in terms of compatible substrate materials. Of these families, GaAs/$Al_xGa_{1-x}As$ is the best controlled, and all compositional Al fractions x are accessible, as the lattice constant does not vary significantly with x. Two other important systems include $Ga_xIn_{1-x}As_yP_{1-y}$ and $Al_xIn_{1-x}As$, which can be deposited onto an InP substrate, and $InAs_{1-x}Sb$ and $Al_xGa_{1-x}Sb$ for GaSb. Finally, the lattice constants available to $Hg_{1-x}Cd_xTe$ are well suited to deposition on CdTe substrates, and allow a large range of accessible bandgaps extending from the visible region down to zero and even negative gaps (with negative bandgaps indicating that the material behaves as a semimetal over this range of compositions).

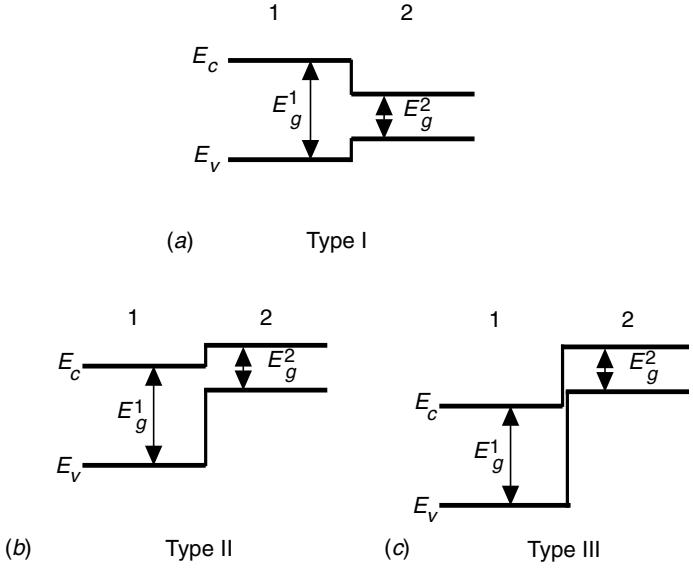


Fig. 8.1. (*a*)–(*c*) The three heterojunction types which may exist between two semiconductors possessing bandgaps E_g^1 and E_g^2. E_c and E_v denote the conduction and valence bands, respectively.

8.2 Envelope function formalism

The electron wavefunctions in a heterojunction can be described using *envelope function* formalism. This approximation is formally identical to that used to describe the propagation of an electromagnetic wave in an inhomogeneous medium, as for instance in the case of a waveguide in (9.27), or in non-linear optics as in (12.21). This section is fairly demanding, and the reader may simply wish to take (8.17) on faith and move on to the next section.

We will now derive the envelope function approximation within the general framework of a semiconductor possessing a periodic potential $V_c(\mathbf{r})$ reflecting the crystal lattice, and a slowly varying potential $V(\mathbf{r})$ with respect to the lattice spacing. We will discuss the simple case of an electron inhabiting the conduction band of a heterostructure – the more complex problem associated with valence band degeneracy will be left to Complement 8.D.

We consider a slowly varying potential $V(\mathbf{r})$ in $\mathbf{r}$ as shown in Fig. 8.3. The Hamiltonian for an electron given this combination of potentials may then be written:

$$H = \frac{p^2}{2m_0} + V_c(\mathbf{r}) + V(\mathbf{r}) \tag{8.1}$$

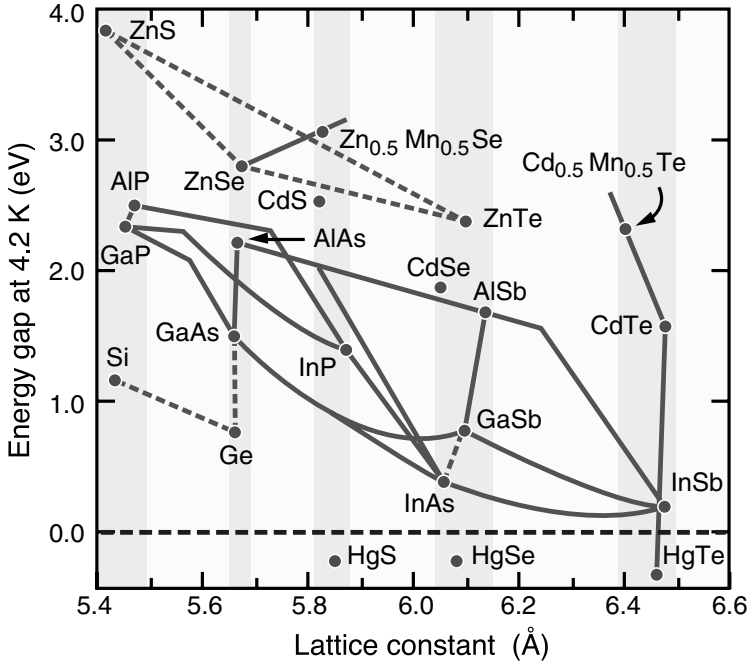


Fig. 8.2. Bandgaps and lattice constants for various bulk semiconductors.

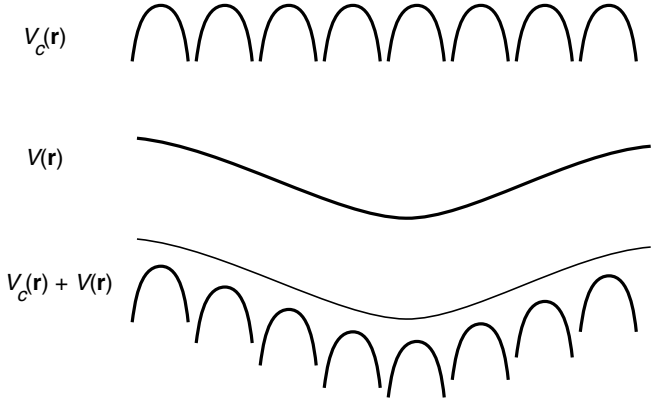


Fig. 8.3. The potential seen by an electron resulting from the superposition of a slowly varying potential and the crystal potential.

For $V(\mathbf{r}) = 0$ the solutions to Schrödinger's equation $H\psi = E\psi$ are the Bloch wavefunctions $|n, \mathbf{k}\rangle$ (see Chapter 5):

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{8.2}$$

where Ω is the crystal volume, n is the band index, and $u_{n\mathbf{k}}(\mathbf{r})$ possesses the periodicity of the lattice. The corresponding energy is $\varepsilon_{n\mathbf{k}}$.

For $V \neq 0$ we expand the solutions to Schrödinger's equation in (8.1) in terms of the basis set formed by the Bloch functions as:

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{n\mathbf{k}} a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{8.3}$$

which must satisfy the equation:

$$\left[\frac{p^2}{2m_0} + V_c(\mathbf{r}) \right] \sum_{n\mathbf{k}} a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) + \sum_{n\mathbf{k}} V(\mathbf{r}) a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) = E \sum_{n\mathbf{k}} a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{8.4}$$

As each Bloch wavefunction is a solution to the 'crystal + null potential' problem, this last expression may then be written:

$$\sum_{n\mathbf{k}} \varepsilon_{n\mathbf{k}} a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) + \sum_{n\mathbf{k}} V(\mathbf{r}) a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) = E \sum_{n\mathbf{k}} a_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{8.5}$$

We then project this expression onto the wave $|N\mathbf{K}\rangle = \Omega^{-1/2} e^{i\mathbf{k}\cdot\mathbf{r}} u_{N\mathbf{K}}(\mathbf{r})$ and, using the orthogonality of the Bloch wavefunctions, obtain an equivalent expression for Schrödinger's equation:

$$\sum_{n\mathbf{k}} a_{n\mathbf{k}} \int_{\Omega} e^{-i\mathbf{K}\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) V(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) d\mathbf{r} = (E - \varepsilon_{N\mathbf{K}}) a_{N\mathbf{K}} \tag{8.6}$$

Up to this point, no shortcuts have been taken in the derivation. It is therefore reasonable to ponder what approximations might be made to simplify this expression. As a guide, we note that if $V(\mathbf{r})$ was constant, then the left term in (8.6) would be the Fourier transform of the quantity $u_{n\mathbf{k}}(\mathbf{r}) u_{N\mathbf{K}}(\mathbf{r})$, and would therefore depend only on the crystal.

The fact that the potential $V(\mathbf{r})$ varies slowly, implies that its Fourier transform $\tilde{V}_{\mathbf{q}}$ given by:

$$\begin{aligned} V(\mathbf{r}) &= \sum_{\mathbf{q}} \tilde{V}_{\mathbf{q}} e^{i\mathbf{q}\cdot\mathbf{r}} \\ \tilde{V}_{\mathbf{q}} &= \frac{1}{\Omega} \int_{\Omega} V(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} \end{aligned} \tag{8.7}$$

is null except for very small $\mathbf{q}$ values with respect to the radius of the Brillouin zone. Inserting the transform (8.7) into (8.6), the following term emerges:

$$\sum_{n\mathbf{k}\mathbf{q}} a_{n\mathbf{k}} \frac{\tilde{V}_{\mathbf{q}}}{\Omega} \int_{\Omega} e^{i(\mathbf{k}+\mathbf{q}-\mathbf{K})\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) u_{n\mathbf{k}}(\mathbf{r}) d\mathbf{r} \tag{8.8}$$

which would be predominantly zero if V was constant. As the functions $u_{N\mathbf{K}}$ and $u_{n\mathbf{k}}$ are periodic, we can set $\mathbf{r} = \mathbf{R}_i + \mathbf{r}'$, where $\mathbf{R}_i$ designates the position of the ith cell, and $\mathbf{r}'$ is restricted to the volume of the primitive cell, which results in the following sum over i:

$$\frac{1}{\Omega}\int_{\Omega} e^{i(\mathbf{k}+\mathbf{q}-\mathbf{K})\cdot\mathbf{r}} u^*_{N\mathbf{K}}(\mathbf{r}) u_{n\mathbf{k}}(\mathbf{r}) d\mathbf{r}$$

$$= \frac{1}{\Omega}\sum_i e^{i(\mathbf{k}+\mathbf{q}-\mathbf{K})\cdot\mathbf{R}_i} \int_{\text{cell}} e^{i(\mathbf{k}+\mathbf{q}-\mathbf{K})\cdot\mathbf{r}'} u^*_{N\mathbf{K}}(\mathbf{r}') u_{n\mathbf{k}}(\mathbf{r}') d\mathbf{r}' \tag{8.9}$$

$$= \Delta_{\mathbf{k}+\mathbf{q}-\mathbf{K}} \frac{1}{\Omega_{\text{cell}}} \int_{\text{cell}} u^*_{N\mathbf{K}}(\mathbf{r}') u_{n\mathbf{K}+\mathbf{q}}(\mathbf{r}') d\mathbf{r}$$

where $\int_{\text{cell}}$ indicates that the integral is taken over the primitive cell and $\Delta_{\mathbf{k}}$ is the Kronecker delta. In this expression, we have supposed that $\mathbf{k}$ and $\mathbf{K}$ lie within the interior of the Brillouin zone. If they were near the edge of the Brillouin zone, then we would have to take into account *Umklapp* processes in which $\mathbf{k} + \mathbf{q} - \mathbf{K} = \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector (including these effects, however, would add little in terms of novel content to the discussion).

We will again make use of the fact that only small vectors $\mathbf{q}$ contribute to the potential, in expanding the overlap integral for the periodic portions of the Bloch wavefunctions in terms of $\mathbf{q}$:

$$\frac{1}{\Omega_{\text{cell}}} \int_{\text{cell}} u^*_{N\mathbf{K}}(\mathbf{r}') u_{n\mathbf{K}+\mathbf{q}}(\mathbf{r}') d\mathbf{r}'$$

$$= \frac{1}{\Omega_{\text{cell}}} \int_{\text{cell}} u^*_{N\mathbf{K}}(\mathbf{r}') u_{n\mathbf{K}}(\mathbf{r}') d\mathbf{r}' + \frac{1}{\Omega_{\text{cell}}} \int_{\text{cell}} u^*_{N\mathbf{K}}(\mathbf{r}') \mathbf{q}\cdot\nabla_{\mathbf{K}} u_{n\mathbf{K}}(\mathbf{r}') d\mathbf{r}' + \cdots \tag{8.10}$$

$$\cong \Delta_{Nn} + \frac{1}{\Omega_{\text{cell}}} \mathbf{q}\cdot \int_{\text{cell}} u^*_{N\mathbf{K}}(\mathbf{r}') \nabla_{\mathbf{K}} u_{n\mathbf{K}}(\mathbf{r}') d\mathbf{r}'$$

The orthogonality *for the same wavevector* $\mathbf{K}$ of the functions $u_{N\mathbf{K}}$ and $u_{n\mathbf{k}}$ results from the orthogonality of Bloch functions for different states. (Note! For different wavevectors, the periodic portions are not orthogonal.) The simplest approximation consists in keeping only the first term, which allows one to decouple each

of the different bands in Schrödinger's equation in (8.6). Given (8.9) and (8.10), Schrödinger's equation in reciprocal space takes the form:

$$\varepsilon_{N\mathbf{K}} a_{N\mathbf{K}} + \sum_{\mathbf{q}} a_{N\mathbf{K}+\mathbf{q}} \tilde{V}_{\mathbf{q}} = E a_{N\mathbf{K}} \tag{8.11}$$

By introducing *the envelope function*:

$$\zeta_N(\mathbf{r}) = \sum_{\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} \tag{8.12}$$

Definition of the envelope function

we may write (8.11) in real space as:

$$\sum_{\mathbf{K}} \varepsilon_{N\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} + \sum_{\mathbf{K}\mathbf{q}} a_{N\mathbf{K}+\mathbf{q}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} \tilde{V}_{\mathbf{q}} = E \sum_{\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} \tag{8.13}$$

or again as:

$$\begin{aligned} &\sum_{\mathbf{K}} \varepsilon_{N\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} + \sum_{\mathbf{q}} \tilde{V}_{\mathbf{q}} \mathrm{e}^{-\mathrm{i}\mathbf{q}\cdot\mathbf{r}} \sum_{\mathbf{K}'} a_{N\mathbf{K}'} \mathrm{e}^{\mathrm{i}\mathbf{K}'\cdot\mathbf{r}} = E \sum_{\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} \\ &\int \frac{\mathrm{d}^3\mathbf{K}}{(2\pi)^3} \varepsilon_{N\mathbf{K}} a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} + V(\mathbf{r})\zeta_N(\mathbf{r}) = E\zeta_N(\mathbf{r}) \end{aligned} \tag{8.14}$$

As only the components close to $\mathbf{K}$ are coupled by (8.11), we may expand the dispersion relation for the band $\varepsilon_{N\mathbf{K}}$ as a series about $\mathbf{K}$. For the more important case where $\mathbf{K} = \mathbf{0}$, we introduce the effective mass m_N^* for the Nth band:

$$\varepsilon_{N\mathbf{K}} = \varepsilon_{N\mathbf{0}} + \frac{\hbar^2 K^2}{2m_N^*} \tag{8.15}$$

which we will assume to be isotropic to simplify matters.

Substituting (8.15) into (8.14), we find:

$$\int \frac{\mathrm{d}^3\mathbf{K}}{(2\pi)^3} \left(\varepsilon_{N\mathbf{0}} + \frac{\hbar^2 K^2}{2m_N^*} \right) a_{N\mathbf{K}} \mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{r}} = \varepsilon_{N\mathbf{0}} \zeta_N(\mathbf{r}) - \frac{\hbar^2}{2m_N^*} \nabla_{\mathbf{r}}^2 \zeta_N(\mathbf{r}) \tag{8.16}$$

which leads us to the reward for our efforts – Schrödinger's equation for the envelope function:

$$\left[\frac{p^2}{2m_N^*} + V(\mathbf{r})\right]\zeta_N(\mathbf{r}) = (E - \varepsilon_{N\mathbf{0}})\zeta_N(\mathbf{r}) \tag{8.17}$$

Schrödinger's equation for the envelope function

This equation is the major finding of this chapter. It shows that the role of the effective mass greatly exceeds that encountered in the context of the semiclassical description in Chapters 5 and 6: *in the case of a slowly varying potential* V(**r**) *with respect to the crystal potential, the wavefunction for an electron in the* N*th band is therefore that for a particle possessing the effective mass associated with the* N*th band and subjected to a potential* V(**r**). Every influence of the material is taken into account through the effective mass m_N^* and the energy $\varepsilon_{N\mathbf{0}}$ of the Nth band at $\mathbf{K} = \mathbf{0}$ in Schrödinger's equation, (8.17).

In the case of the adopted approximation, the complete wavefunction is given by:

$$\psi_N \cong \zeta_N(\mathbf{r})u_{N\mathbf{0}}(\mathbf{r}) \tag{8.18}$$

and the solution corresponding to a potential $V = 0$ is:

$$\psi_{N\mathbf{k}} \cong \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}u_{N\mathbf{0}}(\mathbf{r}) \tag{8.19}$$

which differs from the exact solution by the substitution of $u_{N\mathbf{k}}(\mathbf{r})$ by $u_{N\mathbf{0}}(\mathbf{r})$ in (8.3).

In the case of a heterojunction for which the external potential $V(\mathbf{r})$ is discontinuous over an atomic scale, the derivation we have just carried out is no longer strictly valid. If, however, we seek states close in energy to the conduction band we can approximate the solutions on either side of the heterojunction by (8.17) and (8.18) along with the Bloch function $u_{N\mathbf{0}}(\mathbf{r})$ proper to each material. To obtain the stationary state for an electron in the heterostructure, we must join the two solutions at the interface (see Section 1.3.1). To begin, the wavefunctions must be continuous:

$$\zeta_N(0^-)u_{N\mathbf{0}}(0^-) = \zeta_N(0^+)u_{N\mathbf{0}}(0^+) \tag{8.20}$$

where the '$\pm$' signs refer to the materials on the left and right of the interface at $z = 0$. In this equation, the envelope function varies very little over the cell before the interface, and if we take the average of (8.20) over this last cell, we find:

$$\zeta_N(0^-) = \zeta_N(0^+) \tag{8.21}$$

The exact solution must conserve the probability flux which crosses the interface. As demonstrated in Section 1.3, the probability current is given by:

$$\mathbf{j} = \frac{\hbar}{m_N^*\mathrm{i}}[\zeta_N^*(\mathbf{r})\nabla\zeta_N(\mathbf{r}) - \zeta_N(\mathbf{r})\nabla\zeta_N^*(\mathbf{r})] \tag{8.22}$$

and conservation of current across the interface leads to the continuity equation:

$$\frac{1}{m_N^-}\frac{\partial}{\partial z}\zeta_N(0^-) = \frac{1}{m_N^+}\frac{\partial}{\partial z}\zeta_N(0^+) \tag{8.23}$$

This last condition can be easily incorporated into the formalism by writing Schrödinger's equation for the envelope function as:

$$\left[-\nabla\frac{\hbar^2}{2m_N^*(\mathbf{r})}\nabla + V(\mathbf{r})\right]\zeta_N(\mathbf{r}) = (E - \varepsilon_{N0})\zeta_N(\mathbf{r}) \tag{8.24}$$

where $m_N^*(\mathbf{r})$ takes into account the variation of the effective mass from one material to another.

We must remark that this rather heuristic approach to the heterojunction hides many important problems. Therefore, we have restricted the approximation to a simple band, where the effective mass does not depend on $\mathbf{K}$, i.e. a parabolic band (see (8.15)). If the involved energies E and $\Delta V(\mathbf{r})$ are significant, this assumption is not admissible. We have also supposed that the band N is non-degenerate at $\mathbf{K} = \mathbf{0}$, which is clearly not the case for the valence bands. These two aspects (band parabolicity and degeneracy) are related. The derivation of envelope functions requiring a multi-band treatment forms the (challenging!) topic of Complement 8.D. For the time being, we will proclaim without scruples that the approximation for the envelope function in (8.18) along with its associated Schrödinger equation in (8.24) are sufficient to describe the quantum behaviour of an electron in a heterostructure.

8.3 The quantum well

By employing two heterojunctions, it now becomes possible to implement a one-dimensional potential profile (along the growth direction) for electrons corresponding to the quantum well studied in Chapter 1.

Figure 8.4 shows the most comprehensively studied quantum well structure to date consisting of a well layer of GaAs inserted between two larger bandgap $Al_xGa_{1-x}As$ barrier layers. For $x < 0.4$ the conduction band offset is proportional to the aluminium fraction x, so that $\Delta E_c \approx x \times 836\,\text{meV}$. By adjusting the Al fraction x, and the thickness of the GaAs layer at the time of growth, we can create a quantum well with electronic properties tailored to the user's specifications – this practice is referred to as *quantum engineering*.

For electrons in the conduction band, the AlGaAs layers form potential barriers of height $V_B = \Delta E_c$. Schrödinger's equation for the envelope function then becomes:

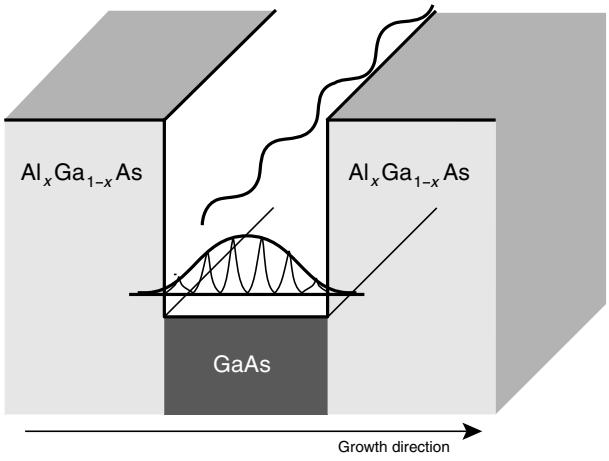


Fig. 8.4. A potential well results when a GaAs well layer is grown between two larger bandgap $Al_xGa_{1-x}As$ barrier layers. When the width of the well layer is sufficiently small, the motion of the electrons in the quantum well is quantized in the growth direction, and the allowed energy levels corresponding to motion in this direction become discrete. In the plane parallel to the interfaces, the motion of these electrons remains unrestricted. As a result, the total electronic wavefunction is given by the product of the envelope function (solution to the one-dimensional Schrödinger's equation) with the periodic Bloch function $u_{n\mathbf{k}}$ (contributed by the periodicity of the crystal lattice), and the plane waves describing the free motion in the plane parallel to the interfaces.

$$-\nabla\frac{\hbar^2}{2m^*(z)}\nabla\zeta(\mathbf{r}) + V(z)\zeta(\mathbf{r}) = E\zeta(\mathbf{r}) \tag{8.25}$$

where $V(z)$ has value of zero in the well and V_B in the barriers. Both the effective mass and the potential depend only on z, while in the x and y directions (parallel to the interfaces) the effective potentials do not vary. Consequently, $\zeta(\mathbf{r})$ can be written in the form:

$$\zeta(\mathbf{r}) = \zeta_{n\mathbf{K}}(z)\exp(\mathrm{i}\mathbf{K}\cdot\mathbf{R}) \text{ or } \mathbf{R} = (x, y) \text{ and } \mathbf{K} = (k_x, k_y) \tag{8.26}$$

where the exponential envelope function represents the free motion of electrons parallel to the interfaces, and $\zeta_{n\mathbf{K}}(z)$ is determined by Schrödinger's equation in one dimension:

$$-\frac{\mathrm{d}}{\mathrm{d}z}\frac{\hbar^2}{2m^*(z)}\frac{\mathrm{d}}{\mathrm{d}z}\zeta_{n\mathbf{K}}(z) + \left[V(z) + \frac{\hbar^2K^2}{2m^*(z)}\right]\zeta_{n\mathbf{K}}(z) = E\zeta_{n\mathbf{K}}(z) \tag{8.27}$$

Labelling the effective mass in the well as m_w and in the barrier as m_B, we find:

$$-\frac{\mathrm{d}}{\mathrm{d}z}\frac{\hbar^2}{2m^*(z)}\frac{\mathrm{d}}{\mathrm{d}z}\zeta_{n\mathbf{K}}(z)+\left\{V(z)+\frac{\hbar^2K^2}{2}\left[\frac{1}{m(z)}-\frac{1}{m_w}\right]\right\}\zeta_{n\mathbf{K}}(z)$$
$$=\left(E-\frac{\hbar^2K^2}{2m_w}\right)\zeta_{n\mathbf{K}}(z) \tag{8.28}$$

For small K, the correction to the 'potential energy' due to the variation in the effective mass is negligible. In almost every case, this correction is dropped for the good reason that the function $\zeta_{n\mathbf{K}}(z)$ then becomes independent of $\mathbf{K}$ obeying the simple equation:

$$\left[-\frac{\mathrm{d}}{\mathrm{d}z}\frac{\hbar^2}{2m^*(z)}\frac{\mathrm{d}}{\mathrm{d}z}+V(z)\right]\zeta_n(z)=\varepsilon_n\zeta_n(z)$$
$$\varepsilon_n=E_n-\frac{\hbar^2K^2}{2m_w} \tag{8.29}$$

If the effective mass did not depend on the different materials, this equation would be same as that obtained for the quantum well studied in Chapter 1. As both the effective masses and potentials are constant over the regions partitioned by the heterojunctions, the solution to (8.29) is no more difficult to obtain.

Let us consider the interface at $z = L/2$ for $\varepsilon < V_B$ in the well. The solution is a combination of two plane waves:

$$\zeta(z)=a\mathrm{e}^{\mathrm{i}k(z-L/2)}-b\mathrm{e}^{-\mathrm{i}k(z-L/2)},\ z<L/2 \tag{8.30}$$

with $\hbar^2k^2/2m_w=\varepsilon$.

In the barrier region, for $z > L/2$, the admissible solution consists of a decaying exponential:

$$\zeta(z)=c\mathrm{e}^{-\kappa(z-L/2)},\ z>L/2 \tag{8.31}$$

with $\hbar^2\kappa^2/2m_B=V_B-\varepsilon$.

Applying the continuity conditions at $z = L/2$, we obtain:

$$\frac{b}{a}=\mathrm{e}^{-2\mathrm{i}\phi}=\frac{(m_w/m_B)\kappa-\mathrm{i}k}{(m_w/m_B)\kappa+\mathrm{i}k} \tag{8.32}$$

Looking at expression (8.30), (8.32) tells us that the electronic wavefunction is phase shifted by 2ϕ as a result of its reflection by the $z = L/2$ interface. ϕ is then given by:

$$\sin\phi=\frac{k}{\{[(m_w/m_B)\kappa]^2+k^2\}^{1/2}} \tag{8.33}$$

Considering the entire well starting from the $z = -L/2$ interface, the phase of the

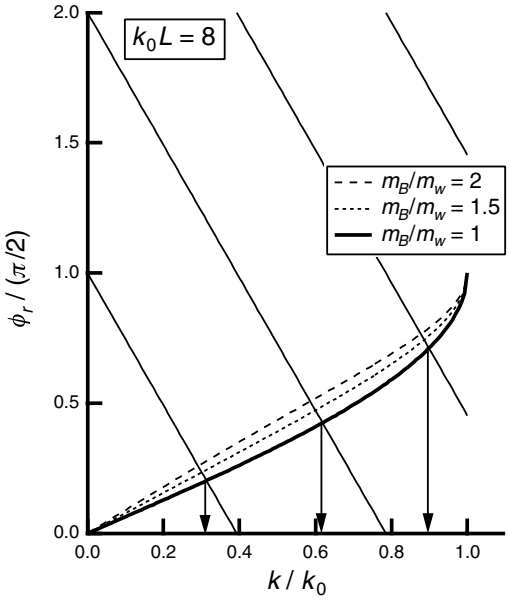


Fig. 8.5. Graphical solution to (8.34). The solutions are given by the intersection points between the sloping lines $n - kL/\pi$ and the arcsine functions.

plane wave travelling from the left to the interface at the right increases by kL, with the reflection adding 2ϕ to the phase, and the return trip of the reflected wave to $-L/2$ adding another kL. Finally, the reflection at $z = -L/2$ contributes a second phase shift of 2ϕ. For a stationary state, the total phase after a complete round-trip must equal $2n\pi$, with n being an integer. The leads to the condition that $2\phi = n\pi - kL$, or again:

$$\frac{2}{\pi}\arcsin\frac{k}{\{[(m_w/m_B)\kappa]^2 + k^2\}^{1/2}} = n - \frac{k_0 L}{\pi}\frac{k}{k_0} \tag{8.34}$$

with $\hbar^2 k_0^2/2m_w = V_B$.

The graphical solution in Fig. 8.5 shows how this condition is met for several values of m_B/m_w. For $m_B = m_w$, the solutions in Fig. 8.5 are equivalent to those in Fig. 1.2.

The energy of a stationary state which we may designate by $|n\mathbf{K}\rangle$, is then according to (8.29):

$$E_{n\mathbf{K}} = \varepsilon_n + \frac{\hbar^2 K^2}{2m_w} \tag{8.35}$$

where ε_n are discrete energies and solutions to (8.29), and $\mathbf{K}$ is continuous within the (K_x, K_y) plane. The states are organized into *subbands* as depicted in Fig. 8.6. Each subband comprises an ensemble of electronic states possessing the *same* state

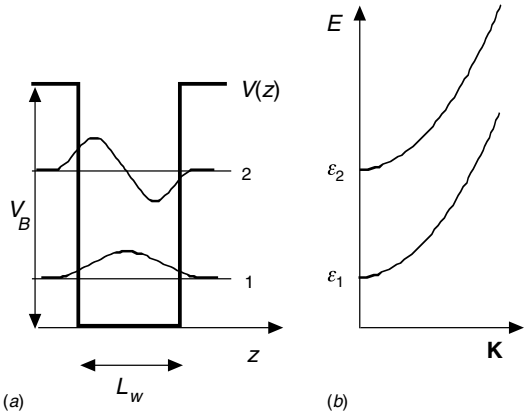


Fig. 8.6. Quantum well structure (*a*) showing two subbands in the well along with their corresponding envelope functions (perpendicular to the interfaces), and (*b*) energy spectra $E_n(\mathbf{K})$.

of quantized motion perpendicular to the interfaces, and continuously varying wavevectors **K** corresponding to their unhindered in-plane motion. The term *two-dimensional electron gas* is used to refer to the population of electrons which occupy such a subband structure.

It is important to note that the wavefunctions remain separable because the potential depends only upon z. The shape of $V(z)$ intervenes only in determining the energy minima of the subbands ε_n. The energy ε_n is also referred to as the *confinement energy*. So the concepts of subbands and confinement energies are valid in many more systems than the simple symmetric quantum well treated here.

8.4 Density of states and statistics in a quantum well

In each subband, the electronic states contain a plane wave component in the directions parallel to the interfaces. As done many times now for three-dimensional continuums (see Complement 1.A, Section 5.3, . . .), we will enclose the two-dimensional quantum well in the plane by a macroscopic rectangular boundary with area $A = L_x L_y$. The limiting conditions at the edges of this rectangular section (i.e. a two-dimensional 'fictitious box') introduce a pseudo-quantization of the allowed momentum values $\hbar\mathbf{K}$. The Born–von Karman cyclic boundary conditions (see Section 5.2) require that:

$$\begin{aligned} \psi(x, y, z) &= \psi(x + L_x, y, z) \\ \psi(x, y, z) &= \psi(x, y + \mathrm{L}_y, z) \end{aligned} \tag{8.36}$$

For the wavefunctions in (8.26), this condition implies that:

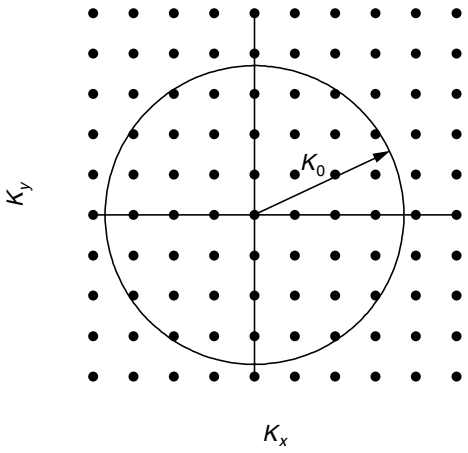


Fig. 8.7. The pseudo-quantized values for $\mathbf{K} = (2\pi n_x/L_x, 2\pi n_y/L_y)$ are represented by points in a reciprocal space plane. The states possessing energies less than E_0 are enclosed by the circle of radius K_0 (Eq. (8.39)).

$$\mathbf{K} = (K_x, K_y) = \left(\frac{2\pi n_x}{L_x}, \frac{2\pi n_y}{L_y}\right) \tag{8.37}$$

where n_x and n_y are integers. If we represent the allowed values for $\mathbf{K}$ as points in a plane (Fig. 8.7), each state then occupies an area $(2\pi)^2/L_xL_y$ and the density of states in this two-dimensional $\mathbf{K}$ space is then constant, i.e. $D(\mathbf{K}) = g_sA/(2\pi)^2$. (In this last expression, the spin degeneracy $g_s = 2$ has been taken into account as each $\mathbf{K}$ state admits two solutions with opposing spins.)

We often need to calculate the sums of functions f which depend only on the energies of the states:

$$\sum_{\mathbf{K}} f[E(\mathbf{K})] \cong \int \frac{Ag_s}{(2\pi)^2} \mathrm{d}^2\mathbf{K} f[E(\mathbf{K})] = \int f(E) D_n(E) \mathrm{d}E \tag{8.38}$$

where $D_n(E)$ is the density of states in the nth subband. Assuming a parabolic dispersion as in Eq. (8.35) the number of states $N_n(E_0)$ having an energy less than E_0 is equal to the number of states having a wavevector $\mathbf{K}$ lying within a circle of radius $K_0 = (2m_w(E_0 - \varepsilon_n)/\hbar^2)^{1/2}$, as indicated in Fig. 8.7, i.e. for $E_0 > \varepsilon_n$:

$$N_n(E_0) = \frac{Ag_s}{(2\pi)^2} \pi K_0^2 = \frac{Ag_s m_w (E_0 - \varepsilon_n)}{\pi\hbar^2} \tag{8.39}$$

The density of states in the subband is therefore constant:

$$D_n(E) = \frac{\mathrm{d}N_n}{\mathrm{d}E} = \frac{Ag_s m_w}{2\pi\hbar^2} \theta(E - \varepsilon_n) \tag{8.40}$$

For $E < \varepsilon_n$, there are no available states and $D_n(E) = 0$, which is reflected by the Heaviside function θ.

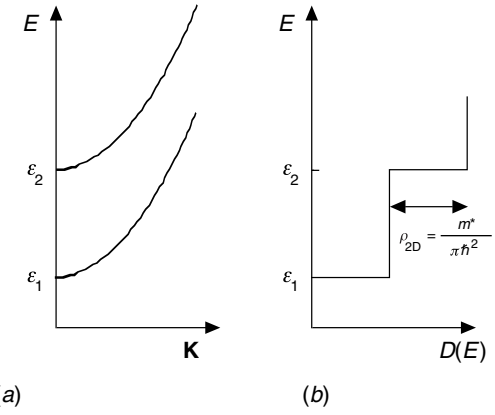


Fig. 8.8. Subbands and density of states for a two-dimensional structure.

The density of states for the system is then given by the sum of the density of states from each subband.

$$D(E) = \sum_n D_n(E) = \frac{Ag_s m_w}{2\pi\hbar^2} \sum_n \theta(E - \varepsilon_n) \tag{8.41a}$$

and has the overall shape of a staircase (Fig. 8.8) with the step height being a constant of the material. The position of the steps, however, depends on the explicit values of ε_n and therefore on the potential which confines the electrons into discrete subbands. Often one makes use of the two-dimensional density of states ρ_{2D} per subband per unit area (in $eV^{-1}\,cm^{-2}$), which is given by:

$$\rho_{2D} = \frac{D_n(E)}{A} = \frac{m_w}{\pi\hbar^2} \tag{8.41b}$$

Two-dimensional density of states

At thermodynamic equilibrium, the probability that an electron occupies a state with energy $E_{n\mathbf{K}}$ is still governed by Fermi statistics:

$$f(E_{n\mathbf{K}}) = \frac{1}{1 + \exp[(E_{n\mathbf{K}} - E_F)/k_B T]} \tag{8.42}$$

where E_F is the Fermi energy. The surface density n_n of electrons in a subband n is then:

$$n_n = \int D_n(E) f(E) \mathrm{d}E = n_c \ln\left[1 + \exp\left(\frac{E_F - \varepsilon_n}{k_B T}\right)\right]$$
$$n_c = \rho_{2D} k_B T = \frac{m_w}{\pi\hbar^2} k_B T \tag{8.43}$$

Critical density

where n_c is the *critical density* of the subband. The total electron surface density in the system is then given by:

$$n_s = \sum_n n_n \tag{8.44}$$

Equations (8.43)–(8.44) give n_s as a function of E_F and T. They implicitly allow one to determine the Fermi level if n_s is known.

If the separation between the lowest levels is larger than the thermal energy (i.e. $\varepsilon_2 - \varepsilon_1 \gg k_B T$), and if the density n_s is not too great, then only one subband will be populated, and we may write for n_s and E_F:

$$\begin{aligned} n_s &= n_c \ln\left[1 + \exp\left(\frac{E_F - \varepsilon_1}{k_B T}\right)\right] \\ E_F &= \varepsilon_1 + k_B T \ln\left[\exp\left(\frac{n_s}{n_c}\right) - 1\right] \end{aligned} \tag{8.45}$$

Equation (8.45) shows that there exists a transition between *degenerate* and *non-degenerate* regimes for a two-dimensional electron gas. In the case where $n_s \gg n_c$, we obtain a degenerate electron gas for which:

$$\begin{aligned} n_s &= \rho_{2D}(E_F - \varepsilon_1) \\ E_F &= \varepsilon_1 + \frac{n_s}{\rho_{2D}} \end{aligned} \tag{8.46}$$

and non-degenerate if $n_s \ll n_c$, in which case:

$$\begin{aligned} n_s &= n_c \exp\left(-\frac{\varepsilon_1 - E_F}{k_B T}\right) \\ E_F &= \varepsilon_1 - k_B T \ln\left(\frac{n_s}{n_c}\right) \end{aligned} \tag{8.47}$$

In Fig. 8.9, we show both linear and logarithmic plots for $n_s(E_F)$:

Example

We will calculate the density of states and the critical density for a GaAs quantum well.

The effective mass m_c in the GaAs well is $0.067 m_0$ resulting in a density ρ_{2D} of:

$$\rho_{2D} = 1.6 \times 10^{-19}\,\mathrm{C}\,\frac{0.067 \times 0.9 \times 10^{-30}\,\mathrm{kg}}{3.14 \times (1.05 \times 10^{-34}\,\mathrm{J\,s})^2} = 2.78 \times 10^{13}\,\mathrm{eV^{-1}\,cm^{-2}}$$

The critical density n_c is then $n_c = \rho_{2D} \times 25.9\,\mathrm{meV}$ or $7.2 \times 10^{11}\,\mathrm{cm^{-2}}$.

A $10^{12}\,\mathrm{cm^{-2}}$ doped quantum well is therefore degenerate. The position of the

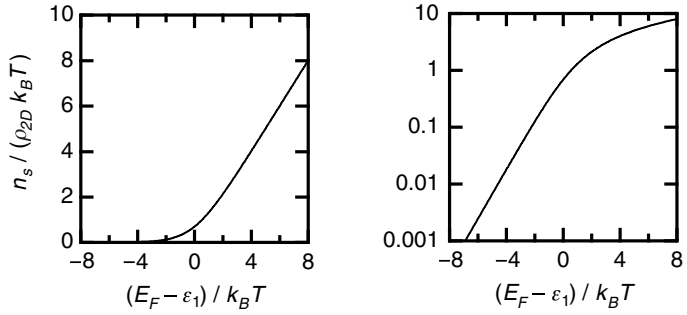


Fig. 8.9. Density of states in a subband as a function of the position of the Fermi level position E_F. (*a*) Linear and (*b*) logarithmic plots.

Fermi level above the confinement energy is given by (8.46) or:

$$E_F - \varepsilon_1 = 10^{12}\,\mathrm{cm}^{-2}/(2.78 \times 10^{13}\,\mathrm{eV}^{-1}\,\mathrm{cm}^{-2}) = 36\,\mathrm{meV}$$

8.5 Optical interband transitions in a quantum well

8.5.1 Hole states in the valence bands

A pair of materials forming a type I heterostructure can be used to implement a quantum well for electrons in the conduction band. In the valence band, the potential profile seen by the electrons is an inverted well. If we make the (very) simplifying approximation that the valence band may be represented by two parabolic bands (one for the heavy holes with effect mass $m_{hh} < 0$, and one for the light holes with $m_{lh} < 0$) we have, using the envelope approximation, the two following Schrödinger equations:

$$\begin{aligned}
&\left[-\frac{\mathrm{d}}{\mathrm{d}z}\frac{\hbar^2}{2m^*_{hh}(z)}\frac{\mathrm{d}}{\mathrm{d}z} + V(z)\right]\zeta_n^{hh}(z) = \varepsilon_n^{hh}\zeta_n^{hh}(z) \\
&\left[-\frac{\mathrm{d}}{\mathrm{d}z}\frac{\hbar^2}{2m^*_{lh}(z)}\frac{\mathrm{d}}{\mathrm{d}z} + V(z)\right]\zeta_n^{lh}(z) = \varepsilon_n^{lh}\zeta_n^{lh}(z)
\end{aligned} \tag{8.48}$$

with the potential:

$$\begin{aligned}
&V(z) = 0,\ -\frac{L}{2} < z < \frac{L}{2} \\
&V(z) = -V_B = -\Delta E_v,\ \frac{L}{2} < |z|
\end{aligned} \tag{8.49}$$

The three-dimensional solutions to these Schrödinger equations are then:

$$
\begin{aligned}
E_{n\mathbf{K}}^{hh} &= E_v + \varepsilon_n^{hh} + \frac{\hbar^2 K^2}{2m_{hh}} \\
E_{n\mathbf{K}}^{lh} &= E_v + \varepsilon_n^{lh} + \frac{\hbar^2 K^2}{2m_{lh}} \\
\psi_{n\mathbf{K}}^{hh} &\cong \frac{1}{\sqrt{A}} \zeta_n^{hh}(z) \exp(\mathrm{i}\mathbf{K}\cdot\mathbf{R}) u_{hh0}(\mathbf{r}) \\
\psi_{n\mathbf{K}}^{lh} &\cong \frac{1}{\sqrt{A}} \zeta_n^{lh}(z) \exp(\mathrm{i}\mathbf{K}\cdot\mathbf{R}) u_{lh0}(\mathbf{r})
\end{aligned}
\tag{8.50}
$$

where E_v is the valence band maximum for the well material, and the masses and eigenenergies ε_n are negative. Clearly (see Fig. 5.13) we may consider the problem as involving positive masses by changing the signs of both the masses and the energies. In Schrödinger's equation, this corresponds to substituting $V(z) \to -V(z)$ and $\varepsilon_n \to -\varepsilon_n$. The well potential is then recognized as a confining potential for holes (the direction of the positive energies is then towards the bottom of Fig. 8.10) and the dispersion of the corresponding subbands for the holes is convex as was the case for conduction electrons in Fig. 8.6.

Using this model, the main effect of the quantum well is to lift the degeneracy between the heavy hole and light hole bands at the zone centre due to their effective mass differences ($|\varepsilon_n^{hh}| \leq |\varepsilon_n^{lh}|$). In actuality, the subband structure is far more complex (we shall return to this matter in Complement 8.D). The simple model used here is analogous to that used in probing the optical properties of bulk semiconductors where we neglected mass anisotropy in the valence band.

8.5.2 Optical transitions between the valence and conduction bands

We are now interested in transitions between different states in a quantum well under the influence of a sinusoidally time-varying perturbation such as an electromagnetic wave. Three types of transitions can be imagined: (i) an *interband* transition in which an electron in a valence subband may be excited to a conduction subband; (ii) an *intersubband* transition, in which an electron moves from one subband to another while remaining in the same band; and (iii) an *intrasubband* transition, in which an electron is promoted to a different **K** state within the same subband n.

This last type of transition may result from a scattering process as discussed in the context of free-carrier absorption in Complement 7.C. In this case, interaction with a phonon or an impurity is required to furnish the necessary exchange of

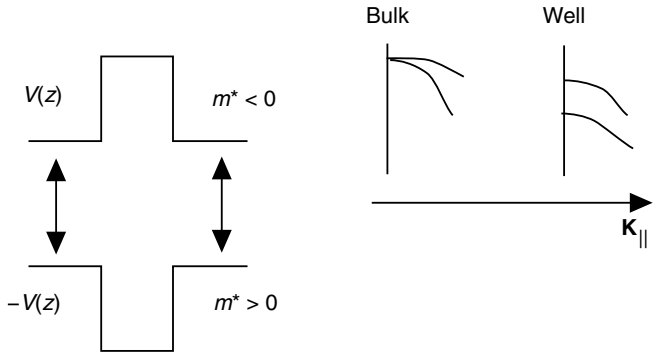


Fig. 8.10. (*a*) The potential seen by an electron in the valence band is equivalently seen as an attractive well by holes. (*b*) The well potential lifts the degeneracy of the heavy hole and light hole bands as a result of the differences in the effective masses.

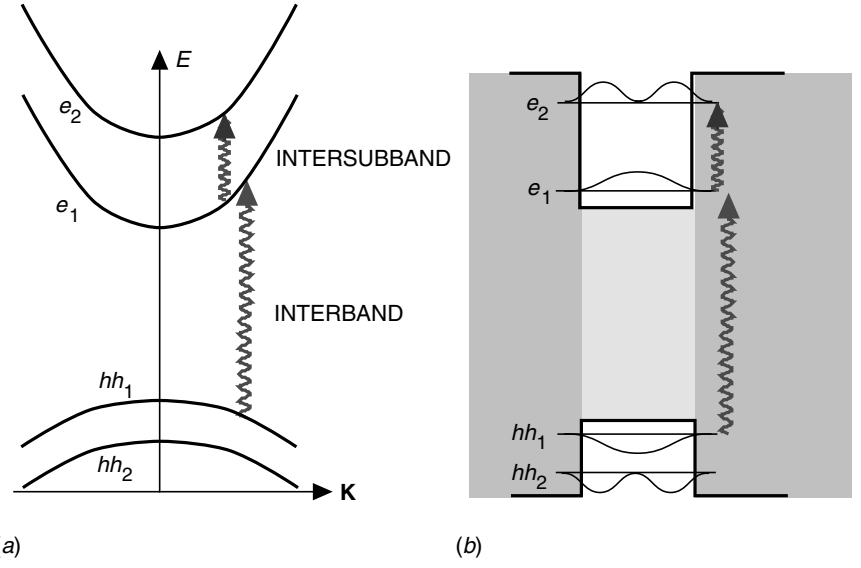


Fig. 8.11. Optical transitions are vertical in **K** space. Interband and intersubband transitions are represented (*a*) in reciprocal space and (*b*) in real space.

momentum between the lattice and the electron (see Fig. 7.C.2). The two other transition types are depicted in Fig. 8.11 for the case involving an interaction with an electromagnetic wave. The novel feature introduced by the two-dimensional aspect of the quantum well system is that *optical intersubband transitions* are allowed to first order. This process is evidently not allowed in a bulk system. We will discuss this important mechanism in the following section.

For interband transitions, we can calculate the transition rate using Fermi's golden rule. We consider a wave travelling across the well:

$$\mathbf{E}(\mathbf{r},t)=\mathbf{E}_{\mathbf{q}}\cos(\mathbf{q}\cdot\mathbf{r}-\omega t)=\frac{1}{2}\mathbf{E}_{\mathbf{q}}\exp[\mathrm{i}(\mathbf{q}\cdot\mathbf{r}-\omega t)]+\frac{1}{2}\mathbf{E}_{\mathbf{q}}^{*}\exp[-\mathrm{i}(\mathbf{q}\cdot\mathbf{r}-\omega t)] \quad (8.51)$$

The electric dipole interaction potential (see Section 3.2 and Complement 3.D) with a quantum well electron is:

$$V_{\mathbf{q}}(\mathbf{r},t)=-e\mathbf{E}\cdot\mathbf{r}=-\frac{e}{2}\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}\exp[\mathrm{i}(\mathbf{q}\cdot\mathbf{r}-\omega t)]+\text{c.c.} \quad (8.52)$$

The transition rate from a state $\psi^{v}_{n\mathbf{K}}$ in the valence band, to a state $\psi^{c}_{m\mathbf{K}'}$ in the conduction band is given according to the golden rule as:

$$S(n\mathbf{K}\rightarrow m\mathbf{K}')=\frac{\pi}{2\hbar}e^{2}|\langle m\mathbf{K}'|\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}|n\mathbf{K}\rangle|^{2}\delta(E^{c}_{m\mathbf{K}'}-E^{v}_{n\mathbf{K}}-\hbar\omega) \quad (8.53)$$

The electric dipole matrix element has changed with respect to the bulk case (see Eq. (7.10)):

$$\langle m\mathbf{K}'|\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}|n\mathbf{K}\rangle \quad (8.54)$$
$$\frac{1}{A}\int\zeta^{c*}_{m}(z)u^{*}_{c0}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\mathbf{K}'\cdot\mathbf{R}}\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}\mathrm{e}^{\mathrm{i}\mathbf{q}\cdot\mathbf{r}}\zeta^{v}_{n}(z)u_{v0}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{K}\cdot\mathbf{R}}\mathrm{d}\mathbf{r}$$

Continuing with the recurring practice in this chapter, given that the envelope function varies slowly in comparison with lattice spacing, we may separate out the slow and rapid changing portions by posing $\mathbf{r}=\mathbf{r}'+(\mathbf{R}_i,z_i)$, where $\mathbf{r}'$ belongs to the primitive cell, and $\mathbf{r}_i=(\mathbf{R}_i,z_i)$ denotes the position of the ith primitive cell. The integral then becomes a sum over i:

$$\sum_{i_z}\zeta^{c*}_{m}(z_i)\zeta^{v}_{n}(z_i)\frac{1}{N_xN_{\mathrm{y}}}\sum_{i_x,i_y}\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{R}_i}$$

$$\times\frac{a_z}{\Omega_{\mathrm{cell}}}\Bigg[\int_{\mathrm{cell}}\mathrm{d}\mathbf{r}'u^{*}_{c0}(\mathbf{r}')\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}'u_{v0}(\mathbf{r}')\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{r}'} \quad (8.55)$$

$$+\mathbf{E}_{\mathbf{q}}\cdot\mathbf{r}_i\int_{\mathrm{cell}}\mathrm{d}\mathbf{r}'u^{*}_{c0}(\mathbf{r}')u_{v0}(\mathbf{r}')\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{r}'}\Bigg]$$

where a_z is the lattice spacing in the z direction. This last equation would be intractable were it not for our ability to make certain simplifying assumptions. To begin, we note within the sum over $\mathbf{R}_i$ a conservation rule for the parallel component of the momentum $\mathbf{K}+\mathbf{q}-\mathbf{K}'=\mathbf{0}$. As the wavevector of light is negligible in comparison to the electronic wavevectors $\mathbf{K}$ and $\mathbf{K}'$ in the Brillouin zone, this

momentum requirement is taken into account by the Kronecker delta $\Delta_{\mathbf{K},\mathbf{K}'}$ and is non-zero only for $\mathbf{K} = \mathbf{K}'$. For this same reason, the sum over z_i can be replaced by an integral, leading to the following expression for the first integral:

$$I_1 = \Delta_{\mathbf{K},\mathbf{K}'}\mathbf{E}_\mathbf{q} \cdot \mathbf{r}_{vc} \int \zeta_m^{c*}(z)\zeta_n^v(z)\mathrm{e}^{\mathrm{i}q_z z}\mathrm{d}z \tag{8.56}$$

wherein we find the matrix element $\mathbf{r}_{vc}$ of $\mathbf{r}$ between the *bulk* valence and conduction bands.

Additionally, if the z extension of the wavefunctions ($\sim$the well thickness) is very small in comparison to the wavelength λ, then, simply, we have:

$$I_1 = \Delta_{\mathbf{K},\mathbf{K}'}\mathbf{E}_\mathbf{q} \cdot \mathbf{r}_{vc}\langle m, c|n, v\rangle \tag{8.57a}$$

with

$$\langle m, c|n, v\rangle = \int \zeta_m^{c*}(z)\zeta_n^v(z)\mathrm{d}z \tag{8.57b}$$

In the second part of Eq. (8.55), the exponential inside the integral is practically a constant since $\lambda \gg a_z$ and $\mathbf{K}$ and $\mathbf{K}'$ are far from the edge of the Brillouin zone (otherwise the envelope function approximation is not valid) with the result being that the orthogonality of the Bloch functions cancels the term out.

The transition rate is therefore:

$$S(n\mathbf{K} \to m\mathbf{K}') = \frac{\pi e^2}{2\hbar}|\mathbf{E}_\mathbf{q} \cdot \mathbf{r}_{vc}|^2|\langle m, c|n, v\rangle|^2\delta(E_s - \hbar\omega)\Delta_{\mathbf{K},\mathbf{K}'} \tag{8.58a}$$

with

$$E_s = \varepsilon_m^c + \frac{\hbar^2K^2}{2m_c} - |\varepsilon_n^v| + \frac{\hbar^2K^2}{2m_v} + E_g = E_{\text{threshold}} + \frac{\hbar^2K^2}{2m_c} + \frac{\hbar^2K^2}{2m_v} \tag{8.58b}$$

Let us note first that the conservation of total energy introduces a threshold for optical absorption $\hbar\omega > E_{\text{threshold}} = E_g + \varepsilon_m^c + |\varepsilon_n^v| > E_g$. The optical absorption threshold in a quantum well is therefore *blueshifted*, with respect to the absorption which takes place in a bulk semiconductor, by an amount equal to the sum of the confinement energies present in the conduction and valence bands. This blueshift is used to characterize the physical parameters of a quantum well (barrier composition, well thickness, . . .). The generation rate of electron–hole pairs is equal to the number of transitions per second:

$$G = \frac{\mathrm{d}n}{\mathrm{d}t} = \frac{1}{A}\sum_{n,m,\mathbf{K},\mathbf{K}'} S(n\mathbf{K} \to m\mathbf{K}')f_n^v(\mathbf{K})[1 - f_m^c(\mathbf{K})] \tag{8.59}$$

where the occupation probabilities assure the presence of an electron in an initial

state $|n,v\rangle$ and an unoccupied final state $|m,c\rangle$. Conservation of energy and parallel momentum determine K as a function of the photon energy:

$$\frac{\hbar^2K^2}{2}\left(\frac{1}{m_c}+\frac{1}{|m_v|}\right)=\frac{\hbar^2K^2}{2m_r}=\hbar\omega-E_g-\varepsilon_m^c-|\varepsilon_n^v| \tag{8.60}$$

where we recognize m_r as the reduced mass. At thermodynamic equilibrium, we have as a consequence:

$$\begin{aligned} f_n^v(\mathbf{K})=f_n^v(\hbar\omega)&=\frac{1}{1+\exp\left(\dfrac{E_v-|\varepsilon_n^v|-\varepsilon_n^v(\hbar\omega)-E_F}{k_BT}\right)}\\ \varepsilon_n^v(\hbar\omega)&=\frac{m_r}{|m_v|}(\hbar\omega-E_g-\varepsilon_m^c-|\varepsilon_n^v|)\end{aligned} \tag{8.61}$$

and

$$\begin{aligned} f_m^c(\mathbf{K})=f_m^c(\hbar\omega)&=\frac{1}{1+\exp\left[\dfrac{E_c+\varepsilon_m^c+\varepsilon_m^c(\hbar\omega)-E_F}{k_BT}\right]}\\ \varepsilon_m^c(\hbar\omega)&=\frac{m_r}{m_c}(\hbar\omega-E_g-\varepsilon_m^c-|\varepsilon_n^v|)\end{aligned} \tag{8.62}$$

For transitions between a valence subband n and a conduction subband m, we may sum over $\mathbf{K}$:

$$\begin{aligned} G_{mn}&=\frac{\pi e^2}{2\hbar}|\mathbf{E_q}\cdot\mathbf{r}_{vc}|^2|\langle n,v|m,c\rangle|^2\times\frac{1}{A}\sum_{\mathbf{K}}f_n^v(\mathbf{K})[1-f_m^c(\mathbf{K})]\delta\left(\frac{\hbar^2K^2}{2m_r}+E_{\text{threshold}}-\hbar\omega\right)\\ &=\frac{\pi e^2}{2\hbar}|\mathbf{E_q}\cdot\mathbf{r}_{vc}|^2|\langle n,v|m,c\rangle|^2\times\int\frac{g_s\mathrm{d}^2\mathbf{K}}{(2\pi)^2}f_n^v(\hbar\omega)[1-f_m^c(\hbar\omega)]\delta\left(\frac{\hbar^2K^2}{2m_r}+E_{\text{threshold}}-\hbar\omega\right)\end{aligned}$$

or

$$G_{mn}=\frac{\pi}{2\hbar}\underbrace{e^2|\mathbf{E_q}\cdot\mathbf{r}_{vc}|^2|\langle n,v|m,c\rangle|^2}_{\text{dipole moment}}\underbrace{\frac{m_r}{\pi\hbar^2}\times\theta(\hbar\omega-E_{\text{threshold}})}_{\text{density of states}}\underbrace{f_n^v(\hbar\omega)[1-f_m^c(\hbar\omega)]}_{\text{occupation probability}} \tag{8.63a}$$

which we write again as:

$$G_{mn}=\frac{\pi e^2}{2\hbar}|\mathbf{E_q}\cdot\mathbf{r}_{vc}|^2|\langle n,v|m,c\rangle|^2\frac{m_r}{\pi\hbar^2}\times\theta(\hbar\omega-E_{\text{threshold}})f_n^v(\hbar\omega)[1-f_m^c(\hbar\omega)] \tag{8.63b}$$

Interband absorption rate for a quantum well ($s^{-1}\,cm^{-2}$)

In this expression, we recognize the product of the dipole moment, with the

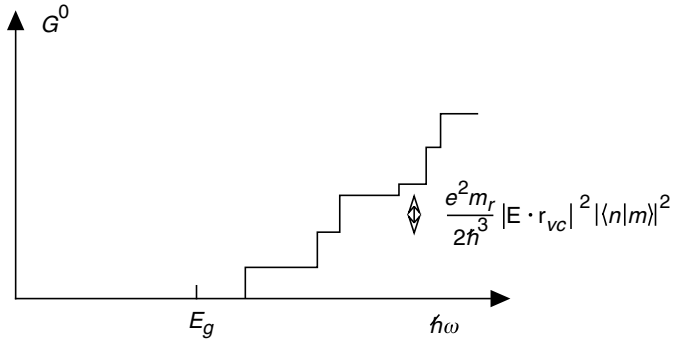


Fig. 8.12. The generation rate in an intrinsic quantum well is a staircase-shaped function. Each time a new subband transition becomes available at a given absorption energy, the generation rate increases a step. The energy threshold for pair generation resides above the bulk equivalent gap of the well material.

overlap element $\langle n, v|m, c\rangle$, the *joint density of states* in two dimensions, $m_r/\pi\hbar^2$, and the appropriate occupation statistics. The generation rate G_{mn} is therefore a product of the statistical factors f_m and f_n which depend on the temperature, the Fermi level, and the generation rate $G^0 = \Sigma G^0_{mn}$ that would occur if $f_m(1 - f_n)$ was equal to unity and dependent uniquely upon the structure. This is illustrated in Fig. 8.12.

The overlap $|\langle m, c|n, v\rangle|^2$ of the envelope functions can introduce selection rules. For example, in a symmetric quantum well, transitions between subbands with different parity are forbidden. For subbands lying deep within a well, the envelope functions for the electrons and holes are almost identical so that only transitions between states sharing identical indices ($m = n$) are possible.

The analysis carried out here for absorption induced transitions between the valence and conduction bands yields exactly the same results for stimulated emission (in this case involving transitions from the conduction to the valence band) with the exception of the occupation statistics. *The stimulated recombination* by the field is:

$$R = \sum_{m,n} G^0_{mn}[1 - f^v_n(\hbar\omega)]f^c_m(\hbar\omega) \tag{8.64}$$

and the net rate of creation–recombination induced by the field is therefore:

$$\left(\frac{\mathrm{d}n}{\mathrm{d}t}\right)_{\mathrm{net}} = \frac{\pi e^2}{2\hbar}|\mathbf{E_q} \cdot \mathbf{r}_{vc}|^2 \frac{m_r}{\pi\hbar^2} \times \sum_{n,m} |\langle m, c|n, v\rangle|^2 \theta(\hbar\omega - E_{\mathrm{threshold}})[f^v_n(\hbar\omega) - f^c_m(\hbar\omega)] \tag{8.65}$$

Net optical generation–recombination rate for a quantum well ($\mathrm{s}^{-1}\,\mathrm{cm}^{-2}$)

We recall finally that for two Bloch functions, with eigenstates sharing identical **k**s in different bands, we have:

$$\langle u_{c\mathbf{k}}|[H,\mathbf{r}]|u_{v\mathbf{k}}\rangle = (E_{c\mathbf{k}} - E_{v\mathbf{k}})\langle u_{c\mathbf{k}}|\mathbf{r}|u_{v\mathbf{k}}\rangle = -\frac{\mathrm{i}\hbar}{m_0}\langle u_{c\mathbf{k}}|\mathbf{p}|u_{v\mathbf{k}}\rangle \tag{8.66}$$

As indicated earlier in Chapter 7, the matrix element $\mathbf{r}_{vc}$ may therefore be expressed by $\mathbf{p}_{vc}$, the Kane matrix element used in Complement 5.C within the framework of the $\mathbf{k}\cdot\mathbf{p}$ method.

$$\mathbf{r}_{vc} = -\frac{\mathrm{i}\hbar}{E_g m_0}\mathbf{p}_{vc} \tag{8.67}$$

We may wonder whether we should replace the term E_g in (8.67) by $\hbar\omega$ in the expressions for the creation rate to take into account photons with energy in excess of the gap. We must not, however, ask too much of the envelope function approximation which can only be applied as long as $|\hbar\omega - E_g|/E_g \ll 1$.

The fact that the degeneracy of the valence band leads to distinct heavy hole and light hole subbands, has as a consequence that the matrix element $\mathbf{r}_{vc}$ depends upon the type of transition. Using the hole wavefunctions in Complement 5.C, the reader may convince himself/herself of the following selection rules given in Table 8.1.

Table 8.1. *Varying the electric field orientation with respect to the quantum well introduces selection rules for the transitions.* $\mathbf{r}_{\mathrm{vc}}$ *in equation* (8.65) *is equal to* $\mathbf{r}_{\mathrm{vc}}$ *in the bulk material multiplied by the appropriate factor*(*s*) *appearing in the table*

	TM ($E\|z$)	TE ($E\perp z$)
$hh \to e$	0	$1/\sqrt{2}$
$lh \to e$	$\sqrt{2}/\sqrt{3}$	$1/\sqrt{6}$

In particular, the heavy hole subband to electron subband transitions are forbidden in the TM configuration (i.e. for the electric field perpendicular to the well interfaces – see Chapter 9).

8.6 Optical intersubband transitions in a quantum well

In an intersubband transition, the initial and final states of the electron belong to the same band. This leads to completely different selection rules and behaviours from those observed for interband transitions. To be specific, let us suppose that

this band is the conduction band (see Fig. 8.11). The initial and final states are then:

$$\psi_{n\mathbf{K}} \cong \frac{1}{\sqrt{A}}\zeta_n(z)\exp(\mathrm{i}\mathbf{K}\cdot\mathbf{R})u_{c0}(\mathbf{r})$$
$$\psi_{n\mathbf{K}'} \cong \frac{1}{\sqrt{A}}\zeta_m(z)\exp(\mathrm{i}\mathbf{K}'\cdot\mathbf{R})u_{c0}(\mathbf{r}) \tag{8.68}$$

which share the same periodic Bloch functions $u_{c0}(\mathbf{r})$.

Fermi's golden rule again gives us the transition rate induced by the electromagnetic field (Eq. (8.53)). For the matrix element, the only difference relative to Eqs. (8.54) and (8.55) is that the initial state sits within the conduction band.

$$\begin{aligned}
&\langle m\mathbf{K}'|\mathbf{E}_\mathbf{q}\cdot\mathbf{r}|n\mathbf{K}\rangle \\
&= \sum_{i_z}\zeta_m^*(z_i)\zeta_n(z_i)\frac{1}{N_xN_y}\sum_{i_x,i_y}\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{R}_i} \\
&\quad\times\frac{a_z}{\Omega_{\mathrm{cell}}}\Bigg[\int_{\mathrm{cell}}\mathrm{d}\mathbf{r}'u_{c0}^*(\mathbf{r}')\mathbf{E}_\mathbf{q}\cdot\mathbf{r}'u_{c0}(\mathbf{r}')\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{r}'} \\
&\quad+\mathbf{E}_\mathbf{q}\cdot\mathbf{r}_i\int_{\mathrm{cell}}\mathrm{d}\mathbf{r}'u_{c0}^*(\mathbf{r}')'u_{c0}(\mathbf{r}')\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{r}'}\Bigg]
\end{aligned} \tag{8.69}$$

where the sum over $\mathbf{R}_i$ clearly relates to momentum conservation in the parallel direction. For the first integral, the matrix element of $\mathbf{r}'$ is now zero, as $u_{c0}(\mathbf{r})$ transforms as 1 under the cubic symmetry operations of the crystal lattice.

The second integral in (8.69) this time yields Ω_{cell} when the variation of the exponential is negligible over the primitive cell. We are then left with the sum:

$$I_2 = \sum_{i_z}\zeta_m^*(z_i)\zeta_n(z_i)\frac{a_z}{N_xN_y}\sum_{i_x,i_y}\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{R}_i}\mathbf{E}_\mathbf{q}\cdot\mathbf{r}_i \tag{8.70}$$

which we can turn into an integral in the z direction:

$$I_2 = \frac{1}{N_xN_y}\sum_{i_x,i_y}\mathrm{e}^{\mathrm{i}(\mathbf{K}+\mathbf{q}-\mathbf{K}')\cdot\mathbf{R}_i}\int\zeta_m^*(z)\mathbf{E}_\mathbf{q}\cdot\mathbf{r}\zeta_n(z)\mathrm{d}z \tag{8.71}$$

As the envelope functions are orthogonal, this last integral introduces an intersubband selection rule indicating that *only the E_{qz} component can induce a transition* (see Fig. 8.13).

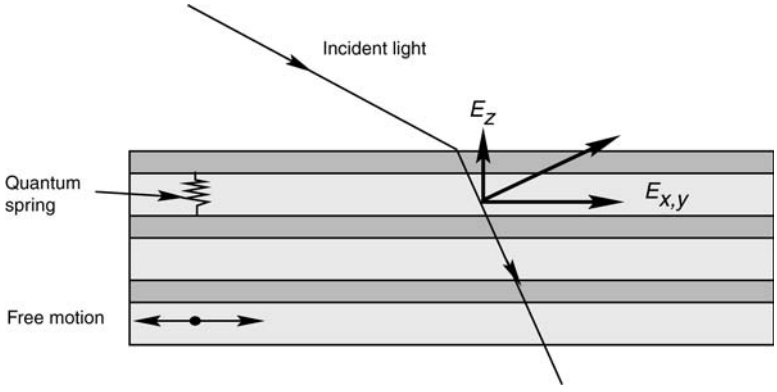


Fig. 8.13. In intersubband transitions, only the component of the electric field normal to the quantum wells can optically couple to the 'quantum oscillators' which result from quantization of the allowed displacement. The parallel components are left to interact (weakly) with a free two-dimensional electron gas.

Integral (8.71) may then be written:

$$I_2 = \Delta_{\mathbf{K},\mathbf{K}'} E_{qz} \int \zeta_m^*(z) z \zeta_n(z) \mathrm{d}z \tag{8.72}$$

leading to the transition rate:

$$S(n\mathbf{K} \to m\mathbf{K}') = \frac{\pi e^2}{2\hbar} |E_{qz}|^2 |\langle m|z|n\rangle|^2 \delta(\varepsilon_m - \varepsilon_n - \hbar\omega) \Delta_{\mathbf{K},\mathbf{K}'} \tag{8.73a}$$

We note that for this last expression, the argument in the Dirac δ *no longer depends on* $\mathbf{K}$ since the subbands are parallel (Fig. 8.14). Therefore, even if the optical transitions bring into play interactions between light and delocalized electronic states within the medium (as indicated by the existence of the energy subband), this interaction is nonetheless resonant *as though the system displayed discrete levels*!

The photon energies resonant with the levels in the quantum well are given by:

$$\hbar\omega = \varepsilon_m - \varepsilon_n \tag{8.73b}$$

To obtain the electronic excitation rate from one subband to another, statistical considerations must be applied, as was the case for *interband* transitions. As a result, the excitation–relaxation rate can be obtained merely by summing over $\mathbf{K}$:

$$G = \left(\frac{\mathrm{d}n_m}{\mathrm{d}t}\right)_{\text{net}} = \frac{1}{A} \sum_{n,\mathbf{K},\mathbf{K}'} S(n\mathbf{K} \to m\mathbf{K}')[f_n(\mathbf{K}) - f_m(\mathbf{K})] \tag{8.74}$$

As $S(n\mathbf{K} \to m\mathbf{K}')$ does not depend on $\mathbf{K}$, the result is particularly simple:

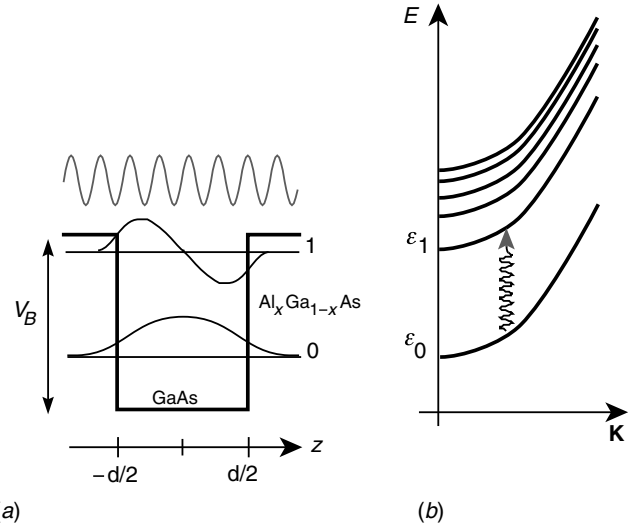


Fig. 8.14. A quantum well with two bound states (*a*), and (*b*) corresponding total energy dispersion curves *E*, resulting in as many subbands. As the subbands are parallel, constant **K** transitions give rise to energy resonances (i.e. constant energy transitions over a broad range of **K** values between pairs of subbands).

$$G = \left(\frac{\mathrm{d}n_n}{\mathrm{d}t}\right)_{\text{net}} = \frac{\pi e^2}{2\hbar}|E_{qz}|^2 \sum_m |\langle n|z|m\rangle|^2 \delta(\varepsilon_m - \varepsilon_n - \hbar\omega)(n_n - n_m)$$
$$\langle n|z|m\rangle = \int \zeta_m^*(z) z \zeta_n(z) \mathrm{d}z \tag{8.75}$$

Optical intersubband generation–recombination rate ($s^{-1}\,cm^{-2}$)

where n_n and n_m are electronic densities in the nth and mth subbands, respectively. The dependence of G on the photon energy is depicted in Fig. 8.15 and is quite different from that shown in Fig. 8.12 for interband transitions. We recognize as well in (8.75), expressions (1.85b) and (3.63) established during the discussion of atomic transitions between discrete levels. Such transitions between bound states in a quantum well are referred to as *bound-to-bound* transitions.

Above the barrier, the subbands do not form a discrete spectrum. The envelope functions in this case are extended and the energy levels are free. They can be classified according to their component wavevector k_z in the barrier:

$$\psi_{k_z\mathbf{K}'} \cong \frac{1}{\sqrt{\Omega}} \zeta_{k_z}(z) \exp(\mathrm{i}\mathbf{K}' \cdot \mathbf{R}) u_{c\mathbf{0}}(\mathbf{r})$$
$$\frac{\hbar^2 k_z^2}{2m_B} = \varepsilon_{k_z} - V_B \tag{8.76}$$

The theory for the excitation rate results from a simple generalization of (8.75) by replacing the sum over m, by an integration over k_z (see Complement 1.A):

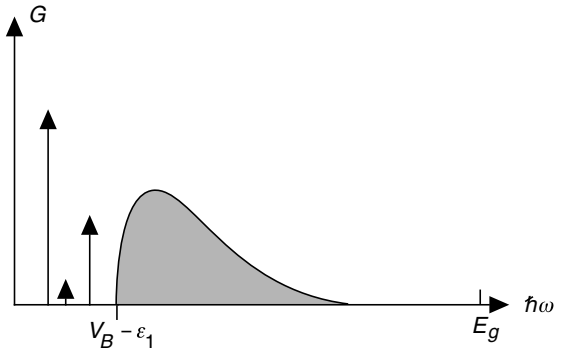


Fig. 8.15. Excitation rate from the fundamental subband as a function of photon energy. The subbands in a quantum well give rise to discrete transitions (*bound-to-bound* transitions). Above the barriers, the subbands form a continuum and may become implicated in *bound-to-continuum* transitions.

$$G = \frac{\pi e^2}{2\hbar} n_n |E_{qz}|^2 \times \int_0^\infty \frac{L_z \mathrm{d}k_z}{\pi} |\langle k_z|z|n\rangle|^2 \delta\left(\frac{\hbar^2 k_z^2}{2m_B} + V_B - \varepsilon_n - \hbar\omega\right) \tag{8.77}$$

We have supposed that the population of states in the barrier n_{kz} is negligible. We note that for all photon energies $\hbar\omega \geq V_B - \varepsilon_n$ a transition from a subband n to a continuum is clearly possible. This situation is analogous to the phenomenon of photoionization which was touched upon in Complement 1.A, and to which we will return in more detail in Chapter 11 during our study of quantum well detectors.

8.7 Optical absorption and angle of incidence

Knowing the optical transition rates for the different mechanisms (interband, intersubband), we can now go about calculating absorption coefficients for quantum wells by applying the principle that each transition corresponds to the absorption of a photon.

We will begin by offering a summary of the characteristics for each of these transition types.

8.7.1 Summary for interband and intersubband transition rates

Figure 8.16 summarizes the qualitative characteristics of interband and intersubband transition rates derived in the two preceding sections. The essential point is that the intersubband transitions lead to resonant absorption spectra for photons with energies equal to the energy spacing between parallel subbands. On

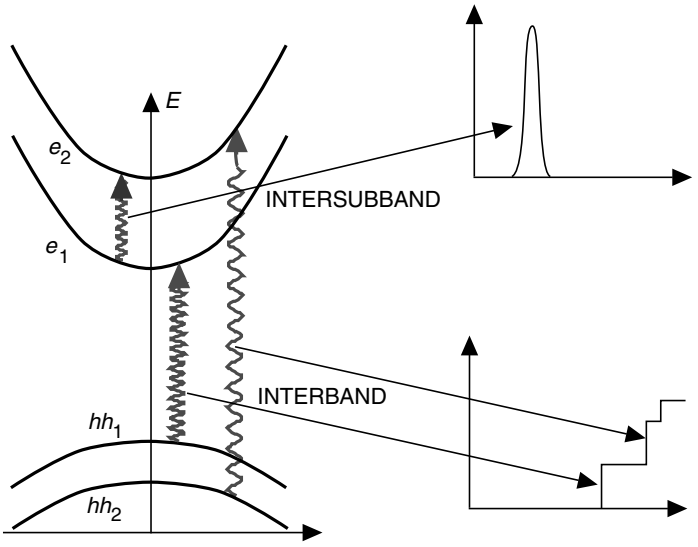


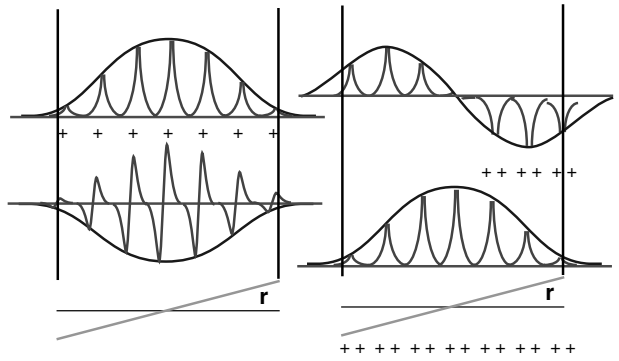
Fig. 8.16. Photon absorption spectra for interband and intersubband transitions are qualitatively different. As electronic subbands are parallel, all transitions between the two electron subbands occur at the same energy. In the case of interband transitions, the absorption spectra take on a staircase appearance, with each step corresponding to the transition energy due to a particular coupling of specific electron and hole subbands.

the other hand, interband transitions lead to staircase like absorption spectra, with each step corresponding to the transition threshold of a paired electron subband and hole subband.

The dipole moments which determine the oscillator strengths for the transitions are also of a different nature, as illustrated in Fig. 8.17. For interband transitions, the dipole moment results from a series of atomic dipoles between Bloch functions (see (8.57)). The effective dipolar matrix element, $r_{\text{eff}} = r_{vc}\langle m, c|n, v\rangle$, is of the order of 0.6 nm in GaAs if $\langle m, c|n, v\rangle = 1$. Alternatively, in intersubband transitions, the dipole results from orthogonality of the envelope functions (see (8.75)). The effective dipolar matrix element, $r_{\text{eff}} = m|z|n\rangle$, is on the order of 5 nm for a very deep 8 nm GaAs/AlAs quantum well. *For a given bulk semiconductor*, the dipole matrix elements for intersubband transitions can therefore be significantly larger than those corresponding to interband transitions.

8.7.2 Influence of the angle of incidence

The principle behind this calculation is identical to that presented for the bulk semiconductor case (see Eq. (7.42)) with a few more subtleties added to take into account the two-dimensional nature of the quantum well system. We may well ask how the absorption coefficient may be expected to change depending on whether the light is normally or perpendicularly incident upon the quantum well. We will



(*a*) Transition $hh_1 \longrightarrow e_1$ (*b*) Transition $e_1 \longrightarrow e_2$

Fig. 8.17. The dipole moments for transitions result either from (*a*) dipoles between different Bloch functions (interband transitions), or from (*b*) dipoles existing between different envelope functions (intersubband transitions).

now present a detailed description for the light absorption mechanisms occurring in a quantum well.

The net transition rate G (in $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) corresponds to an absorbed power W from the electromagnetic wave equal to:

$$W = \hbar\omega G \ [\mathrm{W\,cm}^{-2}] \tag{8.78}$$

The material containing the quantum well is by essence not isotropic. To calculate the absorption experienced by a wave travelling through the material, we must therefore take into account the geometry and the spatial extent of the light ray (see Fig. 8.18).

Figure 8.18 shows a general case. A plane wave with a flux distribution given by $\Phi(r)$, encounters a quantum well which subtends an angle θ to its direction of propagation. The amplitude of the electric field is then given by:

$$\Phi(r) = \frac{1}{2}\varepsilon c|E_q|^2 = \frac{n(\omega)}{2Z_0}|E_q|^2 \tag{8.79}$$

where $n(\omega)$ is the index of refraction and $Z_0 = (\mu_0/\varepsilon_0)^{1/2}$ is the vacuum impedance. The energy which traverses the structure per unit second is:

$$W_{\mathrm{in}} = \int_{-\infty}^{\infty} \Phi(r)\mathrm{d}r = \frac{1}{2}\varepsilon c \int_{-\infty}^{\infty} |E_q(r)|^2 \mathrm{d}r \tag{8.80}$$

The total absorption rate in a quantum well normalized by the incident energy W_{in} becomes:

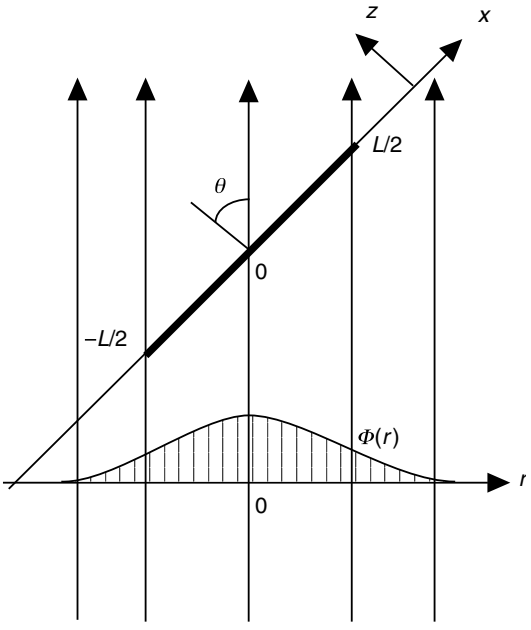


Fig. 8.18. An electromagnetic wave with an intensity distribution $\Phi(r)$ travels across a material and encounters a quantum well of length L oriented at an angle θ to the propagation direction.

$$\frac{W}{W_{\text{in}}} = \frac{\hbar\omega G[\mathbf{E}_q(0)]}{\Phi(0)} \frac{\displaystyle\int_{-L/2}^{L/2} \Phi(r\cos\theta)\mathrm{d}x}{\displaystyle\int_{-\infty}^{\infty} \Phi(r)\mathrm{d}r} = \frac{\hbar\omega G[\mathbf{E}_q(0)]}{\Phi(0)} \frac{\dfrac{1}{\cos\theta}\displaystyle\int_{-L\cos\theta/2}^{L\cos\theta/2} \Phi(r)\mathrm{d}r}{\displaystyle\int_{-\infty}^{\infty} \Phi(r)\mathrm{d}r} \tag{8.81}$$

where the geometry of the problem is contained in the ratio of the integrals. We recall that both the generation rate and the intensity Φ are proportional to $|E_q|^2$. The prefactor therefore depends only upon the polarization of the field relative to the well.

Two limiting cases emerge from this last expression (Fig. 8.19):

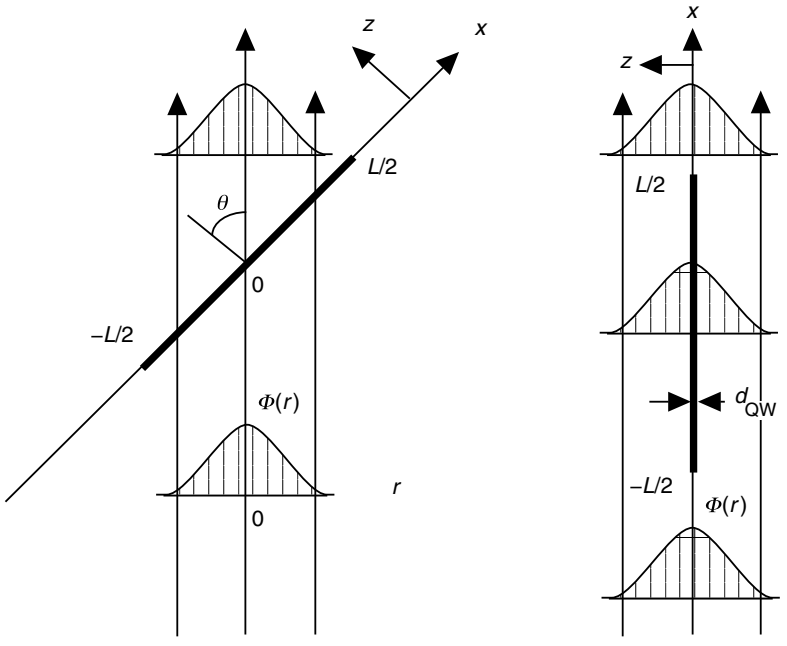


Fig. 8.19. Two limiting geometries for absorption: (i) the absorption does not depend upon the well thickness, (ii) the absorption depends on the length of the well section encountered, and on the overlap of the electromagnetic wave with the well.

1. If the *waist* of the light ray is small enough so that the entire flux traverses the well, the two integrals are equal, and the absorption α_{QW} is:

$$\alpha_{QW} = \frac{W}{W_{in}} = \frac{\hbar\omega G[\mathbf{E}_q(0)]}{\Phi(0)\cos\theta} \text{ [per well]} \tag{8.82}$$

   We note that the width of the well does not explicitly appear in this expression for the absorption. This absorption rate is independent of the well width and is *dimensionless* (expressed as a percentage). On the other hand, the angle θ intervenes in (8.82) and also in the expression for the transition rate G, according to the selection rules given in Table 8.1.
2. If the waist of the light ray is large, which, in particular, is the case for $\theta = \pi/2$, and the intensity does not vary in the well, the absorption is:

$$\frac{W}{W_{in}} = \frac{\hbar\omega G[\mathbf{E}_q(0)]}{\Phi(0)} \frac{\Gamma}{d_{QW}} L \tag{8.83}$$

   where we have introduced the *confinement factor*:

$$\Gamma = \frac{|\mathbf{E}_q(0)|^2 d_{\mathrm{QW}}}{\int_{-\infty}^{\infty} |\mathbf{E}_q(r)|^2 \mathrm{d}r} \tag{8.84}$$

where d_{QW} is the well thickness and Γ is the fraction of the wave which 'sees' the absorbing medium.

In this geometry, it is natural to introduce the absorption coefficient through the expression:

$$\alpha = -\frac{1}{W_{\mathrm{in}}}\frac{\mathrm{d}W_{\mathrm{in}}}{\mathrm{d}x} = \frac{\hbar\omega G[\mathbf{E}_q(0)]}{\Phi(0)}\frac{\Gamma}{d_{\mathrm{QW}}} = \alpha_{\mathrm{QW}}\frac{\Gamma}{d_{\mathrm{QW}}}\ [\mathrm{cm}^{-1}] \tag{8.85}$$

Absorption coefficient for propagation parallel to the well

We have now only to calculate the absorption coefficient for the quantum well, α_{QW}, by using the expressions for the different optical transition rates G which we found in (8.65) for the interband and in (8.75) for intersubband transitions.

Let us begin with the interband transitions. Given expressions (8.65), (8.79), and (8.82), the absorption for a quantum well is given by the contributions from transitions relating to each of the different pairs of subbands involved:

$$\alpha_{\mathrm{QW}} = \sum_{n,m} \alpha_{\mathrm{QW}}^{n\to m}\theta(\hbar\omega - E_g - \varepsilon_m^c - |\varepsilon_n^v|)[f_n^v(\hbar\omega) - f_m^c(\hbar\omega)] \tag{8.86a}$$

with the absorption coefficients for each transition (see Eq. (8.75)):

$$\alpha_{\mathrm{QW}}^{n\to m} = \frac{m_r}{\hbar^3}\hbar\omega\frac{Z_0}{n(\omega)}[e|\eta_q\cdot\mathbf{r}_{vc}\|\langle m,c|n,v\rangle|]^2 \tag{8.86b}$$

Contribution from interband transitions $n \to m$ to quantum well absorption (in %)

In the case of intersubband absorption, absorption also results from transitions between each pair of subbands:

$$\alpha_{\mathrm{QW}} = \sum_{n,m>n} \alpha_{\mathrm{QW}}^{n\to m} \tag{8.87a}$$

giving for each transition (see Eq. (8.75)):

$$\alpha_{\mathrm{QW}}^{n\to m} = \frac{m_r}{\hbar}\hbar\omega\frac{Z_0}{n(\omega)}[e\eta_{qz}|\langle n|z|m\rangle|]^2(n_n - n_m)\delta(\varepsilon_m - \varepsilon_n - \hbar\omega) \tag{8.87b}$$

Contribution from intersubband transitions $n \to m$ to quantum well absorption (%)

In this last expression, we recall that the δ function is in J^{-1} and that it can be replaced by a Lorentzian if the transition is broadened for any particular reason.

Furthermore, we note that $\eta_q = \mathbf{E}_q/|E_q|$ is the polarization vector of the electromagnetic wave and $\eta_{qz} = \sin\theta$ its component along the growth axis of the quantum well. For $n = 1$ and $m = 2$, (8.87b) becomes identical to (3.37) and (3.39) for optical absorption in a two-level system obtained using density matrix formalism. This underlines the profound similarity between intersubband transitions in semiconductors and transition processes in atomic physics.

We see in (8.86b) and (8.87b) a very different angular behaviour between the two types of mechanisms. Most notably, intersubband absorption reveals a characteristic dependence on $\sin^2\theta$, and is therefore null at normal incidence ($\theta = 0$). As we shall see later in Chapter 11, this has very important consequences for quantum-well-based detectors.

Example

1. Interband absorption in a GaAs quantum well

For the fundamental transition hh_1–e_1 in a GaAs quantum well, we may consider the overlap of the envelope functions to be equal to unity. For an incident electromagnetic wave normal to the well, the interband absorption is given by (8.86b) and we find for $\hbar\omega = 1.5\,\text{eV}$:

$$\alpha_{\text{QW}} = \frac{\hbar\omega e^2 m_r}{\hbar^3}\frac{Z_0}{n(\omega)}|\eta_q \cdot \mathbf{r}_{vc}|^2$$

$$= \frac{1.5\,\text{eV} \times 0.059 \times 9.1 \times 10^{-31}\,\text{kg}}{(6.58 \times 10^{-16}\,\text{eV s})^3}\frac{377\,\Omega}{\sqrt{12}}\left(\frac{1}{\sqrt{2}}6 \times 10^{-10}\,\text{m}\right)^2 = 0.55\%$$

where we have used the selection rules in Table 8.1, and the $\mathbf{r}_{cv}$ value for bulk absorption given in Table 7.1.

2. Intersubband absorption in a GaAs quantum well

We consider here a deep GaAs quantum well allowing us to approximate its low lying states by those of an infinite square well of equal width. We will assume a width of 10 nm, a doping level of $10^{12}\,\text{cm}^{-2}$, and that the temperature is sufficiently low so that only the fundamental level is occupied. The separation between the two bottom levels is given by (1.49), or:

$$E_{12} = (4-1)\frac{\hbar^2\pi^2}{2m_c a^2} = \frac{3}{2}\frac{(\pi \times 1.05 \times 10^{-34}\,\text{J s})^2}{0.067 \times 0.9 \times 10^{-30}\,\text{kg} \times (10^{-8}\,\text{m})^2} = 169\,\text{meV}$$

The dipolar matrix element is given by (3.D.24) to be:

$$|\langle 1|z|2\rangle| = \frac{2^4}{(3\pi)^2}a = 0.18a = 1.8\,\text{nm}$$

Finally, the intersubband absorption at resonance for a wave incident at 45° is

given by (8.87b). Assuming a broadening $\hbar\Gamma$ of 10 meV we replace the δ function by $1/(\pi\hbar\Gamma)$ such that:

$$\alpha_{\mathrm{QW}}^{n\to m} = \frac{1}{\hbar}\frac{\hbar\omega}{\hbar\Gamma}\frac{Z_0}{n(\omega)}[e\eta_{qz}|\langle n|z|m\rangle|]^2 n_1$$

$$= \frac{1}{1.05\times 10^{-34}\,\mathrm{J\,s}} \times \frac{0.169}{0.01} \times \frac{377\,\Omega}{3.3}$$

$$\times \frac{(1.6\times 10^{-19}\,\mathrm{C}\times 1.8\times 10^{-9}\,\mathrm{m})^2}{2} 10^{16}\,\mathrm{m}^{-2} \approx 8\times 10^{-3}$$

FURTHER READING

G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Wiley, New York (1991).

S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley Interscience, New York (1995).

C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures*, Academic Press, Boston (1991).

Complement to Chapter 8

8.A Quantum wires and boxes

We saw that semiconductors can be used to create quantum well structures which act to confine the motion of carriers in the z (or growth) direction. If the motion of the carriers is confined in an additional direction (x, for example) the motion of electrons in the structure will remain free only in the single remaining y direction. In this case we will have effectively fashioned an electronic waveguide.

In a technological sense, there are two general approaches for creating such double confinement, see Fig. 8.A.1:

1. Lithography can be used to segment a quantum well laterally into wires.
2. A negative voltage applied to a metallic gate separated by a very short distance can be used. By increasing the electronic potential below the contacts, we create linear (parallel) depletion regions in the quantum well, except in the region lying beneath the space separating the contacts. By increasing the magnitude of the applied voltage, the lateral dimension of the electron waveguide can be decreased beyond *pinch-off*, in which case there is nothing left of the canal in the well.

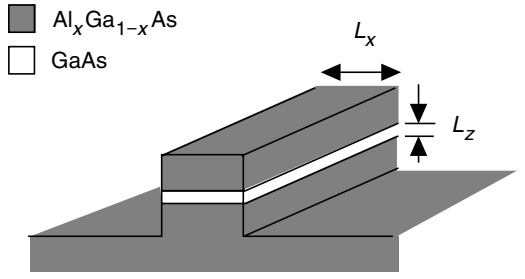
Schrödinger's equation for the envelope function $\psi(\mathbf{r})$ describing the electronic states in a *quantum wire* is clearly:

$$\left(\frac{p_x^2 + p_y^2 + p_z^2}{2m^*} + V(x,z)\right)\psi(x,y,z) = E\psi(x,y,z) \tag{8.A.1}$$

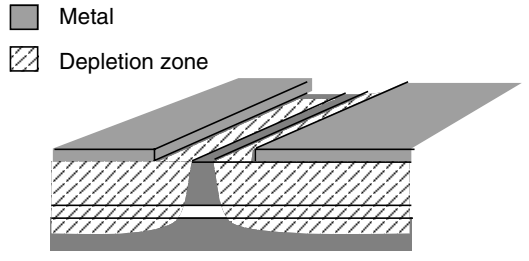
The potential $V(x,z)$ which reigns in such a quantum wire preserves translational symmetry in the y direction. The envelope wavefunction $\psi(\mathbf{r})$ therefore remains separable but with a plane wave in the y direction only, i.e:

$$\psi(x,y,z) = \frac{1}{\sqrt{L_y}}\xi(x,z)\mathrm{e}^{\mathrm{i}k_y y} \tag{8.A.2}$$

The function $\zeta(x,z)$ satisfies the two-dimensional Schrödinger equation:



(*a*)



(*b*)

Fig. 8.A.1. Two methods used to fabricate a quantum wire. (*a*) Definition of a fine mesa by lithography, and (*b*) laterally symmetric depletion of carriers from a quantum well by a potential applied to suitably patterned metal contacts.

$$\left(\frac{p_x^2 + p_z^2}{2m^*} + V(x,z)\right)\zeta_i(x,z) = E_i\zeta_i(x,z) \tag{8.A.3}$$

In general, these equations describe *unidimensional subbands* with discrete bound states $|i\rangle$ in two dimensions and free motion in the y direction. For a given subband i, the energy of the state (i, k_y) is:

$$E_{i,k_y} = E_i + \frac{\hbar^2 k_y^2}{2m^*} \tag{8.A.4}$$

Generalizing the reasoning in Fig. 8.6, the density of states in a subband may be calculated by posing periodic limits in the y direction:

$$D(k_y) = \frac{g_s L_y}{2\pi} \tag{8.A.5}$$

where we recall that g_s is the spin degeneracy coefficient ($g_s = 2$). The density of states in energy may be easily found by writing the density $\mathrm{d}n_i$ over an infinitesimal energy element $\mathrm{d}E$:

$$\mathrm{d}n_i = D_i(E)\mathrm{d}E = 2D_i(k_y)\mathrm{d}k_y \tag{8.A.6a}$$

where the factor of 2 is included since at a given energy E, there are two possible

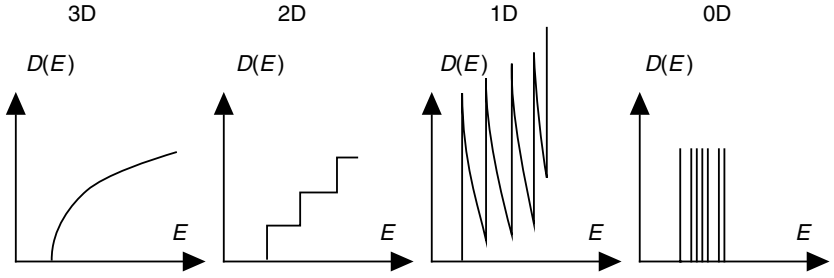


Fig. 8.A.2. Density of states from 3 to 0 dimensions. The overall behaviour of $D(E)$ is seen to depend drastically upon the number of confined dimensions. Units for $D(E)$ are $[\mathrm{eV}^{-1}\,\mathrm{cm}^{-D}]$, where D is the number of unconstrained dimensions.

values for k_y given by (8.A.4). The density of states in energy is then given by:

$$D_i(E) = D_i(k_y)2\left|\frac{\mathrm{d}E}{\mathrm{d}k_y}\right|^{-1} = \frac{g_s L_y}{\pi\hbar}\sqrt{\frac{m^*}{2(E-E_i)}} \tag{8.A.6b}$$

possessing a singularity at $E = E_i$ (see Fig. 8.A.2).

Solution of (8.A.3) generally entails numerical methods or approximations. Nonetheless, if we choose to represent the potential as a rectangular wire with dimensions of $L_x \times L_z$ possessing infinite barriers, the solution may be easily obtained by separating the variables:

$$\zeta_i(x,z) = \zeta_{n_x n_z}(x,z) = \frac{2}{\sqrt{L_x L_z}}\sin\frac{n_x \pi x}{L_x}\sin\frac{n_z \pi z}{L_z} \tag{8.A.7}$$

where $n_x = 1, 2, 3, \ldots, n_z = 1, 2, 3, \ldots$. The quantum confinement energy corresponding to the one-dimensional subband (n_x, n_y) is trivially:

$$\varepsilon_{n_x,n_z} = \frac{\hbar^2\pi^2}{2m^{*2}}\left(\frac{n_x^2}{L_x^2} + \frac{n_z^2}{L_z^2}\right) \tag{8.A.8}$$

We see that if $L_x \gg L_z$, the levels n_x form a ladder with small steps inside the ladder levels resulting from the well separated subbands in n_z. Alternately, if $L_x = L_z$, the two sets of levels cannot be resolved, and many of the levels are degenerate.

We may now extend this reasoning down to zero dimensions. A potential which confines electrons in three dimensions will create completely bound states possessing a discrete spectrum. In this case, we speak of a *quantum box*, a *quantum dot*, or even of an *artificial atom*. Quantum dots can be fabricated with surprising ease by depositing a semiconductor with a very different lattice spacing relative to a substrate of different composition. The required mismatches in lattice constants gives rise to extremely thin *critical thicknesses* as in the case of InAs deposited onto GaAs. In this case, the InAs forms small islands on the GaAs substrate material

leading to pyramidally shaped quantum boxes possessing heights and widths defined on a nanometre scale. Figure 8.A.2 summarizes the results for the density of states in 3, 2, 1, and 0 dimensions.

FURTHER READING

C. Weisbuch and B. Vinter, *Quantum Semiconductor Structures*, Academic Press, Boston (1991).

8.B Excitons

Our treatment of semiconductors up to this point has rested upon the *single electron approximation*. A semiconductor actually consists of a colossal ensemble of interacting atomic nuclei and electrons. As a result, the level of success encountered using this simple approximation can be viewed as nothing less than miraculous. In fact, a semiconductor is a system which presents us with a certain fundamental state. What we observe are *excitations* of this system away from this baseline. What we have described as being 'the energy of an electron' actually reflects the energy difference between a system possessing the extra electron and its fundamental state (without the additional electron). Similarly, the 'energy of a hole' corresponds to the energy difference between the system in its fundamental state and the resulting system obtained by removing an electron.

The justification as to why such excitations can be satisfactorily approximated by single particle Schrödinger equations is a problem best dealt with by *many body theory*. A large part of the success of this approximation results from the form of the crystal potential, which we have only scarcely specified. It turns out that this potential must make allowances for nuclear Coulomb potentials, and screening of the nuclear potential by electrons by including what is known as the *exchange–correlation potential* V_{xc}. This potential describes the energy that the system can acquire through the correlated motion of its electrons. This correlation in motion allows the mean separation between electrons to increase somewhat, leading to a reduction in the Coulomb energy due to electron–electron interaction. The lesson here is that for many types of excitation, the single electron approximation is reasonable, provided the crystal potential has been well chosen.

Implicit in the absorption calculations we carried out in Chapters 7 and 8 for an electromagnetic wave, was the approximation that an excitation of the system which conserves the number of particles may be written as a sum of excitations corresponding to the generation of electron–hole pairs in a system. In the exchange between an electromagnetic wave and a semiconductor, total energy is conserved. A photon disappears and the electronic system is excited (generation of

an electron–hole pair) into a state higher in energy by the amount contained in the photon. In this excited state, however, the correlation between the electronic motions differs from that in the fundamental state. This introduces qualitative changes in the excited states going beyond the quantitative effects included in the single electron description. Again, one must draw upon many body theory to show that this correlation can be represented by the interaction between a hole with positive charge and mass (+e and m_h), and an electron with mass m_e and negative charge. The demonstration of this result, however, would require an enormous investment, and is beyond the scope of this book.

8.B.1 Three-dimensional excitons

We seek to describe the state of an electron–hole pair by taking into account the interaction between their opposite charges. Clearly, we require a two-particle state, i.e. resulting from the tensor product (see Eq. (2.48)) between the electron and hole states. The two-particle wavefunction may be expanded in terms of Bloch functions for the electron and hole $|\mathbf{k}_1, \mathbf{k}_2\rangle$:

$$\begin{aligned}\psi(\mathbf{r}_1, \mathbf{r}_2) &= \frac{1}{\Omega^2} \sum_{\mathbf{k}_1, \mathbf{k}_2} C(\mathbf{k}_1, \mathbf{k}_2) \mathrm{e}^{\mathrm{i}\mathbf{k}_1 \cdot \mathbf{r}_1} u^c_{\mathbf{k}_1}(\mathbf{r}_1) \mathrm{e}^{\mathrm{i}\mathbf{k}_2 \cdot \mathbf{r}_2} u^v_{\mathbf{k}_2}(\mathbf{r}_2) \\ &= \sum_{\mathbf{k}_1, \mathbf{k}_2} C(\mathbf{k}_1, \mathbf{k}_2) |\mathbf{k}_1, \mathbf{k}_2\rangle\end{aligned} \tag{8.B.1}$$

The state of the two particles is then given by the solution to the two-body Schrödinger equation:

$$\left[\frac{p_1^2}{2m_e} + \frac{p_2^2}{2m_h} + V_c(\mathbf{r}_1) + V_c(\mathbf{r}_2) - \frac{e^2}{4\pi\varepsilon |\mathbf{r}_1 - \mathbf{r}_2|} \right] \psi(\mathbf{r}_1, \mathbf{r}_2) = E \psi(\mathbf{r}_1, \mathbf{r}_2) \tag{8.B.2}$$

Let us substitute in this last equation, (8.B.1) and project the result onto $|\mathbf{k}'_1, \mathbf{k}'_2\rangle$, yielding:

$$\begin{aligned}&\left(E_g + \frac{\hbar^2 k_1^2}{2m_e} + \frac{\hbar^2 k_2^2}{2m_h} \right) C(\mathbf{k}_1, \mathbf{k}_2) \\ &+ \sum_{\mathbf{k}'_1, \mathbf{k}'_2} C(\mathbf{k}'_1, \mathbf{k}'_2) \langle \mathbf{k}_1, \mathbf{k}_2 | V(\mathbf{r}_1 - \mathbf{r}_2) | \mathbf{k}'_1, \mathbf{k}'_2 \rangle = E C(\mathbf{k}_1, \mathbf{k}_2)\end{aligned} \tag{8.B.3}$$

Using the fact that the interaction potential varies slowly with respect to the dimensions of the primitive cell, we make use of the reasoning employed in Section .3 and (courageously!) introduce the envelope function to simplify the matrix element:

$$
\begin{aligned}
&\langle \mathbf{k}_1, \mathbf{k}_2 | V(\mathbf{r}_1 - \mathbf{r}_2) | \mathbf{k}_1', \mathbf{k}_2' \rangle \\
&= \frac{1}{\Omega^2} \int d\mathbf{r}_1 d\mathbf{r}_2 e^{i(\mathbf{k}_1' - \mathbf{k}_1)\cdot(\mathbf{r}_1 - \mathbf{r}_2)} V(\mathbf{r}_1 - \mathbf{r}_2) e^{i(\mathbf{k}_2' - \mathbf{k}_2 - \mathbf{k}_1' + \mathbf{k}_1)\cdot \mathbf{r}_2} \\
&= \frac{1}{\Omega} \int d\mathbf{r}_2 \tilde{V}(\mathbf{k}_1 - \mathbf{k}_2) e^{i(\mathbf{k}_2' - \mathbf{k}_2 - \mathbf{k}_1' + \mathbf{k}_1)\cdot \mathbf{r}_2} \\
&= \frac{1}{\Omega} \Delta_{\mathbf{k}_2' - \mathbf{k}_2 - \mathbf{k}_1' + \mathbf{k}_1} \tilde{V}(\mathbf{k}_1 - \mathbf{k}_1') \\
&= \frac{1}{\Omega} \Delta_{\mathbf{k}_2' - \mathbf{k}_2 - \mathbf{q}} \tilde{V}(\mathbf{q})
\end{aligned}
\tag{8.B.4}
$$

where

$$
\tilde{V}(\mathbf{q}) = \frac{1}{\Omega} \int d\mathbf{r} e^{-i\mathbf{q}\cdot\mathbf{r}} V(\mathbf{r}) \tag{8.B.5}
$$

is the Fourier transform of the Coulomb interaction. To arrive at (8.B.4), we have assumed that the periodic portions of the Bloch functions do not depend on $\mathbf{k}$.

We may then write Schrödinger's equation, (8.B.3):

$$
\begin{aligned}
&\left(E_g + \frac{\hbar^2 k_1^2}{2m_e} + \frac{\hbar^2 k_2^2}{2m_h} \right) C(\mathbf{k}_1, \mathbf{k}_2) \\
&+ \sum_{\mathbf{q}} \tilde{V}(\mathbf{q}) C(\mathbf{k}_1 - \mathbf{q}, \mathbf{k}_2 + \mathbf{q}) = E C(\mathbf{k}_1, \mathbf{k}_2)
\end{aligned}
\tag{8.B.6}
$$

As $\mathbf{k}_1 + \mathbf{k}_2$ is conserved in the interaction term, it is in our interest to change the variables $\mathbf{k}_1$ and $\mathbf{k}_2$, into new variables $\mathbf{K}$ and $\mathbf{k}$, where:

$$
\begin{aligned}
&\mathbf{K} = \mathbf{k}_1 + \mathbf{k}_2, \text{ and } \frac{\mathbf{k}}{m_r} = \frac{\mathbf{k}_1}{m_e} - \frac{\mathbf{k}_2}{m_h} \\
&\mathbf{k}_1 = \frac{m_e}{M}\mathbf{K} + \mathbf{k}, \text{ and } \mathbf{k}_2 = \frac{m_h}{M}\mathbf{K} - \mathbf{k}
\end{aligned}
\tag{8.B.7}
$$

where $M = m_e + m_h$ is the total mass and $1/m_r = 1/m_e + 1/m_h$ defines the *reduced mass* m_r. With these new variables we have:

$$
\begin{aligned}
&C(\mathbf{k}_1, \mathbf{k}_2) = C(\mathbf{K}, \mathbf{k}) \\
&C(\mathbf{k}_1 - \mathbf{q}, \mathbf{k}_2 + \mathbf{q}) = C(\mathbf{K}, \mathbf{k} + \mathbf{q})
\end{aligned}
\tag{8.B.8}
$$

and Schrödinger's equation becomes:

$$\left(E_g + \frac{\hbar^2 K^2}{2M} + \frac{\hbar^2 k^2}{2m_r}\right) C(\mathbf{K}, \mathbf{k}) + \sum_{\mathbf{q}} \tilde{V}(\mathbf{q}) C(\mathbf{K}, \mathbf{k} + \mathbf{q}) = E C(\mathbf{K}, \mathbf{k}) \tag{8.B.9}$$

The sum over $\mathbf{q}$ due to the interaction between the electron and the hole only concerns the second argument for C. This signifies that $C(\mathbf{K}, \mathbf{k})$ is separable, and may be written as $C(\mathbf{K}, \mathbf{k}) = f(\mathbf{K}) g(\mathbf{k})$, where $g(\mathbf{k})$ must obey:

$$\left(E_g + \frac{\hbar^2 K^2}{2M} + \frac{\hbar^2 k^2}{2m_r}\right) g(\mathbf{k}) + \sum_{\mathbf{q}} \tilde{V}(\mathbf{q}) g(\mathbf{k} + \mathbf{q}) = E g(\mathbf{k}) \tag{8.B.10}$$

We may now introduce the *envelope function* for the relative motion:

$$g(\mathbf{r}) = \sum_{\mathbf{k}} g(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \tag{8.B.11}$$

where we recognize an equivalent definition to that given in (8.12). The operations are the same as those performed on (8.14) in arriving at (8.17) and they lead to Schrödinger's equation, (8.B.10), in real space:

$$-\frac{\hbar^2 \nabla_{\mathbf{r}}^2}{2m_r} g(\mathbf{r}) + V(\mathbf{r}) g(\mathbf{r}) = \left(E - \frac{\hbar^2 K^2}{2M} - E_g\right) g(\mathbf{r}) \tag{8.B.12}$$

We recognize in this last equation, Schrödinger's equation for a hydrogen atom (or for *positronium* better yet!). The allowed energies are then:

$$E_{n\mathbf{K}} = E_g + \frac{\hbar^2 K^2}{2M} - \frac{\mathrm{Ry}^*}{n^2}, \; n = 1, 2, \ldots$$
$$\mathrm{Ry}^* = \frac{m_r}{m_0} \frac{\varepsilon_0^2}{\varepsilon^2} \mathrm{Ry} \tag{8.B.13}$$

where Ry is the ionization energy of the *s* state of the hydrogen atom. This quantity is known as the Rydberg and is equal to 13.6 eV. As we might have expected, correlation between the electron and the hole creates discrete states below the gap through which the electron and the hole become bound to one another (in terms of their relative motion), while remaining free to move through the crystal as a pair. The term $\hbar^2 K^2/2M$ corresponds to the kinetic energy associated with the free motion of the centre of mass of the two particles. Above $E_g + \hbar^2 K^2/2M$, we have a continuum of allowed energy states. This continuum corresponds to an excitation of the system resulting effectively in release of the electron and hole from one other. Figure 8.B.1 represents the excitation spectrum of a semiconductor.

Interaction of the system with an electromagnetic wave conserves total energy

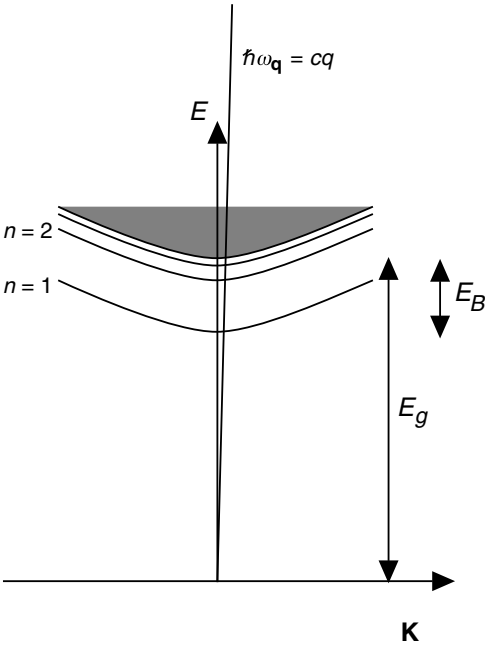


Fig. 8.B.1. The excitation spectrum for a semiconductor displays bound excitonic states below the bandgap. The coupling between a photon with wavevector **q** conserves total momentum and energy. The absorption spectrum therefore consists of a few discrete lines followed by a continuum.

and momentum. This implies that $\mathbf{K} = \mathbf{q} \approx \mathbf{0}$, where **q** is the photon wavevector. The absorption spectrum therefore displays discrete lines below the bandgap and a continuum for $\hbar\omega > E_g$.

The electron–hole correlation has other consequences for the semiconductor aside from apparition of discrete lines in the absorption spectra. The continuum absorption also increases near the gap. As a result, the absorption as a function of the photon energy shows a discontinuity instead of the more gentle dependence $(\hbar\omega - E_g)^{1/2}$ calculated in Chapter 7. This effect (described by the *Sommerfeld factor*) is due to spatial overlap of the electron and hole wavefunctions due to correlations which seek to draw these two particles to each other. Figure 8.B.2 compares the enhanced absorption resulting from the excitonic states with that obtained using the single electron model.

Example

1. Excitons in GaAs

For GaAs, we have $m_e = 0.067m_0$, $m_{hh} = 0.51m_0$, and $\varepsilon = 12\varepsilon_0$. The binding energy for the $n = 1$ exciton is then:

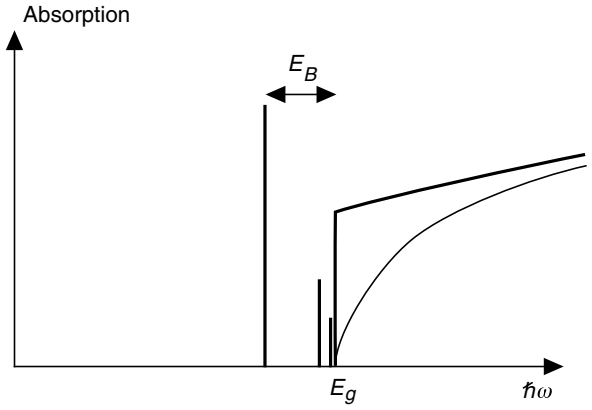


Fig. 8.B.2. Excitonic correlation between an electron and a hole modifies the absorption spectrum. With respect to the 'non-correlated' spectrum (fine line), the absorption shows sharp excitonic resonances below the bandgap and an abrupt increase just above the bandgap.

$$E_B = \frac{m_r}{m_0}\left(\frac{\varepsilon_0}{\varepsilon}\right)^2 \mathrm{Ry} = \frac{0.059}{12^2}\,13.6\,\mathrm{eV} = 5.6\,\mathrm{meV}$$

We see that E_B is weak in comparison to $k_B T$ at room temperature. As a result, the exciton can only remain bound at very low temperatures. The average separation between an electron and a hole in a $n = 1$ exciton is (in analogy to the Bohr radius for a hydrogen atom):

$$a_B^* = \frac{m_0}{m_r}\frac{\varepsilon_0}{\varepsilon}a_B = \frac{12}{0.059}\,0.052\,\mathrm{nm} = 10.6\,\mathrm{nm}$$

and is consistent with our supposition of a slowly varying envelope function in comparison to the primitive cell dimensions.

2. Excitons in GaN
For GaN, a large-gap semiconductor, $E_g = 3.4\,\mathrm{eV}$, $m_e = 0.22m_0$, $m_{hh} \approx m_0$, and $\varepsilon = 9.8\varepsilon_0$. The reduced mass is then $0.18m_0$, $E_B = 26\,\mathrm{meV}$, and $a_B^* = 2.8\,\mathrm{nm}$. As a result, excitonic effects are much more important in large-gap semiconductors. At room temperature a fraction, $1 - \exp(-26/25.9)$, or 64% of generated excitons will remain bound.

8.B.2 Two-dimensional excitons

We saw that, in a quantum well, the motion of electrons can be confined over distances of the order of a few nanometers, i.e. over distances significantly smaller than the Bohr radius of the three-dimensional exciton. A consequence of this confinement is that the bands are quantized into subbands separated in energy by

a few tens of meV and greater than the binding energy of the three-dimensional exciton. For correlations between electrons in the fundamental conduction subband and holes in the fundamental valence band, the Coulomb attraction can only occur along directions parallel to the interfaces. In comparison to the confinement potential, the Coulombic potential does not play a significant role in the perpendicular direction. In this case, the excitonic wavefunction is:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{A^2} \sum_{\mathbf{K}_1, \mathbf{K}_2} C(\mathbf{K}_1, \mathbf{K}_2) \zeta_1^c(z_1) \mathrm{e}^{\mathrm{i}\mathbf{K}_1 \cdot \mathbf{R}_1} u_{c0}(\mathbf{r}_1) \zeta_1^v(z_2) \mathrm{e}^{\mathrm{i}\mathbf{K}_2 \cdot \mathbf{R}_2} u_{v0}(\mathbf{r}_2) \tag{8.B.14}$$

where the vectors **R** and **K** designate motion parallel to the interfaces. Using the same transformations as in the preceding section, we obtain a Schrödinger equation which governs the motion in the plane:

$$-\frac{\hbar^2 \nabla_{\mathbf{R}}^2}{2m_r} g(\mathbf{R}) + V_{\mathrm{eff}}(\mathrm{R}) g(\mathbf{R}) = \left(E - \frac{\hbar^2 K^2}{2M} - \varepsilon_1^c - \varepsilon_1^v - E_g \right) g(\mathbf{R}) \tag{8.B.15}$$

where the effective interaction potential is:

$$V_{\mathrm{eff}}(\mathbf{R}) = \int \mathrm{d}z_1 \mathrm{d}z_2 \frac{e^2 |\zeta_1^c(z_1)|^2 |\zeta_1^v(z_2)|^2}{4\pi\varepsilon \sqrt{R^2 + (z_1 - z_2)^2}} \approx \frac{e^2}{4\pi\varepsilon R} \tag{8.B.16}$$

This last approximation is valid in the limiting case where the envelope functions have a negligible extension in comparison to the excitonic radius. The solution of this equation for a 'two-dimensional hydrogenic atom' yields the following eigenenergies:

$$E_{n\mathbf{K}} = E_g + \varepsilon_1^c + \varepsilon_1^v + \frac{\hbar^2 K^2}{2M} - \frac{\mathrm{Ry}^*}{(n - 1/2)^2}, n = 1, 2, \ldots \tag{8.B.17}$$

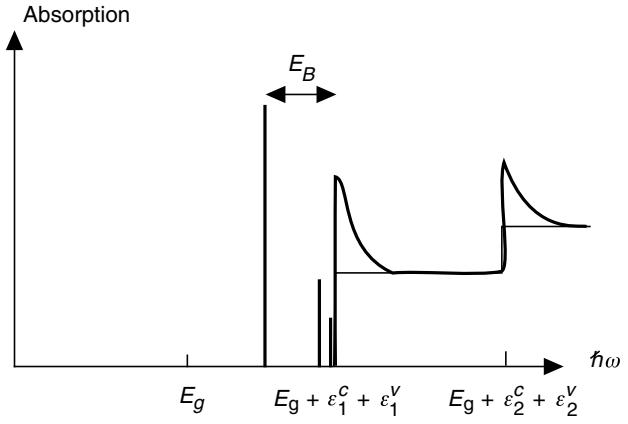
using the same effective Rydberg constant as defined in the last section.

The two-dimensional exciton therefore *has a fundamental binding energy four times that of the equivalent three-dimensional exciton.*

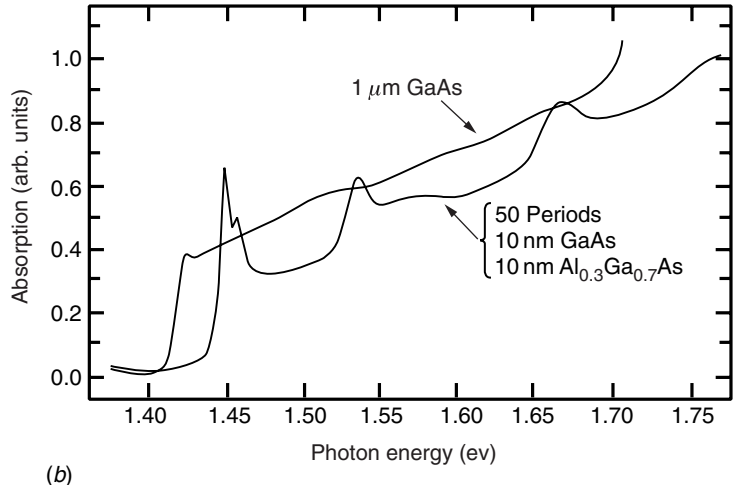
Example

We consider a 5 nm wide GaAs quantum well. In the preceding example we saw that GaAs possesses a three-dimensional excitonic binding energy of 5.6 meV. In this example, the well width is significantly less than the exciton's Bohr radius $a_B^* = 10.6\,\mathrm{nm}$, and the binding energy is $4 \times E_B$ or 22.4 meV. As a result, a significant fraction of excitons in an ensemble, ($1 - \mathrm{e}^{-22.4/25.9}$ or 58%) will remain bound at room temperature. Correspondingly, excitonic features are readily apparent in the absorption and photoluminescence features from these wells (see Fig. 8.B.3).

In the more realistic case where the spatial extent of the wavefunctions $\zeta^c(z)$ and



(*a*)



(*b*)

Fig. 8.B.3. The absorption spectrum for a quantum well is modified by the correlation between electrons and holes, i.e. by the presence of excitons. In comparison to the case involving only free electron–hole pairs (fine lines), excitons cause the emergence of discrete absorption lines below the effective quantum well bandgap, and each step resulting from continuum absorption is deformed by a narrow absorption peak (*a*). This result is particularly clear in experiments involving quantum wells with varying widths (*b*) (after S. Schmitt-Rink, D. S. Chemla, and D. A. B. Miller, *Adv. Phys.* **38**, 89 (1989), reprinted with permission of Taylor & Francis Ltd, http://www.tandf.co.uk/journals).

$\zeta^v(z)$ are not entirely negligible in comparison to a_B^*, the effective interaction in (8.B.16) is weakened with respect to the purely two-dimensional case, and the excitonic binding energy diminishes gradually towards the three-dimensional value as the well width increases beyond the effective Bohr radius. A quantitative treatment of this effect requires that other subbands in addition to the fundamental be taken into account. Only numerical methods afford a reasonable description of this intermediate regime between two and three dimensions.

As in the three-dimensional case, correlation between the electron and the hole profoundly modifies the interband absorption spectrum for quantum wells. Below the effective gap $E_g + \varepsilon_1^c + \varepsilon_1^v$, the spectrum contains discrete lines. Above the gap, there is an increase in the absorption relative to the free electron and hole case. Figure 8.B.3 illustrates this qualitative modification to the absorption spectra.

FURTHER READING

R. J. Elliott, *Phys. Rev.* **108**, 1384 (1957).
H. Haug and S. W. Koch, *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, World Scientific, New York (1990).
S. Schmitt-Rink, D. S. Chemla, and D. A. B. Miller, *Adv. Phys.* **38**, 89 (1989).

8.C Quantum confined Stark effect and the SEED electromodulator

In Complement 1.C we saw how an electric field F applied perpendicularly to a quantum well will displace the energy levels in the well (or levels which we now know correspond to the bottoms of subbands). This *quantum confined Stark effect* can be exploited in an electromodulator. In this case, an applied voltage is used to modulate the intensity of a transmitted light beam.

For this, we choose a frequency for the light beam in such a manner so that at $F = 0$, the photon energy $h\nu_0$ is situated just beneath the effective quantum well bandgap:

$$h\nu_0 < E_g + \varepsilon_1^c(0) + \varepsilon_1^v(0) \tag{8.C.1}$$

Under these conditions, light is not absorbed by the well (see Fig. 8.C.1).

Using the simple model in Complement 1.C, the level ε_1^c decreases according to Eq. (1.C.2):

$$\varepsilon_1^c(F) = \varepsilon_1^c(0) - e^2F^2\frac{|z_{12}|^2}{\varepsilon_1^c(0) - \varepsilon_2^c(0)} \tag{8.C.2}$$

where, for a very deep well (Eq. (1.C.4)):

$$\varepsilon_1^c(F) = \varepsilon_1^c(0) - \frac{2^9}{3^5\pi^6}\frac{e^2L^4m_e}{\hbar^2}F^2 \tag{8.C.3}$$

The fundamental hole level is displaced in the same fashion:

$$\varepsilon_1^v(F) = \varepsilon_1^v(0) - \frac{2^9}{3^5\pi^6}\frac{e^2L^4m_h}{\hbar^2}F^2 \tag{8.C.4}$$

and the effective gap (Fig. 8.C.2) decreases (or *redshifts*) according to:

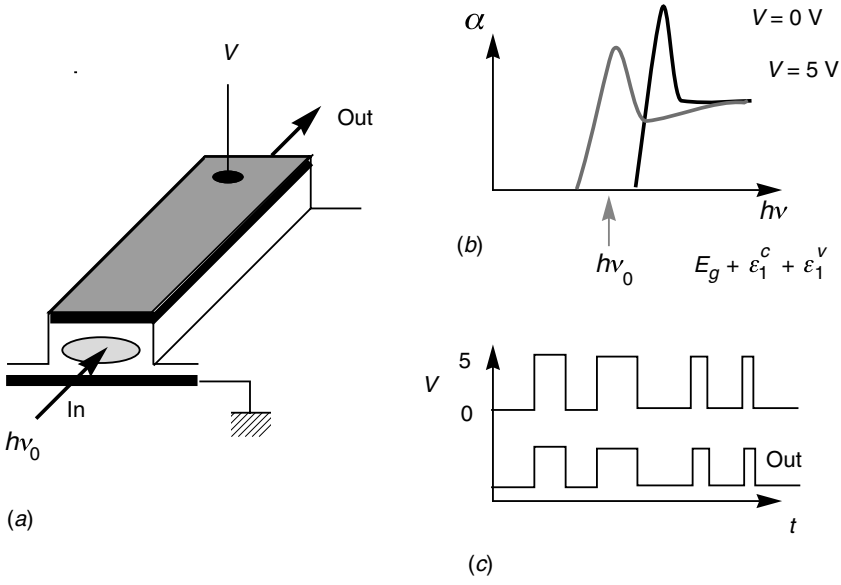


Fig. 8.C.1. Operation of a modulator based on the quantum confined Stark effect.

$$E_{g,\mathrm{eff}}(F) = E_{g,\mathrm{eff}}(0) - \frac{2^9}{3^5\pi^6}\frac{e^2L^4M}{\hbar^2}F^2 \tag{8.C.5}$$

where $M = m_e + m_h$. For a fixed $h\nu_0$, the effective gap can become smaller than $h\nu_0$ and the well absorbs light from the beam. The absorption edge is extremely abrupt (see Figs 8.10 or 8.B.3). As a result, a small modulation in electric field leads to strong modulation in the intensity of the transmitted light. The modulation amplitude can also be increased by adding several quantum wells in series.

Example

For a GaAs quantum well of width $L = 200\,\text{Å}$, and an applied field strength of $F = 10\,\mathrm{kV\,cm^{-1}}$, we find:

$$|\Delta E_{g,\mathrm{eff}}| = \frac{2^9}{3^5\pi^6}\frac{1.6\times10^{-19}\,\mathrm{C}(2\times10^{-8}\,\mathrm{m})^4(0.067+0.57)9.1\times10^{-31}\,\mathrm{kg}}{(1.05\times10^{-34}\,\mathrm{J\,s})^2}$$

$$\times(10^6\,\mathrm{V\,m^{-1}})^2\mathrm{e} = 2.7\,\mathrm{meV}$$

For wavelengths of the order of $1.5\,\mu\mathrm{m}$, this corresponds to a redshift in the wavelength of the absorption edge of $\Delta\lambda/\lambda = \Delta E/E$, or $\Delta\lambda = 5\,\mathrm{nm}$.

We note from Eq. (8.C.5), that the confined Stark effect exhibits a very strong dependence on the quantum well width L. While this effect favours large well widths, the magnitude of the effect cannot be arbitrarily increased by increasing the well width. This limitation occurs because the overlap of the electron and hole

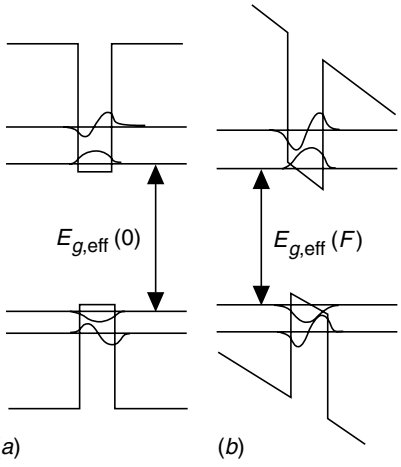


Fig. 8.C.2. Under the effect of an electric field F, the effective quantum well bandgap and the associated absorption edge decrease in energy. As a result, photons with energies lying initially below the effective bandgap at zero field (*a*), will be absorbed by the quantum well when a field is applied (*b*).

wavefunctions (localized at opposing interfaces under the influence of the applied field), become increasingly separated from one another as the well width is increased. This effect leads to a decrease in the absorption (Eq. (8.63b)).

In the large well limit, the subbands in a given band approach one another until they can no longer be differentiated from one another, and the quantum confined Stark effect tends towards the Franz–Keldysh effect studied in Complement 7.A. One might ask then which of the two mechanisms may be expected to lead to superior modulator performance. The advantage gained by quantum wells is that a high level of absorption is guaranteed by the absorbing state as both the electrons and holes are localized within the same well. Alternatively, the Franz–Keldysh effect leads to relatively weak absorption in the 'light blocking' state. This is because exponential tailing of the bandgap always remains quite small (see Fig. 7.A.3). Another advantage afforded by the quantum confined Stark effect is that the well thickness can be optimized (at the time of growth) for the specific wavelength of light we wish to modulate.

An additional phenomenon contributed by the confined Stark effect is that the quantum well behaves as a photoconductor under the influence of the electric field. The electrons and holes photogenerated during absorption can escape the well either by thermionic emission above the barriers or by tunnelling through the potential barriers (see Fig. 8.C.3*a*). The electric field then sweeps the carriers out of the structure creating a photocurrent which is to a first approximation proportional to the product of the absorption by the light flux. The combined use of the quantum confined Stark effect and photoconductivity effect leads to a bistable device referred to by the acronym 'SEED', i.e. *Self-Electro-optic-Effect Device*.

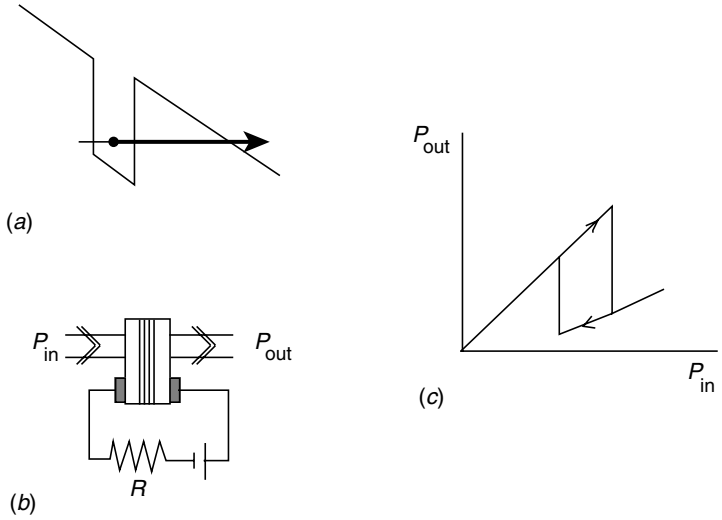


Fig. 8.C.3. The operation of a SEED device. (*a*) Under the effect of an electric field, carriers are released from the quantum well by (for example) tunnelling. (*b*) The structure containing the quantum wells is polarized by a circuit possessing a series resistance *R*. (*c*) In response to the electric circuit, the light output intensity exhibits an abrupt decrease when the input power rises above a certain threshold value.

The operation of a SEED is illustrated in Fig. 8.C.3. We introduce the electromodulator into a biasing circuit possessing a series resistance R. The modulator is then polarized by the voltage so that the photon energy exceeds the excitonic peak associated with the fundamental transition leading to strong absorption. The SEED is therefore in a *strong absorption state* (see Fig. 8.C.1). For a low incident power level P_{in}, the photocurrent I remains weak as well. As the incident optical power is increased, however, so does the photocurrent, leading to a potential drop across the resistance R. As a result, the voltage on the modulator V_{SEED} decreases, leading to a reduction in the field strength F_{SEED} across the well, and a decrease in the absorption (as the reduced field leads to a blueshift of the absorption edge). Once the voltage across the SEED drops sufficiently, the device switches to a *weak absorption state*. This, however, leads to a decrease in the photocurrent and tends to re-establish the strong absorption state in the SEED. This cycle is summarized below.

$$\alpha_{high} \Rightarrow I\uparrow \Rightarrow V_{SEED}\downarrow \Rightarrow F_{SEED}\downarrow \Rightarrow \text{blueshift} \Rightarrow \alpha_{low} \Rightarrow I\downarrow \Rightarrow V_{SEED}\uparrow \Rightarrow F_{SEED}\uparrow \Rightarrow \text{redshift}$$

Therefore, the SEED possesses two accessible stable states for a given incident power level whose occupation depends upon the history of previous incident power levels. The bistability regime over which the SEED manifests this behaviour can be used, for example, as an optical memory element.

Additionally, by replacing the external resistance by a second SEED, it is

possible to assemble ensembles of such elements into circuits capable of executing logical operations on combined light beams.

FURTHER READING

S. Schmitt-Rink, D. S. Chemla, and D. A. B. Miller, *Adv. Phys.* **38**, 89 (1989).

8.D Valence subbands

We saw in Section 8.3 how electron confinement by a potential well creates a shift in the conduction band minima and leads to a quantum confinement energy. In the valence band, we cannot use the same type of approximation. The levels at the top of the valence band are degenerate with respect to the heavy and light hole bands at the Brillouin zone centre. We saw in Complement 1.B that a perturbation can lift degenerate levels. In that case, we were able to solve the problem by establishing the vector subspace spanned by the degenerate states, and by studying the effect of the perturbation on this subspace. We will now show how to extend this approach to include the envelope function approximation for several bands.

The fundamental idea behind this approximation is to suppose that the hole wavefunction can be expanded as:

$$\psi(\mathbf{r}) = \sum \zeta_n(\mathbf{r}) u_{n\mathbf{0}}(\mathbf{r}) \tag{8.D.1}$$

where $u_{n\mathbf{0}}$ are the wavefunctions at the Brillouin zone centre for the different hole bands (e.g. $n = 1$ for heavy holes, $n = 2$ for light holes) and where ζ_n is an envelope function which varies over a significantly greater length scale than the lattice spacing. Over many primitive cells, Eq. (8.D.1) specifies that we find ourselves in a vector subspace generated by the degenerate hole bands. Schrödinger's equation for ψ, which can always be expressed in the form:

$$\left[\frac{p^2}{2m_0} + V_c(\mathbf{r}) + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{8.D.2}$$

where V_c is the crystal potential and $V(\mathbf{r})$ is the external potential, then becomes:

$$\sum_n u_{n\mathbf{0}}(\mathbf{r})\left[\frac{p^2}{2m_0} + \varepsilon_{n0} + V(\mathbf{r})\right]\zeta_n(\mathbf{r}) + \frac{(\mathbf{p}\zeta_n)\cdot[\mathbf{p}u_{n\mathbf{0}}(\mathbf{r})]}{m_0} = E\sum_n u_{n\mathbf{0}}(\mathbf{r})\zeta_n(\mathbf{r}) \tag{8.D.3}$$

where we have used the fact that $u_{n,\mathbf{0}}$ are stationary states of the crystal for $\mathbf{k} = 0$ with energy $\varepsilon_{n\mathbf{0}}$. As was common practice over the course of this chapter, we project this equation onto a known basis. To do so, we multiply this equation by $u^*_{N\mathbf{0}}$ and

integrate over all space. The variation of ζ_n over the distance of a lattice constant being negligible, we may reduce the integrals following the method employed in (8.55). The orthogonality of the Bloch functions over each cell leads to the following equation for ζ_n:

$$\left[\frac{p^2}{2m_0} + V(\mathbf{r})\right]\zeta_N(\mathbf{r}) + \sum_n \frac{\mathbf{P}_{Nn}}{m_0}\cdot[\mathbf{p}\zeta_n(\mathbf{r})] = (E - \varepsilon_{N0})\zeta_N(\mathbf{r}) \tag{8.D.4}$$

where $\mathbf{P}_{Nn} = \langle u_{N0}|\mathbf{p}|u_{n0}\rangle$ is the matrix element which couples the Nth and nth bands in $\mathbf{k}\cdot\mathbf{p}$ theory (see Complement 5.C).

We see that (8.D.4) is the same as the matrix equation obtained over the course of $\mathbf{k}\cdot\mathbf{p}$ theory (see (5.C.10)) if we replace $\mathbf{k}$ by the $\mathbf{p}$ operator and add the external potential to the matrix diagonal. For a potential $V(z)$ which varies in only one direction, $\mathbf{K}_{\|}$ will remain a good quantum number, and the envelope function may be put into the form $\zeta_n(\mathbf{r}) = (1/\sqrt{A})\zeta_{n\mathbf{K}}(z)\exp(\mathrm{i}\mathbf{K}_{\|}\cdot\mathbf{R})$ which reduces the equation to a series of coupled differential equations for $\zeta_{n\mathbf{K}}(z)$:

$$\left[-\frac{\hbar^2}{2m_0}\frac{\mathrm{d}^2}{\mathrm{d}z^2} + V(z)\right]\zeta_N(z) + \sum_n \frac{P^z_{Nn}}{m_0}\left[\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}z}\zeta_n(z)\right] = [E - \varepsilon_N(\mathbf{K}_{\|}, 0)]\zeta_N(z) \tag{8.D.5}$$

where we have used the fact that the plane wave in the direction perpendicular to z diagonalizes the $\mathbf{k}\cdot\mathbf{p}$ matrix with eigenenergies $\varepsilon_N(\mathbf{K}_{\|}\,0)$.

In a square quantum well (see Fig. 8.D.1), P_{Nn} and $\varepsilon_N(\mathbf{K}_{\|}, 0)$ are constant within each material. The effect of the potential $V(z)$ is to introduce band offsets which we take into account by imposing the continuity conditions on the wavefunction and on the probability flux at the interfaces (see Section 8.2). In each material, we see that the solution to (8.D.5) corresponding to an energy E is of the form:

$$\zeta_n = \alpha_n \mathrm{e}^{\mathrm{i}k_n z} + \beta_n \mathrm{e}^{-\mathrm{i}k_n z} \tag{8.D.6}$$

where k_n is a complex number and is one of the two solutions to the equation:

$$\varepsilon_n(\mathbf{K}_{\|}, k_n) = E \tag{8.D.7}$$

i.e. k_n is the transverse wavevector at which band n has energy E.

It is worth noting that for an energy within the bandgap and above the top of the valence bands, Eq. (8.D.7) has a purely imaginary solution $k_n = \mathrm{i}\kappa_n$ (see, for example, Eq. (5.C.25)). In bulk, such a solution is not admissible as it diverges towards infinity. In the case of the tunnel effect, however, such a solution can emerge as an evanescent wave over a finite, or semi-infinite region of space. Figure 8.D.2 shows the principle for the two-hole bands in two materials for an energy close to the valence band.

In the barrier material, $E > \varepsilon_N^{(b)}(\mathbf{K}_{\|}, 0)$ for the two valence bands and only an evanescent solution is possible. If the small bandgap material is situated between $-L/2 < z < L/2$, the solution for $z < -L/2$ must have $\alpha_n^{(1)} = 0$ in the expansion of

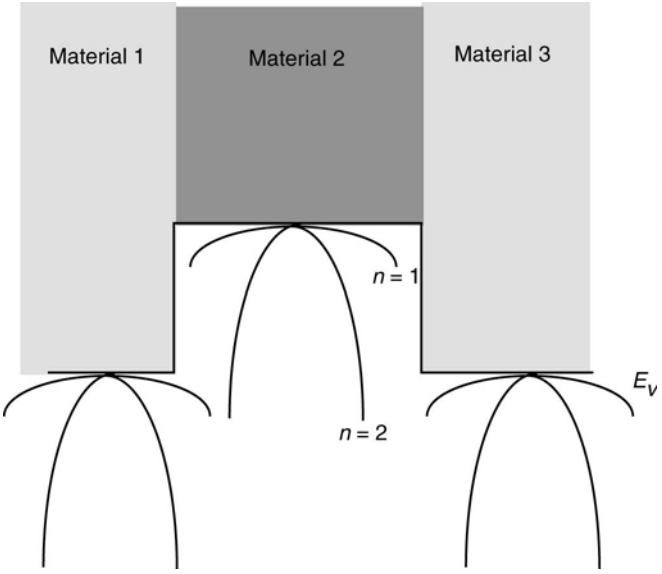
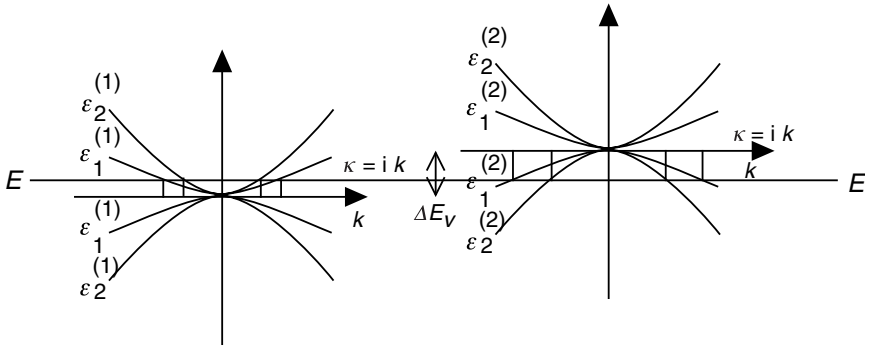


Fig. 8.D.1. Valence bands in a quantum well.



(*a*) Barrier material (*b*) Well material

Fig. 8.D.2. In the vicinity of the valence band, the k values for which $\varepsilon(k) = E$ are real for E lying below the valence band maximum, and imaginary for E lying above. This figure shows how the wavevectors which enter into Eqs. (8.D.6)–(8.D.9) are determined for a given energy E. $\mathbf{K}_{\|} = \mathbf{0}$.

(8.D.6). For $z > L/2$, only $\beta_n^{(3)} = 0$ yields a possible solution. Alternatively, in the well material, $E < \varepsilon_N^{(w)}(\mathbf{K}_{\|}, 0)$ without any restrictions on the coefficients. At the interfaces, the continuity conditions from (8.D.5) are, for example, at $z = z_- = -L/2$:

$$\alpha_N^{(1)} e^{\kappa_N z_-} = \alpha_N^{(2)} e^{ik_N z_-} + \beta_N^{(2)} e^{ik_N z_-} \tag{8.D.8}$$

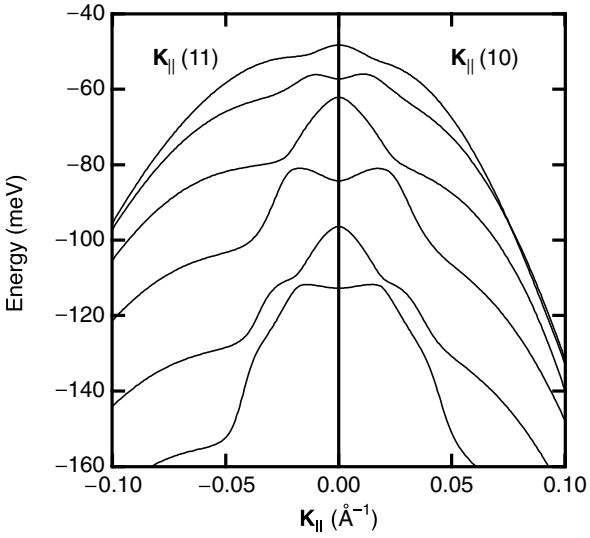


Fig. 8.D.3. Valence subbands for a quantum well corresponding to motion parallel to the interfaces. The dispersion curves are seen to deviate significantly from the simple form of a parabola. In this calculation, the quantum well width was taken to be 12 nm, the bottom of the well is at −44 meV and the height of the barrier is 70 meV. (Courtesy of A. Fily @ THALES/LCR.)

for the continuity of the wavefunction and:

$$-\frac{\hbar^2}{2m_0}\left(\frac{\mathrm{d}}{\mathrm{d}z}\alpha_N^{(1)}\mathrm{e}^{\kappa_N z}\right)_{z_-} + \sum_n \frac{P_{Nn}^{(1)}}{m_0}\left(\frac{\hbar}{i}\alpha_n^{(1)}\mathrm{e}^{\kappa_n z_-}\right)$$
$$= -\frac{\hbar^2}{2m_0}\left[\frac{\mathrm{d}}{\mathrm{d}z}(\alpha_N^{(2)}\mathrm{e}^{\mathrm{i}k_N z} + \beta_N^{(2)}\mathrm{e}^{\mathrm{i}k_N z})\right]_{z_-} + \sum_n \frac{P_{Nn}^{(2)}}{m_0}\left[\frac{\hbar}{i}(\alpha_n^{(2)}\mathrm{e}^{\mathrm{i}k_n z_-} + \beta_n^{(2)}\mathrm{e}^{-\mathrm{i}k_n z_-})\right] \tag{8.D.9}$$

for the continuity condition on the wavefunction's derivative. We see in (8.D.9) that the effect of *the interface is to mix the heavy hole and light hole components*, even though the components are independent in each material. The four relationships (8.D.8)–(8.D.9) for $N = 1$ (heavy hole) and $N = 2$ (light hole) determine the four coefficients in the quantum well material as a function of the two coefficients in the barrier material for $z > L/2$. The limiting conditions at the interface at $z_+ = L/2$ determine the four coefficients for the evanescent waves in the barrier material. We must then seek the energies and the coefficients $\alpha_n^{(1)}$ which cancel out the coefficients $\alpha_n^{(3)}$. After this, we need only determine a global factor by normalizing the wavefunction.

While the methodology for the calculation is clear, it must be performed numerically. We show a few illustrative results in Fig. 8.D.3. It is evident (but not surprising) that the subband dispersion curves deviate significantly from simple parabolas.

9 Waveguides

9.1 Introduction

One of the main goals of optoelectronics is to use photons as elementary bits for information transport and processing. Before electrons can be usurped from their privileged role in integrated electronic circuits, however, a means must be developed which will allow photons to be routed to specific locations where they can participate in logical or analogue operations. Waveguides provide a method of trapping light in conduits which can act as optical interconnections. A second advantage of waveguides is that they provide a means of intensifying light–matter interactions by concentrating the light energy density into narrow channels. This is of primary concern in the operation of devices based on non-linear optical interactions with matter (e.g. electro-optic modulators, frequency conversion optics, . . .) and in maximizing laser diode efficiencies.

Example
Ten milliwatts of continuous-wave light can easily be guided in a 3 μm wide by 0.2 μm thick waveguide. This results in a power per unit surface of:

$$P = 10^{-2}\,\mathrm{W}/(3 \times 10^{-4} \times 2 \times 10^{-5}\,\mathrm{cm}^2) = 1.6\,\mathrm{MW\,cm}^{-2}$$

which is a significant power density. We can compare this value with the maximum electrical power levels which can be transported in electron-based circuits. This typically corresponds to a current density of $1\,\mathrm{MA\,cm}^{-2}$ (this limit arises due to electromigration of atoms in the metallic interconnects) under a voltage of 1 V, leading to a power density of $1\,\mathrm{MW\,cm}^{-2}$. It is also worth noting that the peak optical density carried by waveguides can increase up to $1\,\mathrm{GW\,cm}^{-2}$ under pulsed conditions.

9.2 A geometrical approach to waveguides

We will begin by interpreting light waveguiding in terms of propagating light rays. This approach will allow us to develop an intuitive grasp of the concepts involved in this phenomenon quickly. We consider a *core layer* formed by a dielectric material of thickness d, with an index of refraction n_1, sandwiched between two very thick (and ideally infinite) *confining layers* with an index of refraction n_2. A

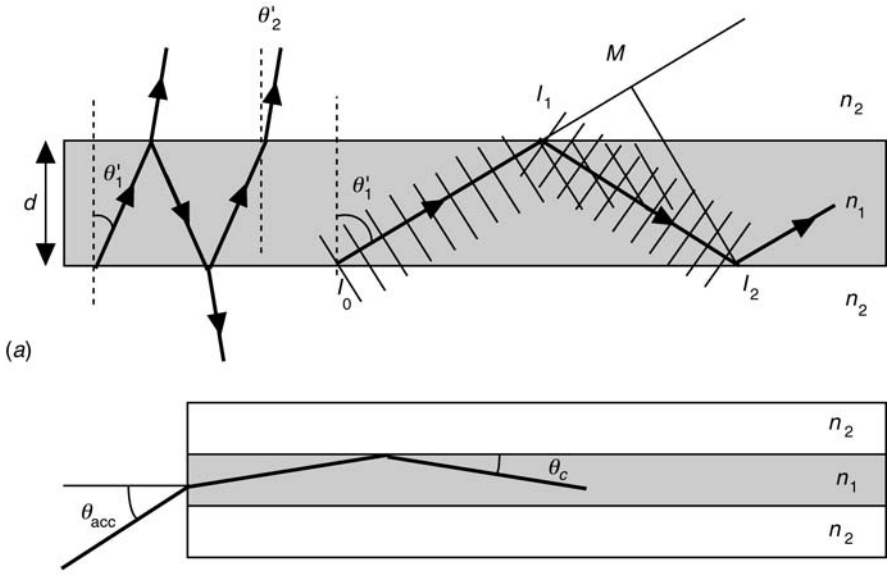


Fig. 9.1. (*a*) Light ray representation of conditions leading to the optical confinement of light by a waveguide. (*b*) The numerical aperture of a waveguide is given by the sine of the maximum acceptance angle for light incident from air.

light ray inside medium 1 makes an angle θ'_1 to the normal of the plane which separates the layers. This ray is then transmitted by refraction into medium 2 making an angle θ'_2 with the normal (see the left-hand side of Fig. 9.1*a*). Recalling the principal results from the theories of Descartes–Snell and Fresnel:

- the angles θ'_1 and θ'_2 are related by:

$$n_1 \sin\theta'_1 = n_2 \sin\theta'_2 \tag{9.1}$$

- if $n_1 > n_2$, the lightwave in medium 1 is totally reflected if it is incident at a critical angle θ'_c (*total internal reflection*) given by:

$$\theta'_c = \arcsin\left(\frac{n_2}{n_1}\right) \tag{9.2}$$

- if the polarization of the wave is transverse electric (TE), i.e. if the polarization of the electric field is parallel to the layer planes, the wave is reflected and its phase shifts relative to the incident wave by an angle ϕ_r lying between 0 and π:

$$\tan\frac{\phi_r}{2} = \frac{(\sin^2\theta'_1 - \sin^2\theta'_c)^{1/2}}{\cos\theta'_1} = \left(\frac{\sin^2\theta_c}{\sin^2\theta_1} - 1\right)^{1/2} \tag{9.3}$$

where $\theta_1 = \pi/2 - \theta'_1$. We note that as a general rule, we reserve the prime symbol $'$ to label the complement of an angle θ, i.e. $\theta' = \pi/2 - \theta$.

The right-hand side of Fig. 9.1*a* depicts the situation corresponding to total internal reflection, where the wave becomes guided once $\theta'_1 > \theta'_c$. We must not believe, however, that any electromagnetic wave satisfying the latter condition will be guided. In fact, each reflection at the interface between the two different media will dephase the wave causing it to interfere with itself destructively *unless the reflected wave remains in phase with waves produced by preceding reflections.* This last condition is apparent in Fig. 9.1*a*. We follow two successive reflections over two interfaces and three impact points I_0, I_1, and I_2. The wave at I_2 clearly results from at least two other waves: the wave which has not been reflected and which corresponds to the displacement I_0M, and the other which has experienced two reflections and has travelled $I_0I_1 + I_1I_2$ in addition to receiving a phase shift of $(2\phi_r)$ from the two reflections. These two contributions have a path difference of $I_1I_2 - I_1M = I_1I_2 - I_1I_2\cos 2\theta_1 = 2\sin^2\theta_1 I_1I_2 = 2\sin^2\theta_1 \times d/\cos\theta'_1 = 2d\sin\theta_1$ and a phase shift of $2\phi_r$ leading to a total dephasing of $2kd\sin\theta_1 - 2\phi_r$, where k is the wavevector $2\pi n_1/\lambda_0$ and λ_0 is the vacuum wavelength of the electromagnetic wave. The phase shifts resulting from multiple reflections will tend to interfere destructively unless they are multiples of 2π, whereby we obtain the *self-consistent condition* for a confined wave:

$$\frac{2\pi n_1}{\lambda_0} 2d\sin\theta_1 - 2\phi_r = 2\pi m \tag{9.4}$$

We need now only substitute this last condition into Fresnel's relation, (9.3), to obtain:

$$\tan\left(\frac{\pi n_1 d}{\lambda_0}\sin\theta_1 - \frac{m\pi}{2}\right) = \left(\frac{\sin^2\theta_c}{\sin^2\theta_1} - 1\right)^{1/2} \tag{9.5}$$

Equation (9.5) with the unknown $\sin\theta_1$ allows one to calculate the different angles θ_m that lead to guided waves in the waveguide, without forgetting that the phase shift ϕ_r has a value between 0 and π. Figure 9.2 shows a graphical method of obtaining the allowed solutions to this equation. When $\sin\theta_1$ sweeps the interval $0 \to \sin\theta_c$, the right-hand term periodically sweeps through $+\infty \to 0$, whereas the left-hand term periodically sweeps through $0 \to +\infty$ every $\lambda_0/2n_1d$. Figure 9.2 also allows one to calculate the number N_m of guided modes allowed in the guide. There will be as many as the number of periods in the interval $0 \to \sin\theta_c$ or $\sin\theta_c/(\lambda_0/2n_1d)$. We find therefore:

$$N_m = \mathrm{Int}\left[2\frac{d}{\lambda_0}\mathrm{NA}\right] + 1 \tag{9.6}$$

Number of allowed TE modes in a symmetric waveguide

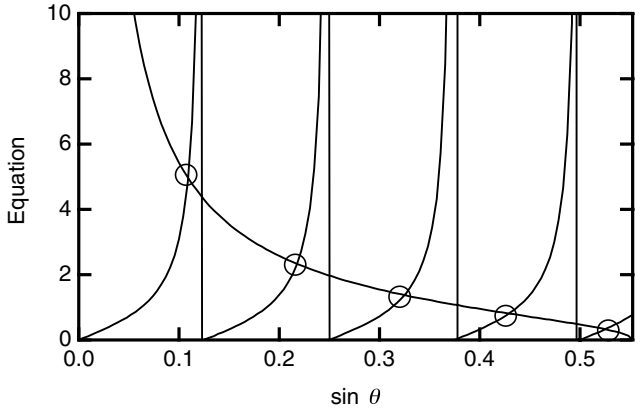


Fig. 9.2. Graphical solution to Eq. (9.5) for conditions given in the example below.

where Int is the integer function and NA is the *numerical aperture* of the guide given using (9.2), by:

$$\mathrm{NA} = (n_1^2 - n_2^2)^{1/2} \tag{9.7}$$

Numerical aperture for a waveguide

Figure 9.1*b* shows that the numerical aperture is the sine of the acceptance angle of the guide for incident rays originating in air. If the numerical aperture is greater than 1, the acceptance angle is $\pi/2$. Equations (9.6) and (9.7) show that, in analogy to the one-dimensional well which always possesses a bound state, a symmetric waveguide always has at least one allowed TE mode. Expression (9.6) also allows one to obtain a simple requirement for a *single mode waveguide*:

$$\frac{d}{\lambda_0} < \frac{1}{2\mathrm{NA}} \tag{9.8}$$

An identical line of reasoning can be followed for the TM modes (along with the appropriate set of Fresnel equations) and can be generalized to asymmetric waveguides. We prefer, however, to present the more powerful approach founded on Maxwell's equations.

Example

A waveguide has been fabricated out of a layer of InGaAs ($n_1 = 3.9$) sandwiched between two layers of AlGaAs ($n_2 = 3.0$). The numerical aperture of the guide $\mathrm{NA} = (3.9^2 - 3.0^2)^{1/2} = 2.49$. If the InGaAs layer thickness is 1 μm, there would be five allowed modes at 0.9 μm.

9.3 An oscillatory approach to waveguides

We recall that the electric field of an electromagnetic wave in a neutral medium ($\mathbf{j} = 0$ and $\rho = 0$) possessing an index of refraction $n(\mathbf{r})$ is a solution to the wave equation obtained from Maxwell's equations (see Chapter 2):

$$\nabla^2\mathbf{E}(\mathbf{r},t) - \mu_0\varepsilon_0 n^2(\mathbf{r})\frac{\partial^2}{\partial t^2}\mathbf{E}(\mathbf{r},t) = 0 \tag{9.9}$$

We suppose that the wave possesses a radial frequency ω, i.e. that the field $\mathbf{E}$ is given by:

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}[\mathbf{E}(\mathbf{r})\mathrm{e}^{\mathrm{i}\omega t}] \tag{9.10}$$

which leads to the Helmholtz equation for the amplitude $\mathbf{E}(\mathbf{r})$:

$$\nabla^2\mathbf{E}(\mathbf{r}) + k^2 n^2(\mathbf{r})\mathbf{E}(\mathbf{r}) = 0 \tag{9.11}$$

where k is the norm of the wavevector $\mathbf{k}$ and relates to ω through the vacuum *dispersion relation* $k = \omega/c$. We consider the particular geometry (see Fig. 9.3) where the wave travels along the Oz axis. We may then decompose the field as $\mathbf{E}(\mathbf{r},t) = \mathbf{E}(x,y)\mathrm{e}^{-\mathrm{i}\beta z}$; and Eq. (9.11) can then be written:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\mathbf{E}(x,y) + [k^2 n^2(\mathbf{r}) - \beta^2]\mathbf{E}(x,y) = 0 \tag{9.12}$$

We note that taking the *propagation constant* β to be the same in media 1 and 2 corresponds to the *self-consistency condition* set forth in Section 9.2. Finally, we will suppose that the wave does not vary along the y direction, i.e. $\partial/\partial y = 0$. Equation (9.12) can then be separated into three:

$$\begin{aligned}
&\frac{\partial^2}{\partial x^2}E(x) + (k^2 n_2^2 - \beta^2)E(x) = 0, \text{for } x > 0 \\
&\frac{\partial^2}{\partial x^2}E(x) + (k^2 n_1^2 - \beta^2)E(x) = 0, \text{for } -d < x < 0 \\
&\frac{\partial^2}{\partial x^2}E(x) + (k^2 n_2^2 - \beta^2)E(x) = 0, \text{for } x < -d
\end{aligned} \tag{9.13}$$

where we recall that $n_1 > n_2$. We may at this point analyse the different types of solutions to (9.13) by examining the effect of different values for kn_i and β (see Fig. 9.4).

- If $\beta > kn_1, k^2 n_i^2 - \beta^2$ is negative regardless of the value of x, and the solutions to (9.13) diverge exponentially (clearly an unphysical solution).

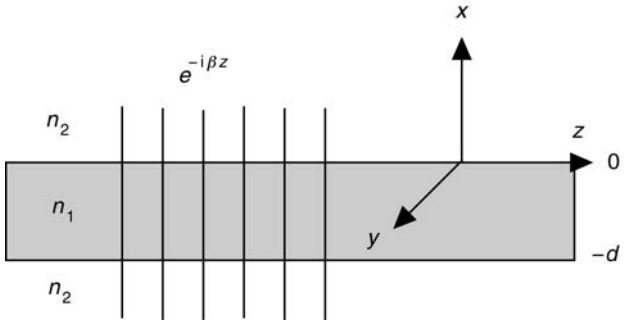


Fig. 9.3. Geometry for a planar waveguide.

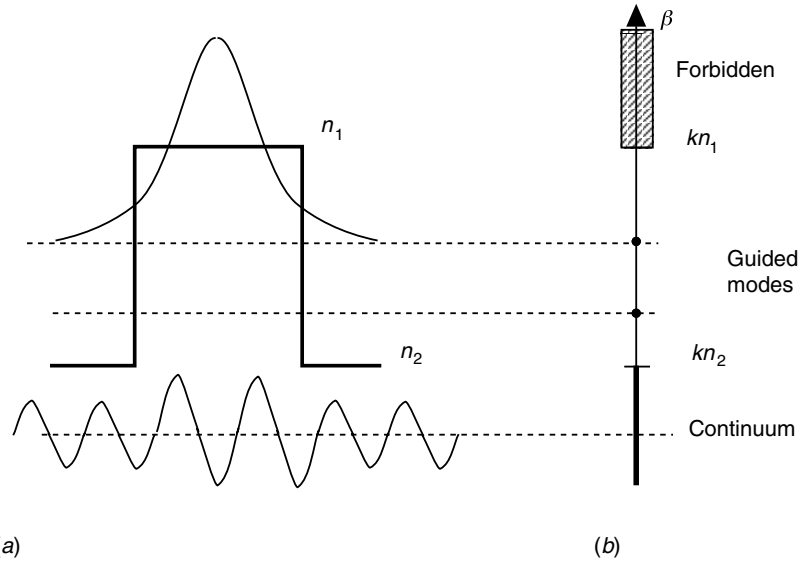


Fig. 9.4. The various electromagnetic regimes in a wave guide are analogous to the different electronic states in a quantum well: (*a*) shows the spatial variation of the index of refraction $n(\mathbf{r})$ as well as the variations in the field amplitude; (*b*) shows the allowed values for the propagation constant β.

- If $kn_1 > \beta > kn_2$, then the amplitude $E(x)$ is sinusoidal within the guide and decreases exponentially outside. These are the guided modes, which we will determine later on by introducing the relevant boundary conditions.
- Lastly, if $\beta < kn_2$, the solutions are sinusoidal everywhere and the waves escape from the guide. These are referred to as *leaky modes* and correspond to the situation where the incident light beam enters the waveguide in such a manner that the effective angle within the waveguide exceeds the critical angle for total internal reflection. We will ignore for the time being these leaky modes, which in fact allow a certain number of interesting applications. Rather, we will examine in more detail the solutions to (9.12) for the case of TE and TM waves.

Transverse electric (TE) waves

In this case, the electric field is oriented along the Oy axis. We denote its amplitude as $E_y(x)$, i.e. $E(x,z) = \mathrm{Re}[E_y(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta z)}]$. The magnetic field B is then determined by the Lenz–Maxwell equation ($\nabla \times \mathbf{E} = -\partial\mathbf{B}/\partial t$), which gives:

$$B_x(x) = -\frac{\beta}{\omega}E_y(x)$$
$$B_z(x) = \frac{\mathrm{i}}{\omega}\frac{\partial}{\partial x}E_y(x) \tag{9.14}$$

From Maxwell's equations we know that the electric field $E_y(x)$ and the magnetic field $B_z(x)$ (i.e. the first derivative of E_y) are continuous at both interfaces $x = 0$ and $x = -d$. We note that these propagation equations, derived from (9.12) together with these boundary conditions, are formally identical to Schrödinger's equations which lead to the quantization of electronic energy states in quantum wells described by a potential $V(x)$ (Section 1.4). The correspondences to be drawn are then:

$$\omega^2 c^2 n(\mathbf{r})^2 \leftrightarrow -\frac{2m_0}{\hbar^2}V(\mathbf{r})$$

$$\beta^2 \leftrightarrow -\frac{2m_0}{\hbar^2}E$$

$$E(\mathbf{r}) \leftrightarrow \psi(\mathbf{r})$$

Figure 9.4 illustrates this parallel between a quantum well and a waveguide. The leaky modes correspond to the continuum of unbound states above a quantum well, whereas the guided modes are conceptually equivalent to the bound states in a quantum well. We note, however, that the shape (depth) of the well depends on the frequency ω as shown in the correspondence relations above.

The solution to (9.13) given the boundary conditions:

$$E_y(x), \text{ continuous at } x = 0 \text{ and } x = -d$$
$$\frac{\partial}{\partial x}E_y(x), \text{ continuous at } x = 0 \text{ and } x = -d \tag{9.15}$$

follows in a similar fashion to the quantum well case (*Section* 1.4). *Anticipating the spatial dependence of the wave amplitude, we pose*:

$$E_y = \begin{cases} A\exp(-\kappa x), & \text{for } x > 0 \\ B\cos(\alpha x) + C\sin(\alpha x), & \text{for } -d < x < 0 \\ D\exp[\kappa(x+d)], & \text{for } x < -d \end{cases} \tag{9.16}$$

where κ is the *optical attenuation constant* outside of the guide. This is equivalent to the *tunnelling attenuation length* in quantum mechanics. We will begin by introducing the self-consistency condition represented in (9.13), which gives:

$$\begin{aligned}\beta^2 + \alpha^2 &= n_1^2 k^2 \\ \beta^2 - \kappa^2 &= n_2^2 k^2 \\ k &= \omega/c\end{aligned} \qquad (9.17a–c)$$

We still need to write the boundary conditions (9.15):
At $x = 0$:

$$A = B$$

$$C = -\frac{\kappa}{\alpha}$$

At $x = -d$:

$$D = A\left[\cos(\alpha d) + \frac{\kappa}{\alpha}\sin(\alpha d)\right]$$

$$\alpha \sin(\alpha d) - \kappa \cos(\alpha d) = \kappa\left[\cos(\alpha d) + \frac{\kappa}{\alpha}\sin(\alpha d)\right]$$

The last condition furnishes the parameters for the guided wave:

$$\begin{aligned}\tan(\alpha d) &= \frac{2\kappa\alpha}{\alpha^2 - \kappa^2} \\ \beta^2 + \alpha^2 &= n_1^2 k^2 \\ \beta^2 - \kappa^2 &= n_2^2 k^2 \\ k &= \omega/c\end{aligned} \qquad (9.18)$$

System of equations governing the TE modes in a symmetric waveguide

The problem is therefore to find, for a given radial frequency ω and waveguide thickness d, the propagation constant β which simultaneously satisfies all conditions (9.18) and leads to an implicit equation in β completely equivalent to (9.5). This approach can be generalized to any shape of waveguide. We list below, a MATHEMATICA program that can be used to obtain the solutions for an asymmetric waveguide.

Example

The generalization of (9.18) for the case involving an asymmetric waveguide immediately leads to:

$$\tan(\alpha d) = \frac{\kappa_2 + \kappa_3}{\alpha[1 - (\kappa_2\kappa_3/\alpha^2)]}$$

$$\beta^2 + \alpha^2 = n_1^2 k^2$$

$$\beta^2 - \kappa_2^2 = n_2^2 k^2$$

$$\beta^2 - \kappa_3^2 = n_3^2 k^2$$

using the usual notation. So we consider a waveguide consisting of a GaAs core layer 0.3 μm thick ($n_1 = 3.3$) sandwiched between a confining AlAs layer ($n_2 = 2.7$) on one side and air ($n_3 = 1$) on the other. We will neglect the intrinsic dispersion in these materials.

```
Lambda=0.9;t=.3;Bet=.;
n3=1;n1=3.3;n2=2.7;
k=2*N[Pi]/Lambda;
kt=k*t;
Bett=Bet*t;
ht= Sqrt[(n1^2*kt^2-Bett^2)];
qt= Sqrt[(-n3^2*kt^2+Bett^2)];
pt= Sqrt[(-n2^2*kt^2+Bett^2)];
f[Bet__]:=Tan[ht]-(qt+pt)/(ht-pt*qt/ht);
Plot[f[Bet],{Bet,10,50}];
sol=FindRoot[f[Bet]==0,{Bet,21}];
neff=Bet*Lambda/(2*N[Pi])/.sol
```

from which we obtain a propagation constant of 21.3 μm^{-1} and an effective index of refraction $n_{eff} = 3.1267$.

It is interesting in the particular case of the symmetric guide to relate the geometrical approach in Section 9.1 to the oscillatory approach used in this section. To do so, it is sufficient to note that the propagation vector in the waveguide is $(0, \pm n_1 k \cos\theta_m, n_1 k \sin\theta_m)$. Then, taking (9.18) into account, we have:

$$\beta_m = n_1 k \cos\theta_m$$

$$\alpha_m = n_1 k \sin\theta_m \tag{9.19}$$

$$\kappa_m = n_1 k(\cos^2\theta_m - \cos^2\theta_c)^{1/2}$$

Equivalence between the geometric and oscillatory approaches

which is nothing else than (9.5). Equation (9.19) is well suited to graphical interpretation, as shown in Fig. 9.5. We see that the 0th order (minimum θ_m) leads to a maximum attenuation coefficient κ_m, i.e. the most heavily confined mode.

We now need to normalize the amplitude of the electromagnetic wave in the guide, i.e. to find a value for the constant A in (9.16). There exists a certain degree of arbitrariness in this determination as the lateral dimension Oy has not been specified. The custom is to choose A in such a fashion that the normalized field E_y corresponds to a power of 1 W over 1 m of guide length (along Oy). The normalization condition may then be written:

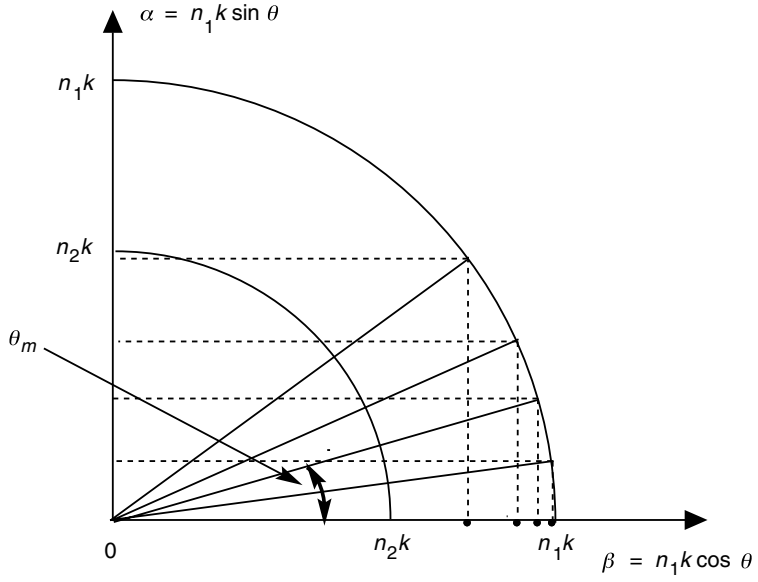


Fig. 9.5. Graphical representation of conditions leading to waveguiding.

$$S = \frac{1}{\mu_0}\overline{\int_{-\infty}^{+\infty} \mathbf{E} \times \mathbf{B} \mathrm{d}x} = -\frac{1}{2\mu_0}\int_{-\infty}^{+\infty} E_y B_x \mathrm{d}x$$
$$= \frac{\beta_m}{2\omega\mu_0}\int_{-\infty}^{+\infty} [E_y^m(x)]^2 \mathrm{d}x = p_0 \tag{9.20}$$

where in this last equation, we utilized relation (9.14) (the horizontal bar indicates that the quantity is time averaged; p_0 is the *unit normalized power* corresponding to 1 W m^{-1} and m is the TE mode index in the guide). We recommend the use of p_0 to keep the equations homogeneous, which is of prime importance when one deals with non-linear optics, for instance (see Complement 9.C)! Using expression (9.16) for the field and the continuity conditions, Eq. (9.20) leads to:

$$A_m = 2\alpha_m \left[\frac{\omega\mu_0}{|\beta_m|\left(d + \frac{2}{\kappa_m}\right)(\alpha_m^2 + \kappa_m^2)}\right]^{1/2} \sqrt{p_0} \tag{9.21}$$

As a concluding remark, since the functions $E_y^m(x)$ are solutions to linear differential equations (i.e. eigenvectors of the differential operator (9.12)), they are orthog-

onal to each other. More precisely, taking into account the normalization condition (9.20), the functions $E_y^m(x)$ satisfy:

$$\int_{-\infty}^{+\infty} E_y^m(x)E_y^n(x)\mathrm{d}x = \frac{2\omega\mu_0 p_0}{\beta_m}\delta_{m,n} \tag{9.22}$$

Therefore, if the guide is *infinitely confining* (i.e. if $n_1/n_2 \to \infty$), the modes may be written:

$$E_y^m(x) = 2\sqrt{\frac{\omega\mu_0}{d\beta_m}p_0}\sin\frac{m\pi x}{d} \tag{9.23a}$$

*m*th mode in an infinitely confining waveguide

Let us note for this case some very useful approximations for the coefficients α and κ ($m = 1$):

$$\begin{aligned}\alpha &\approx \frac{\pi}{d}\\ \kappa &\approx (n_1^2 - n_2^2)^{1/2}\frac{2\pi}{\lambda_0}\end{aligned} \tag{9.23b}$$

Transverse magnetic (TM) waves

This time, the components of the electromagnetic field are in part

$$B_y(x,z,t) = \mathrm{Re}[B_y(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta z)}] \tag{9.24}$$

for the transverse magnetic field; and the electric field $\mathbf{E}$ given by $\nabla \times \mathbf{B} = 1/(n_i c)^2 \partial\mathbf{E}/\partial t$, is:

$$\begin{aligned}E_x(x,z,t) &= \frac{\beta c^2}{n_i^2\omega}B_y(x,z,t)\\ E_z(x,z,t) &= -\frac{ic^2}{n_i^2\omega}\frac{\partial}{\partial x}B_y(x,z,t)\end{aligned} \tag{9.25}$$

where n_i is the optical index of medium i. The TM eigenmodes in the guide are obtained by imposing the continuity conditions for the fields B_y and E_z at each of the two interfaces (we leave this as an exercise to the reader). This yields:

$$\tan(\alpha d) = \frac{2\varepsilon_{12}\kappa\alpha}{\alpha^2 - (\varepsilon_{12}\kappa)^2}$$
$$\beta^2 + \alpha^2 = n_1^2 k^2$$
$$\beta^2 - \kappa^2 = n_2^2 k^2 \qquad (9.26)$$
$$k = \omega/c \text{ and } \varepsilon_{12} = \left(\frac{n_1}{n_2}\right)^2$$

System of equations which determine the TM modes in a symmetric waveguide

For each combination of values for the frequency ω and the guide thickness d, we obtain multiple values for the allowed propagation constant β for the guide modes. A commonly used representation for the various modes in a waveguide consists in plotting the dispersion curve $n_{\text{eff}}(\lambda_0)$, where n_{eff} is the *effective optical index* for the guide given by $n_{\text{eff}} = \beta c/\omega$.

Figure 9.6 shows this dispersion curve for a theoretical waveguide with $n_1 = 2$, $n_2 = 1.5$, and $n_3 = 1$ as a function of d/λ_0. As is apparent from the figure, there are several possible solutions, i.e. several modes are possible, and waveguides are therefore, in general, *multimode* in nature. We will see later on that *single mode* guides are preferred in most applications. In this case, single mode waveguiding is guaranteed when condition (9.8) is satisfied, i.e. when the ratio d/λ_0 is less than NA/2. We can also conclude from Fig. 9.6 that a waveguide will naturally induce a dispersion as there is a variation in the effective optical index between the extrema n_1 and n_2. This is referred to as the *modal dispersion* of the waveguide. This dispersion adds to the natural dispersion contributed by the materials comprising the guide. Figure 9.7 shows the result of a complete calculation for the GaAs/AlGaAs system using Afromowitz' dispersion relations given in Complement 7.B. We will see that this dispersion leads to negative effects in non-linear waveguides.

9.4 Optical confinement

It is interesting to estimate the quantity of energy which is effectively guided or trapped within the guide between 0 and $-d$. This is described by a *confinement factor* Γ, which is defined as:

$$\Gamma = \frac{\int_{-d}^{0} |E(x)|^2 \mathrm{d}x}{\int_{-\infty}^{+\infty} |E(x)|^2 \mathrm{d}x} \qquad (9.27)$$

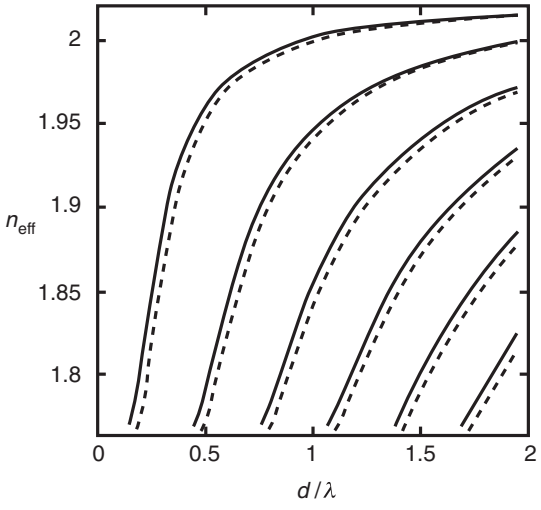


Fig. 9.6. Dispersion curves for a waveguide in a non-dispersive medium with $n_1 = 2$, $n_2 = 1.5$, and $n_3 = 1$. The TE (TM) modes are shown in solid (dashed) lines.

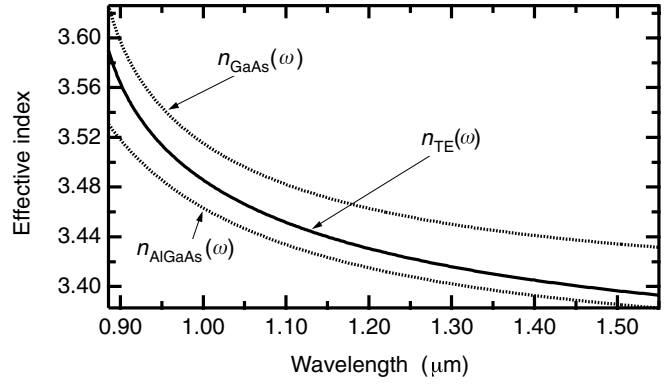


Fig. 9.7. Dispersion in a waveguide consisting of a 0.5 μm thick core layer sandwiched between two 1.0 μm $Al_{0.1}Ga_{0.9}As$ layers. The naturally occurring dispersion in the constituent materials adds to that resulting from the waveguide geometry. (Courtesy of A. Fiore @LCR/THALES.)

So that we can familiarize ourselves with the concept, we will calculate the confinement factor for the mth mode of a symmetric waveguide. By symmetry, (9.27) can be written:

$$\Gamma_m = 1 - \frac{2\int_0^{\infty} |E_m(x)|^2 \mathrm{d}x}{\int_{-\infty}^{+\infty} |E_m(x)|^2 \mathrm{d}x} \tag{9.28}$$

or, given (9.16) and (9.22):

$$\begin{aligned}\Gamma_m &= 1 - \frac{A_m^2 \beta_m}{2\kappa_m \omega \mu_0} \\ &= 1 - \left(\frac{\alpha_m}{k\mathrm{NA}}\right)^2 \frac{1}{(1 + \kappa_m \mathrm{d}/2)}\end{aligned} \tag{9.29}$$

We then substitute into (9.29) the various trigonometric expressions for A_m, β_m, and κ_m obtained in (9.19) and (9.21), to find:

$$\begin{aligned}\Gamma_m &= 1 - \frac{\sin^2 \theta_m}{\sin^2 \theta_c} \frac{1}{1 + \pi \dfrac{d}{\lambda_0/n_1} (\sin^2 \theta_c - \sin^2 \theta_m)^{1/2}} \\ &= 1 - \frac{X^2}{1 + \dfrac{\pi}{2} N_m \sqrt{1 - X^2}}\end{aligned} \tag{9.30a}$$

where X is the ratio $\sin \theta_m / \sin \theta_c$ and we recall that N_m is the number of allowed modes in the guide. Figure 9.8 represents the variation in the confinement factor for a typical waveguide as a function of $\sin \theta_m / \sin \theta_c$. This latter quantity increases as the mode index m increases. We note in this figure that the confinement is at a maximum when the order is at a minimum. *The TE_0 and TM_0 modes of a waveguide therefore experience maximum electromagnetic waveguiding*. This fact is incorporated into the design of many optoelectronic devices (e.g. laser diodes, modulators, . . .).

A waveguide designer will therefore prefer to make use of single mode waveguides. A limited expansion of (9.30a) leads to an extremely useful approximation for TE_0:

$$\Gamma_0 = \frac{D^2}{2 + D^2} \text{ with } D = \frac{2\pi d}{\lambda_0} \sqrt{n_1^2 - n_2^2} \tag{9.30b}$$

A demonstration of this is left as an exercise to the reader. The example below illustrates the use of this formula.

Example

We consider a waveguide consisting of a 0.2 μm thick core layer of GaSb and confining AlGaSb layers. We will neglect the natural dispersion of the materials, and take $n_1 = 3.837$ for the optical index of GaSb and $n_2 = 3.589$ for AlGaSb. The numerical aperture NA of the guide is therefore 1.36. The number of modes at 1.55 μm is given by (9.6) or $N = \mathrm{Int}(2 \times 0.2 \times 1.36/1.55) + 1 = 1$. The guide is therefore single mode. We can therefore apply expression (9.30b) to obtain the

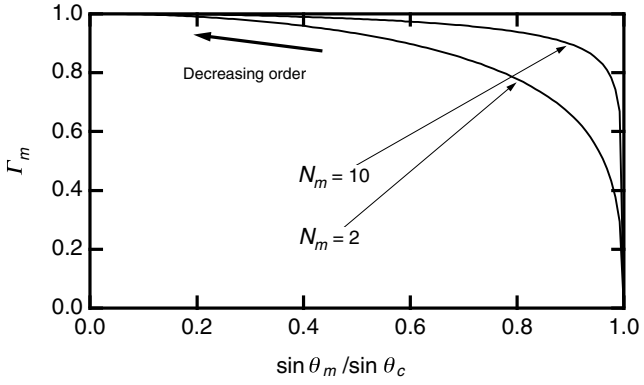


Fig. 9.8. The confinement factor of an optical waveguide as a function of $\sin\theta_m/\sin\theta_c$. Waveguide modes with weak indices ($\sin\theta_m$ small) offer better optical confinement.

confinement factor for the TE_0 mode in the waveguide. We obtain $D = 1.10$ and $\Gamma_0 = 0.38$.

The MATHEMATICA program below compares the values obtained for the confinement factors using the exact method (9.30a) and the approximation (9.30b) for the system described above. The results are shown in Fig. 9.9 and confirm the validity of the approximation (9.30b).

```
Lambda=1.55;t=.
n1=3.837;n2=3.589;ON=Sqrt[n1^2-n2^2];
ninit=(n1+n2)/2;
k= 2*N[Pi]/Lambda;
Bet=2*N[Pi]*neff/Lambda;
alpha= Sqrt[(n1^2*k^2-Bet^2)];
kapa= Sqrt[(-n2^2*k^2+Bet^2)];
fct=Tan[alpha*t]-2*kapa*alpha/(alpha^2-kapa^2);
tab=Table[{t,sol=FindRoot[fct,{neff,ninit}];
gamma= Abs[1-(n1^2-neff^2)/(n1^2-n2^2)/
(1+k*t/2*Sqrt[neff^2-n2^2])]/.sol},{t,.01,.4,.01}];
plot1=ListPlot[tab]
dis= 2*N[Pi]*t*ON/Lambda;
plot2=Plot[dis^2/(2+dis^2),{t,0,.4}]
Show[plot1,plot2]
```

9.5 Coupling between guided modes: coupled mode theory

Guided modes are eigenstates of the electromagnetic field in the waveguide. They form a complete basis (along with the radiative states which can propagate outside the guide and which we shall continue to ignore for the present) in which any

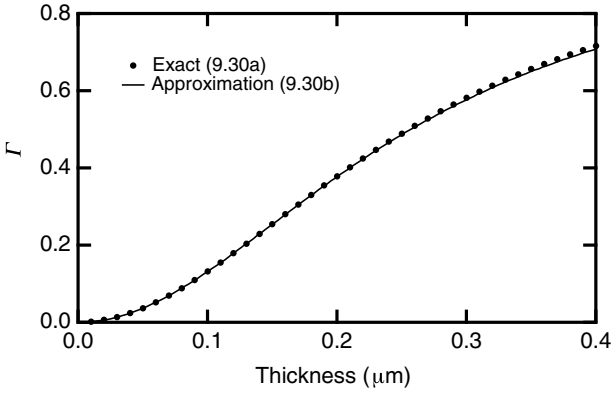


Fig. 9.9. Comparison between the confinement factor Γ obtained using the exact method in (9.30a) and the approximation in (9.30b) for the 1.55 μm GaSb/AlGaSb waveguide described in the example.

perturbation can be expanded. Such perturbations can originate from many possible sources: corrugation of the guide (random or periodic), non-linear interactions (second harmonic generation, for example), or possibly from the presence of another waveguide close enough to allow coupling by photon tunnelling. All these perturbations are treated within the framework of a very powerful formalism known as *coupled mode theory*. This theory is quite similar to perturbation theory in quantum mechanics.

We will begin with Maxwell's equations in matter:

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{r},t) &= -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \\ \nabla \times \mathbf{B}(\mathbf{r},t) &= \mu_0 \frac{\partial}{\partial t}\mathbf{D}(\mathbf{r},t) \end{aligned} \tag{9.31}$$

where $\mathbf{D}$ is the displacement vector:

$$\begin{aligned} \mathbf{D}(\mathbf{r},t) &= \varepsilon_0 \mathbf{E}(\mathbf{r},t) + \varepsilon_0 \chi_1(\mathbf{r})\mathbf{E}(\mathbf{r},t) + \mathbf{P}_{\text{per}}(\mathbf{r},t) \\ &= \varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r},t) + \mathbf{P}_{\text{per}}(\mathbf{r},t) \end{aligned} \tag{9.32}$$

In this last definition, $\varepsilon(\mathbf{r}) = n^2(\mathbf{r})$ is the dielectric constant for the different materials composing the guide $(1, 2, \ldots)$ and $\mathbf{P}_{\text{per}}$ is the perturbative polarization, which excludes the dielectric response of the different layers (as they are included in $\varepsilon(\mathbf{r})$). Equations (9.31) then become:

$$\nabla^2 \mathbf{E}(\mathbf{r}) - \mu_0 \varepsilon(\mathbf{r}) \frac{\partial^2}{\partial t^2}\mathbf{E}(\mathbf{r}) = \mu_0 \frac{\partial^2}{\partial t^2}\mathbf{P}_{\text{per}}(\mathbf{r},t) \tag{9.33}$$

Propagation equation with a perturbative polarization source

This equation forms the basis for all perturbative treatments of propagation effects in guides (scattering, diffraction, . . .) and plays a central role in optoelectronics (see, for example, Chapter 12). To keep the required notation manageable, we will not consider variations in y, but will interest ourselves with the amplitude of the electric field over $\mathbf{e}_y$ (which we will label as $E(\mathbf{r}, t)$). The differential operator ∇^2 is then $\partial^2/\partial x^2 + \partial^2/\partial z^2$. The field $E(\mathbf{r}, t)$ can be expanded in the basis of the guided modes, $\mathbf{E}_m(\mathbf{r}, t) = \mathbf{e}_y E_m(x) \mathrm{e}^{\mathrm{i}(\omega t - \beta_m z)}$, or:

$$E(\mathbf{r}, t) = \frac{1}{2} \sum_m A_m(z) E_m(x) \mathrm{e}^{\mathrm{i}(\omega t - \beta_m z)} + \text{c.c.} \tag{9.34}$$

where we recall that each mode must satisfy:

$$\frac{\partial^2}{\partial x^2} E_m(x) + [k^2 \varepsilon(\mathbf{r}) - \beta_m^2] E_m(x) = 0 \tag{9.35}$$

The terms A_m are the amplitudes of the guided wave varying along z owing to the perturbation. Taking into account the normalization (9.20), the square of the norm $|A_m|^2$ is *the optical power per unit width of the waveguide due to the* m*th mode.* Without perturbations, the terms A_m are independent of z. Substituting (9.34) into (9.33), we obtain:

$$\begin{aligned} &\sum_m \left\{ \frac{A_m(z)}{2} \left[\frac{\partial^2}{\partial x^2} E_m(x) - \beta_m^2 E_m(x) + \mu_0 \varepsilon(\mathbf{r}) \omega^2 E_m(x) \right] \right. \\ &\left. + \frac{1}{2} \left[\frac{\partial^2}{\partial z^2} A_m(z) - 2\mathrm{i}\beta_m \frac{\partial}{\partial z} A_m(z) \right] E_m(x) \right\} \mathrm{e}^{\mathrm{i}(\omega t - \beta_m z)} + \text{c.c.} = \mu_0 \frac{\partial^2}{\partial t^2} P^y_{\text{per}}(\mathbf{r}, t) \end{aligned} \tag{9.36}$$

where P^y_{per} is the component of the perturbing polarization vector in the Oy direction. We note that the first sum is null given (9.36). Also, we make the approximation that the amplitude of the wave is slowly varying (the envelope function approximation). This allows us to write:

$$\left| \frac{\partial^2}{\partial z^2} A_m \right| \ll \beta_m \left| \frac{\partial}{\partial z} A_m \right| \tag{9.37}$$

The envelope function approximation

Equation (9.36) then becomes:

$$\sum_m -\mathrm{i}\beta_m \frac{\partial}{\partial z} A_m(z) E_m(x) \mathrm{e}^{-\mathrm{i}(\omega t - \beta_m z)} + \text{c.c.} = \mu_0 \frac{\partial^2}{\partial t^2} P^y_{\text{per}}(\mathbf{r}, t) \tag{9.38}$$

We then multiply by a mode $E_l(z)$ and take the integral which is equivalent to projecting (9.33) onto the eigenmode basis of the guide. Making use of the orthonormality conditions (9.22), we obtain after grouping similar terms:

$$\frac{\partial}{\partial z}A_l^- e^{i(\omega t+\beta_l z)} - \frac{\partial}{\partial z}A_l^+ e^{i(\omega t-\beta_l z)} - \text{c.c.}$$
$$= -\frac{i}{2\omega p_0}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{+\infty}[P^y_{\text{per}}(x,t)E_l(x)]dx \tag{9.39}$$

Coupled mode equation for a perturbing polarization

where we recall that p_0 is a normalization constant ($p_0 = 1\,\text{W m}^{-1}$). In this last equation, A_l^+ and A_l^- are the amplitudes for the lth mode propagating in the $+z$ and $-z$ directions, respectively. Equation (9.39) describes how a perturbing polarization $P_{\text{per}}(\mathbf{r}, t)$ can induce a transfer of energy between different cavity eigenmodes. Often, the perturbing polarization is a function of other perturbed modes and (9.39) (i.e. the *coupled mode equation*) acts to couple all these modes. This is a very powerful formalism with a broad range of applicability. We will use it to describe the function of optical couplers (Complement 9.A), Bragg waveguides (Complement 9.B), distributed feedback lasers (Complement 13.A), and optical frequency conversion in non-linear waveguides (Complement 9.C).

FURTHER READING

S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley Interscience, New York (1995).
D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, New York (1974).
B. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley Interscience, New York (1991).
A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

Complement to Chapter 9

9.A Optical coupling between guides: electro-optic switches

We saw in Section 9.3 that an electromagnetic wave trapped in the core layer of a waveguide possesses an evanescent component in the confining layer (see Fig. 9.4). If a second guide capable of confining a wave of the same frequency is placed close enough to the first waveguide so that the overlap with the evanescent portion of the wave is significant, the confined wave will move from the first waveguide through the confining layer to the second waveguide. This phenomenon is analogous to *electronic tunnelling* through barriers in quantum mechanics. Coupled mode formalism is particularly well suited to describing this form of coupling between optical waveguides.

We will consider two waveguides (see Fig. 9.A.1). The first, the *right* waveguide, consists of a core layer with an index n_R placed between two confining layers having an index n. The second waveguide, situated a distance D to the *left* of the first, possesses a core layer with index n_L and is placed between the same barrier material as the right waveguide. The two guides were fabricated so that if they were separated from one another by an infinite distance, they would each possess a single fundamental eigenmode possessing electric field components along Oy, with $E_R(x,z,t)$ and $E_L(x,z,t)$, each, respectively, being solutions to equations of type (9.13) and of the form:

$$E_R(x,z,t) = \frac{1}{2}[E_R(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_R z)} + \text{c.c.}]$$
$$E_L(x,z,t) = \frac{1}{2}[E_L(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_L z)} + \text{c.c.}] \tag{9.A.1}$$

When brought into close proximity, each waveguide acts on the other to produce a perturbation.

Following the principles set forth in Section 9.4, identical to those obtained within the framework of time-independent perturbation theory in quantum mechanics (see, for example, Complement 1.B), we may write the new eigenmodes for the coupled system as linear combinations of the fundamental eigenmodes of each

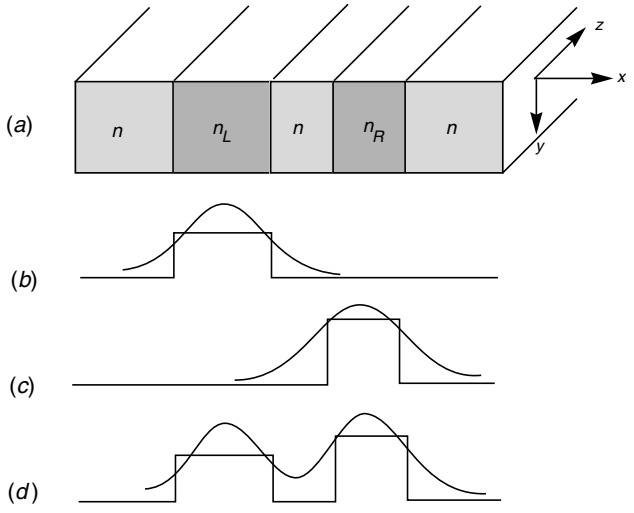


Fig. 9.A.1. (*a*) Geometry for coupled waveguides. (*b*) Wave confined in the left waveguide only. (*c*) Wave confined in the right waveguide only. (*d*) Wave coupled between both waveguides.

isolated guide (as was the case with quantum wells, we will neglect the superior modes):

$$E(x,z,t) = \frac{1}{2}[A_R(z)E_R(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_R z)} + A_L(z)E_L(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_L z)} + \text{c.c.}] \tag{9.A.2}$$

If the guides are infinitely far apart, the amplitudes $A_R(z)$ and $A_L(z)$ are constants. Bringing both guides into proximity will induce energy transfer between the waveguides causing variations in the amplitudes $A_R(z)$ and $A_L(z)$.

For the waveguide on the right, the perturbation P^R_{per} is the polarization in the left guide due to the synchronous excitation of the new field $E(x,z,t)$ given by (9.A.2), or:

$$P^R_{\text{per}}(x,z,t) = \varepsilon_0[n_L^2(x) - n^2]E(x,z,t) \tag{9.A.3}$$

where $n_L(x)$ is a function which equals n_L in the left guide and is n everywhere else.

Considering the problem for a single wave propagation direction, the coupled wave equation, (9.39), may be written:

$$-\frac{\partial}{\partial z}A_R\mathrm{e}^{\mathrm{i}(\omega t-\beta_R z)} = \frac{\mathrm{i}\omega\varepsilon_0}{4p_0}\left\{\int_{\substack{\text{left}\\\text{guide}}}[n_L^2(x) - n^2][A_R(z)E_R^2(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_R z)}]\mathrm{d}x\right. \tag{9.A.4}$$

$$\left. + \int_{\substack{\text{left}\\\text{guide}}}[n_L^2(x) - n^2][A_L(z)E_L(x)E_R(x)\mathrm{e}^{\mathrm{i}(\omega t-\beta_L z)}]\mathrm{d}x\right.$$

In (9.A.4), the first term in the integral on the right-hand side has no important effect. It acts only as a small correction to the propagation constant β_R. The second term, however, acts to couple the two waveguides. Equation (9.A.4) and its equivalent for the left guide then become:

$$\begin{aligned}\frac{\partial}{\partial z}A_R &= -\mathrm{i}g_{LR}A_L\mathrm{e}^{-\mathrm{i}\Delta\beta z}\\ \frac{\partial}{\partial z}A_L &= +\mathrm{i}g_{RL}A_R\mathrm{e}^{\mathrm{i}\Delta\beta z}\end{aligned} \tag{9.A.5}$$

where we have introduced the *phase mismatch* $\Delta\beta$ between the two waveguides given by:

$$\Delta\beta = \beta_L - \beta_R \tag{9.A.6}$$

and the *overlap integrals* for the two waves are given by:

$$\begin{aligned}g_{LR} &= \frac{\omega\varepsilon_0}{4p_0}\int_{\substack{\text{left}\\ \text{guide}}}[n_L^2(x) - n^2]E_L(x)E_R(x)\mathrm{d}x\\ g_{RL} &= \frac{\omega\varepsilon_0}{4p_0}\int_{\substack{\text{right}\\ \text{guide}}}[n_R^2(x) - n^2]E_L(x)E_R(x)\mathrm{d}x\end{aligned} \tag{9.A.7}$$

Equations (9.A.5) are the classic equations for coupled pendulums and are entirely equivalent to the Rabi equations discussed in Complement 1.E. These equations may be solved by introducing trial functions of the type $\mathrm{e}^{\mathrm{i}\lambda t}$. We assume that the amplitudes at the entrance of each waveguide are $A_L(0) = A_0$ and $A_R(0) = 0$.

Anticipating symmetric waveguides we will suppose that $g_{RL} = g_{LR} = g$. The solutions (9.A.5) may be written:

$$\begin{aligned}A_R(z) &= A_0\frac{g}{\mathrm{i}\gamma}\mathrm{e}^{\mathrm{i}(\Delta\beta z/2)}\sin\gamma z\\ A_L(z) &= A_0\mathrm{e}^{-\mathrm{i}(\Delta\beta z/2)}\left(\cos\gamma z + \mathrm{i}\frac{\Delta\beta}{2\gamma}\sin\gamma z\right)\end{aligned} \tag{9.A.8}$$

with:

$$\gamma^2 = g^2 + \left(\frac{\Delta\beta}{2}\right)^2 \tag{9.A.9}$$

The optical power levels in the guides are then:

$$\begin{aligned} I_R(z) &= P_0 \frac{g^2}{\gamma^2} \sin^2(\gamma z) \\ I_L(z) &= P_0 - P_R(z) \end{aligned} \tag{9.A.10}$$

where P_0 is the incident power at the entrance of the left waveguide. Equation (9.A.10) is rich in information. First, it indicates that the wave energy oscillates between the two guides over the entire distance of travel with a spatial period of λ_{trans} referred to as the *transfer distance* (see Fig. 9.A.2) given by:

$$\lambda_{\text{trans}} = \frac{2\pi}{\sqrt{g^2 + \left(\frac{\Delta\beta}{2}\right)^2}} \tag{9.A.11}$$

Additionally, the maximum fraction of power which can be transferred from one guide to the next is:

$$\frac{I_R^{\max}}{I_0} = \frac{g^2}{g^2 + (\Delta\beta/2)^2} \tag{9.A.12}$$

As was the case with Rabi oscillations, we are not surprised to find that power can be completely transferred between the waveguides over the course of an oscillation as long as the phases are in agreement, i.e. as soon as:

$$\Delta\beta = \beta_L - \beta_R = 0 \tag{9.A.13}$$

Condition for phase matching between two waveguides

Waveguides can therefore act as optical switches (all the more so if they are identical). Let us calculate the transfer constant κ for two TE_0 waves for the case where $n_R = n_L = n_1$. The fields $E_L(x)$ and $E_R(x)$ are given by (9.16). The overlap integrals (9.A.7) can be easily calculated:

$$g = \frac{\omega\varepsilon_0(n_1^2 - n_2^2)}{4p_0} \int_{-d/2}^{d/2} B_L \cos(\alpha x) A_R e^{-\kappa(D-x)} dx \tag{9.A.14}$$

B_L is the wave amplitude in the left guide, where the guide centre has been taken as the origin (explaining the integral over $-d/2$ to $d/2$), and A_R is the amplitude of the evanescent wave stemming from the right guide. We recall that the electric fields

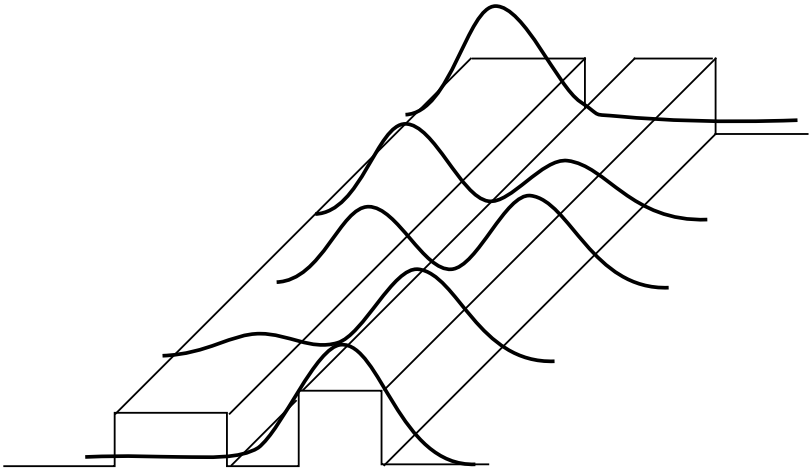


Fig. 9.A.2. The power contained by the wave oscillates between the two coupled guides. The distance required for this oscillation to complete a full cycle is referred to as the transfer distance λ_{trans}.

$E(x)$ vary as $\cos(\alpha x)$ in the two guides and as $e^{-\kappa x}$ in the confining layers. These amplitudes are given in (9.21) yielding:

$$g = \frac{\lambda_0}{\pi n_{\text{eff}}} \frac{\alpha^2}{\alpha^2 + \kappa^2} \frac{\kappa}{d + 2/\kappa} e^{-\kappa D} \tag{9.A.15}$$

The variation in $e^{-\kappa D}$ justifies the interpretation of the constant κ in terms of a *photonic tunnelling constant* between the two guides.

Example

We consider two guides with InGaAs core layers ($n_1 = 3.5$) and confining InAlAs barrier layers ($n_2 = 3.3$). The width of the guides is 0.6 μm. The coefficients α and κ are obtained using the MATHEMATICA program listed below:

```
Lambda=1.55;t=0.6;(*micron*)
n1=3.5;n2=3.3;
k= 2*N[Pi]/Lambda;
Bet=2*N[Pi]*neff/Lambda;
alpha= Sqrt[(n1^2*k^2-Bet^2)];
kapa= Sqrt[(-n2^2*k^2+Bet^2)];
fct=Tan[alpha*t]-2*kapa*alpha/(alpha^2-kapa^2);
Plot[fct,{neff,n2,n1}]
sol=FindRoot[fct==0,{neff,3.43}]
alpha/.sol(*micron^-1*)
kapa/.sol(*micron^-1*)
```

from which we obtain $\alpha = 2.97\,\mu\text{m}^{-1}$ and $\kappa = 3.67\,\mu\text{m}^{-1}$. If the two guides are

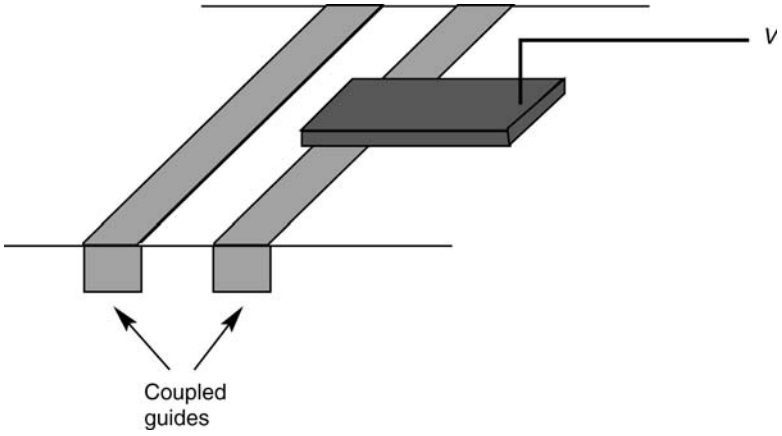


Fig. 9.A.3. Schematic for an electro-optic switch based on coupled waveguides.

spaced apart by a distance of 1 μm, the coupling coefficient g is given by (9.A.16) to be $4.6 \times 10^{-3}\,\mu\text{m}^{-1}$. The transfer length λ_{trans} is therefore of the order of $2\pi/g = 1.3\,\text{mm}$.

This energy transfer mechanism between coupled waveguides is put to use in electro-optic switches, illustrated in Fig. 9.A.3. In this case, two identical waveguides are fabricated with electro-optic cores having as a result optical indices n which vary as a function of the applied field F according to:

$$\Delta n = n^3 r F \tag{9.A.16}$$

where r is the electro-optical index and is typically equal to a few pm V^{-1} in semiconductors ($r = 1.6\,\text{pm V}^{-1}$ and $n^3 r = 59\,\text{pm V}^{-1}$ for GaAs), with the result being that a variation Δn of 5×10^{-4} is easily obtained for a field modulation of $100\,\text{kV cm}^{-1}$. The guide lengths are selected so that under a null electric field (the two guides are identical and $\Delta\beta = 0$) all the energy in the left guide is transferred into the right guide at the end of the guides, i.e. $gL = \pi/2$. A metallic electrode is deposited on the left waveguide so that when an electric field F is applied, the index of the left guide is altered along with the propagation constant β_g. This leads to a phase difference between the two guides of $\Delta\beta(F)$. The output power at the exit of the right waveguide is then a function of the electric field given by (9.A.10), or:

$$\frac{I_R(L)}{I_L(L)} \approx \sin^2\left[\frac{\pi}{2}\left\{1 + \left[\frac{L\Delta\beta(F)}{\pi}\right]^2\right\}^{1/2}\right] \tag{9.A.17}$$

The optical power will be almost completely transferred to the left guide ($I_R(L) = 0$) when a sufficiently strong electric field F has been applied so that:

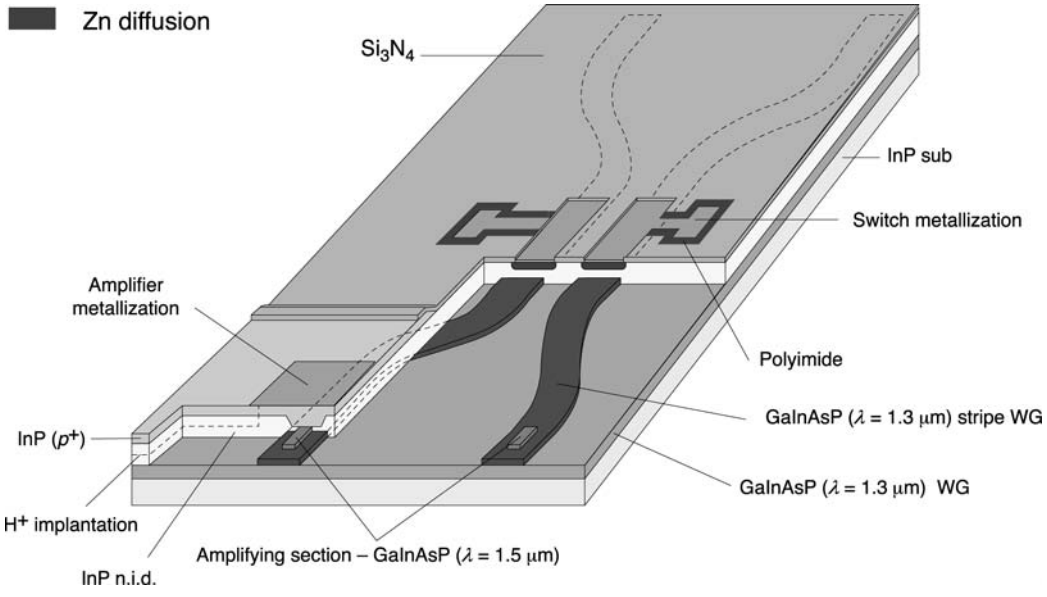


Fig. 9.A.4. Implementation of a coupled waveguide switch. The schematic view shows a 2×2 switch integrated monolithically with an optical amplifier. (Courtesy of N. Vodjdani @LCR/THALES.)

$$L\Delta\beta(F) = \sqrt{3}\pi \tag{9.A.18}$$

To estimate the required field strengths, we may crudely assume that a variation in the index Δn_L (given by (9.A.16)) affects the propagation constant as $\beta_L = 2\pi n_{\text{eff}}/\lambda_0$ (we recall that λ_0 is the vacuum wavelength and that n_{eff} is the effective optical index in the guide) so that $\Delta\beta = 2\pi\Delta n/\lambda_0$. Condition (9.A.18) then implies a *switching voltage* given by the condition that $\Delta n_{\text{eff}}(F) = (\sqrt{3}/2)(\lambda_0/L)$. For a guide length of 2 mm and a wavelength of 1 μm, we need to be able to impart a change Δn of 4×10^{-4}, i.e. corresponding to a switching field of $100\,\text{kV cm}^{-1}$ or 10 V for a 1 μm thick guide. Figure 9.A.4 shows an actual implementation of such an electro-optical modulator. As no bulk displacement of charges is necessary to operate this device, capacitive effects are virtually null and the switching times are extremely short. Maximum modulation frequencies for these types of modulators are in excess of 10 GHz.

FURTHER READING

A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

9.B Bragg waveguides

We will see how the coupled mode equation, (9.39), allows us to study the behaviour of a guided wave in a corrugated optical guide with a spatial period of L (see Fig. 9.B.1). When the grating period equals the wavelength of the propagating light (i.e. $2\pi/\beta$), the grating portion of the guide acts as a Bragg reflector (see Complement 9.D). This situation with respect to photon propagation is similar to that encountered in Chapter 5 for the case of electrons in crystalline solids (and which led to the formation of allowed energy bands). The Bragg grating can be represented by a periodic modulation in the relative permittivity. To begin, we shall assume that this modulation varies sinusoidally as:

$$\Delta\varepsilon(\mathbf{r}) = \varepsilon_M(x)\sin\left(\frac{2\pi}{\Lambda}z\right) \tag{9.B.1}$$

The perturbative polarization contributed by this modulation is then given by:

$$P_{\text{pert}} = \varepsilon_0\Delta\varepsilon(\mathbf{r})E(\mathbf{r},t) \tag{9.B.2}$$

The electromagnetic field $E(\mathbf{r},t)$ in the guide can be decomposed into the basis of the waveguide eigenmodes as described in (9.34). Employing (9.B.1) and (9.B.2), the perturbative polarization may be written:

$$P_{\text{pert}} = \frac{\varepsilon_0\varepsilon_M}{4p_0}\mathrm{e}^{\mathrm{i}\omega t}\sum_m A_m(z)E_m(x)\mathrm{e}^{-\mathrm{i}(\beta_m\pm(2\pi/\Lambda))z} + \text{c.c.} \tag{9.B.3}$$

We then substitute (9.B.3) into the coupled mode equation (9.39), by replacing the derivative $\partial^2/\partial t^2$ by $-\omega^2$ to obtain:

$$\begin{aligned}
&\frac{\partial}{\partial z}A_l^-\mathrm{e}^{\mathrm{i}\beta_l z} - \frac{\partial}{\partial z}A_l^+\mathrm{e}^{-\mathrm{i}\beta_l z} - \text{c.c.} \\
&= -\frac{\mathrm{i}\omega\varepsilon_0}{8p_0}\sum_m A_m^+(z)\mathrm{e}^{-\mathrm{i}(\beta_m\pm(2\pi/\Lambda))z}\int_{-\infty}^{+\infty}\varepsilon_M(x)[E_m(x)E_l(x)]\mathrm{d}x \\
&\quad -\frac{\mathrm{i}\omega\varepsilon_0}{8p_0}\sum_m A_m^-(z)\mathrm{e}^{\mathrm{i}(\beta_m\pm(2\pi/\Lambda))z}\int_{-\infty}^{+\infty}\varepsilon_M(x)[E_m(x)E_l(x)]\mathrm{d}x + \text{c.c.}
\end{aligned} \tag{9.B.4}$$

We took care to preserve the two *counterpropagative modes* A^+ and A^-, as we anticipate their coupling by the grating. Equation (9.B.4) couples the different amplitudes $A_m(z)$ through the coupled differential equations. We can avoid a lot of

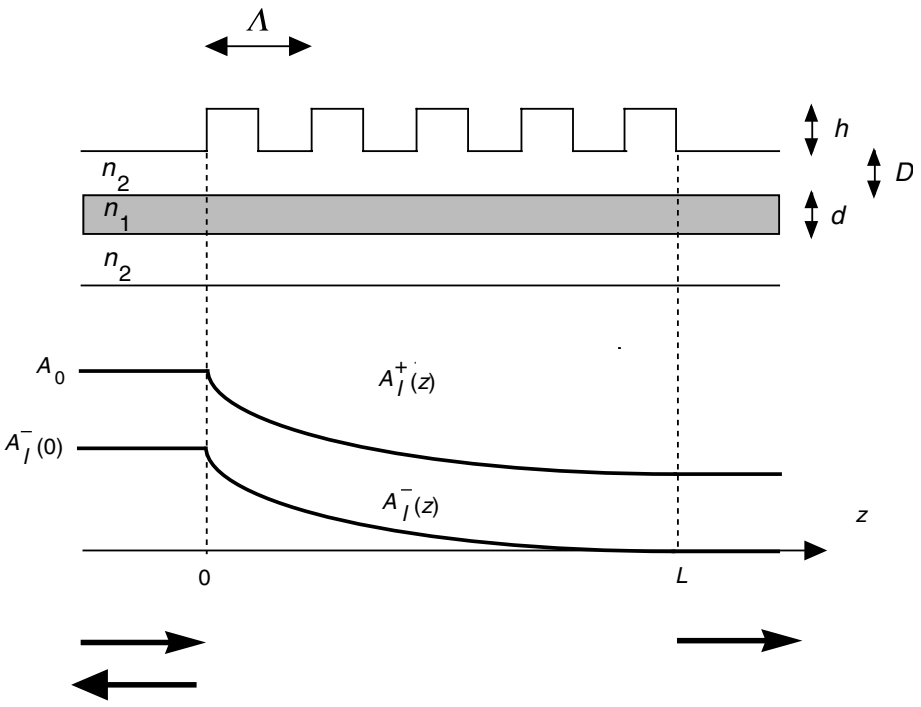


Fig. 9.B.1. Geometry for a Bragg waveguide.

needless work by inspecting this equation. Two types of propagation constants are introduced into (9.B.4):

$$\beta_m - \beta_l \pm \frac{2\pi}{\Lambda}$$

which couples $\partial A_l^-/\partial z$ to A_m^- and:

$$\beta_m + \beta_l \pm \frac{2\pi}{\Lambda}$$

which couples $\partial A_l^-/\partial z$ to A_m^+.

The first case does not interest us but does provide a means of dealing with near-field Fraunhofer-type diffraction problems. We are concerned, however, with the second case which couples an incident wave with index l from the right to a second with index m on the left. More specifically, we will deal with the special case $l = m$, i.e. corresponding to reflection. We keep in (9.B.4) only those terms in which the phase mismatch $2\Delta\beta$ is close to zero, since only those terms will eventually interfere constructively. This mismatch is given by:

$$\Delta\beta = \beta_l - \frac{\pi}{\Lambda} \tag{9.B.5}$$

Phase mismatch in a Bragg waveguide

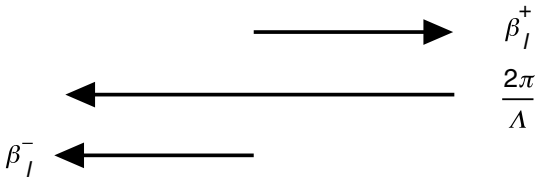


Fig. 9.B.2. Interpretation of the phase matching condition in a Bragg waveguide in terms of wavevector conservation.

This last condition can be interpreted in terms of the conservation of total wavevector. The periodic lattice supplies a wavevector $-2\pi/\Lambda$, which subtracts from the wavevector of the incident wave β_l^+ and generates the wavevector corresponding to the reflected wave β_l^- (see Fig. 9.B.2). Equation (9.B.4) can then be separated into two coupled equations:

$$\frac{\mathrm{d}}{\mathrm{d}z}A_l^-(z) = -\mathrm{i}gA_l^+(z)\mathrm{e}^{-2\mathrm{i}\Delta\beta z}$$
$$\frac{\mathrm{d}}{\mathrm{d}z}A_l^+(z) = +\mathrm{i}gA_l^-(z)\mathrm{e}^{2\mathrm{i}\Delta\beta z} \tag{9.B.6a}$$

Equations for coupled modes in a Bragg waveguide

where the coupling constant g is given by:

$$g = \frac{\omega\varepsilon_0}{8p_0}\int_{-\infty}^{+\infty}\varepsilon_M(x)E_l(x)^2\mathrm{d}x \tag{9.B.6b}$$

Starting from (9.B.6), it can readily be shown that:

$$\frac{\mathrm{d}}{\mathrm{d}z}(|A_l^-|^2 + |A_l^+|^2) = 0 \tag{9.B.7}$$

which implies that combined optical power from the two modes remains constant.

The coupled equations, (9.B.6), can be solved in standard fashion: i.e. we substitute one of the two equations into the other giving, for example:

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}A_l^+ - 2\mathrm{i}\Delta\beta\frac{\mathrm{d}}{\mathrm{d}z}A_l^+ - g^2A_l^+ = 0 \tag{9.B.8}$$

The amplitudes A_l^+ and A_l^- are therefore linear combinations of the exponentials of the arguments:

$$\lambda = \mathrm{i}\Delta\beta \pm \delta \tag{9.B.9a}$$

where δ is the discriminant of (9.B.8):

$$\delta = \sqrt{g^2 - \Delta\beta^2} \tag{9.B.9b}$$

To obtain the complete solution, we impose boundary conditions upon the amplitudes in the guide. The wave is incident from the left with an amplitude A_0 at the grating entrance ($z = 0$) (see Fig. 9.B.1), whereas no wave from the left is incident upon guide exit at $z = L$. The two boundary conditions are then that $A_l^+(0) = A_0$ and $A_l^-(L) = 0$, and the solutions to (9.B.6) are easily found to be[1]:

$$\begin{aligned} A_l^-(z) &= A_0 \frac{g}{\mathrm{i}\Delta\beta\,\mathrm{sh}(\delta L) + \delta\,\mathrm{ch}(\delta L)} \mathrm{e}^{-\mathrm{i}\Delta\beta z}\mathrm{sh}[\delta(z-L)] \\ A_l^+(z) &= A_0 \frac{\mathrm{e}^{+\mathrm{i}\Delta\beta z}}{\mathrm{i}\Delta\beta\,\mathrm{sh}(\delta L) + \delta\,\mathrm{ch}(\delta L)} \{\delta\,\mathrm{ch}[\delta(z-L)] - \mathrm{i}\Delta\beta\,\mathrm{sh}[\delta(z-L)]\} \end{aligned} \tag{9.B.10a}$$

The behaviour of the two waves is particularly simple when the Bragg condition is satisfied ($\Delta\beta = 0$), i.e. when the wavevector π/Λ of the grating equals that of the guided wave β_l. The solutions (9.B.10a) then take the simple form:

$$\begin{aligned} A_l^-(z) &= A_0 \frac{\mathrm{sh}[\delta(z-L)]}{\mathrm{ch}(\delta L)} \\ A_l^+(z) &= A_0 \frac{\mathrm{ch}[\delta(z-L)]}{\mathrm{ch}(\delta L)} \end{aligned} \tag{9.B.10b}$$

Figure 9.B.1 shows in this last case the behaviour of the square of the amplitudes (i.e. the optical power per unit width in the guide) as a function of distance. The incident wave A_l^+ decays exponentially in the grating region giving rise to a reflected wave A_l^-. If the product δL is sufficiently large, i.e. if the grating is sufficiently effective, $A_l^-(0)$ is of the order of A_0, whereas $A_l^+(L) \approx 0$. The transfer of energy between the two waves is then complete and the grating region behaves as a Bragg reflector.

This Bragg reflector possesses an associated *spectral bandwidth* which is set by the requirement that δ be real, i.e. by taking into account (9.B.9b):

$$\frac{\pi}{\Lambda} - g < \beta_l(\omega) < \frac{\pi}{\Lambda} + g \tag{9.B.11}$$

Bandwidth of a Bragg reflector

or a bandwidth of $2g$ in wavenumber. Figure 9.B.3 shows a typical example for transmission in a Bragg waveguide as described by:

$$T = \frac{|A_l^+(L)|^2}{|A_l^+(0)|^2} = \left| \frac{\delta}{\delta\,\mathrm{ch}(\delta L) + \mathrm{i}\Delta\beta\,\mathrm{sh}(\delta L)} \right|^2 \tag{9.B.12a}$$

[1] Where sh = sinh, ch = cosh.

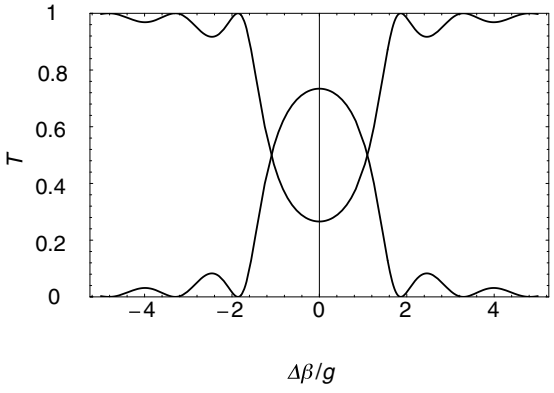


Fig. 9.B.3. Transmission and reflection coefficients for a Bragg waveguide as a function of normalized phase mismatch $\Delta\beta/g$ for $gL = 2$.

Condition (9.B.11), as well as the expression for the transmission in the waveguide (9.B.12a), shows that the product gL controls the maximum efficiency of the Bragg guide. At resonance, $\Delta\beta = 0$ and the transmission is given by:

$$T_{\text{res}} = \left| \frac{1}{\text{ch}(\delta L)} \right|^2 \approx \text{e}^{-2gL} \tag{9.B.12b}$$

If the coupling coefficient g is weak, the Bragg guide must be long enough to keep $gL \gg 1$. In order to understand the roles of the various parameters (and how they should be combined to yield an efficient Bragg waveguide) we will seek to derive an expression for the efficiency of a highly confined waveguide.

We consider a waveguide of thickness d, with optical indices $\sqrt{\varepsilon_1}$ and $\sqrt{\varepsilon_2}$ for the core and confinement layers, respectively (see Fig. 9.B.1). Assuming TE_0 wave propagation in a highly confining guide, (9.13) yields for the transverse wavelength α and the tunnelling penetration depth κ the following approximations (see (9.23b)):

$$\begin{aligned} \alpha &\approx \frac{\pi}{d} \\ \frac{\kappa}{\alpha} &\approx (\varepsilon_1 - \varepsilon_2)^{1/2} \frac{2d}{\lambda_0} \end{aligned} \tag{9.B.13}$$

and the normalization coefficient of the guided wave (9.21) is then:

$$A^2 = 4p_0 \left(\frac{\alpha}{\kappa} \right)^2 \frac{\mu_0 c}{n_{\text{eff}} d} \tag{9.B.14}$$

If the grating is created at a distance D from the core, with a modulation amplitude h (see Fig. 9.B.1), the coupling constant g is then (according to (9.B.6b)):

$$g = \frac{\omega \varepsilon_0 \varepsilon_M}{8 p_0} \int_D^{D+h} A^2 \mathrm{e}^{-2\kappa x} \mathrm{d}x \tag{9.B.15}$$

or, given (9.B.13) and (9.B.14):

$$g = \frac{\pi}{4} \frac{\varepsilon_M}{n_{\mathrm{eff}}(\varepsilon_1 - \varepsilon_2)} \left(\frac{\lambda_0 h}{d^2} \right) \frac{e^{-2\kappa D}}{d} \tag{9.B.16}$$

This last expression shows that the efficiency of a Bragg waveguide is dominated by the photon tunnelling effect (i.e. $\propto \mathrm{e}^{-2\kappa D}$).

Example

We seek the product gL for a GaAs/AlGaAs waveguide having the following parameters:

$n_1 = 3.3$, $n_2 = 3$, and $\varepsilon_M = 3$
$\lambda_0 = 1\,\mu\mathrm{m}$, $d = 0.5\,\mu\mathrm{m}$, $D = 0.2\,\mu\mathrm{m}$, and $h = 0.1\,\mu\mathrm{m}$

The transverse wavenumber α is π/d, or $6.28\,\mu\mathrm{m}^{-1}$, and the photon tunnelling constant κ is given by (9.B.13) to be $(3.3^2 - 3^2)^{1/2}\alpha = 8.63\,\mu\mathrm{m}^{-1}$. The product gL is then found using (9.B.16) or:

$$gL = \frac{\pi}{4} \frac{3}{3(3.3^2 - 3^2)} \left(\frac{1 \times 0.1}{0.5^2} \right) \mathrm{e}^{-2 \times 8.6 \times 0.2} \frac{L}{d} = 5 \times 10^{-3} \frac{L}{d}$$

To obtain a value of gL of the order of 1, we would need a guide length of $200d$ or $100\,\mu\mathrm{m}$.

FURTHER READING

S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley Interscience, New York (1995).
A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

9.C Frequency conversion in non-linear waveguides

As we shall see in Chapter 12, semiconductors possess second-order non-linear susceptibilities which are larger than most other materials. We recall that in such a medium, a non-linear second harmonic polarization $P^{2\omega}$ is generated by the interaction of two waves of frequency ω according to:

$$P_k^{2\omega} = \varepsilon_0 \chi_{kij}^{2\omega} E_i^{\omega} E_j^{\omega} \tag{9.C.1}$$

This formula can be understood as follows: an electromagnetic field polarized along the i direction interacts with a second wave polarized along the j axis resulting in a non-linear polarization along the k axis. χ is therefore a tensor quantity. Furthermore, the real physical polarization *is obtained by taking the real part of (9.C.1).*

As we shall see in this complement, there is significant interest in using waveguides to perform optical frequency conversion by exploiting the optical non-linearities in semiconductors. The first reason for doing so is that the optical confinement in waveguides leads to significant electromagnetic field amplitudes even for relatively modest optical power levels. The second reason is that it is possible to manage problems associated with phase mismatch with the help of current microfabrication technology.

Utilizing both the geometry and the notation used for waveguides in Chapter 9 (see Fig. 9.3), we will examine the two following cases:

9.C.1 TE mode in–TE mode out

In this case, we will suppose that a single guided TE wave interacts with itself, so that the amplitude is given by:

$$E_{\mathrm{TE}}^{\omega}(x,z,t) = \frac{1}{2}\left[A_y^{\omega}(z) E_y^{\omega}(x) \mathrm{e}^{-\mathrm{i}(\omega t - \beta_{\omega,\mathrm{TE}} z)} + \text{c.c.}\right] \tag{9.C.2}$$

where E_y^{ω} is a solution to the waveguide equation (9.13), $A_y^{\omega}(z)$ anticipates the variation of the amplitude of the pump beam (with frequency ω) along the guide, and $\beta_{\omega,\mathrm{TE}}$ is the propagation constant of the TE wave with frequency ω in the guide given by the solution to (9.13). Eliminating the terms in $\omega - \omega$ (which correspond to optical rectification), the non-linear polarization at 2ω in (9.C.1) becomes:

$$P_{yyy}^{2\omega}(x,z,t) = \frac{1}{2}\left[P_{yyy}^{2\omega}(x,z) \mathrm{e}^{-\mathrm{i}(2\omega t - 2\beta_{\omega,\mathrm{TE}} z)} + \text{c.c.}\right] \tag{9.C.3a}$$

and using (9.C.2):

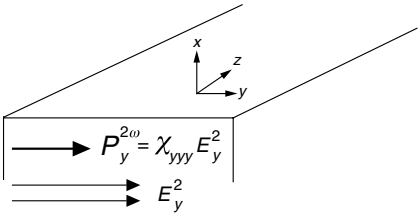


Fig. 9.C.1. Configuration leading to second harmonic generation TE–TE.

$$P_{yyy}^{2\omega}(x,z) = \varepsilon_0 \chi_{yyy}^{2\omega} [A_y^{\omega}(z) E_y^{\omega}(x)]^2 \tag{9.C.3b}$$

Inspecting the consequences associated with the introduction of non-linear polarization in the coupled mode equation, (9.23), we see that this 2ω source term will act to generate waves with a frequency of 2ω. Equation (9.39) allows one to calculate the variation in amplitude of the 2ω wave in the guide. Considering only those waves which propagate towards the right, (9.39) becomes:

$$\begin{aligned} &-\frac{\mathrm{d}}{\mathrm{d}z} A_y^{2\omega} \mathrm{e}^{\mathrm{i}(2\omega t - \beta_{2\omega,\mathrm{TE}} z)} - \text{c.c.} \\ &= -\frac{\mathrm{i}}{2\omega p_0} \frac{\partial^2}{\partial t^2} \int_{-\infty}^{+\infty} \frac{1}{2} \{\varepsilon_0 \chi_{yyy}^{2\omega} f(x) [A_y^{\omega}(z) E_y^{\omega}(x)]^2 \mathrm{e}^{\mathrm{i}(2\omega t - 2\beta_{\omega,\mathrm{TE}} z)} + \text{c.c.}\} E_y^{2\omega}(x) \mathrm{d}x \end{aligned} \tag{9.C.4}$$

$\beta_{2\omega,\mathrm{TE}}$ is the propagation constant of the 2ω wave within the guide. Using the notation for the effective index, we see that $\beta_{2\omega,\mathrm{TE}} = n_{\mathrm{eff}}(2\omega) 2\omega/c$. $E_y^{2\omega}(x)$ is the guided mode at a frequency of 2ω. The function $f(x)$ represents the variation in non-linear susceptibility throughout the waveguide. For example, if only the core material is non-linear, then $f(x)$ is a boxcar function. Equation (9.C.4) then takes the simplified form:

$$-\frac{\partial}{\partial z} A_y^{2\omega}(\mathrm{z}) = \frac{\mathrm{i}\omega\varepsilon_0}{p_0} \chi_{yyy}^{2\omega} S_{yy} (A_y^{\omega})^2 \mathrm{e}^{\mathrm{i}\delta\beta z} \tag{9.C.5}$$

where S_{yy} is the overlap integral of the modes E_y with themselves, given by:

$$S_{yy} = \int_{-\infty}^{+\infty} f(x) [E_y^{\omega}(x)]^2 E_y^{2\omega}(x) \mathrm{d}x \tag{9.C.6}$$

and $\delta\beta$ is the phase mismatch:

$$\delta\beta = \beta_{2\omega,\mathrm{TE}} - 2\beta_{\omega,\mathrm{TE}} = \frac{4\pi}{\lambda_0} [n_{\mathrm{eff}}(2\omega) - n_{\mathrm{eff}}(\omega)] \tag{9.C.7}$$

Equation (9.C.5) describes the energy transfer between the fundamental and the harmonic waves. To simplify the upcoming discussion, we will assume that the conversion efficiency is low. The amplitude $A_y^{\omega}(z)$ can then be assumed to be constant across the guide, and equal to its value at the guide entrance (A_0^{ω}). Equation (9.C.5) can then be integrated:

$$A_y^{2\omega}(L) = -\frac{\omega\varepsilon_0}{p_0}\chi_{yyy}^{2\omega}S_{yy}L(A_0^{\omega})^2\left[\frac{e^{i\delta\beta L}-1}{\delta\beta L}\right] \tag{9.C.8}$$

Recalling that the normalization conditions (9.20)–(9.22) stipulate that the square of the norm of the amplitude $|A|^2$ of the modes in the waveguide is the power injected into this mode per unit width in the guide, we obtain a conversion efficiency in the guide of:

$$\frac{P_y^{2\omega}(L)}{(P_0^{\omega})^2} = \frac{1}{4}(\omega\varepsilon_0\chi_{yyy}^{2\omega}S_{yy}L)^2\mathrm{sinc}^2\left(\frac{\delta\beta L}{2}\right) \tag{9.C.9}$$

In this last formula, the power levels P_y are in W per unit guide width. Let us observe for a moment the change in the conversion efficiency as a function of guide length L. The non-linear medium draws energy from the pump beam and transfers it into the second harmonic, with the inverse process occurring periodically along the length of the guide. As a result, the maximum efficiency that can be obtained occurs at the end of an interaction length, which is referred to (however improperly) as a *coherence length*, and which we shall call a *phase matching length*

$$\lambda_{\mathrm{pm}} = \frac{\lambda_0}{4[n_{\mathrm{eff}}(2\omega) - n_{\mathrm{eff}}(\omega)]} \tag{9.C.10}$$

Phase matching length in a waveguide

This length increases (as does the maximum efficiency) as the dispersion in the guide decreases. We will see later that certain strategies can be used to obtain phase matching artificially. In this case, the sinus-cardinal (*sinc*) function in (9.C.9) equals unity, and frequency conversion can take place constructively over the entire length of the guide leading to a parabolic dependence (in L^2) of the conversion efficiency.

We will now give a more physical expression for (9.C.9). To do so, we must calculate the overlap integral S_{yy}. We will make the following simplifying approximations:

- Only the core material is non-linear and the function $f(x)$ is a boxcar function.
- The guides are *infinitely confining* and the modes are sinusoidal as in (9.23), i.e:

$$E_y^{\omega}(x) = 2\sqrt{\frac{\omega\mu_0}{d\beta_{\omega}}p_0}\sin\frac{\pi x}{d} \tag{9.C.11}$$

The overlap integral can then be written:

$$S_{yy} = 8\sqrt{2}\left(\frac{\omega\mu_0 p_0}{d}\right)^{3/2} \frac{1}{\beta_\omega \beta_{2\omega}^{1/2}} \int_0^d \sin^3\left(\frac{\pi x}{d}\right) \mathrm{d}x$$

$$= \frac{32\sqrt{2}}{3\pi} \frac{1}{\sqrt{d}} \frac{(\omega\mu_0 p_0)^{3/2}}{\beta_\omega \beta_{2\omega}^{1/2}} \tag{9.C.12a}$$

or:

$$S_{yy}^2 = \left(\frac{32}{3\pi}\right)^2 \frac{1}{d} \frac{\mu_0^{3/2} p_0^3}{n_{\text{eff}}^3} \tag{9.C.12b}$$

supposing $n_{\text{eff}}(\omega) \approx n_{\text{eff}}(2\omega)$. Substituting this last equation into (9.C.9), we finally obtain the value for the frequency conversion efficiency for an infinitely confining non-linear waveguide without pump *depletion* (i.e. with $A^\omega = \text{constant}$):

$$\frac{P_y^{2\omega}(L)}{P_0^\omega} = \left(\frac{16}{3\pi}\right)^2 Z_0^3 \frac{(\omega\varepsilon_0 \chi_{yyy}^{2\omega} L)^2}{n_{\text{eff}}^3} \left(\frac{P_0^\omega}{dl}\right) \text{sinc}^2\left(\frac{\delta\beta L}{2}\right) \tag{9.C.13a}$$

Conversion efficiency for a non-linear waveguide

where Z_0 is the vacuum impedance of 377 Ω. Caution! In this last equation, the power levels P are in *watts* (and no longer in W m^{-1} of guide width), and l is then the width of the guide. In the phase matching case, the conversion efficiency is given by:

$$\frac{P_y^{2\omega}(L)}{P_0^\omega} = 4.3 \times 10^{-20} \left(\frac{L}{\lambda_0}\right)^2 \frac{(\chi_{yyy}^{2\omega})^2_{\text{pm V}^{-1}}}{n_{\text{eff}}^3} \left(\frac{P_0^\omega}{\text{d}l}\right)_{\text{W m}^{-2}} \tag{9.C.13b}$$

This equation reveals a significant barrier (of the order of 10^{-20} for typical $\chi^{2\omega}$s and power levels) to obtaining reasonable efficiencies and explains the attraction of waveguides for use in frequency conversion. Thanks to optical confinement in waveguides, electromagnetic fields can be significant even for relatively low incident power levels resulting in sizeable conversion efficiencies.

Example

Supposing that the phase matching problem has been resolved, we consider a GaAs waveguide ($\chi^2 \approx 100\,\text{pm V}^{-1}$ and $n_{\text{eff}} \approx 3$) 2 mm in length, 0.5 μm thick, and 2 μm wide. If we couple 10 mW of optical power at $\lambda_0 = 2\,\mu\text{m}$ into the guide, the conversion efficiency would be $4.3 \times 10^{-20}(2 \times 10^{+3}\,\mu\text{m}/2\mu\text{m})^2(10^2\,\text{pm V}^{-1})^2(10^{-2}\,\text{W}/(0.5 \times 10^{-6} \times 2 \times 10^{-6}\,\text{m}^2))/3^3$, i.e. 16%.

Quasi-phase matching

We saw that the phase matching conditions are fulfilled when the dispersion

between the pump wave at ω and the harmonic at 2ω is null (i.e. $\beta_\omega = \beta_{2\omega}$). We know that in semiconductors, however, the band structure induces a significant amount of dispersion, and that this dispersion is enhanced in guides (see Fig. 9.7). Phase matching, it would seem, is inherently impossible in a bulk semiconductor. It is, however, possible to make astute use of technology to modulate the properties of these materials spatially to achieve a state referred to as *quasi-phase matching*.

In a first intuitive approach, we will suppose that we can modulate the value of the non-linear susceptibility in a periodic fashion along the waveguide as:

$$\chi^{2\omega}_{yyy}(z) = \chi^{2\omega}_{yyy} \sin\left(\frac{2\pi}{\Lambda} z\right) \tag{9.C.14}$$

In this case, the coupled mode equation, (9.C.4), changes only slightly: the term in the exponential under the integral becomes $\mathrm{i}(2\omega t - (2\beta_{\omega,\mathrm{TE}} \pm 2\pi/\Lambda)z)$. All the theoretical derivations for the conversion efficiency remain as before except that the phase mismatch term becomes:

$$\delta\beta = \beta_{2\omega,\mathrm{TE}} - 2\beta_{\omega,\mathrm{TE}} \pm \frac{2\pi}{\Lambda} \tag{9.C.15}$$

The conditions leading to quasi-phase matching may then be written as:

$$\beta_{2\omega,\mathrm{TE}} - 2\beta_{\omega,\mathrm{TE}} = \frac{2\pi}{\Lambda} \tag{9.C.16}$$

Quasi-phase matching condition

Figure 9.C.2 offers an interpretation for this condition. The wavevector of the lattice ($2\pi/\Lambda$) adds to that of the pump wave leading to equivalence with the wavevector of the harmonic wave.

Technologically, this can be carried out by spatially destroying the optical non-linearity in a periodic fashion using ion implantation, growth techniques, impurity diffusion, etc. In this case, the non-linearity can be written:

$$\chi^{2\omega}_{yyy}(z) = \frac{\chi^{2\omega}_{yyy}}{2} + \sum_{n\,\mathrm{odd}} \frac{2\chi^{2\omega}_{yyy}}{n\pi} \sin\left(\frac{2n\pi}{\Lambda} z\right) \tag{9.C.17}$$

In the coupled mode equation, (9.C.4), only the term corresponding to the quasi-phase matched condition will yield a non-zero contribution:

$$\beta_{2\omega,\mathrm{TE}} - 2\beta_{\omega,\mathrm{TE}} = \frac{2n\pi}{\Lambda} \tag{9.C.18}$$

Thus, phase matching may be obtained for a corrugation length equal to Λ ($n = 1$, first order), 3Λ ($n = 3$, third order), etc. A very simple calculation shows that for

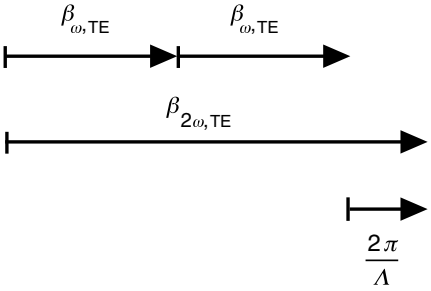


Fig. 9.C.2. Graphical interpretation of the quasi-phase matching condition in terms of wavevectors.

first-order phase matching the frequency conversion efficiency remains the same as that given by (9.C.13a) but with a slightly diminished effective non-linear susceptibility with respect to $\chi^{2\omega}$ given by:

$$\chi^{2\omega}_{\mathrm{eff}} = \frac{\chi^{2\omega}_{yyy}}{\pi} \tag{9.C.19}$$

Higher-order phase matching corrugations lead to even smaller effective non-linear susceptibilities.

Figure 9.C.3 illustrates the physical principle behind quasi-phase matching. Over successive intervals each spanning λ_{pm}, the non-linear dipoles are eradicated and the mechanism responsible for destructive interference along the waveguide is disabled.

9.C.2 TE mode in–TM mode out

In this case, non-linear polarization of the guide is written:

$$P^{2\omega}_{xyy}(x,z) = \varepsilon_0 \chi^{2\omega}_{xyy}[A^{\omega}_{y}(z)E^{\omega}_{y}(x)]^2 \tag{9.C.20}$$

This term is non-zero only if the non-linear tensor element is non-zero (determined by the symmetry of the semiconductor crystal). Figure 9.C.4 shows how non-linear interaction occurs. The calculation can be carried out exactly as in the preceding section allowing us to write:

$$\frac{P^{2\omega}_{\mathrm{TM}}(L)}{P^{\omega}_{\mathrm{TE}}} = \left(\frac{16}{3\pi}\right)^2 Z_0^3 \frac{(\omega\varepsilon_0\chi^{2\omega}_{xyy}L)^2}{n^3_{\mathrm{eff}}}\left(\frac{P^{\omega}_{\mathrm{TE}}}{dl}\right)\mathrm{sinc}^2\left(\frac{\delta\beta L}{2}\right) \tag{9.C.21}$$

This time the phase mismatch is given by:

$$\delta\beta = \beta_{2\omega,\mathrm{TM}} - 2\beta_{\omega,\mathrm{TE}} = \frac{4\pi}{\lambda_0}[n^{\mathrm{TM}}_{\mathrm{eff}}(2\omega) - n^{\mathrm{TE}}_{\mathrm{eff}}(\omega)] \tag{9.C.22}$$

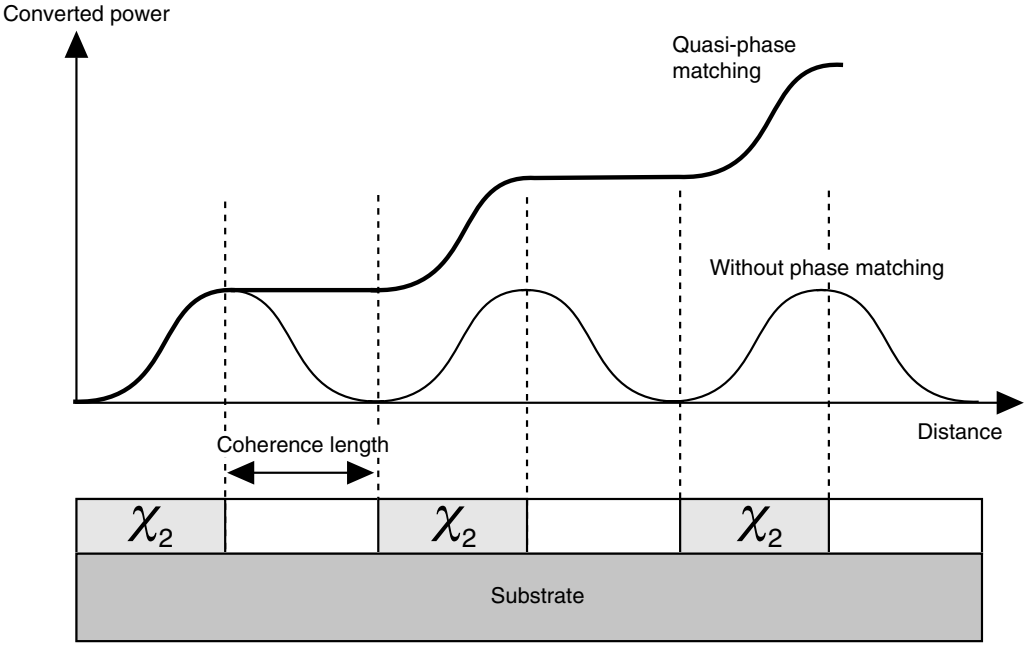


Fig. 9.C.3. Quasi-phase matching technique for optical frequency conversion in a non-linear waveguide. The optical non-linearity is alternately destroyed or left unchanged over successive intervals (with each of these spanning one coherence length).

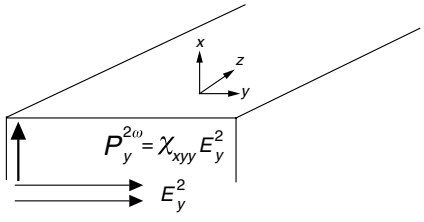


Fig. 9.C.4. Configuration for TE–TM second harmonic generation.

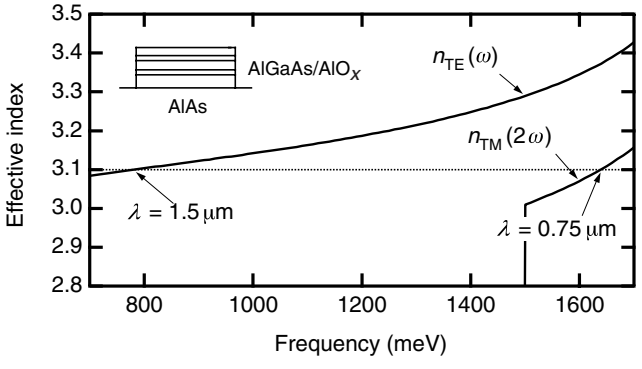


Fig. 9.C.5. Modal phase matching in a waveguide implemented using a GaAs/AlGaAs/Al_2O_3 heterostructure. (Courtesy of A. Fiore @LCR/THALES.)

Equation (9.C.22) shows another scenario based upon the birefringence in the guide. In this case, it is sufficient to achieve in the guide:

$$n_{\text{eff}}^{\text{TM}}(2\omega) = n_{\text{eff}}^{\text{TE}}(\omega) \tag{9.C.23}$$

Modal phase matching condition in a non-linear waveguide

Figure 9.C.5 shows such a phase matched waveguide implemented using a GaAs/AlGaAs/Al_2O_3 heterostructure.

FURTHER READING

A. Fiore, V. Berger, E. Rosencher, P. Bravetti, and J. Nagle, *Nature* **391**, 463 (1998).
S. Someckh and A. Yariv, *Appl. Phys. Lett.* **21**, 140 (1972).
J. P. van der Ziel, *Appl. Phys. Lett.* **26**, 60 (1975).

9.D Fabry–Pérot cavities and Bragg reflectors

The optical feedback provided by mirrors is an indispensable ingredient in establishing laser oscillations within an optical cavity. We saw that two conditions are required before a cavity can achieve laser threshold: one on the gain (4.21a) and another on the phase (4.24). This latter condition meant that the laser modes had to be eigenmodes of the optical cavity. Two types of mirrors can be used to define the cavity – *metallic mirrors* and *dielectric mirrors*. Dielectric mirrors are particularly important for semiconductor lasers – most notably in the case of *vertical cavity surface emitting lasers* (VCSELs). The formalism used in describing wave propagation in a laminar medium is identical to that describing wave propagation in a laminar guide. This complement is therefore well situated in a chapter devoted to waveguides.

We will first consider an electromagnetic wave $\omega = 2\pi\nu$ propagating in two media (1 and 2) which share a common interface at $z = 0$ (see Fig. 9.D.1). The wave amplitudes in each of the media are given by the real part of:

$$E_i(z) = E_{Ri}\mathrm{e}^{-\mathrm{i}k_i z} + E_{Li}\mathrm{e}^{+\mathrm{i}k_i z} \tag{9.D.1}$$

The real waves are certainly of the form $\mathrm{Re}(E_i(z)\mathrm{e}^{\mathrm{i}\omega t})$. The index R indicates that the amplitude corresponds to the wave moving towards the right ($\mathrm{e}^{\mathrm{i}(\omega t - kz)}$), whereas the index L corresponds to a wave moving towards the left ($\mathrm{e}^{\mathrm{i}(\omega t + kz)}$). We then write the *Fresnel equations*, which state that the electric and magnetic fields must be continuous at the interfaces. As $\mathbf{B} = \mathbf{k} \times \mathbf{E}/\omega$ (see (2.17a–d)) and $k_i = 2\pi n_i/\lambda_0$, where λ_0 is the wavelength in the vacuum and n_i is the optical index in medium i, these two conditions may be written at $z = 0$:

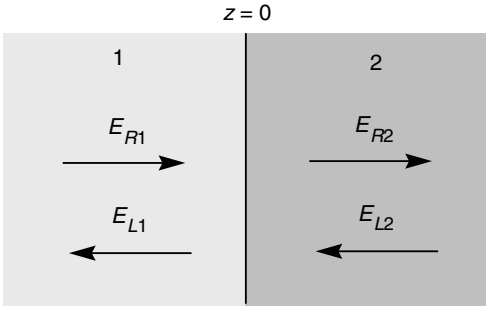


Fig. 9.D.1. Wave propagation across an interface.

$$E_{R1} + E_{L1} = E_{R2} + E_{L2}$$
$$k_1 E_{R1} - k_1 E_{L1} = k_2 E_{R2} - k_2 E_{L2} \tag{9.D.2}$$

We will assume that the wave is incident from the left, i.e. that $E_{L2} = 0$, and the system (9.D.2) can be immediately solved to obtain:

$$E_{R2} = \tau_{12} E_{R1}$$
$$E_{L1} = \rho_{12} E_{R1} \tag{9.D.3}$$

with:

$$\tau_{12} = \frac{2n_1}{n_1 + n_2}$$
$$\rho_{12} = \frac{n_1 - n_2}{n_1 + n_2} \tag{9.D.4}$$

The interpretation of (9.D.3) is very simple. τ_{12} is the transmission coefficient of the wave from medium 1 to medium 2, whereas ρ_{12} is the reflection coefficient of the wave of medium 1 off medium 2. We note several important points. First, $\tau_{12} \neq \tau_{21}$, whereas $\rho_{12} = -\rho_{21}$. Also, we notice immediately that:

$$\tau_{12}\tau_{21} + \rho_{12}^2 = 1 \tag{9.D.5}$$

which is none other than a statement of the conservation of the Poynting vector flux across the interface.

We will now generalize our notation to the description of two interfaces $i - 1$, and i present in a multilayered dielectric stack with the interfaces being located at z_i. We will label as $E_{R,i}$ and $E_{L,i}$ the amplitudes just to the right of the interface $i - 1/i$, and $E'_{R,i}$ and $E'_{L,i}$ the amplitudes just to the left of the $i/i + 1$ interface (see Fig. 9.D.2). The continuity equations, (9.D.2), can be placed in matrix form:

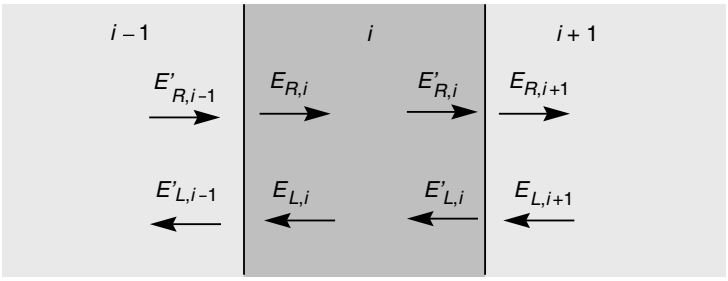


Fig. 9.D.2. Formalism for the propagation matrix *S*.

$$\begin{bmatrix} E'_{L,i-1} \\ E'_{R,i-1} \end{bmatrix} = P_{i-1,i} \begin{bmatrix} E_{L,i} \\ E_{R,i} \end{bmatrix} \tag{9.D.6}$$

with the *transfer matrix P* given by:

$$P_{i-1,i} = \frac{1}{\tau_{i-1,i}} \begin{bmatrix} 1 & \rho_{i-1,i} \\ \rho_{i-1,i} & 1 \end{bmatrix} \tag{9.D.7}$$

We remark that, given (9.D.4) and (9.D.5), the determinant of $P_{i-1,i} = n_i/n_{i-1}$. The wave propagation in the ith layer is taken into account by using the propagation constant k_i or:

$$\begin{bmatrix} E'_{L,i} \\ E'_{R,i} \end{bmatrix} = D_i \begin{bmatrix} E_{L,i} \\ E_{R,i} \end{bmatrix} \tag{9.D.8}$$

where D_i is the *dephasing matrix*:

$$D_i = \begin{bmatrix} \mathrm{e}^{-\mathrm{i}\phi_i} & 0 \\ 0 & \mathrm{e}^{+\mathrm{i}\phi_i} \end{bmatrix} \tag{9.D.9}$$

and where $\phi_i = k_i l_i = 2\pi n_i l_i/\lambda_0$ is the dephasing of the wave in the ith layer with thickness l_i.

Therefore, for a series of N arbitrary dielectric layers, the electric fields in media 1 and N are related through:

$$[E'_1] = P_{12} D_2 P_{23} \cdots D_{N-1} P_{N-1,N} [E_N] = S_{1,N} [E_N] \tag{9.D.10}$$

Propagation matrix *S*

The matrix $S_{1,N}$ is the propagation matrix. This representation is also referred to as the *S matrix* formalism. As it is a product of a series of matrices, the determinant of the *S* matrix is the product of their determinants and is found to be n_N/n_1. The determinant is therefore independent of the different media separating medium 1 from medium N. The notation in (9.D.10) is also interesting as it allows one to 'propagate' the boundary conditions through the structure. This formalism is identical with that developed in Chapter 1 to obtain the eigenstates for a quantum well.

If we seek the transmission coefficient τ_{1N} and the reflection coefficient ρ_{1N} for the entire structure, as well as the transmittance T_{1N} and the *reflectance* R_{1N}, we need only impose the boundary condition $E_{L,N} = 0$, which signifies that no other wave comes from the left, which leads to:

$$\rho_{1N} = \frac{S_{12}}{S_{22}} \text{ and } R_{1N} = |\rho_{1N}|^2 \tag{9.D.11}$$

$$\tau_{1N} = \frac{1}{S_{22}} \text{ and } T_{1N} = \frac{n_N}{n_1}|\tau_{1N}|^2 \tag{9.D.12}$$

We note that $R_{1N} + T_{1N} = (|S_{12}|^2 + n_N/n_1)/|S_{22}|^2 = 1$, since $\det(S) = n_N/n_1$. This last equality specifies that light flux is conserved, and this is rather reassuring!

9.D.1 The Fabry–Pérot cavity

We will now apply this formalism to the simple problem involving a Fabry–Pérot cavity (see Fig. 9.D.3). It consists of a dielectric film of thickness l_{cav} coated with two thin metallic layers. Rigorously, this system is composed of five media (vacuum → metal → dielectric → metal → vacuum). This problem could be solved in its entirety by introducing a complex index of refraction for the metal and by making its real portion tend towards $-\infty$. This will be left as an exercise to the reader. Rather, we will consider the system to consist of three media (vacuum = 1 → dielectric = 2 → vacuum = 3) with the metallic layers being taken into account by the transfer matrices P_{12} and P_{21} given by:

$$P_{12} = \frac{1}{\tau_{12}}\begin{bmatrix} 1 & \rho_0 e^{i\theta} \\ \rho_0 e^{i\theta} & 1 \end{bmatrix} \text{ and } P_{21} = \frac{1}{\tau_{21}}\begin{bmatrix} 1 & \rho_0 e^{-i\theta} \\ \rho_0 e^{-i\theta} & 1 \end{bmatrix} \tag{9.D.13}$$

One of the numerous properties of metallic films that the reader can derive as an exercise is that (contrary to the case of an interface between dielectrics) the coefficients τ_{12} and τ_{21} are equal. The term $e^{i\theta}$ represents the dephasing contributed by the metallic mirror ($\theta = \pi$ in the case of a perfect mirror) and the reflection coefficient ρ_0 is close to unity and takes into account the high reflectivity of the mirrors. The propagation matrix $S_{13} = P_{12}D_2P_{23}$ can be easily calculated:

$$S_{13} = \frac{1}{\tau_{12}\tau_{21}}\begin{bmatrix} e^{-i\phi} + \rho_0^2 e^{i\phi} & \rho_0[e^{-i(\phi-\theta)} + e^{i(\phi-\theta)}] \\ \rho_0[e^{-i(\phi-\theta)} + e^{i(\phi-\theta)}] & e^{i\phi} + \rho_0^2 e^{-i\phi} \end{bmatrix} \tag{9.D.14}$$

which leads to the *transmittance* T_{FP} for the Fabry–Pérot cavity (see (9.D.12)):

$$T_{\text{FP}} = \frac{|\tau_{12}\tau_{21}|^2}{|1 - \rho_0^2 e^{-2i(\phi-\theta)}|^2} \tag{9.D.15}$$

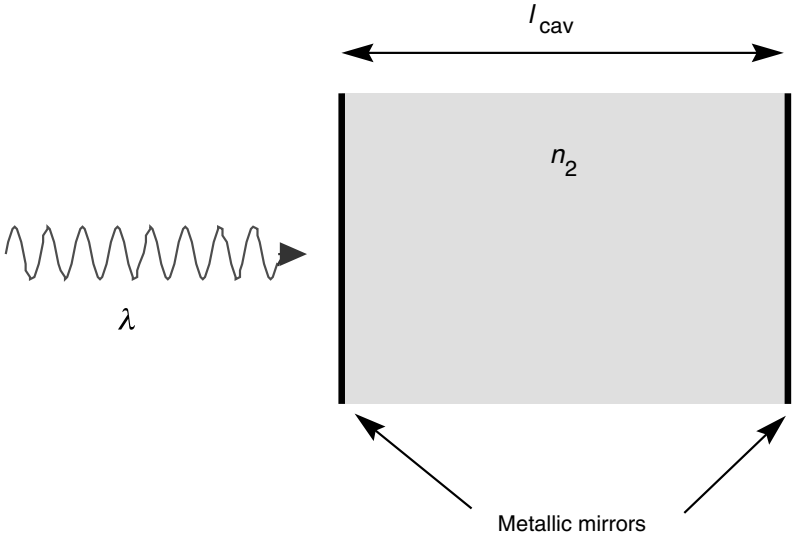


Fig. 9.D.3. A Fabry–Pérot cavity.

We thus introduce the total dephasing coefficient $\delta = 2(\phi - \theta)$, the mirror *transmittance* ($T = |\tau_{12}|^2$), and *reflectance* ($R = \rho_0^2$). The total cavity transmittance, i.e. the portion of the light intensity incident from the left which crosses the cavity, is therefore:

$$T_{FP} = \frac{T^2}{1 + R^2 - 2R\cos\delta} \tag{9.D.16}$$

where we recall that phase δ is given by:

$$\delta = 2\left(\frac{2\pi n_{op} l_{cav}}{\lambda_0} - \theta\right) \tag{9.D.17}$$

When the wavelength of the incident light upon the cavity is varied, the transmittance oscillates between the minimum T_{min} and maximum T_{max} values given by (9.D.16):

$$T_{min} = \frac{T^2}{(1 + R)^2}$$

$$T_{max} = \frac{T^2}{(1 - R)^2} \tag{9.D.18}$$

If the mirrors were perfect ($R + T + A = 1$ with the mirror absorption coefficient $A = 0$), we would have $T_{max} = 1$, i.e. all the energy would be transmitted across the cavity. This might seem surprising as if one approaches two mirrors possessing a weak transmittance T, the total transmittance of the combined mirrors (when

separated by the proper distance) turns out to be 1 instead of the T^2 we might have simply expected! This results from a constructive interference effect typical of wave phenomena. A similar phenomenon exists in the case of *resonant tunnelling* of electrons across double quantum barriers.

Substituting (9.D.18) into the expression for cavity transmittance (9.D.16), we obtain:

$$T_{\mathrm{FP}} = T_{\max}\frac{1}{1 + C\sin^2(\delta/2)} \tag{9.D.19}$$

where C is the cavity *contrast* given by:

$$C = \frac{T_{\max}}{T_{\min}} - 1 = \frac{4R}{(1-R)^2} \tag{9.D.20}$$

Near the cavity resonance ($\delta \approx 2m\pi$, where m is an integer), (9.D.19) takes the form of a Lorentzian:

$$\begin{aligned} T_{\mathrm{FP}} &\approx T_{\max}\frac{1}{1 + (2F/\pi)^2\left[\left(2n_{\mathrm{op}}\dfrac{l_{\mathrm{cav}}}{\lambda_0} - m\right)\pi - \theta\right]^2} \\ &\approx T_{\max}\frac{1}{1 + (2F/\pi)^2\left[\left(\dfrac{\nu}{\Delta\nu} - m\right)\pi - \theta\right]^2} \end{aligned} \tag{9.D.21}$$

Fabry–Pérot cavity transmittance

where $\Delta\nu$ is the separation between the optical modes and F is the *finesse* of the cavity given by:

$$F = \pi\frac{\sqrt{R}}{1-R} \tag{9.D.22}$$

Cavity finesse

which is of the order of π/T for $R \approx 1$. Figure 9.D.4 shows T_{FP} as a function of light frequency. The cavity displays transmission maxima separated by:

$$\Delta\nu = \frac{c}{2n_{\mathrm{op}}l_{\mathrm{cav}}} \tag{9.D.23}$$

which is the cavity mode spacing (or *free spectral range*) with a full width at half maximum given by:

$$\delta\nu = \frac{\Delta\nu}{F} \tag{9.D.24}$$

from which the quantity F derives its interpretation as a measure of *finesse*. For a

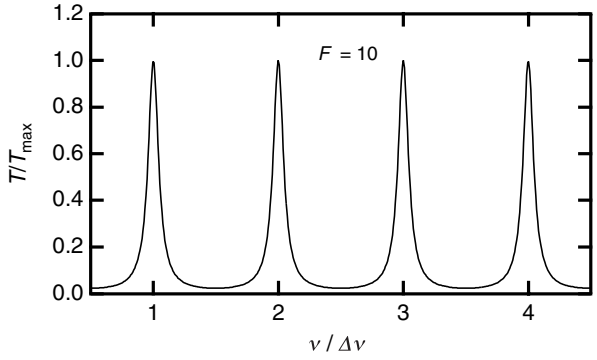


Fig. 9.D.4. Transmission spectrum for a Fabry–Pérot cavity with a finesse, $F = 10$.

perfect mirror ($\theta = \pi$), the cavity resonances occur for cavity thicknesses of:

$$l_{\text{cav}} = m\frac{\lambda_0}{2n_{\text{op}}} \tag{9.D.25}$$

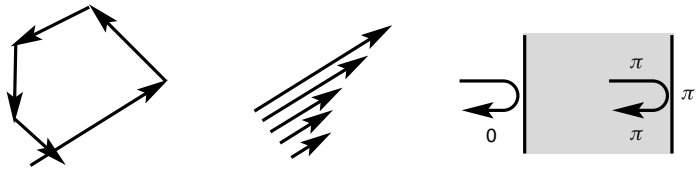
where m is an integer. A resonant Fabry–Pérot cavity is therefore a *half-wave* type cavity. An interpretation of Fabry–Pérot resonance is given in Fig. 9.D.5. In this diagram, the magnitudes and relative phases of the various propagating electric fields are represented as Fresnel vectors, or *phasors*. Off resonance, the vectors corresponding to the reflected waves add together (in a circular fashion) to zero. Everything takes place as though the two metallic mirrors behaved independently of one another, with each mirror possessing a transmittance T and the total transmittance for the pair being T^2.

On the other hand, as the dielectric film thickness approaches the 'half-wave' criterion, the phase difference between the individual phasors are multiples of 2π (i.e. they are aligned), and they sum (interfere) constructively.

We will now calculate the complex amplitude of the electric field inside a cavity assuming perfect mirrors. This is obtained by using $[E_2] = P_{23}[E_3]$ along with (9.D.1), i.e:

$$\frac{E_2(z)}{E_{R1}} = \frac{\tau_{12}}{1 - \rho_0^2 \mathrm{e}^{-2\mathrm{i}\phi}}[\mathrm{e}^{-\mathrm{i}k_2 z} - \rho_0 \mathrm{e}^{+\mathrm{i}k_2 z}\mathrm{e}^{-2\mathrm{i}\phi}] \tag{9.D.26a}$$

The field *amplitude* inside the cavity is the norm of (9.D.26a) (this is not a trivial result and is left to the reader as an exercise). Figure 9.D.6 shows the distribution of the electric field inside the cavity. We notice a resonance effect which acts to increase the maximum field strength inside the cavity. We can easily find the maximum field strength by supposing that at resonance (i.e. $\phi = \pi$ and $\rho_0 \approx 1$), the electric field (9.D.26a) approaches a sinusoidal stationary wave which drops to zero at the mirrors at $z = 0$ and $z = l_{\text{cav}}$, with $k_2 = \pi/l_{\text{cav}}$. The maximum field



(*a*) (*b*) (*c*)

Fig. 9.D.5. (*a*) When the Fabry–Pérot cavity is off resonance, the phasors add together in a circular fashion. (*b*) When the cavity approaches resonance, the phase difference between each phasor is a multiple of 2π and the individual phasors align with one another to produce a transmittance which approaches unity. (*c*) This latter condition occurs in a half-wave cavity.

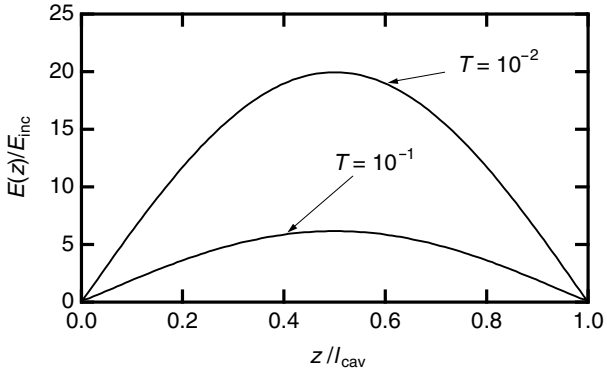


Fig. 9.D.6. Increase in the internal cavity field resulting from a decreased transmission coefficient T.

strength then occurs when $z = l_{\text{cav}}/2$, or:

$$\frac{E_{2,\max}}{E_{R1}} = \frac{\tau}{1-\rho_0^2}(\mathrm{e}^{-\mathrm{i}(k_2 l_{\text{cav}}/2)} - \rho_0 \mathrm{e}^{\mathrm{i}(k_2 l_{\text{cav}}/2)}) = \frac{\tau}{1-\rho_0} \tag{9.D.26b}$$

i.e. taking (9.D.5) and (9.D.21) into account:

$$\frac{E_{2,\max}}{E_{R1}} = 2\sqrt{\frac{F}{\pi}} = \frac{2}{\sqrt{T}} \tag{9.D.26c}$$

The light intensity in the middle of the cavity is therefore enhanced by a factor $4/T$ compared with the incoming flux; the factor $1/T$ stemming from the mirror reflectivity and the factor 4 from the constructive interference of the two counter-propagating waves. The energy stored in the cavity is obtained by integrating over the sinusoidal distribution, i.e:

$$\mathscr{E}_{\text{cav}} \approx \frac{1}{2}\varepsilon_0 n_2^2 S \left| \frac{2}{\sqrt{T}} E_{R1} \right|^2 \int_0^{l_{\text{cav}}} \sin^2 k_2 z \mathrm{d}z = \varepsilon_0 n_2^2 S l_{\text{cav}} \frac{|E_{R1}|^2}{T} \tag{9.D.27}$$

where S is the surface of the cavity. This is to be compared with the energy lost per cycle, which corresponds to the integrated photon flux per cycle. The ratio between the two quantities, the cavity *quality factor* Q, given (9.D.27) and (9.D.23) may be written as:

$$Q = \frac{\text{energy stored}}{\text{energy lost per cycle}} = F\frac{v}{\Delta v} \tag{9.D.28}$$

Quality factor and finesse of a cavity for a wave of frequency v

This quantity relates to the photon lifetime through:

$$Q = 2\pi v \tau_c \tag{9.D.29}$$

which fixes the relationship between τ_c and F.

Example

1. We consider a Fabry–Pérot cavity assembled using a dielectric rod 50 cm long with an index of refraction of 1.5. Metallic mirrors are deposited at either end possessing reflectances $R = 99.9\%$ and transmittances $T = 1 - R = 10^{-3}$. The cavity finesse is then $F = \pi R/T = 3140$. The mode spacing is $3 \times 10^8\,\text{m s}^{-1}/(3 \times 0.5\,\text{m})$ or $2 \times 10^8\,\text{Hz}$. For a wavelength of $1\,\mu\text{m}$ ($v = 3 \times 10^{14}\,\text{Hz}$), the quality factor $Q = 3.14 \times 10^3 \times 3 \times 10^{14}\,\text{Hz}/2 \times 10^8\,\text{Hz}$, or $Q = 5 \times 10^9$! Such values are in reality difficult to obtain as it is difficult to achieve the degree of parallelism required between both mirrors given the large separation distance.
2. This time we consider a cavity implemented using a $0.25\,\mu\text{m}$ thick layer of AlGaAs with $n = 3$. The wavelength of the resonant wave in this cavity is then $\lambda_0 = 2n_{\text{op}}l_{\text{cav}}$ or $1.5\,\mu\text{m}$. Two metallic mirrors are deposited on either side of the AlGaAs layer possessing a reflectance $R = 99.9\%$ and a transmittance $T = 1 - R = 10^{-3}$. The cavity finesse is then as obtained in the preceding example, $F = \pi R/T = 3140$. This time, the mode spacing is however, $3 \times 10^8\,\text{m s}^{-1}/(2 \times 3 \times 0.25 \times 10^{-6}\,\text{m})$, or $1.5 \times 10^{14}\,\text{Hz}$. The quality factor is then $3.14 \times 10^3 \times 2 \times 10^{14}\,\text{Hz}/1.5 \times 10^{14}\,\text{Hz}$, or $Q = 4200$.

9.D.2 Bragg mirrors

We now consider a succession of bilayers formed by alternating layers of media 1 and 2. We will interest ourselves for the time being in the propagation matrix S defined between the right edges of medium 1 over two successive bilayers (see Fig. 9.D.7).

$$S = P_{12}D_2P_{21}D_1 \tag{9.D.30}$$

with:

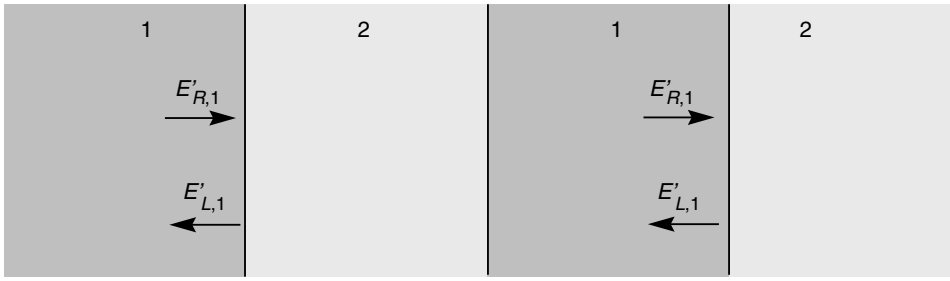


Fig. 9.D.7. Geometry and the associated elementary propagation matrix for a Bragg mirror.

$$P_{12} = \frac{1}{\tau_{12}}\begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix};\ P_{21} = \frac{1}{\tau_{21}}\begin{bmatrix} 1 & -\rho_{12} \\ -\rho_{12} & 1 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} e^{-i\phi_1} & 0 \\ 0 & e^{i\phi_1} \end{bmatrix};\ D_2 = \begin{bmatrix} e^{-i\phi_2} & 0 \\ 0 & e^{i\phi_2} \end{bmatrix} \tag{9.D.31}$$

where τ_{12}, τ_{21}, and $\rho_{12} = -\rho_{21}$ are as defined in (9.D.4). The S matrix is then easily found to be:

$$S = \frac{1}{\tau^2}\begin{bmatrix} e^{-i\phi_1}(e^{-i\phi_2} - \rho^2 e^{i\phi_2}) & 2i \sin\phi_2 \rho e^{-i\phi_1} \\ -2i \sin\phi_2 \rho e^{i\phi_1} & e^{i\phi_1}(e^{i\phi_2} - \rho^2 e^{-i\phi_2}) \end{bmatrix} \tag{9.D.32}$$

with $\tau^2 = \tau_{12}\tau_{21}$. This S matrix takes a particularly simple form when layers 1 and 2 are of the 'quarter-wave' type, i.e. when $\phi_1 = \phi_2 = \pi/2$ or alternately $l_{\text{cav},i} = m\pi\lambda/4n_i$, where m is an integer:

$$S = -\frac{1}{\tau^2}\begin{bmatrix} 1 + \rho^2 & 2\rho \\ 2\rho & 1 + \rho^2 \end{bmatrix} \tag{9.D.33}$$

This matrix possesses the eigenvalues:

$$X_+ = (1 + \rho)^2$$

$$X_- = (1 - \rho)^2 \tag{9.D.34}$$

and eigenvectors:

$$V_+ = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } V_- = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{9.D.35}$$

which readily leads to the diagonalization of (9.D.33):

$$S = T^{-1}MT, \text{ with } M = \frac{1}{\tau^4}\begin{bmatrix} (1 + \rho)^2 & 0 \\ 0 & (1 - \rho)^2 \end{bmatrix} \text{ and } T = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \tag{9.D.36}$$

A series of N successive bilayers then yields a propagation matrix given by:

$$S^N = T^{-1}M^N T = \frac{1}{\tau^{4N}} T^{-1} \begin{bmatrix} (1+\rho)^{2N} & 0 \\ 0 & (1-\rho)^{2N} \end{bmatrix} T \tag{9.D.37}$$

with a reflectance given by (9.D.12):

$$R_N = \left| \frac{(1+\rho)^{2N} - (1-\rho)^{2N}}{(1+\rho)^{2N} + (1-\rho)^{2N}} \right|^2 \tag{9.D.38}$$

or, given (9.D.4):

$$R_N = \left[\frac{1 - \left(\frac{n_2}{n_1}\right)^{2N}}{1 + \left(\frac{n_2}{n_1}\right)^{2N}} \right]^2 \tag{9.D.39}$$

Reflectance for a sequence of *N* dielectric quarter-wave bilayers

where we have arbitrarily taken $n_1 > n_2$. We see that as the number of layers grows in size, the reflectance of the system tends towards 1 resulting in a *Bragg mirror*.

We are well within our rights to imagine that such a Bragg reflection constitutes a highly resonant phenomenon, and that the conditions on wavelength for obtaining $R = 1$ must have to be Draconian at best. Nothing in fact could be further from the truth! In fact, forbidden bands (or stop bands) for light propagation form in complete analogy with the band structure results obtained in Chapter 5 for electrons in crystalline solids.

We may generalize the approach used above to describe conditions outside of resonance (albeit a considerable task). We will only indicate the number of required bilayers N to obtain a *stop band* $\Delta\lambda$ such that $1 - R < 10^{-n}$:

$$\frac{\Delta\lambda}{\lambda} = \left[\frac{8 \ln 10}{\pi^2} \frac{n_1 - n_2}{n_1 + n_2} \left(2 \operatorname{Log} \frac{n_1}{n_2} - \frac{n + \operatorname{Log} 4}{N} \right) \right]^{1/2} \tag{9.D.40}$$

When the number of bilayers N tends towards infinity, the stop band tends towards the limit given by:

$$\frac{\Delta\lambda}{\lambda} = \frac{4}{\pi} \left(\frac{n_1 - n_2}{n_1 + n_2} \ln \frac{n_1}{n_2} \right)^{1/2} \tag{9.D.41}$$

Width of a *stop band*

Example

We wish to create a Bragg mirror out of GaAs/AlAs bilayers, which reflects at a peak wavelength of $\lambda = 1\ \mu\text{m}$. The optical indices for GaAs and AlAs are 3.30 and 2.9, respectively. The '$\lambda/4$' thicknesses required are then $l_{\text{GaAs}} = 757\ \text{Å}$ and

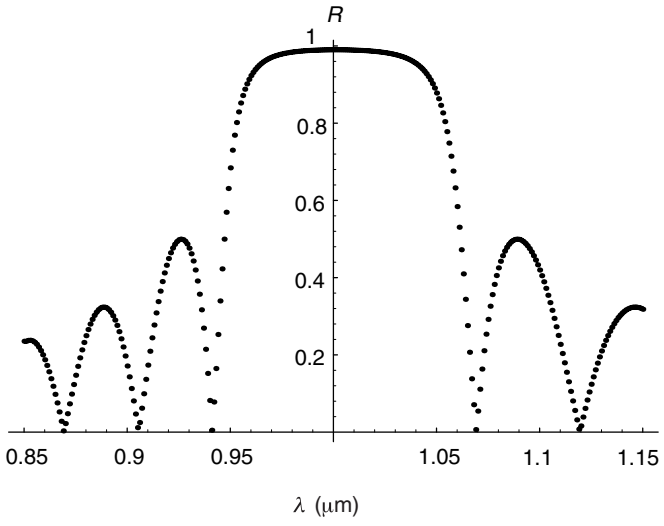


Fig. 9.D.8. Reflectance spectrum for the Bragg mirror specified in the example.

$l_{\mathrm{AlAs}} = 862\,\text{Å}$ for the two materials. For a series of 20 bilayers, the mirror reflectivity will be:

$$R_{15} = [(1 - 0.878^{40})/(1 + 0.878^{40})]^2 = 97.8\%$$

This would require a total epitaxial deposition of $20 \times 1578\,\text{Å}$, or $3.1\,\mu\text{m}$. Assuming typically epitaxial deposition rates of $\approx 1\,\mu\text{m}\,\text{h}^{-1}$, growing such a layer would require a little over 3 hours.

The width of the stop band that would be obtained is given by (9.D.41) to be:

$$\Delta\lambda/\lambda = 4/\pi[0.44/6.36\ln(3.40/2.96)]^{1/2}, \text{ or } \Delta\lambda = 0.116\,\mu\text{m}.$$

Figure 9.D.8 shows the reflectance spectrum $R(\lambda)$ calculated using the MATHEMATICA program below for this 20 bilayer mirror. We see that the width of the stop band is close to the value calculated above.

```
x=.;l=1. (*μm*)
n1=3.3;n2=2.9;l1=l/(4*n1);l2=l/(4*n2);
r12=(n1-n2)/(n1+n2);t12=2n1/(n1+n2);
r21=(n2-n1)/(n1+n2);t21=2n2/(n1+n2);
A=Array[a,{2,2}];a[1,1]=1.;a[1,2]=r12;a[2,1]=r12;a[2,2]=1.;M12=A/t12;
B=Array[b,{2,2}];b[1,1]=1.;b[1,2]=r21;b[2,1]=r21;b[2,2]=1.;M21=B/t21;
R = Table[{x,D1=Array[d1,{2,2}];D2=Array[d2,{2,2}];
d1[1,2]=0;d1[2,1]=0;d1[1,1]=N[Exp[-I*2*Pi*n1*l1/x]];d1[2,2]=N[Exp[I*2*Pi*n1*l1/x]];
d2[1,2]=0;d2[2,1]=0;d2[1,1]=N[Exp[-I*2*Pi*n2*l2/x]];d2[2,2]=N[Exp[I*2*Pi*n2*l2/x]];
S = D2.M21.D1.M12;Sp=MatrixPower[S,20];Spp= M12.Sp;
```

```
r=Abs[Spp[[1,2]]/Spp[[2,2]]]},{x,.85,1.15,.001}];
ListPlot[R]
```

FURTHER READING

G. Bruhat, *Optique*, 6th Edn, Masson, Paris (1992).
M. V. Klein and T. E. Furtak, *Optics*, John Wiley, New York (1986).
P. Yeh, *Optical Waves in Layered Media*, Wiley Interscience, New York (1988).

10 Elements of device physics

10.1 Introduction

Semiconductors are materials that are extremely sensitive to external perturbations (e.g. illumination, electric fields, thermal gradients, . . .). This characteristic is put to use in a variety of semiconductor devices. The typical response of a semiconductor to such a perturbation will be to screen it out. Up to this point, we have derived two *characteristic screening lengths*. One of these was the *depletion length* (6.68):

$$L = \sqrt{\frac{2\varepsilon V}{eN_D}} \tag{10.1}$$

which measures the distance over which a potential V is screened by a fixed space charge of density N_D (typically consisting of the ionized dopant atoms) and the *Debye length* (6.25):

$$l_D = \sqrt{\frac{\varepsilon k_B T}{e^2 n_0}} \tag{10.2}$$

which measures the distance over which a potential is screened by mobile charges with density n_0. We recall that e is the electron charge, ε is the permittivity of the semiconductor, k_B is Boltzmann's constant, and T the temperature. The range of accessible doping concentrations in semiconductors is limited at the low end by the background density of electrically active impurities introduced during sample growth and processing (for Si and GaAs these lower limits are nowadays $\sim 10^8\,\mathrm{cm}^{-3}$ and $\sim 10^{14}\,\mathrm{cm}^{-3}$, respectively). The upper doping limits are set by the metallic threshold for $N_D > N_c$ or N_v depending on whether the material is n or p type (see Chapter 5). We see that these screening length scales range from 0.1 to 10 μm for doping levels which vary between $10^{18}\,\mathrm{cm}^{-3}$ and $10^{14}\,\mathrm{cm}^{-3}$, respectively. This lower limit of 0.1 μm determines the typical dimensions for *microelectronic components*. In this chapter, we will study the electronic behaviour of a few structures which, in addition to their relevance to optoelectronics, form the basic building blocks of modern electronics.

10.2 Surface phenomena

At the surface of a semiconductor, the periodicity of the crystal potential is broken. Each atom at the surface is bound on the semiconductor side to underlying atoms within the bulk, whereas the electronic wavefunctions on the vacuum side of the interface are unterminated and free to spill out into empty space. Such unterminated *surface states* are referred to as *dangling bonds*. These two phenomena lead to the formation of electronic states within the forbidden gap of the bulk material near the surface (see Complement 5.D). In fact, these localized surface states are distributed in a continuous fashion across the gap and form a continuum within the bandgap. To characterize this continuum, we introduce a characteristic level ϕ_0 above the valence band which possesses the following property: when the surface states are occupied up to the level ϕ_0, the surface is electrically neutral. In the language of semiconductors, the states below ϕ_0 behave as donors (as they are neutral when occupied by electrons) whereas the states above ϕ_0 behave as acceptors (as they are negative when occupied by electrons). In practice, the surface is generally exposed to the ambient atmosphere and it is highly probable that impurities adsorbed at the surface will contribute to the population of localized states within the gap. This contribution can, however, be incorporated into the definition of ϕ_0.

At the surface of an *n*-type semiconductor, for example, the Fermi level is determined by the dopant concentration inside the material. At thermodynamic equilibrium, the Fermi level is constant throughout the structure. At the surface, the surface states must be occupied up to the Fermi level thereby creating a surface charge:

$$\sigma = -eN_{ss}[E_F - \phi_0 - E_v(0)] \tag{10.3}$$

where $N_{ss}\,(\mathrm{cm}^{-2}\,\mathrm{eV}^{-1})$ is the density of surface states (for simplicity assumed constant in the gap here) and $E_v(0)$ is the position of the valence band at the surface (see Fig. 10.1).

As the system is globally neutral, this surface charge must be compensated by a space charge in the semiconductor distributed over the depletion region (see Fig. 10.1). In contrast to the surface charge (which has a thickness of the order of an atomic monolayer), the depletion depth L is macroscopic. The drop in the electrostatic potential in this space charge region was described in Chapter 6 and is given by Eq. (6.69) to be:

$$V_d = \frac{1}{2}\frac{e}{\varepsilon}N_D L^2 = \frac{1}{2\varepsilon}\sigma L \tag{10.4}$$

where σ is the charge in the depletion layer. The position of the conduction band at

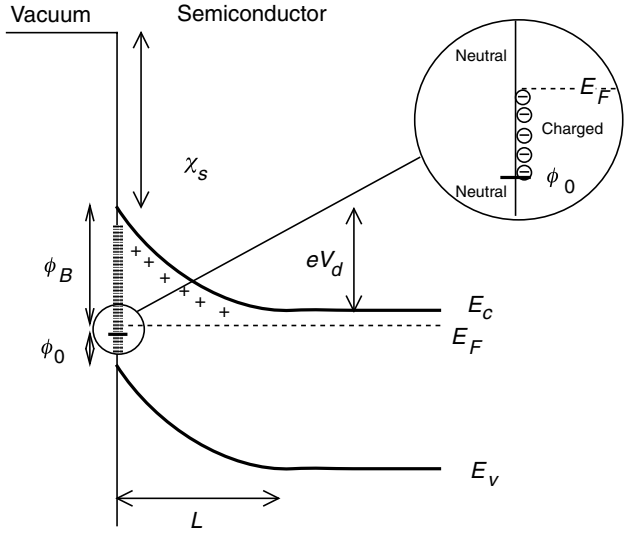


Fig. 10.1. The surface states between the Fermi level and ϕ_0 create a surface charge which is compensated by the space charge in the depletion region.

the surface is therefore $E_c(0) = E_c(\infty) + eV_d$, and represents an internal potential barrier against electron escape from the bulk semiconductor. The potential drop V_d is obtained by substituting (10.4) into (10.3):

$$V_d = \frac{1}{2}\frac{e}{\varepsilon} L N_{ss}[E_F - \phi_0 - eV_d - E_v(\infty)] \tag{10.5}$$

or

$$V_d\left(1 + \frac{1}{2}\frac{e^2}{\varepsilon} L N_{ss}\right) = \frac{1}{2}\frac{e}{\varepsilon} L N_{ss}[E_F - \phi_0 - E_v(\infty)] \tag{10.6}$$

In the limit of very large surface state densities, V_d no longer depends on N_{ss}. The Fermi level at the surface therefore remains pinned at $E_v(0) + \phi_0$, and the position of the Fermi level is determined by:

$$E_c(0) - E_F = eV_d + E_c(\infty) - E_F = E_g - \phi_0 \tag{10.7}$$

and the thickness of the depletion region is equal to the corresponding depletion length for a potential V_d.

We will now analyse the electronic behaviour of a metal layer deposited on a semiconductor. First we note that before an electron can escape the semiconductor and move into the vacuum, an electron at the bottom of the conduction band at the surface must overcome a barrier χ_s. This is the *electronic affinity of the semiconductor*, and it represents the energy which holds the electrons in the material (see Fig. 10.2).

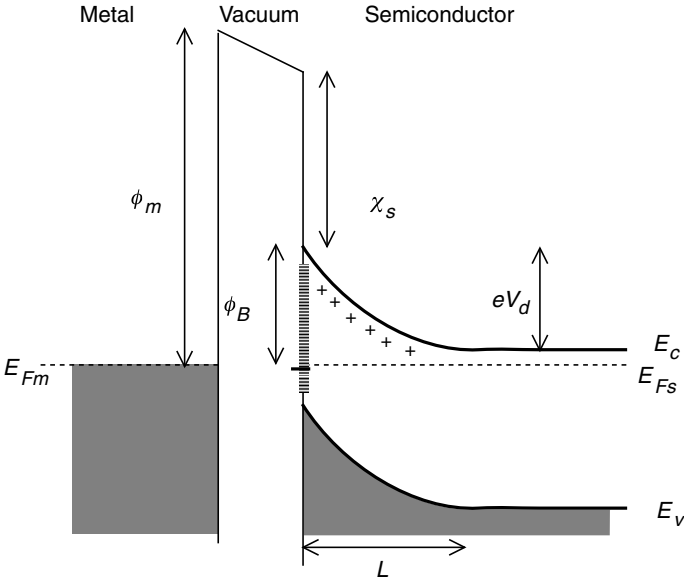


Fig. 10.2. Between a metal and a semiconductor in equilibrium, a transfer of charge takes place allowing the Fermi levels to become aligned. The distribution of this charge determines the height of the Schottky barrier ϕ_B.

As a thought experiment, we will bring the metal nearer to the semiconductor surface. As the metal does not have a gap, its affinity is equal to the work function ϕ_m of the metal (i.e. the energy difference between its Fermi level and the vacuum). If we short the two materials by an external conducting wire, the Fermi levels will become aligned. This results from a charge transfer between the semiconductor and the metal thereby creating a potential energy difference $\phi_m - \chi_s - (E_c(\infty) - E_F)$ between the internal volumes of these two materials. In a metal, which possesses a colossal number of free electrons (10^{22}–$10^{23}\,\mathrm{cm}^{-3}$), the transferred charges are confined to a surface layer with a thickness given by the Debye length (referred to as the Thomas–Fermi length in metals). As this quantity is extremely small in metals, this charge layer can be represented as a surface charge.

On the side of the semiconductor, the transferred charge will distribute itself according to two possible models.

1. The Schottky model. If there are no surface states, the charge is created only within the depletion layer. As the separation between the two materials tends towards 0, the conduction band is then fixed by:

$$E_c(0) = \phi_m - \chi_s \tag{10.8}$$

This equation makes explicit that there are no charges at the interface. The potential drop across the depletion layer is:

$$eV_d = E_c(0) - E_c(\infty) = \phi_m - \chi_s - E_c(\infty) \tag{10.9}$$

where $E_c(\infty)$ is determined by the doping concentration. This is the *Schottky model.*

2. The Bardeen model. In the opposite limit, where the density of surface states is large, the transferred charge can easily be supplied by emptying a sufficient number of surface states without significantly affecting the semiconductor bands. In this case, eV_d is completely determined by ϕ_0 and is independent of the type of metal deposited. This is the *Bardeen model.*

In both models, the energy:

$$e\phi_B = E_c(0) - E_F \tag{10.10}$$

is referred to as the *Schottky barrier*. In the first model, ϕ_B depends on the metal, but not upon the doping in the semiconductor. In the Bardeen model, ϕ_B is independent of both metal and semiconductor doping.

In reality, experimental results sit somewhere between these two models, but more often than not, the results are consistent with the Bardeen model. Furthermore, once a metal is in contact with a semiconductor, the surface states may be modified by chemical bonding between the materials with the result being that both ϕ_0 and ϕ_B become dependent upon the metal. To summarize, the Schottky barrier depends on the metal/semiconductor pair, but this dependence is much smaller than that predicted by the Schottky model. In what follows, we will assume that ϕ_B depends weakly on the metal/semiconductor pair, and is unaffected by the doping concentration. Values for ϕ_B will be taken to follow from experiment.

The metal/semiconductor pair is said to form a *Schottky contact* to the semiconductor. We will see that this type of contact acts as a *rectifier* (i.e. allowing preferential current flow in one direction), an important characteristic exploited in Schottky diodes. Such devices are also useful as optical detectors (see Chapter 11).

10.3 The Schottky junction

When a metal and a semiconductor are in contact, the barrier between each of the materials and the vacuum no longer exists, and an electron with energy in excess of the barrier $E_c(0)$ can move between the metal and the semiconductor. At thermodynamic equilibrium, the electron density at the interface originating from the semiconductor is classically given by:

$$n(0) = N_c \exp\left(-\frac{E_c(0) - E_F}{k_B T}\right) = N_c \exp\left(-\frac{e\phi_B}{k_B T}\right) \tag{10.11}$$

For a large barrier height ϕ_B with respect to k_BT, this density is naturally very small in comparison with the carrier density inside the semiconductor:

$$\frac{n(0)}{n(\infty)} = \frac{n(0)}{N_D} = \exp\left(-\frac{eV_d}{k_BT}\right) \tag{10.12}$$

where $eV_d = E_c(0) - E_c(\infty)$.

Among these $n(0)$ surface electrons, half of them have a velocity component directed towards the metal and the other half towards the semiconductor. The average velocity of the electrons directed towards the metal is:

$$\langle v_z \rangle = \frac{\int_0^\infty v_z \exp\left(-\frac{m_c v_z^2}{2k_BT}\right) \mathrm{d}v_z}{\int_0^\infty \exp\left(-\frac{m_c v_z^2}{2k_BT}\right) \mathrm{d}v_z} = \sqrt{\frac{2k_BT}{m_c\pi}} \tag{10.13}$$

with a corresponding current density of:

$$j_{s\to m} = e\frac{n(0)}{2}\langle v_z \rangle = e\sqrt{\frac{k_BT}{2\pi m_c}} N_c \exp\left(-\frac{e\phi_B}{k_BT}\right) \tag{10.14}$$

As long as the system is in equilibrium, this current is perfectly compensated by a current flowing in the opposite direction from the metal to the semiconductor.

$$j_{s\to m} = -j_{m\to s} \tag{10.15}$$

If we apply a voltage ϕ_{app}, we now have (see Fig. 10.3):

$$E_{F,m} = E_{F,s} - e\phi_{\text{app}} \tag{10.16}$$

where the Fermi levels are defined far from the junction. In the metal, the conductivity is so great that no drop in potential is possible and $E_c(0)$ therefore remains at $E_{F,m} + e\phi_B$. The applied potential is therefore confined to the semiconductor, which must alter its depletion region so that the potential drop becomes $E_c(0) - E_c(\infty) = eV_d - e\phi_{\text{app}}$.

Note: This modulation of the depletion length L with the applied bias leads to a capacitance C given by the incremental charge $\mathrm{d}Q$ due to a variation in the applied voltage $\mathrm{d}\phi_{\text{app}}$:

$$\frac{C}{A} = \frac{1}{A}\frac{\mathrm{d}Q}{\mathrm{d}\phi_{\text{app}}} = eN_D\frac{\mathrm{d}L}{\mathrm{d}\phi_{\text{app}}} = \sqrt{\frac{e\varepsilon N_D}{2(V_d - \phi_{\text{app}})}} = \frac{\varepsilon}{L} \tag{10.17}$$

This capacitance is therefore a function of the applied voltage – a property exploited largely in ultrahigh frequency applications. The Schottky barrier therefore acts as a *varactor diode*, offering a variable reactance. If the applied voltage is the sum of a small amplitude signal, with a frequency ω_s and amplitude V_s, and a local oscillator, with frequency $\omega_0 = \omega_s - \Delta\omega$ and amplitude $V_0 \gg V_s$, the reactive current will possess a component at the difference frequency $\Delta\omega$. This current may be written ($Q = N_DL$):

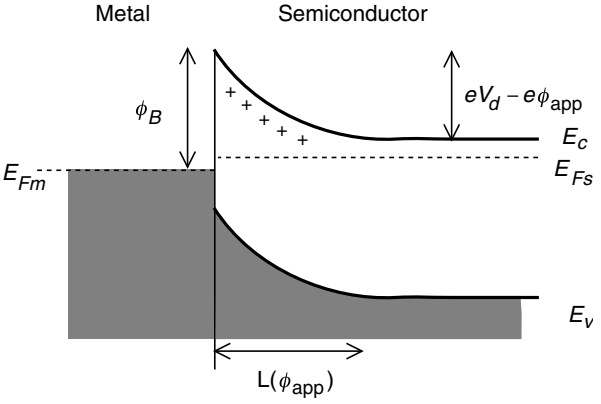


Fig. 10.3. A Schottky junction under an applied forward voltage ϕ_{app}. The depletion region shrinks and the electron density on the semiconductor side increases, causing an increase in the electron flux from the semiconductor to the metal.

$$\begin{aligned}\frac{dQ}{dt} &= \frac{dQ}{d\phi_{app}}\frac{d\phi_{app}}{dt} = C(0)\sqrt{\frac{V_d}{V_d - \phi_{app}}}\frac{d\phi_{app}}{dt} \\ &\cong -C(0)\left[1 + \frac{V_0\cos(\omega_0 t) + V_s\cos(\omega_s t)}{2V_d}\right][V_0\omega_0\sin(\omega_0 t) + V_s\omega_s\sin(\omega_s t)]\end{aligned} \tag{10.18}$$

which contains a current with frequency $\Delta\omega$ of:

$$\frac{dQ}{dt} = \frac{dQ}{d\phi_{app}}\frac{d\phi_{app}}{dt} = \frac{C(0)}{2}\frac{V_0 V_s}{V_d}\frac{\Delta\omega}{2}\sin(\Delta\omega t) \tag{10.19}$$

In general, all the frequencies $\omega_0 \pm \omega_0, \omega_0 \pm \omega_s, \omega_s \pm \omega_s$, are contained in the diode current. The varactor therefore translates the high frequency signal at ω_s to an intermediate or low frequency signal at $\omega_s - \omega_0$. It is therefore a useful element in *heterodyne detection*. As with other non-linear devices, the varactor also performs frequency doubling and rectification functions $\omega_s - \omega_s$. We will return to these concepts later within the context of non-linear optics (Chapter 12).

On the metallic side of the interface, nothing has changed. The electrons still see the same barrier to entry into the semiconductor, and the metal to semiconductor current density remains:

$$j_{m\to s} = -e\frac{n(0)}{2}\langle v_z\rangle = -e\sqrt{\frac{k_B T}{2\pi m_c}}N_c\exp\left(-\frac{e\phi_B}{k_B T}\right) \tag{10.20}$$

On the semiconductor side, however, the potential barrier does change. It is this asymmetric response between the metal and the semiconductor which leads to the rectifying behaviour of the diode. To simplify the calculations, we will suppose that the diffusivity (and hence mobility) is very large in the semiconductor. We can then assume that the electron transfer will only slightly perturb the semiconductor's equilibrium state. The carrier density on the semiconducting side of the junction is then:

$$n(0) = N_c\exp\left[-\frac{E_c(0) - E_{F,s}}{k_B T}\right] = N_c\exp\left(-\frac{e\phi_B - e\phi_{app}}{k_B T}\right) \tag{10.21}$$

and the current directed towards the metal is:

$$j_{s\to m} = \mathrm{e}\sqrt{\frac{k_B T}{2\pi m_c}} N_c \exp\left(-\frac{e\phi_B}{k_B T}\right)\exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) \tag{10.22}$$

The net current crossing the structure is then:

$$j(\phi_{\mathrm{app}}) = j_{\mathrm{sat}}\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) - 1\right] \tag{10.23}$$

Current–voltage characteristics for a Schottky diode

with the *saturation current* j_{sat} given by:

$$\begin{aligned} j_{\mathrm{sat}} &= e\sqrt{\frac{k_B T}{2\pi m_c}} N_c \exp\left(-\frac{e\phi_B}{k_B T}\right) \\ &= \frac{4ek_B^2\pi m_c}{h^3} T^2 \exp\left(-\frac{e\phi_B}{k_B T}\right) \\ &= A^* T^2 \exp\left(-\frac{e\phi_B}{k_B T}\right) \end{aligned} \tag{10.24}$$

We call A^* the *Richardson constant*. Its value, assuming a mass of $m_c = m_0$, is $A = 120\,\mathrm{A\,cm^{-2}\,K^{-2}}$. A characteristic plot for Eq. (10.23) is shown in Fig. 10.4.

Example
For an Al–GaAs junction, $\phi_B = 0.7\,\mathrm{V}$. The room temperature saturation current is then:

$$\begin{aligned} j_{\mathrm{sat}} &= A\frac{m_c}{m_0} T^2 \exp\left(-\frac{e\phi_B}{k_B T}\right) \\ &= 120 \times 0.067 \times 300^2 \times \exp\left(-\frac{0.7}{0.025}\right)\mathrm{A\,cm^{-2}} = 0.5\,\mu\mathrm{A\,cm^{-2}} \end{aligned}$$

For a diode with a $100 \times 100\,\mu\mathrm{m}^2$ area, the maximum reverse current is $|I_{\mathrm{sat}}| = 0.05\,\mathrm{nA}$. For an applied voltage $\phi_{\mathrm{app}} = 0.2\,\mathrm{V}$, the forward current is:

$$I = I_{\mathrm{sat}}\left[\exp\left(\frac{0.2}{0.025}\right) - 1\right] = 0.15\,\mu\mathrm{A}$$

and for $\phi_{\mathrm{app}} = 0.4\,\mathrm{V}$, $I = 0.44\,\mathrm{mA}$. The diode resistance R_0 at $\phi_{\mathrm{app}} = 0$ equals:

$$R_0 A = \left(\frac{\partial j}{\partial \phi_{\mathrm{app}}}\right)^{-1}_{\phi_{\mathrm{app}}=0} = \frac{k_B T}{e j_{\mathrm{sat}}} = \frac{0.025}{0.5 \times 10^{-6}}\,\Omega\,\mathrm{cm}^2 = 50\,\mathrm{k}\Omega\,\mathrm{cm}^2$$

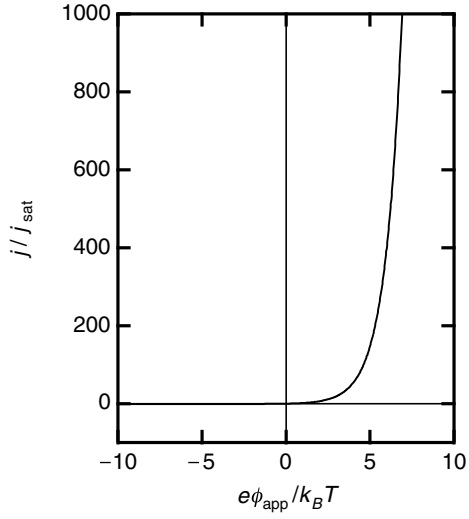


Fig. 10.4. Current–voltage characteristic for a Schottky diode.

To take into account the detailed characteristics of actual Schottky diodes, a number of corrections must be added to this idealized model. To begin with, electron transport in the depletion region can be limited by the capacity of the semiconductor to transport electrons if the diffusivity is low. In other words, under forward bias, the density $n(0)$ will be lower than the equilibrium value for the semiconductor (Eq. (10.21)) and the injected flux into the metal will be reduced. This results in a modification to the exponential prefactor j_{sat} in the I–V characteristics, whereas the dependence on ϕ_{app} remains unchanged.

Furthermore, the barrier ϕ_B may be reduced by the *image charge*. An electron exiting the metal will polarize the electron gas in the metal. The resulting affinity of the electron to the metal, known from electrostatics, can be represented by an attraction of this electron to a single virtual image charge $V_{\text{im}}(z) = -e^2/(4\varepsilon(2z))$, where z is the distance between the electron and the metallic surface. This potential then adds $-eFz$ to the depletion region potential, with the effect being that the maximum of this latter potential is reduced with respect to ϕ_B: $V_{\max} = \phi_B - (eF/4\varepsilon)^{1/2}$. The overall barrier is therefore lowered, increasing j_{sat} and inferring upon it a dependence on the field strength F (and hence on the applied voltage). Also, if the field F is very strong (as is the case for a heavily doped semiconductor even under low applied voltages) the electrons can cross the barrier by *tunnelling* and the I–V characteristics become linear. In this case, the junction is referred to as an *ohmic contact*.

Lastly, several parasitic processes (electron–hole pair production in the space charge region, leakage currents near the perimeter of the contact, . . .) can be sources of additional current. These processes depend strongly upon the explicit fabrication 'recipes' used to produce the contacts. In practice, these parasitic effects

are taken into account by an *ideality factor n*, which is always ≥ 1 – a manifestation of Murphy's law in Nature – obtained from the measured characteristics of the contact:

$$j(\phi_{\text{app}}) = j_{\text{sat}}\left[\exp\left(\frac{e\phi_{\text{app}}}{nk_BT}\right) - 1\right] \tag{10.25}$$

The Schottky diode is ideal when $n = 1$.

10.4 The *p–n* junction

A *p–n* junction is composed of a doped *p*-type semiconductor on one side of a junction and a doped *n*-type semiconductor on the other. The *I–V* characteristic for the current crossing such a junction as a function of the applied voltage exhibits rectification as was the case for the Schottky diode; however, the processes involved are quite different.

We will begin by studying an abrupt *p–n* junction at thermodynamic equilibrium (i.e. a junction over which the doping type changes over a very small distance in comparison with the spatial extent of the depletion region – see Fig. 10.5). For $z > 0$, the semiconductor is doped *p* type with a constant density of acceptors N_A. Far from the junction, the Fermi level is therefore near the valence band, and the hole density in the valence band is $p_p = N_A$, where the subscript '*p*' designates the *p* side of the junction. Also, since at thermal equilibrium we have $np = n_i^2$, the electron density is $n_p = n_i^2/N_A$, with electrons playing the role of minority carriers given that $n_p \ll N_A$. On the *n* side of the junction (for $z < 0$), the semiconductor is doped with a constant density N_D of donors. Far from the junction, the Fermi level therefore resides near the conduction band, and the majority electron and minority hole carrier densities are, respectively, $n_n = N_D$ and $p_n = n_i^2/N_D \ll n_n$.

Across the junction, the hole density cannot discontinuously change from p_n to p_p. There is therefore a zone about $z = 0$, where $p(z) < p_p = N_A$. Similarly in the case of the electrons, we also have over this region $n(z) < n_n = N_D$. At equilibrium, $n(z)p(z) = n_i^2$ is always true, but the sum of the mobile charge densities will be much smaller than the densities of the fixed ionized dopant atoms. The depletion region is therefore characterized by a charge density $-eN_A$ on the *p* side of the junction and by a positive charge eN_D on the *n* side. This dipolar layer is responsible for creating an electrostatic potential on either side of the junction. At equilibrium, the potential allows the electron and hole Fermi levels to line up. This potential is referred to as the internal or *built-in potential* as it is already present at equilibrium. It is given by:

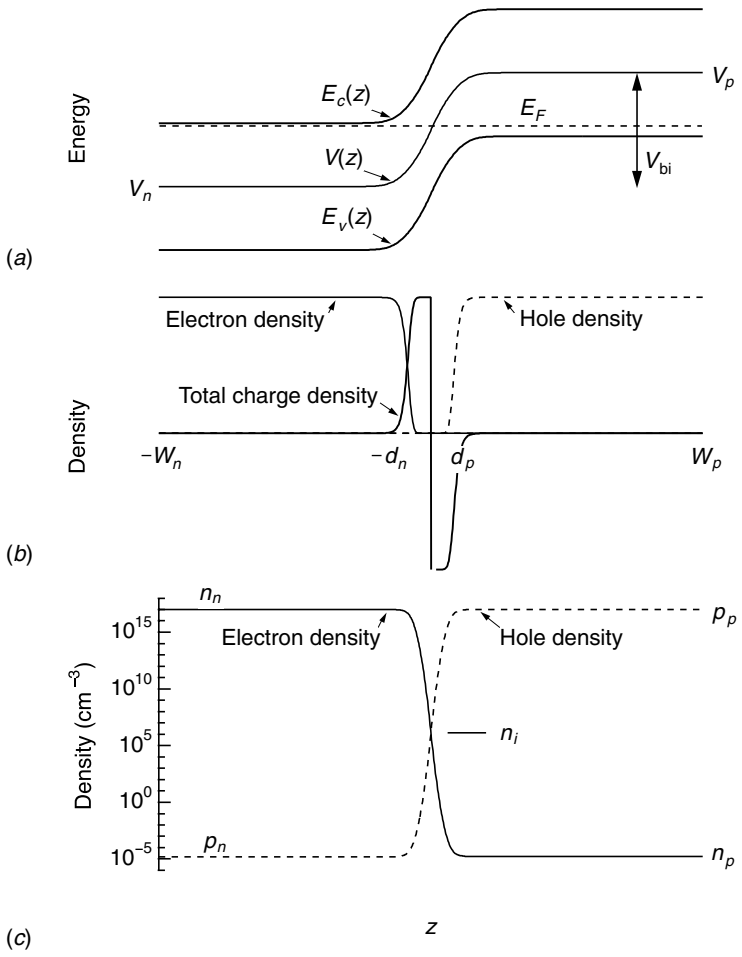


Fig. 10.5. A *p–n* diode at thermodynamic equilibrium. The potential created by the charge dipoles in the depletion region between $-d_n$ and d_p lifts the conduction and valence bands on the *p* side relative to the bands on the *n* side. This built-in potential acts to confine the electrons to $z < 0$ and the holes to $z > 0$.

$$\begin{aligned} V_{\text{bi}} &= (E_F - E_v)_n - (E_F - E_v)_p \\ &= E_g - k_B T \ln\left(\frac{N_v N_c}{N_A N_D}\right) \end{aligned} \tag{10.26}$$

where we have used (5.52) and (5.54) to calculate the Fermi levels as a function of doping concentrations.

The physical origin for the potential barrier is as follows. If we imagine that we bring the *n*- and *p*-type regions of the semiconductor into contact with one another, the majority carriers from either side of the junction will diffuse into each

other (i.e. into regions where they existed previously in minority concentrations) as would be expected by Fickian diffusion. In the absence of any opposing mechanism, both electrons and holes would continue to diffuse until their concentrations became constant across the entire structure. The charged dopant atoms left behind, however, act to oppose this mechanism. As more carriers cross to the other side, the magnitude of the space charge formed by the residual ionized dopant atoms increases, creating an energy barrier which impedes and eventually stops any further net carrier diffusion. We therefore see that carrier diffusion will play a dominant role in the operation of a *p–n* diode. It is for this reason that the built-in field is at times referred to as a *diffusion potential*. We note, in passing, the power of the Fermi level concept. This quantity defines the electronic affinity in each of the materials and takes into account, chemical potentials, electrostatic potentials, etc. The final distribution of the potential will be *that which maintains a constant Fermi level across the material.*

We can now go ahead and calculate the electrostatic potential $\phi(z)$ or, simpler yet, the potential energy of the electrons $V(z) = -e\phi(z)$. We will make the depletion region approximation as it considerably simplifies the calculations – i.e. we will neglect the influence of the carriers on the potential distribution in the space charge region.

Therefore, over $-d_n < z < 0$, the solution to Poisson's equation:

$$\frac{\mathrm{d}^2 V}{\mathrm{d}z^2} = \frac{e^2}{\varepsilon} N_D \tag{10.27}$$

gives:

$$V(z) = V_n + \frac{e^2}{2\varepsilon} N_D (z + d_n)^2 \tag{10.28}$$

and on the p side for $0 < z < d_p$:

$$\frac{\mathrm{d}^2 V}{\mathrm{d}z^2} = -\frac{e^2}{\varepsilon} N_A \tag{10.29}$$

which yields:

$$V(z) = V_p - \frac{e^2}{2\varepsilon} N_A (z - d_p)^2 \tag{10.30}$$

At the junction situated at $z = 0$, the physical continuity of the potential and its first derivative, i.e. the electric field, leads to the conditions:

$$\begin{aligned} V_n + \frac{e^2}{2\varepsilon} N_D d_n^2 &= V_p - \frac{e^2}{2\varepsilon} N_A d_p^2 \\ N_D d_n &= N_A d_p \end{aligned} \tag{10.31}$$

with the solution (since $V_p - V_n = V_{\text{bi}}$):

$$d_n^2 = \frac{2\varepsilon V_{\text{bi}}}{e^2 N_D} \frac{1}{1 + (N_D/N_A)}$$
$$d_p^2 = \frac{2\varepsilon V_{\text{bi}}}{e^2 N_A} \frac{1}{1 + (N_A/N_D)} \tag{10.32}$$

The semiconductor bands become shifted in energy by the presence of the potential since the energy of the electrons increases along each point by $-e\phi(z) = V(z)$:

$$E_c(z) = E_c(-W_n) + V(z) - V_n$$
$$E_v(z) = E_c(z) - E_g = E_v(W_p) + V(z) - V_p \tag{10.33}$$

From this energy band configuration for $E_c(z)$ and $E_v(z)$, we may calculate the electron and hole densities:

$$n(z) = N_c \exp\left[-\frac{E_c(z) - E_F}{k_B T}\right] = N_D \exp\left[-\frac{V(z) - V_n}{k_B T}\right]$$
$$p(z) = N_v \exp\left[-\frac{E_F - E_v(z)}{k_B T}\right] = N_A \exp\left[-\frac{V_p - V(z)}{k_B T}\right] \tag{10.34}$$

and in particular:

$$n(d_p) = N_D \exp\left(-\frac{V_{\text{bi}}}{k_B T}\right) = n_p$$
$$p(-d_n) = N_A \exp\left(-\frac{V_{\text{bi}}}{k_B T}\right) = p_n \tag{10.35}$$

We have therefore determined at thermodynamic equilibrium: the form of the potential $V(z)$, the *band diagram* for $E_c(z)$ and $E_v(z)$, the widths of the space charge regions, and the carrier densities at each location.

Example

We consider a GaAs *p–n* junction with each region doped with $N_D = N_A = 10^{17}\,\text{cm}^{-3}$. We seek the diffusion potential V_{bi} and the electric field at $z = 0$ in the junction. We have from Eq. (10.26):

$$V_{\text{bi}} = 1.42\,\text{eV} - 0.025\,\text{eV} \ln\left(\frac{4.7 \times 10^{17} \times 7.0 \times 10^{18}}{10^{17} \times 10^{17}}\right) = 1.27\,\text{eV}$$

and $d_n = d_p = 92\,\text{nm}$. The depletion layer is therefore very thin. The maximum field strength occurs at $z = 0$ and is equal to:

$$F(0) = \frac{1}{e}\frac{\text{d}V}{\text{d}z} = \frac{e}{\varepsilon} N_D d_n = \frac{1.6 \times 10^{-19}}{12 \times 8.84 \times 10^{-12}} 10^{23} \times 9.2 \times 10^{-8} \frac{V}{m} = 139\,\text{kV}\,\text{cm}^{-1}$$

Clearly, at equilibrium, no current will flow in the structure. Outside of the depletion region, there is neither an electric field nor carrier gradient. On the other hand, in the depletion region, there is a strong electric field, accompanied by significant carrier gradients. *The constant value of the Fermi level across the structure guarantees that all current components will cancel out if the density profile is as described in* (10.34). Before applying a voltage to the structure, it is instructive to keep in mind the different length scales which characterize the junction. The zone defined by $z < 0$ represents a colossal reservoir of free electrons held at bay by the depletion region potential. Only a very low concentration of minority electrons can tail into the p side of the junction. Similarly, the $z > 0$ region constitutes a hole reservoir, held in check by the depletion layer potential. Finally, in the neutral regions, the elevated majority carrier densities ensure that the conductivity remains high permitting the electric field to remain null. Therefore, the potential remains constant on either side of the depletion region, even in the presence of not too large currents.

The application of a weak forward bias ϕ_{app} (i.e. one which decreases the barrier height) introduces a difference between the Fermi levels on either side of the depletion region, see Fig. 10.6:

$$E_{F,p} = E_{F,n} - e\phi_{\text{app}} \tag{10.36}$$

In the neutral regions, the potential remains constant, and so the only location where the potential drop can vary is in the narrowing depletion layer. This effect can be accounted for by replacing V_p in (10.31) by $V_p - e\phi_{\text{app}}$, yielding:

$$\begin{aligned} d_n(\phi_{\text{app}}) &= d_n(0)\sqrt{\frac{V_{\text{bi}} - e\phi_{\text{app}}}{V_{\text{bi}}}} \cong d_n(0)\left(1 - \frac{e\phi_{\text{app}}}{2V_{\text{bi}}}\right) \\ d_n + d_p &\cong [d_n(0) + d_p(0)]\left(1 - \frac{e\phi_{\text{app}}}{2V_{\text{bi}}}\right) \end{aligned} \tag{10.37}$$

In the electron reservoir, the potential at d_p is lowered to $V_p - e\phi_{\text{app}}$. *As the depletion region is narrow, we can assume that the electron reservoir will remain in equilibrium up to* $\mathrm{d_p}$. The electron density is then from (10.34) and (10.35):

$$\begin{aligned} n(d_p) &= n_n \exp\left[-\frac{V(d_p) - V_n}{k_B T}\right] = n_n \exp\left(-\frac{V_{\text{bi}} - e\phi_{\text{app}}}{k_B T}\right) \\ &= n_p \exp\left(\frac{e\phi_{\text{app}}}{k_B T}\right) \end{aligned} \tag{10.38}$$

Shockley condition

where we recall that $n_p = n_i^2/N_A$ is the minority equilibrium carrier density. Equation (10.38) is fundamental in describing p–n diodes and is referred to as the

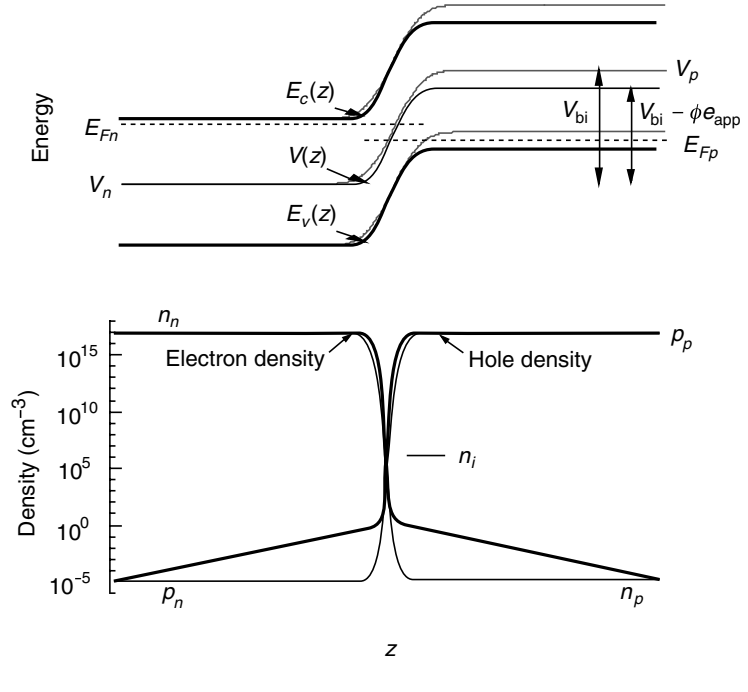


Fig. 10.6. Under the influence of an applied forward bias ϕ_{app}, the depletion region shrinks, the potential barrier between the *n* and *p* regions decreases, and the concentration of minority carriers increases (dark lines). These carriers recombine with the majority carriers, and account for the ensuing current flow in the structure.

Shockley condition. At $z = d_p$ there is a surplus of electrons with respect to the equilibrium value n_p. This surplus must be removed by the *p* contact (at $z = W_p$ in Fig. 10.5) or by recombining with holes in the $d_p < z < W_p$ region. In both cases, a current is created between the junction and the *p* contact. It is important to remember that as long as $e\phi_{app} < V_{bi}$, the electrons will remain the minority carriers by several orders of magnitude in comparison with the hole densities, and so charge neutrality in the *p* region is relatively unperturbed by the injection of electrons. As a result, there is no field to sweep the electrons out of this region (it is in the wrong direction!) and so the only means of transport available is diffusion. The transport equation for the electrons (see (6.60)) is then quite simply:

$$\frac{\partial n}{\partial t} - \frac{\partial}{\partial z} D_n \frac{\partial n}{\partial z} = -\frac{n(z) - n_p}{\tau_n} \tag{10.39}$$

especially in the stationary state, where:

$$D_n \frac{\partial^2 n}{\partial z^2} - \frac{n(z) - n_p}{\tau_n} = 0 \tag{10.40}$$

Here, D_n is the diffusion constant for electrons in the p region and τ_n is their lifetime prior to recombining with holes. By introducing the diffusion length:

$$L_n = \sqrt{D_n \tau_n} \tag{10.41}$$

Diffusion length

the equation may be written, with $\Delta n(z) = n(z) - n_p$:

$$\frac{\partial^2 \Delta n}{\partial z^2} - \frac{\Delta n(z)}{L_n^2} = 0 \tag{10.42}$$

The Shockley condition (Eq. (10.38)) allows one to calculate Δn at the edge of the depletion region:

$$\Delta n(d_p) = n_p \left[\exp\left(\frac{e\phi_{\text{app}}}{k_B T} \right) - 1 \right] \tag{10.43a}$$

For the second limiting condition, we will model the contact at $z = W_p$ by a lifetime that is so short that the electrons and holes are in thermodynamic equilibrium with the Fermi level of the contact, so that:

$$\Delta n(W_p) \approx 0 \tag{10.43b}$$

The solution for $\Delta n(z)$ is therefore:

$$\Delta n(z) = \Delta n(d_p) \frac{\sinh\left(\frac{W_p - z}{L_n} \right)}{\sinh\left(\frac{W_p - d_p}{L_n} \right)} \tag{10.44}$$

with the two limiting cases:

$$\begin{aligned} \Delta n(z) &= \Delta n(d_p) \exp\left(-\frac{z - d_p}{L_n} \right), \; L_n \ll W_p \\ \Delta n(z) &= \Delta n(d_p) \left(1 - \frac{z - d_p}{W_p - d_p} \right), \; W_p \ll L_n \end{aligned} \tag{10.45}$$

The first limiting case corresponds to a situation where no injected electrons can reach the contact because they have all already recombined with holes. In this case, given that the total current flux must be preserved, as we move away from the junction (i.e. as z increases from d_p towards W_p) the electron flow is gradually replaced by holes to fuel the recombination processes. In the second limiting case, the diffusion length is sufficiently great so that all the electrons injected at $z = d_p$ arrive at W_p without recombining with holes.

In both cases, the total current generated by the injection of electrons is the

diffusion current at $z = d_p$ (positive if directed from the p side to the n side), i.e:

$$j = j_n(d_p) = -eD_n \left.\frac{\mathrm{d}n}{\mathrm{d}z}\right|_{z=d_p} = \frac{eD_n}{L_n}\Delta n(d_p)\frac{\cosh\left(\frac{W_p - d_p}{L_n}\right)}{\sinh\left(\frac{W_p - d_p}{L_n}\right)}$$

$$= \frac{eD_n n_p}{L_n}\coth\left(\frac{W_p - d_p}{L_n}\right)\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) - 1\right] \tag{10.46}$$

$$= j_{ns}\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) - 1\right]$$

where j_{ns} is the *diffusion current limit for electrons* and is given by:

$$j_{ns} = \begin{cases} \dfrac{n_p eD_n}{L_n} = \dfrac{eD_n}{L_n}\dfrac{n_i^2}{N_A},\ L_n \ll W_p \\ \dfrac{n_p eD_n}{W_p - d_p} \cong \dfrac{eD_n}{W_p}\dfrac{n_i^2}{N_A},\ W_p \ll L_n \end{cases} \tag{10.47}$$

A condition required for this theory to hold is that the depletion region be sufficiently narrow so that neither pair-creation nor recombination occur. This implies, among other things, that $L_n \gg d_n + d_p$.

The reasoning applied to the injection of minority electrons into the p region, similarly applies to holes in the n region. The potential barrier for the holes at $z = -d_n$ decreases in such a fashion that the hole density is:

$$p(-d_n) = p_n \exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) = \frac{n_i^2}{N_D}\exp\left(\frac{e\phi_{\mathrm{app}}}{k_B T}\right) \tag{10.48}$$

and the diffusion equation for the density of minority holes in the n-type region $(-W_n < z < -d_n)$ is this time:

$$\frac{\mathrm{d}^2\Delta p}{\mathrm{d}z^2} - \frac{\Delta p(z)}{L_p^2} = 0 \tag{10.49}$$

where $L_p^2 = D_p\tau_p$ is the square of the diffusion length for holes in the n region. The solution of this last equation with the limiting conditions at $z = -W_n$ and $z = -d_n$ is analogous to (10.44) and leads to the following equation for density of excess holes:

$$\Delta p(z) = \Delta p(-d_n)\frac{\sinh\left(\frac{z + W_n}{L_p}\right)}{\sinh\left(\frac{-d_n + W_n}{L_p}\right)} \tag{10.50}$$

The current engendered by the injection of holes into the *n*-type region is consequently:

$$j_p = j_p(-d_n) = j_{ps}\left[\exp\left(\frac{e\phi_{\text{app}}}{k_B T}\right) - 1\right] \tag{10.51}$$

with the *diffusion current limit for holes*:

$$j_{ps} = \frac{p_n e D_p}{L_p}\coth\left(\frac{-d_n + W_n}{L_p}\right) = \begin{cases} \dfrac{p_n e D_p}{L_p} = \dfrac{eD_p}{L_p}\dfrac{n_i^2}{N_D}, & L_p \ll W_n \\ \dfrac{p_n e D_p}{W_n - d_n} \cong \dfrac{eD_p}{W_n}\dfrac{n_i^2}{N_D}, & W_n \ll L_p \end{cases} \tag{10.52}$$

The total current at z, in the n region is therefore the sum of three components: an electron current which supplies electrons injected into the p region, an electron current which recombines with holes injected into the n region, and a hole current injected into the n region. At any given location within the p-type region, complementary versions of these three processes also apply.

The *total current crossing the structure* is then the sum of (10.46) and (10.51) for the currents injected on either side of the depletion region:

$$j(\phi_{\text{app}}) = j_s\left[\exp\left(\frac{e\phi_{\text{app}}}{k_B T}\right) - 1\right], j_s = j_{ns} + j_{ps} \tag{10.53}$$

Current–voltage characteristic for a *p–n* diode

We discussed the $e\phi_{\text{app}} > 0$ case, in which we injected minority carriers. For $e\phi_{\text{app}} < 0$, i.e. under reverse bias conditions, the equations remain the same, except that instead of a *surplus* of minority carriers, there is a *deficit* which promotes the generation of electron–hole pairs in the neutral regions. The current–voltage characteristics remain governed by Eq. (10.53) but with $e\phi_{\text{app}} < 0$.

Example

For the GaAs junction studied above, we have $N_A = N_D = 10^{17}\,\text{cm}^{-3}$. If the electron mobility in the p region is $\mu_n = 5000\,\text{cm}^2\,\text{V}^{-1}\text{s}^{-1}$ and that of the holes in the n region is $\mu_p = 800\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$, we find using the Einstein relation, $D_n = \mu_n k_B T/e = 5000 \times 0.025\,\text{cm}^2\,\text{s}^{-1} = 125\,\text{cm}^2\,\text{s}^{-1}$, and $D_p = 20\,\text{cm}^2\,\text{s}^{-1}$. For carrier lifetimes of $\tau_n = \tau_p = 1$ ns, the diffusion lengths are $L_n = 3.5\,\mu\text{m}$ and $L_p = 1.4\,\mu\text{m}$ – clearly greater than the thickness of the depletion layer. The intrinsic carrier density is given by (5.49) to be $n_i = 1.8 \times 10^6\,\text{cm}^{-3}$. A thick diode therefore possesses the following diffusion current limits:

$$j_{ns} = \frac{n_i^2 e D_n}{N_A L_n} = \frac{(1.8 \times 10^{12})^2 \times 1.6 \times 10^{-19} \times 125 \times 10^{-4}}{10^{23} \times 3.5 \times 10^{-6}}\frac{\text{A}}{\text{m}^2} = 1.8 \times 10^{-18}\frac{\text{A}}{\text{cm}^2}$$

$$j_{ps} = \frac{n_i^2 e D_p}{N_D L_p} = \frac{(1.8 \times 10^{12})^2 \times 1.6 \times 10^{-19} \times 20 \times 10^{-4}}{10^{23} \times 1.4 \times 10^{-6}} \frac{\text{A}}{\text{m}^2} = 0.7 \times 10^{-18} \frac{\text{A}}{\text{cm}^2}$$

The inverse current is therefore $j_s = 2.5 \times 10^{-18}\,\text{A cm}^{-2}$ and the resistance at $\phi_{\text{app}} = 0$ is:

$$R_0 A = \left(\frac{\partial j}{\partial \phi_{\text{app}}} \right)^{-1}_{\phi_{\text{app}}=0} = \frac{k_B T}{e j_s} = \frac{0.025}{2.5 \times 10^{-18}} \Omega\,\text{cm}^2 = 10^{16}\,\Omega\,\text{cm}^2$$

These values for the current limits under reverse bias are very weak. In fact, under these conditions, the measured currents are dominated by leakage currents which originate at electronic defects (see Complement 10.B).

The expression for the current–voltage characteristic (Eq. (10.53)) is worthy of a few final comments:

1. The dependence on ϕ_{app} is the same as for the Schottky diode, but the difference in the prefactors reflects the fact that different transport mechanisms are at play. The Schottky diode is a unipolar device in which the current flow is established by the transport of majority carriers. The *p–n* diode, however, is a bipolar device in which carrier transport is controlled by the minority carriers in each region.
2. A *p–n* diode behaves as a better rectifier when the diffusion currents are minimal. Large gap semiconductors (with small n_i^2 values), with high doping levels, and long lifetimes help to decrease the diffusion currents.
3. Using a highly asymmetric doping scheme ($N_A \gg N_D$, for example), all the current results from the injection of electrons into the *p*-type region, while the injection of holes is minimized. A similar effect can also be obtained by utilizing different gap materials to form a *p–n* heterojunction in which n_i^2 varies on either side of the junction (see Complement 10.A).

FURTHER READING

Kwok K. Ng, *Complete Guide to Semiconductor Devices*, McGraw Hill, New York, (1995).
S. M. Sze, *Physics of Semiconductor Devices*, Wiley Interscience, New York (1981).

Complement to Chapter 10

10.A A few variants of the diode

10.A.1 *p–n* heterojunction diode

The *heterojunction diode* is a *p–n* diode in which the *p*- and *n*-type regions are defined using semiconducting materials with different bandgaps. Let us begin by considering the structure in Fig. 10.A.1 before thermodynamic equilibrium establishes itself in both regions. In this case, each material possesses its own Fermi level. The semiconductor heterojunction was studied in Chapter 8. In such a structure, the bandgap discontinuity was found to be shared between the valence and conduction band offsets ΔE_v and ΔE_c and spatially distributed across atomic-scale dimensions. As these discontinuities are chemical in origin, they do not change as the materials approach equilibrium. As is the case with any semiconductor junction, at thermodynamic equilibrium majority carriers diffuse into regions where they are in a minority, until the Fermi levels become aligned. In Fig. 10.A.1, the alignment of the Fermi levels on both sides of the structure leads to the creation of an internal potential given by:

$$\begin{aligned} V_{\text{bi}} &= E_F^{(1)} - E_F^{(2)} = E_c^{(1)} - k_B T \ln\left(\frac{N_c}{N_D}\right) - E_v^{(2)} - k_B T \ln\left(\frac{N_v}{N_A}\right) \\ &= E_g^{(2)} + \Delta E_c - k_B T \ln\left(\frac{N_c N_v}{N_D N_A}\right) \\ &= E_g^{(1)} - \Delta E_v - k_B T \ln\left(\frac{N_c N_v}{N_D N_A}\right) \end{aligned} \tag{10.A.1}$$

Poisson's equation and the results obtained for the depletion layer thicknesses ((10.26)–(10.31)) remain unchanged. The internal potential V_{bi}, however, is given by (10.A.1).

The bands again follow the potential shown for the equilibrium structure in Fig.

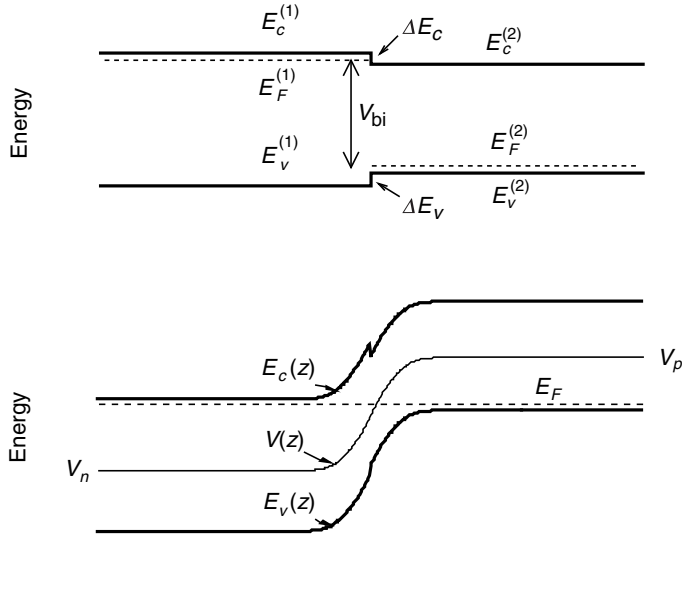


Fig. 10.A.1. A *p–n* heterojunction diode. Material (1) is doped *n* type and material (2) is doped *p* type. (*a*) Prior to onset of thermodynamic equilibrium, and (*b*) at equilibrium.

10.A.1. Clearly, the application of an external potential on this structure is described by the same minority carrier transport theory as in (10.44) and (10.49) with the important difference this time being that:

$$p_n = \frac{n_{i1}^2}{N_D} \text{ and } n_p = \frac{n_{i2}^2}{N_A} \tag{10.A.2}$$

where the *intrinsic densities* are different in each material. For a symmetrically doped structure, $N_D = N_A$, the minority carrier density in the large gap semiconductor is smaller than in the small gap material (i.e. $n_{i1}^2 \ll n_{i2}^2$, as $n_i^2 \propto e^{-E_g/k_BT}$). As a result, the minority carrier current injected into the large gap semiconductor (the holes in material 1) is greatly diminished in comparison with the homojunction case.

10.A.2 The *p–i–n* diode

Another type of diode which is very important in optoelectronics is the *p–i–n* diode, in which a lightly doped region is inserted between highly doped *p*- and *n*-type regions (see Fig. 10.A.2). The built-in potential is as always determined by the position of the Fermi levels in the *n*- and *p*-type regions, i.e:

$$V_{bi} = E_g - k_B T \ln\left(\frac{N_c N_v}{N_D N_A}\right) \tag{10.A.3}$$

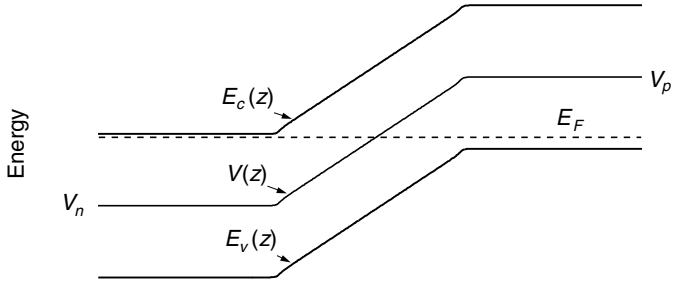


Fig. 10.A.2. Band diagram for a *p–i–n* diode at thermodynamic equilibrium.

The potential $V(z)$ is determined by Poisson's equation, which in the depletion layer approximation (see Fig. 10.A.2) is:

$$\begin{aligned}
&\frac{\mathrm{d}^2V}{\mathrm{d}z^2} = \frac{e^2}{\varepsilon}N_D, && d_1 - d_n < z < d_1 \\
&\frac{\mathrm{d}^2V}{\mathrm{d}z^2} = 0, && d_1 < z < d_2 \\
&\frac{\mathrm{d}^2V}{\mathrm{d}z^2} = -\frac{e^2}{\varepsilon}N_A, && d_2 < z < d_2 + d_p
\end{aligned} \tag{10.A.4}$$

With the solution:

$$V(z) = \begin{cases}
V_n + \dfrac{e^2N_D}{2\varepsilon}(z - d_1 + d_n)^2, & d_1 - d_n < z < d_1 \\
V_n + \dfrac{e^2N_D}{2\varepsilon}d_n^2 + \dfrac{e^2N_D}{\varepsilon}d_n(z - d_1), & d_1 < z < d_2 \\
V_p - \dfrac{e^2N_A}{2\varepsilon}(d_2 + d_p - z)^2, & d_2 < z < d_2 + d_p
\end{cases} \tag{10.A.5}$$

where the depletion lengths d_n and d_p are determined by $V_{\mathrm{bi}} = V_p - V_n$ or:

$$\begin{aligned}
&N_Dd_n = N_Ad_p \\
&V_{\mathrm{bi}} = \frac{e^2N_D}{2\varepsilon}d_n^2\left(1 + \frac{N_D}{N_A} + 2\frac{d_2 - d_1}{d_n}\right) \cong \frac{e^2N_D}{\varepsilon}d_n(d_2 - d_1)
\end{aligned} \tag{10.A.6}$$

In the limit where $d_2 - d_1 \gg d_n, d_p$ the solution is simply:

$$V(z) = \begin{cases}
V_n, & z < d_1 \\
V_n + \dfrac{V_{\mathrm{bi}}}{d_2 - d_1}(z - d_1), & d_1 < z < d_2 \\
V_p, & d^2 < z
\end{cases} \tag{10.A.7}$$

which is nothing else but a linear potential drop between the two highly doped *n*- and *p*-type regions. The system then behaves as a simple metal plate capacitor with a constant electric field in the intrinsic region given by $F = V_{bi}/(d_2 - d_1)$.

The application of a reverse bias to this structure simply acts to increase the field F:

$$F = \frac{V_{bi} + e|\phi_{app}|}{d_2 - d_1} \tag{10.A.8}$$

There are many applications for *p–i–n* diodes. As this structure possesses a constant electric field across the intrinsic region, it serves as a host structure for optoelectronic components which require the modulation of a uniform field (e.g. Stark, (Complement 8.C); or Franz–Keldysh, (Complement 7.A), effect electro-optic modulators). By far, the most important of these consists of the use of the *p–i–n* diode as a high frequency photodetector. The fast response of these *photodiodes* has a dual origin. First, the dynamic capacitance of a *p–i–n* diode ($C = \varepsilon A/(d_2 - d_1)$) is much smaller than a comparable *p–n* junction ($C = \varepsilon A/L$, where L is the width of the depletion region). The RC time constant is thus much smaller in the case of the *p–i–n* device and allows for higher frequency signal detection. Additionally, the intrinsic region can be made sufficiently large to absorb the near totality of the light. As a result, the absorption occurring in the neutral regions is kept to a minimum. This in turn reduces the photogenerated contribution to the diffusion current and effectively removes the contribution of the minority carrier lifetime to the overall dynamic response of the photodiode (see Section 11.4).

A little bit of thought will allow us to determine the limits of this 'constant field' picture for the electric field in the intrinsic region of a *p–i–n* diode. The residual doping in the intrinsic region is never completely null; and if the depletion length L_{res} corresponding to this residual doping level is smaller than $d_2 - d_1$, the field cannot remain constant and will be screened by the space charge associated with the ionized dopants. This depletion length L_{res} essentially depends on the quality of the material. Another limit comes from the intrinsic density n_i of the semiconductor. The relevant parameter here is the Debye length corresponding to n_i: if $d_2 - d_1 > \ell_D$, the free carriers can screen the field and the structure will degenerate into two junctions with a zero field region in the centre. Since ℓ_D depends on temperature, especially through n_i, it becomes impossible to maintain a field in a small gap semiconductor over important distances at ambient temperature.

Example

1. We seek the maximum tolerable residual doping level which would allow us to implement a *p–i–n* diode which operates at a voltage $V \gg V_{bi}$. The thickness of the intrinsic layer is D. From (10.1), we must have:

$$N_D \leq \frac{2\varepsilon V}{eD^2} \tag{10.A.9}$$

For Si, with a voltage of 100 V and a depletion width of 50 μm ($\varepsilon = \varepsilon_0\varepsilon_r$ with $\varepsilon_r = 12$), we must have $N_D < 5 \times 10^{13}\,\text{cm}^{-3}$.

2. We seek the maximum operating temperature for an InAs p–i–n diode with an intrinsic layer thickness D. In such a small gap semiconductor ($E_g = 0.36\,\text{eV}$), the limitation results from the intrinsic carrier population at room temperature whose density is given by (5.49). The associated Debye length is then:

$$\ell_D = \sqrt{\frac{\varepsilon k_B T}{e^2\sqrt{N_c N_v}}}\, e^{E_g/4k_B T} \tag{10.A.10}$$

The effective state densities N_c and N_v are given by (5.46) for effective masses of $m_c/m_e = 0.023$ and $m_v/m_e = 0.4$, or $N_c = 8.7 \times 10^{16}(\text{T}/300)^{1.5}\,\text{cm}^{-3}$ and $N_v = 6.3 \times 10^{18}\,\text{cm}^{-3}\ (\text{T}/300)^{1.5}$. The requirement that $\ell_D > D$ for $D = 5\,\mu\text{m}$ leads to $T_{\max} = 159\,\text{K}$.

10.B Diode leakage current

The theory behind the I–V characteristics of p–n diodes presented in Section 10.3 makes two assumptions:

1. The depletion region is narrow with respect to the diffusion lengths.
2. Neither recombination nor pair production occur in the depletion region.

The results from this theory, i.e. Eqs. (10.45), (10.50), and (10.51) are in satisfactory agreement with experimental data obtained from semiconductor diodes with moderate bandgaps (<1 eV) under low injection conditions, where minority carrier concentrations are non-negligible at room temperature. For large gap materials, there is a significant contribution to the diffusion current from generation–recombination processes in the depletion region and, in many cases, this component is dominant.

We saw that the Shockley condition (10.38) (which assumes that the carriers in the depletion region remain in equilibrium with their respective regions) determines the carrier densities in the depletion region:

$$\begin{aligned} n(z) &= n_n \exp\left[-\frac{V(z) - V_n}{k_B T}\right] \\ p(z) &= p_p \exp\left[-\frac{V_p - V(z)}{k_B T}\right] \end{aligned} \tag{10.B.1}$$

with $V_p - V_n = V_{\text{bi}} - e\phi_{\text{app}}$. We recall that n_n and p_p are the majority carrier concentrations on each side of the junction.

We will now suppose that trapping centres within the material lie near the centre of the bandgap and give rise to Shockley–Read–Hall-type generation and recombination (see Section 6.5). The net rate of electron–hole pair generation is then given by (6.56):

$$\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = -\frac{np - n_i^2}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \tag{10.B.2}$$

where n_1 and p_1 are the densities determined by the trapping energy E_T, with $n_1 \cong p_1 \cong n_i$ for a mid-gap trapping centre. Expression (10.B.2) indicates that as soon as $np \neq n_i^2$ (i.e. typically in the space charge region of a biased diode), there is generation ($np < n_i^2$) or recombination ($np > n_i^2$) of electron–hole pairs.

Under reverse bias conditions ($np < n_i^2$ – see Fig. 10.B.1), we have practically everywhere in the depletion region $p(z) < n_i$ and $n(z) < n_i$, and Eq. (10.B.2) leads to a *generation current*:

$$j_g = -e \int_{-d_n}^{d_p} \frac{n_i^2}{\tau_{n0} p_1 + \tau_{p0} n_1} \mathrm{d}z \cong \frac{e n_i (d_p + d_n)}{\tau_{n0} + \tau_{p0}} \tag{10.B.3}$$

as the electrons and holes are rapidly swept from the depletion region by the strong electric field which dominates there.

This current is to be compared with the diffusion current limit (10.51):

$$j_s = \frac{e n_i^2 D_n}{N_A L_n} + \frac{e n_i^2 D_p}{N_D L_p} \tag{10.B.4}$$

The ratio j_g/j_s is proportional to n_i^{-1}, by which we see the importance of the generation current in large gap semiconductors compared with the diffusion current. The sum of these two currents constitutes the reversed bias diode leakage current (see Fig. 10.B.2). Since the intrinsic carrier density n_i is proportional to $\exp(-E_g/2kT)$, we see that $j_s \propto \exp(-E_g/kT)$ while $j_g \propto \exp(-E_g/2kT)$. The generation–recombination leakage current therefore prevails at high temperature, whereas the diffusion current is dominant at low temperature.

Example

For the GaAs diode encountered in the last example of Section 10.2, we find for a reverse bias of -1 V, a depletion layer thickness of:

$$d_n + d_p = 92\,\text{nm} \sqrt{\frac{1.27 + 1}{1.27}} = 123\,\text{nm}$$

The generation current given $\tau_{0n} + \tau_{0p} = 10$ ns is then:

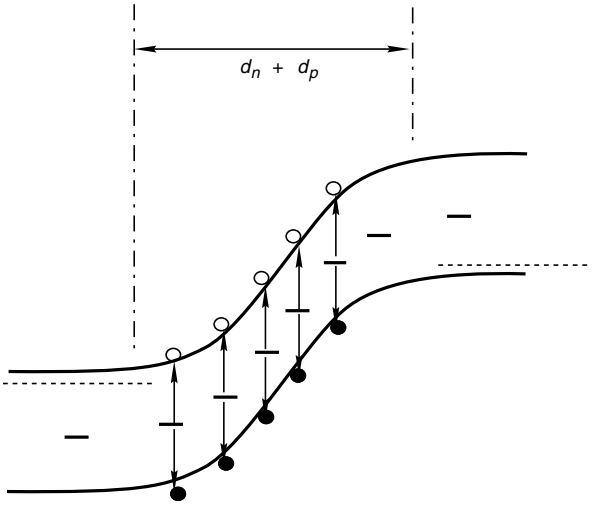


Fig. 10.B.1. Electron–hole generation mechanism by defects in the space charge region.

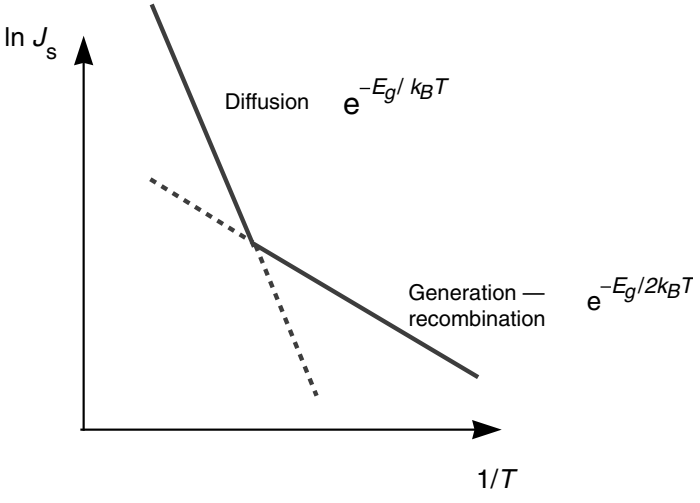


Fig. 10.B.2. Leakage current in a diode as a function of temperature.

$$j_g = \frac{1.6 \times 10^{-19} \times 1.8 \times 10^{12} \times 123 \times 10^{-9}}{10 \times 10^{-9}} \frac{\text{A}}{\text{m}^2} = 3.5 \times 10^{-10}\,\text{A cm}^{-2}$$

which dominates the diffusion current by several (eight!) orders of magnitude.

In a *p–i–n* diode, the generation current therefore becomes more important as the thickness of the intrinsic region increases. This *dark current* in the photodiode plays a central role in the overall performance characteristics of these devices (see Chapter 11).

In the more general case involving an arbitrary bias, we have:

$$j_{g-r} = e\int_{-d_n}^{d_p} \frac{n(z)p(z) - n_i^2}{\tau_{n0}[p(z) + p_1] + \tau_{p0}[n(z) + n_1]}\mathrm{d}z$$
$$= e\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{k_BT}\right) - 1\right]\int_{-d_n}^{d_p} \frac{n_i^2\mathrm{d}z}{\tau_{n0}p(z) + \tau_{p0}n(z) + \tau_{n0}p_1 + \tau_{p0}n_1} \tag{10.B.5}$$

We can find an upper limit approximation to this integral by evaluating the maximum of the integrand: the denominator then possesses a minimum at $V = V_{\mathrm{min}}$ such that:

$$\exp\left(\frac{V_{\mathrm{min}}}{k_BT}\right) = \sqrt{\frac{\tau_{p0}n_n}{\tau_{n0}p_p}}\exp\left(\frac{V_p + V_n}{k_BT}\right) \tag{10.B.6}$$

which substituted into the integrand, yields:

$$j_{g-r} = e\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{k_BT}\right) - 1\right]\int_{-d_n}^{d_p} \frac{n_i^2\mathrm{d}z}{2\sqrt{\tau_{n0}\tau_{p0}}n_i\exp\left(\frac{e\phi_{\mathrm{app}}}{2k_BT}\right) + \tau_{n0}p_1 + \tau_{p0}n_1}$$
$$= \frac{en_i(d_n + d_p)}{2\sqrt{\tau_{n0}\tau_{p0}}}\exp\left(\frac{e\phi_{\mathrm{app}}}{2k_BT}\right) \tag{10.B.7}$$

This last approximation is only valid provided $e\phi_{\mathrm{app}} > k_BT$, i.e. under forward bias conditions.

We see that the characteristics of a diode dominated by generation–recombination currents has the same form as the classic case with the following important differences: (i) the prefactor, and (ii) the *generation–recombination current limit* is different since the denominator in the exponential is $2k_BT$, instead of k_BT. This then justifies that we write for the I–V characteristics of a real diode:

$$j = j_s\left[\exp\left(\frac{e\phi_{\mathrm{app}}}{nk_BT}\right) - 1\right] \tag{10.B.8}$$

where j_s, the *saturation current*, is the sum of the limiting currents, and the *ideality factor* n sits between 1 and 2 (near 1 when the diffusion current is dominant and near 2 when the generation–recombination current dominates).

To conclude, we summarize the three principal mechanisms that determine the carrier lifetime in a semiconductor, and hence the leakage current exhibited by diodes fabricated from these materials:

1. *Radiative recombination.* The radiative recombination of electrons and holes is very efficient in a direct gap material. An expression for this time was given in

(7.65) and can be deduced from Fig. 7.5. For typical doping levels, this value ranges from approximately 10 ns (in GaAs) to 100 ns (in InSb).

2. *Impurity recombination* (*Shockley–Read*). This mechanism was described in Section 6.5 and an expression for it was given in (6.59):

$$\frac{1}{\tau} = \sigma_{\mathrm{SR}} v_{\mathrm{th}} N_t \tag{10.B.9}$$

where N_t is the defect density and v_{th} is the *thermal velocity* of the carriers given by $\langle m_c v^2/2 \rangle = m_c v_{\mathrm{th}}^2 = kT/4\pi$ with (for example) $v_{\mathrm{th}} = 10^7 \times (T/300)^{1/2}$ cm s^{-1} in GaAs. The *capture cross-sections* σ_{SR} generally correspond to the area covered by the impurity wavefunctions (i.e. of the order of 10^{-15} cm^2). Therefore, in GaAs, a defect concentration of 10^{15} cm^{-3} leads to a lifetime of 100 ns.

3. *Auger recombination.* The origin of this mechanism was largely described in Complement 6.D, and an expression for the lifetime τ_{Ai} in an intrinsic material was given in (6.D.21). In a material doped with a concentration N_D, the Auger lifetime can be obtained from (6.D.20):

$$\tau_A = 2\tau_{Ai}\left(\frac{n_i}{N_D}\right)^2 \tag{10.B.10}$$

This mechanism dominates the diode leakage current in small gap semiconductors at room temperature (e.g. InAs, InSb, HgCdTe, . . .).

11 Semiconductor photodetectors

11.1 Introduction

As early as 1850, Antoine Cesar Becquerel discovered that certain materials generate an electrical current when exposed to a flux of light. It took, however, until about 1935 before a quantum theory of condensed matter could be developed to give a satisfactory account of this phenomenon. In spite of a lack of any firm theoretical understanding for these empirical observations, photodetectors were fashioned from these materials and put to work in photography and in military infrared detection applications.

The basic general principles behind the operation of semiconductor detectors are illustrated in Fig. 11.1. In the absence of photoexcitation, the carriers in these materials do not conduct electricity either because: (a) they are in a band where they cannot participate in conduction (e.g. a full valence band), (b) they are blocked by a potential barrier (as in a Schottky detector), or (c) they are trapped in quantum bound states (e.g. extrinsic photoconductors or quantum well detectors).

Optically driven transitions between two ensembles of quantum levels (one conducting and the other insulating), are at the origin of photodetection. For this reason, semiconductor detectors are sometimes referred to as *quantum detectors*.

Sections 11.3 and 11.4 describe (with reference to Fig. 11.1) type (*a*) photodetectors (photoconducting and photovoltaic), Section 11.5 describes type (*b*) internal emission Schottky photodetectors, and Section 11.6 describes type (*c*) quantum well photodetectors. We will see that all these detectors share a common feature. Their performance (or detectivity D^*, which will be defined later) evolves as a function of operation temperature as the square root of $e^{-h\nu/kT}$, where $h\nu$ is the minimum energy of the detected photons. We will see that this can be interpreted as stating, 'if photons of energy $h\nu$ are detected, then, alas, so is a combination of phonons with an average energy of $h\nu$'.

11.2 Distribution of carriers in a photoexcited semiconductor

The complexity of the photodetection mechanisms results from the multitude of different length scales which govern electron–photon interaction in matter. First, there is the issue of the absorption length, which is of the order of a few α^{-1} (see

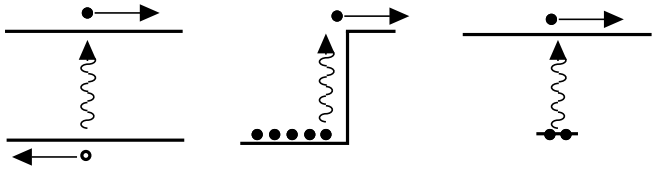


Fig. 11.1. Three types of quantum detection based on: (*a*) band to band transitions which generate electron–hole pairs, (*b*) internal photoemission above a potential barrier, and (*c*) transition from a bound state to a continuum.

Sections 7.3 and 7.4). Then there are the minority carrier diffusion lengths (Chapter 10), which determine the distances over which these carriers can effectively participate in detection mechanisms. Finally, there are the dimensions of the space charge regions, of particular significance to photovoltaic detectors. All these length scales are of about the same order – a few micrometres to a few tens of micrometres. None of these mechanisms can therefore be neglected in favour of others. In most cases, it is necessary to resort to the complete equations for semiconductors (Eqs. (6.60) and (6.61)) to describe the photodetection mechanisms. In this section, we will describe these equations without solving them in their complete generality – such work would be feasible, but tedious, and provide results of limited usefulness. They will, however, be of use to the reader provided the appropriate simplifications are made for a given problem. We will limit ourselves to a one-dimensional treatment (along the Oz axis of light propagation) and neglect problems involving non-uniform illumination, etc.

To begin, the light flux in a semiconductor decays exponentially (see Chapters 3 and 4) as:

$$\Phi(z) = \Phi_0 e^{-\alpha z} \tag{11.1}$$

where Φ_0 is the incident photon flux upon the surface. The absorption coefficient is itself a function of the photon energy as explained in detail in Chapters 7 and 8. Figure 11.2 shows the spectral response $\alpha(h\nu)$ for several semiconductors of particular relevance to optoelectronics. We note the difference in spectral behaviour between *direct gap* (GaAs, InAs, InSb) and *indirect gap* semiconductors like Si, which exhibit a less abrupt absorption edge than their direct gap counterparts.

The photons absorbed by the semiconductors generate electron–hole pairs with density $\Delta n = n - n_0$ and $\Delta p = p - p_0$ with a generation rate G_{op} ($cm^{-3}\,s^{-1}$) given by:

$$G_{op}(z) = \alpha \Phi_0 e^{-\alpha z} \tag{11.2}$$

One of two possible fates await the carriers after creation. Either they will recombine with each other with a characteristic lifetime τ (τ_n for electrons, τ_p for holes), or

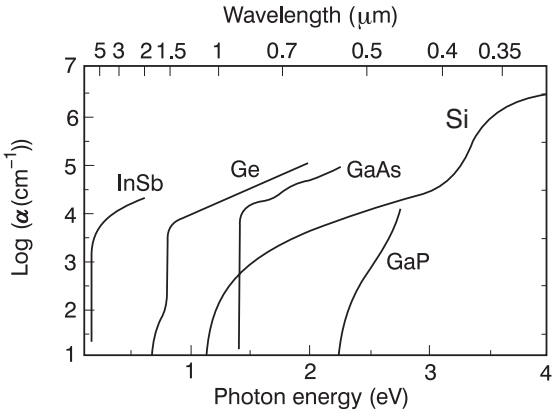


Fig. 11.2. Spectral dependence of absorption for a few important semiconductors.

they will be swept out by the electric field or diffuse towards the ohmic contacts. The equations which describe this mechanism are given in (6.60) and (6.61):

$$\frac{\partial}{\partial t}n = G_{\text{op}}(z) - \frac{\Delta n}{\tau_n} + \frac{1}{q}\frac{\partial}{\partial z}J_n \tag{11.3a}$$

$$\frac{\partial}{\partial t}p = G_{\text{op}}(z) - \frac{\Delta p}{\tau_p} - \frac{1}{q}\frac{\partial}{\partial z}J_p \tag{11.3b}$$

where the electron (J_n) and hole (J_p) current densities are:

$$\frac{J_n}{q} = D_n\frac{\partial}{\partial z}n + n\mu_n E \tag{11.4a}$$

$$\frac{J_p}{q} = -D_p\frac{\partial}{\partial z}p + p\mu_p E \tag{11.4b}$$

and the coefficients μ and D are the mobilities and diffusivities for their respective carrier types. Table 11.1 gives the transport properties of these materials: i.e. typical recombination times (which clearly may depend on illumination), mobilities, bandgaps, and dielectric constants.

To these equations, we must add Poisson's equation which allows us to calculate the electric field as a function of the permanent charge densities (dopants) and photogenerated carriers, i.e:

$$\frac{\mathrm{d}E}{\mathrm{d}z} = \frac{q}{\varepsilon}(p + N_d^+ - n - N_a^-) \tag{11.5}$$

where ε is the dielectric permittivity of the material (with $\varepsilon = \varepsilon_0\varepsilon_r$, where ε_r is the relative permittivity) and N_d^+ and N_a^- are the ionized dopant concentrations. At equilibrium (i.e. without illumination) and in the bulk $N_d^+ + p_0 = N_a^- + n_0$.

Table 11.1. *Values for the recombination lifetimes* τ, *mobilities* μ_n *and* μ_p, *300 K bandgaps, and dielectric constants for important semiconductors*

Material	τ (s)	μ_n (cm^2 V^{-1} s^{-1})	μ_p (cm^2 V^{-1} s^{-1})	E_g (eV)	ε_r
Si	10^{-4}	1350	480	1.12	11.8
Ge	10^{-2}	3900	1900	0.67	16
GaAs	10^{-6}	8500	400	1.42	13.2
InAs	10^{-7}	33 000	460	0.36	14.6
InSb	10^{-7}	10^5	1700	0.18	17.7

Equation (11.5) allows one to understand how photodiode detection efficiencies vary under elevated illumination conditions where for example $\Delta n \approx N_d$. It is important to note that the characteristic time, called the *dielectric relaxation time*, over which the material preserves charge neutrality is quite short ($\tau_{\text{diel}} = \varepsilon/\sigma \approx 10^{-12}$ s, where σ is the conductivity of the material). The medium will therefore remain electrically neutral, i.e:

$$\Delta n(z,t) = \Delta p(z,t) \tag{11.6}$$

This last equation, in spite of its simple form, has very important consequences for the detector response. First, as carriers recombine in pairs, this means that the recombination lifetimes must also be equal ($\tau_n = \tau_p = \tau$). This is only possible if the recombination mechanisms depend in some manner on n and p. Another consequence of equality (11.6) is that an internal field must be established before this relation can be imposed on (11.3), (11.4), and (11.5). We could further complicate the table by introducing capture mechanisms for a single carrier type, spatial inhomogeneities, etc.

Let us begin by simplifying the system we are about to study. We will assume that the detector material is dominantly p type ($p_0 \gg n_0$, we will see later that this is an interesting case for applications) and that no external electric field $E(z)$ is applied. We can always neglect all phenomena tied to the conservation of charge neutrality, such as the appearance of an internal field – this neutrality is guaranteed by the majority dopants. Substituting (11.4a) into (11.3a), we obtain:

$$\underbrace{\frac{\Delta n}{\tau}}_{\text{recombination}} - \underbrace{D_n \frac{\mathrm{d}^2}{\mathrm{d}z^2}\Delta n}_{\text{diffusion}} = \underbrace{\alpha \Phi_0 \mathrm{e}^{-\alpha z}}_{\text{source}} \tag{11.7}$$

The solution to this equation is the sum of a particular solution including a source term and a general solution without a source term, i.e:

$$\Delta n(z) = A\mathrm{e}^{-z/L_D} + B\mathrm{e}^{z/L_D} + \frac{\alpha\tau\Phi_0}{1-(\alpha L_D)^2}\mathrm{e}^{-\alpha z} \tag{11.8}$$

where L_D is the electron diffusion length, $L_D = \sqrt{(D\tau)}$. The constants A and B are determined by the boundary conditions specific to the problem. These conditions can be of many possible types:

Surface recombination S (in $\mathrm{cm\,s^{-1}}$) at $z = z_0$ imposed as:

$$S\Delta n|_{z=z_0} = D\left(\frac{\mathrm{d}}{\mathrm{d}z}\Delta n\right)_{z=z_0} \tag{11.9}$$

Ohmic contact, i.e. an interface where the recombination of minority carriers is instantaneous (an ohmic p^+ contact on p-type material imposes a density of minority carriers n given by $np = n_i^2$, i.e. $n \approx 0$) such that:

$$\Delta n|_{z=z_0} = 0 \tag{11.10a}$$

Space charge region starting at $z = z_0$, i.e. a region out of which all carriers are swept, and again leading to:

$$\Delta n|_{z=z_0} = 0 \tag{11.10b}$$

Let us analyse, for example, the case involving an infinitely thick sample ($d \gg L_D$ and $1/\alpha$). As $z \to +\infty$, $\Delta n \to 0$. The coefficient A is then obtained by imposing the boundary conditions at $z = 0$. We will suppose that no ohmic contact has been made and that the electrons are free to recombine at the surface ($z = 0$) with velocity S_0. Equations (11.8) and (11.9) immediately yield the distribution of photogenerated carriers $\Delta n(z)$:

$$\Delta n(z) = \frac{\alpha\tau\Phi_0}{(\alpha L_D)^2 - 1}\left(\frac{S_0/D_n + \alpha}{S_0/D_n + 1/L_D}\mathrm{e}^{-z/L_D} - \mathrm{e}^{-\alpha z}\right) \tag{11.11}$$

As might be expected, the carrier concentration at the surface becomes null as the recombination velocity tends towards infinity (see Fig. 11.3). If, on the other hand, the semiconductor surface is passivated (i.e. if the density of surface states is null), then $S_0 = 0$ and the number of generated carriers Δn_{tot} in the entire volume is given by integral (11.11) to be:

$$\Delta n_{\mathrm{tot}} = \int_0^\infty \Delta n(z)\mathrm{d}z = \tau\Phi_0 \tag{11.12}$$

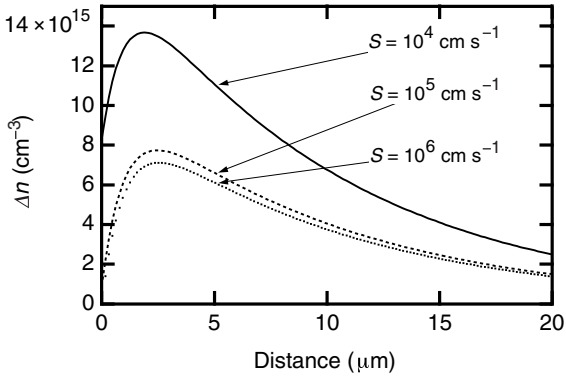


Fig. 11.3. Distribution of photogenerated carriers for a photon flux $\Phi_0 = 10^{18}\,\mathrm{cm}^{-2}$, a diffusion coefficient of $1\,\mathrm{cm}^2\,\mathrm{s}^{-1}$, a diffusion length of 10 μm, and an absorption coefficient of $10^4\,\mathrm{cm}^{-1}$, for various surface recombination velocities S.

This last expression can be read as $\Delta n_{\mathrm{tot}}/\tau = \Phi_0$, i.e. that the entire flux of incident photons Φ_0 is transformed into a flux of carriers $\Delta n_{\mathrm{tot}}/\tau$.

Another very simple case is that involving a sample of thickness d in which no currents can flow through the light entrance and exit surfaces (for instance, a photoconductor in which the electrical current flow is perpendicular to the photon flux): this means, supposing that no external electric field is applied, that $J_n = D_n \partial \Delta n/\partial z = 0$ at $z = 0$ and $z = d$. Integration of (11.7) leads immediately to a total surface density Δn_{tot} of photocarriers generated in the volume of:

$$\Delta n_{\mathrm{tot}} = \Phi_0 \tau (1 - \mathrm{e}^{-\alpha d}) \tag{11.13}$$

The ratio between the flux of created carriers $\Delta n_{\mathrm{tot}}/\tau$ and the flux of incident photons Φ_0 expresses the conversion efficiency of the photon flux into a carrier flux. This is the *internal quantum efficiency* η_i of the detector and is given in this case by:

$$\eta_i = \frac{\Delta n_{\mathrm{tot}}/\tau}{\Phi_0} = 1 - \mathrm{e}^{-\alpha d} \tag{11.14a}$$

Internal quantum efficiency for a detector

To this internal quantum efficiency, may be added other losses. The most important of these is generally due to the loss of photon flux by optical reflection occurring at the surface, i.e. $\eta_{\mathrm{op}} = 1 - R$. This optical efficiency is particularly weak for untreated semiconductor–air interfaces as the semiconductor index of refraction n_{sc} is considerable, and $R = (n_{\mathrm{sc}} - 1)^2/(n_{\mathrm{sc}} + 1)^2$. Thus, for silicon, $n_{\mathrm{sc}} = 3.4$ leading to $R = 0.3$ and $\eta_{\mathrm{op}} = 1 - R = 0.7$. The total efficiency η is therefore generally close to:

$$\eta = \eta_i \eta_{\mathrm{op}} = (1 - R)(1 - \mathrm{e}^{-\alpha d}) \tag{11.14b}$$

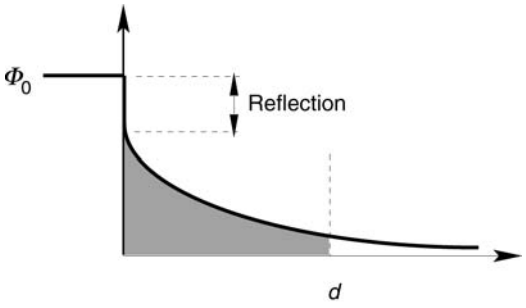


Fig. 11.4. Of the incident flux, Φ_0, only the grey portion transmitted by the surface and absorbed over the detector thickness d can participate in photocarrier generation.

Anti-reflection coatings are often deposited to reduce the losses due to reflection by the semiconductor–air interface significantly over a wide range of wavelengths. Figure 11.4 summarizes the various contributions to the total detector efficiency.

Example

We seek the carrier distribution in a semiconductor with an absorption coefficient α of $10^4\,\mathrm{cm}^{-1}$. We will suppose as well that the material possesses a mobility of $40\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ and a recombination time of $10^{-6}\,\mathrm{s}$. This leads to a diffusion coefficient given by the Einstein relation $D_n = (kT/q)\mu_n$ of $40\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1} \times 25.9\,\mathrm{mV}$ or $1\,\mathrm{cm}^2\,\mathrm{s}^{-1}$ and a diffusion length $(D_n\tau)^{1/2}$ of $10\,\mu\mathrm{m}$. It is therefore clear that α dominates over $1/L$ in (11.11) and that the distribution $\Delta n(z)$ depends above all on the relative magnitudes of S_0/D_n and α. Figure 11.3 shows the distributions $\Delta n(z)$ for different surface recombination velocities S_0.

11.3 Photoconductors

11.3.1 Photoconduction gain

We will now focus upon the behaviour of an ideal photoconductor fashioned from a homogeneous semiconductor material of a given type (say p type), of thickness d, width w, and length l (see Fig. 11.5). As in the immense majority of cases (with the exception of quantum well photoconductors – see Section 11.5), the photon and carrier fluxes are perpendicular to one another. A flux of photons is incident upon the semiconductor surface in the Oz direction. The photons in the beam of light have an energy hv greater than the semiconductor bandgap E_g and are absorbed as indicated by the particular dependence of α on hv. Two electrical contacts are made on either edge (see Fig. 11.5) allowing one to:

1. apply an electric field E parallel to the surface (along Ox);
2. extract carriers 'on demand', i.e. sufficiently and without limitation to be able to maintain charge neutrality in the material.

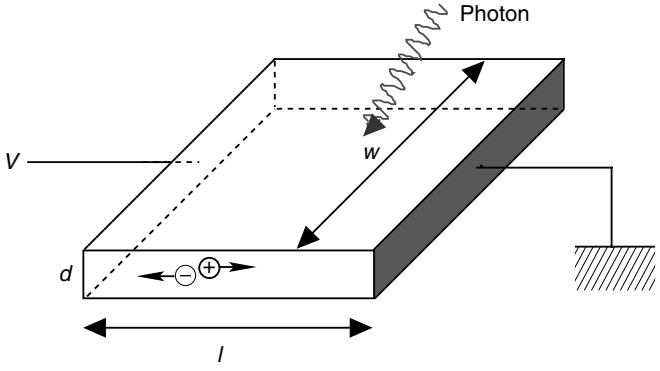


Fig. 11.5. Geometry for a photoconducting detector.

As we just saw, by neglecting the flux of carriers across the two surfaces (top and bottom), the carrier surface density is determined by the equality of the two fluxes $\Delta n_{\text{tot}}/\tau = \eta\Phi_0$ (in $\text{cm}^{-2}\,\text{s}^{-1}$), i.e. by supposing that the diffusion length is sufficiently large so that the inhomogeneities in $\Delta n(z)$ are negligible, and the volumetric concentration $\Delta n(z) = \Delta n =$ constant is given by:

$$\Delta n = \frac{\eta\tau\Phi_0}{d} \tag{11.15}$$

Under the influence of the transverse electric field E due to the potential V ($E = V/l$), a photocurrent begins to circulate within the structure, having a density of $j_{\text{ph}} = \Delta n q \mu_n E$ and a total magnitude given by $I_{\text{ph}} = j_{\text{ph}} wd$, *as the surface crossed by the circulating electric current is* wd, so that:

$$I_{\text{ph}} = \eta q \mu_n \tau \frac{w}{l} \Phi_0 V \tag{11.16}$$

From another point of view (this more subtle approach to photoconduction was developed by A. Rose), *the photon flux crosses the surface* wl. It is therefore reasonable to normalize the current I_{ph} by the incident power $P_{\text{inc}} = h\nu\Phi_0 wl$ over the surface wl giving rise to the *detector responsivity* $\mathscr{R}$:

$$\mathscr{R} = \frac{I_{\text{ph}}}{P_{\text{inc}}} = \eta \frac{\mu_n \tau}{l^2} \frac{V}{h\nu/q} \tag{11.17}$$

This formula shows the importance of the product $\mu_n\tau$ in determining the responsivity of a photodetector. This is in fact the only occurrence of any of the intrinsic properties of the semiconductor in the formula for the photocurrent. The product $\mu_n\tau$ (in $\text{cm}^2\,\text{V}^{-1}$) is used as a figure of merit for detector material. Expression (11.17) can also be put into the form:

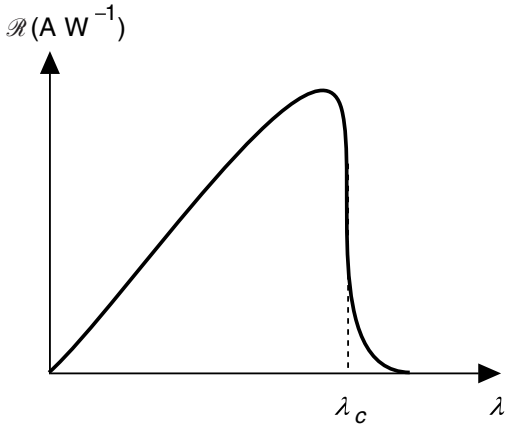


Fig. 11.6. Spectral responsivity for a photoconductor. The linear variation of the responsivity with λ results from the fact that the number of photons necessary to make up 1 W increases with λ. This overall lineshape is the spectral signature of a broad spectrum quantum detector.

$$\mathscr{R} = \eta g \frac{l}{h\nu/q} \tag{11.18}$$

Photoconductor responsivity (A W^{-1})

where g is the *photoconductive gain* given by:

$$g = \frac{\tau}{\tau_{tr}};\ \tau_{tr} = \frac{l}{\mu_n E} = \frac{l^2}{\mu_n V} \tag{11.19}$$

and τ_{tr} is the *transit time* for electrons travelling between the two contacts. We note that the responsivity of a detector as a function of wavelength may be put into the form:

$$\mathscr{R} = \eta g \frac{\lambda(\mu m)}{1.24} \tag{11.20}$$

Figure 11.6 shows the spectral responsivity $\mathscr{R}(\lambda)$ characteristic of broad spectrum quantum detectors.

The expression for the photoconductive gain is easy to understand. It is the ratio of the electronic flux created in the Ox direction divided by photon flux in the Oz direction. The first surprising observation is that this term can be greater than unity, i.e. it gives a false(!) impression that a single photon can give rise to several electron–hole pairs without drawing upon avalanche effects, etc. Several interpretations have been advanced to dispel this apparent paradox. The first of these is attributed to A. Rose, and consists of imagining that an electron can travel through the circuit several times before disappearing by recombination. The same

electron can therefore participate several times in the electrical current. Another interpretation is that the semiconductor medium stockpiles electron–hole pairs during illumination and that the contacts take care of providing all necessary carriers to satisfy charge neutrality.

Example
We consider a 3 μm thick layer of InSb, 1 mm in length (l), and 100 μm in width (w). We wish to know the responsivity of this detector at 6 μm for an applied voltage of 10 V (or a field of $10^2\,\mathrm{V\,cm^{-1}}$) assuming a quantum efficiency η of 60%. The mobility for InSb is given in Table 11.1 to be $10^5\,\mathrm{cm^2\,V^{-1}}$, leading to a transit time τ_{tr} of $10^{-1}\,\mathrm{cm}/(10^5 \times 10^2\,\mathrm{cm\,s^{-1}})$, or $10^{-8}\,\mathrm{s}$. As the lifetime is $10^{-7}\,\mathrm{s}$, we find a photoconductive gain g of $10^{-7}/10^{-8} = 10$. The responsivity $\mathscr{R}$ is then $0.6 \times 10 \times 6\,\mu\mathrm{m}/(1.24\,\mu\mathrm{m\,eV^{-1}})$, or $30\,\mathrm{A\,W^{-1}}$.

11.3.2 Photoconductor detectivity

As discussed in Complements 11.A and 11.B, three mechanisms form the basis of noise in photoconductors. These are: thermal noise i_R; generation–recombination noise i_{GR}; and photon noise (or shot noise associated with random fluctuations in the number of photons arriving at the detector), as given in (11.A.27), (11.A.38), and (11.B.6). We will assume the case where the detector response is not limited by the photon noise, i.e. that the detector is not in the BLIP regime (see Complement 11.B). The noise generated by the detector itself is then:

$$\begin{aligned} i_N^2 &= i_R^2 + i_{GR}^2 = 4\frac{kT}{R}\Delta v + 4qgI_0\Delta v \\ &= \frac{4q}{R}\left(\frac{kT}{q} + gV\right)\Delta v \approx 4qgI_0\Delta v \end{aligned} \tag{11.21}$$

once the applied voltage exceeds kT/q. We recall that I_0 is the *dark current*, R is the photoconductor resistance, Δv is the frequency integration bandwidth given by $\Delta v = 1/2t_{int}$, where t_{int} is the integration time, and g is the photoconductive gain. The signal-to-noise ratio i_S/i_N is then given by:

$$S/N = \frac{i_S}{i_B} = \frac{\mathscr{R}P_{inc}}{\sqrt{4qgI_0\Delta v}} \tag{11.22}$$

We introduce the *noise equivalent power* (NEP), which is the power P_{inc} at the detection limit corresponding to a signal-to-noise ratio of 1. From (11.22), we obtain the NEP for a photoconductor.

$$\mathrm{NEP} = \frac{\sqrt{4qgI_0\Delta v}}{\mathscr{R}} \tag{11.23}$$

Noise equivalent power (NEP)

This last formula is clearly not independent of the measurement system, i.e. it does not allow one to classify detector materials according to some *figure of merit*. In order to normalize (11.23), we note that the NEP is a function of bandwidth as $\sqrt{(\Delta v)}$ and (given $I_0 = J_0 A$) a function of the square root of the surface area of the sample $\sqrt{A}$. We thus introduce the *detectivity* D^*, defined in a general fashion as:

$$D^* = \frac{\sqrt{A\Delta v}}{\mathrm{NEP}} \tag{11.24}$$

Detectivity (cm $\mathrm{Hz}^{1/2}\,\mathrm{W}^{-1}$) and NEP

The (barbaric) measurement unit for detectivity is $\mathrm{cm\,Hz^{1/2}\,W^{-1}}$ which in the civilized world is also called the *jones*. We are now in a position to calculate the detectivity of a photoconductor:

$$D^* = \frac{\mathscr{R}}{\sqrt{4qgJ_0}} \tag{11.25}$$

Note: $\mathrm{J_0}$ *is the dark current density flowing between the two electrodes normalized by the surface area* A *seen by the incident flux of light*! This current density is therefore given (for a *p*-type semiconductor) by:

$$J_0 = qp_0\mu_p E\frac{wd}{wl} \tag{11.26}$$

where p_0 is the carrier density in the dark and E is the electric field between the two contacts. Expression (11.25) can be put into the form:

$$D^* = \frac{\eta}{\sqrt{d}}\frac{l}{hv}\sqrt{\frac{\tau}{4p_0}\frac{\mu_n}{\mu_p}} \tag{11.27}$$

Had we taken an *n*-type semiconductor, the ratio μ_n/μ_p would be replaced by $\mu_p/\mu_n \ll \mu_n/\mu_p$, decreasing the detectivity. The large contrast between the electron minority mobility and the hole majority mobility is an essential ingredient for high detectivity photoconductor devices. As $\eta = 1 - \mathrm{e}^{-\alpha d}$, we see that there exists an optimum value for the ratio $\eta/\sqrt{d}$ close to $0.66\sqrt{\alpha}$. The maximum optical efficiency of a semiconductor in terms of detectivity is therefore 66%. Consequently, the maximum detectivity is given by:

$$D^* \approx 0.66 \frac{1}{h\nu} \sqrt{\frac{\alpha\tau}{4p_0} \frac{\mu_n}{\mu_p}} \tag{11.28}$$

Maximum detectivity for a photoconductor

This expression is interesting as it is only a function of the material and the wavelength. We see, therefore, that the detectivity of the material increases as the dark carrier density p_0 decreases. However, infrared detectors which work over the atmospheric transparency windows between 3–5 and 8–12 μm (see Complement 2.B), necessarily have bandgaps of the order of 250 and 120 meV, respectively. Equation (5.49) tells us that at usual temperatures (i.e. for $kT/q > 25$ meV), the intrinsic carrier densities n_i, due to the thermal excitation of electrons above the gap, are considerable ($n_i \gg 10^{16}\,\text{cm}^{-3}$) and result in low detectivities. Quantum detectors must therefore be cooled to lower temperatures if their intended detection spectrum is to be extended to longer wavelengths. Figure 11.B.4 compares the detectivities for several materials, and clearly demonstrates this behaviour.

The example below illustrates the use of expression (11.28) for detectivity.

Example

We wish to obtain the detectivity for an InSb photoconductor. At room temperature, the intrinsic carrier density calculated from (5.49) is $10^{16}\,\text{cm}^{-3}$. In order to improve the detector performance, it will be operated at a low enough temperature to keep the free carrier density (p_0) at $10^{14}\,\text{cm}^{-3}$. Assuming an absorption coefficient α of $10^4\,\text{cm}^{-1}$ and the values given in Table 11.1, we obtain a photon detectivity at 6.6 μm (0.2 eV) of:

$$D^* \approx 0.66 \frac{1}{0.2 \times 1.6 \times 10^{-19}\,\text{J}} \sqrt{\frac{10^4\,\text{cm}^{-1} \times 10^{-7}\,\text{s}^{-1}}{4 \times 10^{14}\,\text{cm}^{-3}} \times \frac{10^5}{1700}}$$

$$= 2.5\ 10^{11}\,\text{cm Hz}^{1/2}\,\text{W}^{-1}$$

We then calculate the noise equivalent power (NEP) corresponding to a 50×50 μm detection area with a 10 ms integration time, i.e. $\Delta\nu = 50$ Hz:

$$\text{NEP} = 50 \times 10^{-4} \sqrt{50}\,\text{Hz}^{1/2}/2.5 \times 10^{11}\,\text{jones} = 0.14\,\text{pW}$$

11.3.3 Time response of a photoconductor

The time response of a photoconductor subject to a time varying optical signal $G_{\text{op}}(t)$ can be derived from (11.3a,b). The essential characteristics can be easily obtained by assuming uniform illumination. Integrating (11.3a) over the thickness of the material (from 0 to d as in (11.14a)) and by taking into account the quantum efficiency, we immediately obtain the dynamic equation:

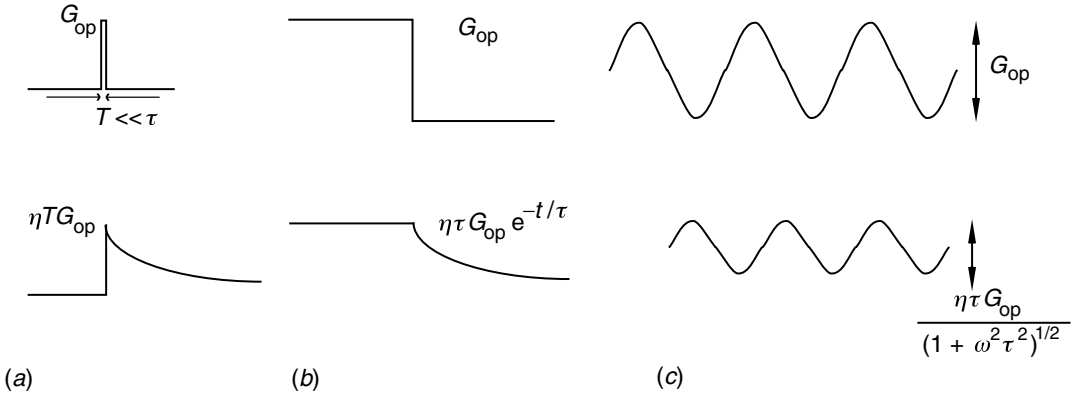


Fig. 11.7. Time response for a photoconductor subject to (*a*) abrupt pulse of illumination, (*b*) abrupt cessation of long pulse of illumination, and (*c*) sinusoidally varying illumination.

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta n_{\mathrm{tot}} + \frac{\Delta n_{\mathrm{tot}}}{\tau} = \eta G_{\mathrm{op}}(t) \tag{11.29}$$

Several possible cases are illustrated in Fig. 11.7:

- *Square light pulse starts at* $\mathrm{t} = 0$ *and ends at* $\mathrm{t} = \mathrm{T} \ll \tau$ *(Fig. 11.7*a*).* Equation (11.29) shows that the response of the material is immediate, but without gain: a density Δn_{tot} instantaneously arises in the material, with the density given by the total number of photons absorbed, or $\Delta n_{\mathrm{tot}} = \eta G_{\mathrm{op}} T$. The density subsequently decays exponentially with the characteristic lifetime τ.
- *Long square light pulse ends at* $\mathrm{t} = 0$ *(Fig. 11.7*b*).* The carrier density given by (11.29) is:

$$\Delta n_{\mathrm{tot}} = \eta \tau G_{\mathrm{op}} \mathrm{e}^{-t/\tau} \tag{11.30}$$

  The time response of the photoconductor is therefore equal to the lifetime in the material.
- *Sinusoidally varying illumination* $\mathrm{G}_{op}(\mathrm{t}) = \mathrm{G}_{op} \sin \omega \mathrm{t}$. Integration of (11.29) is immediate and leads to a sinusoidal response $\Delta n_{\mathrm{tot}}(t) = \Delta n_{\mathrm{tot}}(\omega)\sin(\omega t + \phi)$ with:

$$\Delta n_{\mathrm{tot}}(\omega) = \frac{\eta G_{\mathrm{op}} \tau}{\sqrt{1 + \omega^2 \tau^2}} \tag{11.31}$$

  The photoconductor therefore behaves as a low-pass filter with a frequency cut-off of $1/2\pi\tau$.

These last three results demonstrate the particular dilemma faced by photoconductors. Materials with greater detectivities also have long lifetimes and simultaneously suffer from poor temporal responsivity. Figure 11.8 shows the frequency

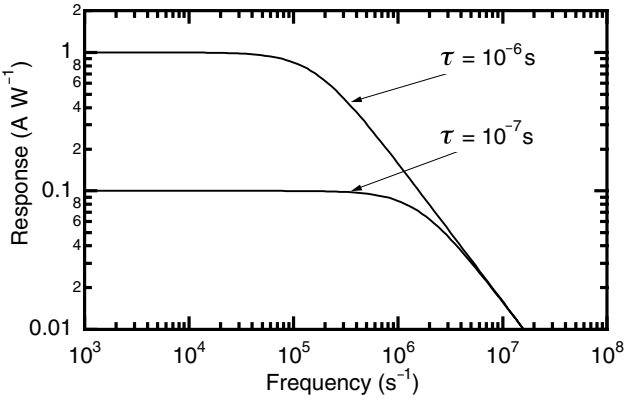


Fig. 11.8. Frequency response for photoconductors with different lifetimes. The product of the gain × bandwidth is constant for photoconductors.

response $\mathscr{R}$ of a photoconductor with two different carrier lifetimes. This result is an illustration of the maxim: *gain* × *bandwidth* = *constant*.

11.4 Photovoltaic detectors

One of the difficulties governing the use of photoconductive detectors is their low impedance. A photoconductor is fundamentally a light-sensitive resistor. As electrical amplification of a signal preferably involves a high input impedance, the combined detector and amplifier elements form an impedance bridge, placing the detector at a net disadvantage. This is not the case with a photodiode.

A *photovoltaic detector* exploits the rectifying characteristics of a *p–n* diode or a Schottky junction to obtain high impedances and to separate electron–hole pairs produced by the light absorption processes. It is therefore the internal field in the diode which is responsible for circulating the current. Figure 11.9 illustrates the operation of a photodiode. We have taken a p^+–n diode in which the bulk of the built-in voltage drop occurs in the n-type material (see Section 10.4). The light enters through the top surface and crosses the p^+ region (assumed to be sufficiently thin to neglect light absorption in this layer). The photons are absorbed in the n region, and give rise to electron–hole pairs in the space charge region (SCR) and in the bulk of the material. The photogenerated pairs in the SCR are immediately separated by the internal electric field which sweeps the holes towards the surface (this constitutes the *generated photocurrent* – J_G). The pairs produced outside the SCR can diffuse within the structure, with some portion of them being able to reach the SCR and undergo charge separation giving rise to a *diffusion current density* J_{diff}.

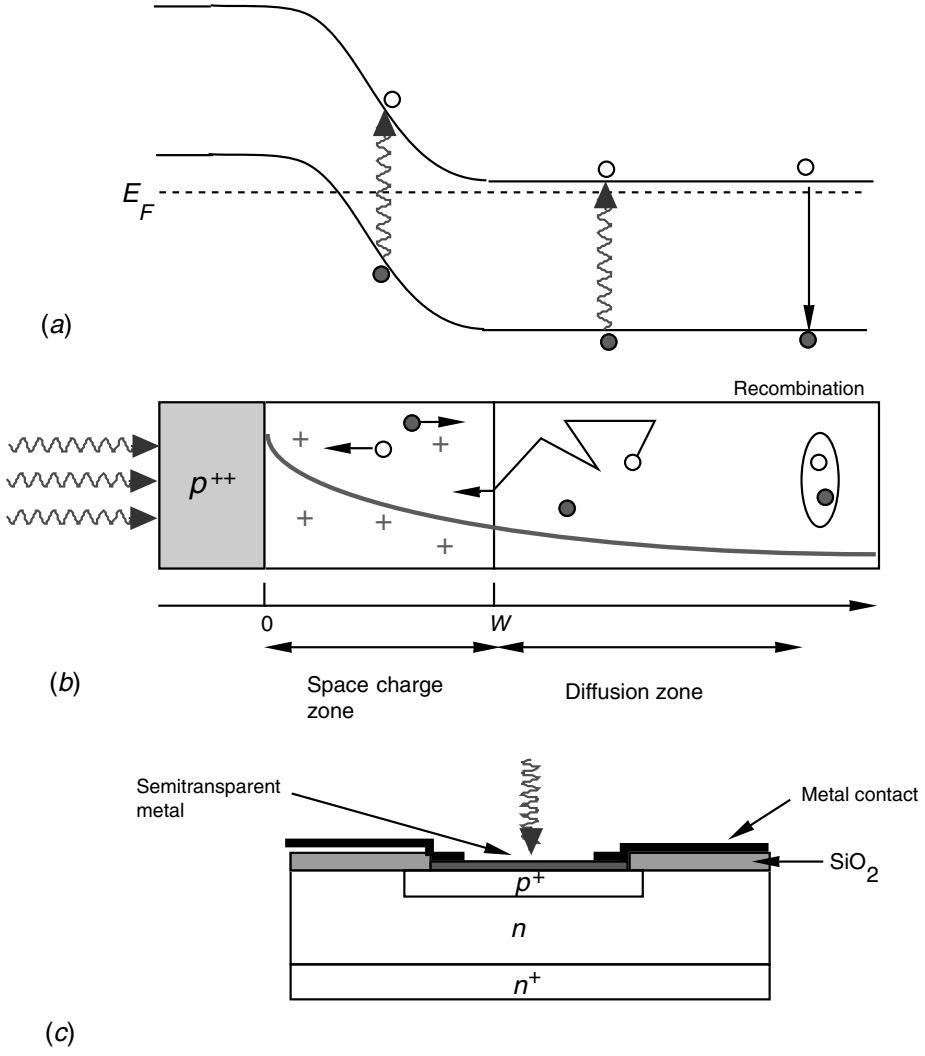


Fig. 11.9. Illustrated operation of a photodiode: (*a*) band diagram, (*b*) absorption geometry, and (*c*) typical device structure.

These two contributions create a total photocurrent density J_{ph} which adds to the *dark current*:

$$J = J_{\text{sat}}(\text{e}^{qV/kT} - 1) - J_{\text{ph}} \tag{11.32}$$

where J_{sat} is the diode saturation current given in (10.24) for a Schottky diode and the *hole diffusion limited current* given in (10.52) for a p^+–n diode. V is the applied potential to the diode. Figure 11.10 shows the characteristics for the diode under illumination. Two detection modes are possible:

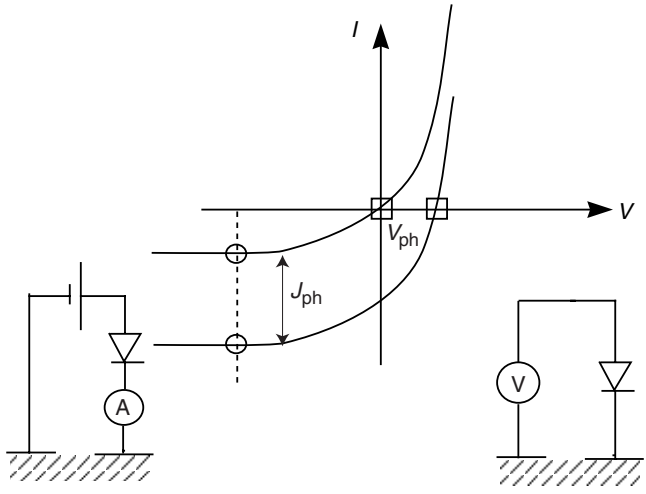


Fig. 11.10. Two operation modes for a photovoltaic detector: (*a*) photocurrent mode, and (*b*) photovoltage mode.

- Photocurrent mode. A diode can be introduced into a circuit with a very low impedance (e.g. placed in series with an ammeter or in series with a low resistor with a voltmeter hooked to its terminals – as in Fig. 11.10*a*). The current measured in this case is then given by (11.32).
- Photovoltage mode. Here, the diode is introduced into a high impedance circuit ($I = 0$). In this case, (11.32) shows that a voltage V_{ph} appears at the diode terminals. The voltage V_{ph} is given in the inverse form as:

$$V_{ph} = \frac{kT}{q} \log\left(1 + \frac{J_{ph}}{J_{sat}}\right) \tag{11.33}$$

The calculation of the photocurrent density J_{ph} follows from the equations in Section 11.3. This is a good exercise which allows one to become familiar with the application of the various boundary conditions (11.10). Let us begin by calculating the current J_G in the space charge region, i.e. between $z = 0$ and $z = W$. In the SCR, the density of excess charges Δn is zero, as all the carriers are instantaneously swept out of this region, and the dynamic equation (11.3a) becomes:

$$\frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z} J_G = -\alpha \Phi_0 \mathrm{e}^{-\alpha z} \tag{11.34}$$

We obtain the current density J_G by integrating (11.34) between 0 and W and by noticing that no electron current can flow from the p^+ region as the electron

concentration can only be zero ($np^+ = n_i^2$), so that $J_G(0) = 0$ and:

$$J_G = -q\Phi_0(1 - e^{-\alpha W}) \tag{11.35}$$

This is the electron current that crosses the boundary between the SCR and the neutral material at $z > W$.

The contribution of the diffusion current is obtained from the differential equations (11.3a) or (11.7) which at stationary state admit as a solution expression (11.8), which we rewrite here as:

$$\Delta n(z) = Ae^{-z/L_D} + Be^{z/L_D} + \frac{\alpha\tau\Phi_0 e^{-\alpha W}}{1 - (\alpha L_D)^2} e^{-\alpha z} \tag{11.36}$$

given that the photon flux at $z = W$ is now only $\Phi_0 e^{-\alpha W}$. We may suppose that the sample is sufficiently thick to assume $B = 0$. Furthermore, every electron at $z = W$ is immediately swept by the electric field imposed by boundary condition (11.10b) $\Delta n(W) = 0$ such that:

$$\Delta n(z) = \frac{\alpha\tau\Phi_0 e^{-\alpha W}}{1 - (\alpha L_D)^2} (e^{-\alpha(z-W)} - e^{-(z-W)/L_D}) \tag{11.37}$$

The diffusion current density at $z = W$ is given by:

$$J_{\text{diff}} = +qD_n \frac{d}{dz} \Delta n|_{z=W} \tag{11.38}$$

that is

$$J_{\text{diff}} = -q \frac{\alpha L_D}{1 + \alpha L_D} \Phi_0 e^{-\alpha W} \tag{11.39}$$

For the relatively common condition where $\alpha W \approx 1$, the contributions made by J_G and J_{diff} ((11.35) and (11.39), respectively) are comparable so that neither of the two terms can be neglected. The total photocurrent is then given by:

$$J_{\text{ph}} = -q\Phi_0 \left(1 - \frac{e^{-\alpha W}}{1 + \alpha L_D}\right) \tag{11.40}$$

This formula indicates that the low quantum efficiency of the space charge region $e^{-\alpha W}$ can be regained once the diffusion length becomes large, i.e. what the space charge region lets through, is recuperated by the diffusion region. Expression (11.40) can, as in the photoconductor case, be put into the form:

$$\mathscr{R} = \eta \frac{1}{h\nu/q} \tag{11.41}$$

or alternatively as:

$$\mathscr{R} = \eta \frac{\lambda(\mu\text{m})}{1.24} \tag{11.42}$$

with a total *quantum efficiency* η given by:

$$\eta = (1 - R)\left(1 - \frac{e^{-\alpha W}}{1 + \alpha L_D}\right) \tag{11.43}$$

This formula is therefore quite similar to (11.18). We note, however, the absence of photoconductive gain, which is equivalent to saying that the gain g in a photodiode is unity.

Example
In *n*-type silicon doped $10^{16}\,\text{cm}^{-3}$, the SCR extends across 1 μm for a junction potential $V_{\text{bi}} = 1$ V. Figure 11.2 shows an absorption coefficient $\alpha = 10^4\,\text{cm}^{-1}$ at 2 eV. The efficiency of the SCR $1 - e^{-\alpha W}$ is then 64%. On the other hand, the diffusion length is $\sqrt{((kT/q)\mu_n\tau)}$ or 500 μm (see Table 11.1), which leads to a quantum efficiency dominated by the reflectivity $(1 - R)$.

11.4.1 Photodiode detectivity

As discussed in Complements 11.A and 11.B, two mechanisms are responsible for noise in photovoltaic detectors: generation noise (without recombination) i_G (11.A.38) and photon arrival fluctuation noise (11.B.6). Again, we will make the assumption that the detector noise is not dominated by photon noise (i.e. that the detector is not in the BLIP regime – see Complement 11.B). The noise due to the detector itself is then:

$$i_N^2 = 2qI_0\Delta v = 2qI_{\text{sat}}(1 + e^{qV/kT})\Delta v \tag{11.44}$$

where I_{sat} is the saturation current of the photodiode. The general idea behind (11.44) is that the 'plus' sign in (11.44) results from the fact that the forward and reverse contributions cancel out at $V = 0$, while their respective contributions to the noise are additive. The signal-to-noise ratio (S/N) is given by:

$$\frac{i_S}{i_N} = \frac{\mathscr{R}P_{\text{inc}}}{\sqrt{2qI_{\text{sat}}(1 + e^{qV/kT})\Delta v}} \tag{11.45}$$

and the *noise equivalent power* (NEP) takes the form of:

$$\text{NEP} = \frac{\sqrt{2qI_{\text{sat}}(1 + e^{qV/kT})\Delta v}}{\mathscr{R}} \tag{11.46}$$

We reintroduce the *detectivity* D^* as it allows us to get rid of parameters that are not intrinsic to the system: (i) the area A, which relates the current I_{sat} to the

current density J_{sat} ($I_{sat} = AJ_{sat}$), and the flux Φ_0 to the incident power P_{inc} ($P_{inc} = A\Phi_0 h\nu$); and (ii) the frequency measurement bandwidth $\Delta\nu$:

$$D^* = \frac{\mathscr{R}}{\sqrt{2qJ_{sat}(1 + e^{qV/kT})}} \tag{11.47a}$$

We insist again on the fact that this time, the detection area and the transport area are one and the same. The photon and electron fluxes are parallel.

We may think it profitable to bias the diode ($V < 0$) to decrease the factor $e^{qV/kT}$ and increase the photodiode detectivity. This bias, however, often adds additional noise (peripheral leakage current, . . .) to the detection signal. Under zero applied bias ($V = 0$), (11.47a) becomes:

$$D^* = \frac{\eta}{h\nu/q} \frac{1}{\sqrt{4qJ_{sat}}} \tag{11.47b}$$

This expression can be written in another form, by noting that:

$$\left.\frac{dj}{dV}\right|_{V=0} = \frac{1}{RA} = \frac{q}{kT} J_{sat} \tag{11.48}$$

where RA ($\Omega\,cm^2$) is the product of the junction resistance with its area such that:

$$D^* = \frac{\eta}{h\nu/q} \sqrt{\frac{RA}{4kT}} \tag{11.49}$$

Photodiode detectivity

Figure 11.11 shows how detectivity D^* varies as a function of the product RA at different temperatures for a detection wavelength of 5 μm ($h\nu = 250$ meV) and a quantum efficiency of 50%.

This last formula is quite predictive when coupled with the expression for the current density j_{sat} in a junction (10.B.4), i.e:

$$j_{sat} = \frac{qD_p}{L_{Dp}} \frac{n_i^2}{N_D} = q\sqrt{\frac{D_p}{\tau_p}} \frac{n_i^2}{N_D} = q\sqrt{\frac{D_p}{\tau_p}} \frac{N_c N_v}{N_D} e^{-E_g/kT} \tag{11.50}$$

where we recall that D_p, L_{Dp}, and τ_p are the diffusivity, diffusion length, and lifetime of the minority carriers (holes) on the n side of the junction; N_D is the doping concentration; n_i is the intrinsic carrier density given in (5.49); E_g is the detector bandgap; and N_c and N_v are the effective state densities in the conduction and valence bands given in (5.46). Therefore the detectivity outside of the detector's BLIP regime varies as $e^{-E_g/2kT}$, which emphasizes the interest in operating these detectors at low temperature. We can also use (10.B.3) if the diode leakage current is dominated by generation–recombination occurring at impurity sites.

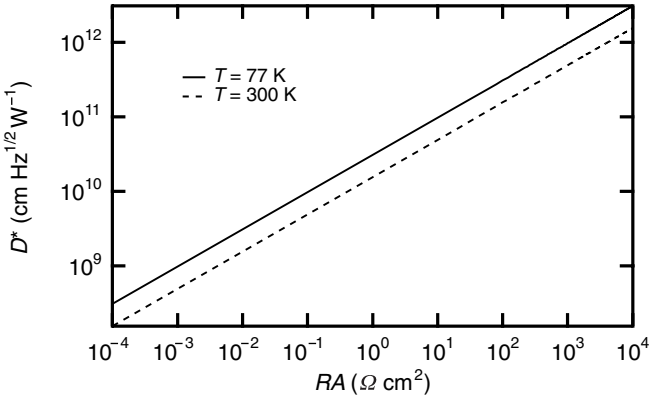


Fig. 11.11. Variation in photodiode detectivity at 5 μm as a function of RA (a quantum efficiency of 50% is assumed).

Example

We wish to calculate the RA product and detectivity for an InSb photodiode doped n type ($10^{16}\,\mathrm{cm}^{-3}$) at a temperature of 150 K. The temperature dependence of the bandgap will be neglected. Table 11.1 gives us the diffusion coefficient at 150 K ($kT/q = 13\,\mathrm{meV}$) or $1700 \times 13 \times 10^{-3} = 22\,\mathrm{cm}^2\,\mathrm{s}^{-1}$. Equations (5.46) give the state densities N_c and N_v for $m_c = 0.0145$ and $m_v = 0.4$, i.e. $N_c = 1.5 \times 10^{16}\,\mathrm{cm}^{-3}$ and $N_v = 2.2 \times 10^{18}\,\mathrm{cm}^{-3}$. The saturation current for an ideal diode is given by (11.50), or:

$$j_{\mathrm{sat}} = 1.6 \times 10^{-19}\,\mathrm{A\,s^1}\sqrt{\frac{22\,\mathrm{cm^2\,s^{-1}}}{10^{-7}\,\mathrm{s}}}\,\frac{1.5 \times 10^{16} \times 2.2 \times 10^{18}}{10^{16}}\,\mathrm{cm^{-3}}\,\mathrm{e}^{-0.18/0.013}$$

$$= 7.6 \times 10^{-3}\,\mathrm{A\,cm^{-2}}$$

and the RA product from $(kT/q)i_{\mathrm{sat}}$ is $1.7\,\Omega\,\mathrm{cm}^2$. For an overall quantum efficiency $\eta = 0.5$, the detectivity is then given by:

$$D^* = \frac{0.5}{0.25\,\mathrm{eV}}\,\frac{1}{\sqrt{4 \times 1.6 \times 10^{-19}\,\mathrm{C}\,7.6 \times 10^{-3}\,\mathrm{A\,cm^{-2}}}}$$

$$= 2.8 \times 10^{10}\,\mathrm{cm\,Hz^{1/2}\,W^{-1}}$$

Figure 11.12 shows the variation of detectivity D^* for this device as a function of temperature. From this graph, we see that it is clearly advantageous to operate infrared detectors at low temperature.

11.4.2 Time response of a photodiode

The time response of a photodiode can be obtained by solving the dynamic

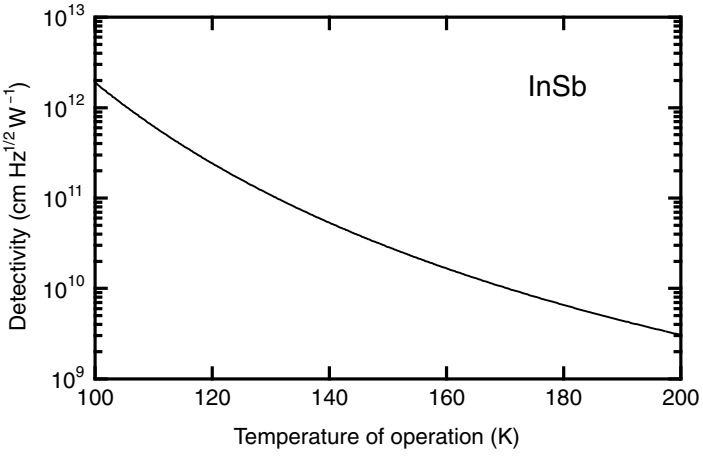


Fig. 11.12. Detectivity as a function of temperature for a *p–n* InSb photodiode.

equations (11.3) and (11.4). Whereas the calculation itself is demanding, the result is quite easy to guess. Three time constants contribute to the overall temporal response of a photodiode, with each resulting from a distinct physical effect:

- *Time constant due to the diffusion current*. This constant is clearly the lifetime τ. It varies between 10^{-4} s in silicon and 10^{-8} s in GaAs (corresponding to bandwidths ranging between 10 kHz and 100 MHz). In many cases (for instance, in telecommunications applications), a time constant in this range is unacceptable.
- *Capacitive time constant*. This same space charge region gives rise to a capacitance C_d given by (10.17):

$$C_d = A\sqrt{\frac{q\varepsilon N_D}{2(V_{\text{bi}} - V)}} = \frac{A\varepsilon}{W} \tag{11.51}$$

where we recall that $\varepsilon = \varepsilon_0\varepsilon_r$ is the permittivity of the medium, V_{bi} is the junction potential ($V_{\text{bi}} = V_D = \phi_B - (E_c - E_F)$ for a Schottky junction as in Fig. 10.2 or as given by (10.25) in the case of a *p–n* junction), W is the thickness of the depletion layer, and N_D is the doping concentration. For rapid operation, this capacitance in series with a matched circuit with impedance $R = 50\,\Omega$ leads to an upper frequency limit given by $1/2\pi RC$.

Example

For a typical detector with an area of 1 mm^2, doped 10^{15} cm^{-3}, and with a potential $V_{\text{bi}} - V = 10$ V, we obtain a 3.6 μm ($= W$) wide SCR and a capacitance of 30 pF. Loaded with a 50-Ω resistance, the cut-off frequency for this detector is $1/2\pi RC = 110$ MHz.

- *Time constant due to carrier transport across the space charge region.* The SCR extends over a distance of the order of 1 μm for an applied potential of ~1 V or a

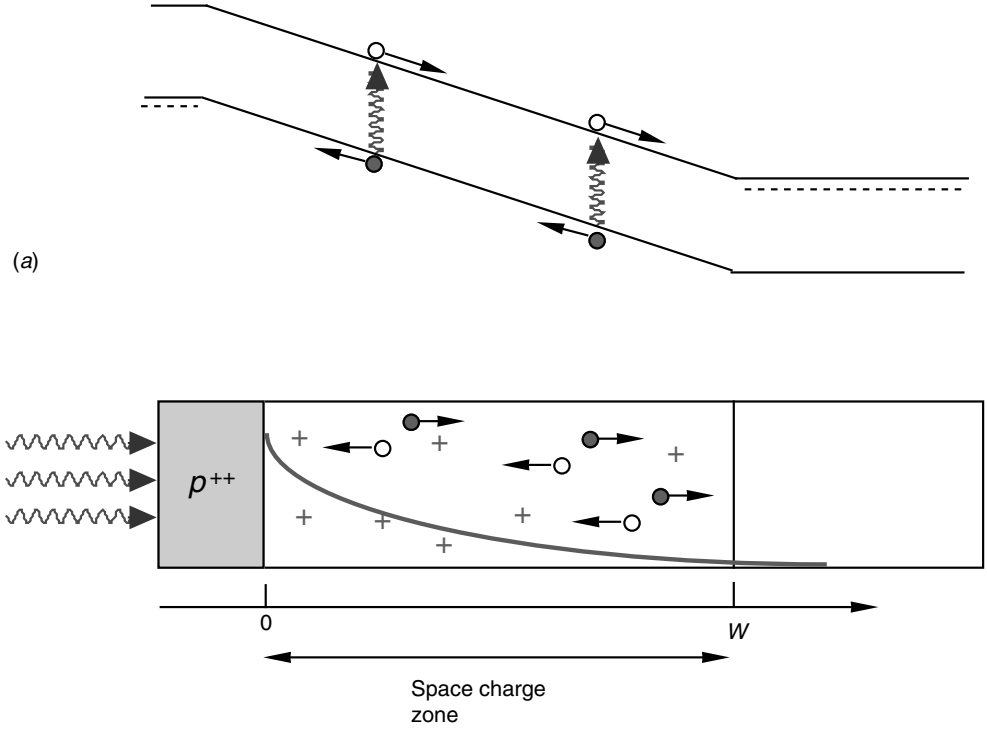


Fig. 11.13. Operation of a *p–i–n* junction.

mean electric field of $10^4\,\mathrm{V\,cm^{-1}}$. Even for a material with a relatively low mobility (say $\mu = 10^2\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$), this corresponds to a velocity of $10^6\,\mathrm{cm\,s^{-1}}$ and therefore a carrier transit time of 100 ps and cut-off frequency of 10 GHz.

The first two limitations are unacceptable in most telecommunications applications which require bandpass frequencies of the order of tens of GHz or more. In these applications, *p–i–n* diodes, see Fig. 11.13, possess an intrinsic region with a large width W, sandwiched between two highly doped n and p regions. These diodes were examined in Section 10.A.2 (see band diagram in Fig. 10.A.2). This device structure presents the following advantages:

- It possesses a low capacitance $C_d = A\varepsilon/W$, and therefore an elevated cut-off frequency $(1/2\pi RC)$.
- The contribution made by the diffusion current J_{diff} is negligible ($\alpha W \gg 1$ in (11.40)), as is the influence of the time constant τ on the response time of the diode. The temporal response of the device is dominated by the generation current in the SCR.

Example
For a typical detector with a detection area of 0.01 mm^2, doped $10^{14}\,cm^{-3}$, and with a potential $V_{bi} - V = 100$ V, we obtain a SCR width W of 36 μm. Contenting ourselves with an intrinsic region width of 10 μm, we obtain a device capacitance of 0.1 pF. In series with a circuit impedance of 50 Ω, such a detector would possess a frequency cut-off of $1/2\pi RC$ or 30 GHz. The average electric field is $10^2/10$ μm or $10^5\,V\,cm^{-1}$ leading to a velocity of the order of $10^7\,cm\,s^{-1}$ assuming a mobility $\mu = 100\,cm^2\,V^{-1}\,s^{-1}$. This results in a transit time of 10 μm per $10^7\,cm\,s^{-1}$ or 100 ps, and a frequency cut-off of 10 GHz.

11.5 Internal emission photodetector

The operation of an *internal emission Schottky photodiode* is depicted in Fig. 11.14. As discussed at length in Section 10.3, a potential barrier ϕ_{ms} naturally resides between the Fermi level of the metallic layer and the bottom of the semiconductor conduction band. We will now consider a beam of light travelling across the semiconductor, composed of photons with energy hv less than the semiconductor bandgap (making the semiconductor essentially transparent to the electromagnetic radiation). The light is absorbed by the metallic layer, transferring its energy to the electrons in the Fermi sea. Electrons can then be emitted into the semiconductor above the potential barrier. This leads immediately to a cut-off wavelength for this device of:

$$\lambda_c = \frac{1.24}{\phi_{ms}(\text{eV})}\,\mu\text{m} \tag{11.52}$$

The quantum yield is rather low in these structures, as in the metal, the electrons lose their energy over very short distances. Indeed, the *ballistic mean free path* L_B is only between 50 and 100 Å in most metals. The absorption coefficient for metals is given by the inverse of the *skin depth* δ (see (7.C.6)): for a metal with a resistivity of the order of 10 μΩ cm, this leads to a skin depth δ of 200 Å at infrared wavelengths, and an absorption coefficient $\alpha = 1/\delta = 5 \times 10^5\,cm^{-1}$. We must therefore expect a photon conversion rate into hot electrons in the metal of $1 - e^{-\alpha L_B}$, i.e. between 10 and 20%. To increase the absorption, the metal/semiconductor structure is usually placed inside an optical cavity (see Fig. 11.14).

Furthermore, all the hot electrons excited above the barrier are not necessarily destined to be recuperated by the semiconductor, as they must of course be travelling in an appropriate direction. This is illustrated in Fig. 11.15.

We will consider those electrons which have a total energy E and relative momentum $P \sim E^{1/2}$ above the Fermi level of the metal. The hot electrons whose

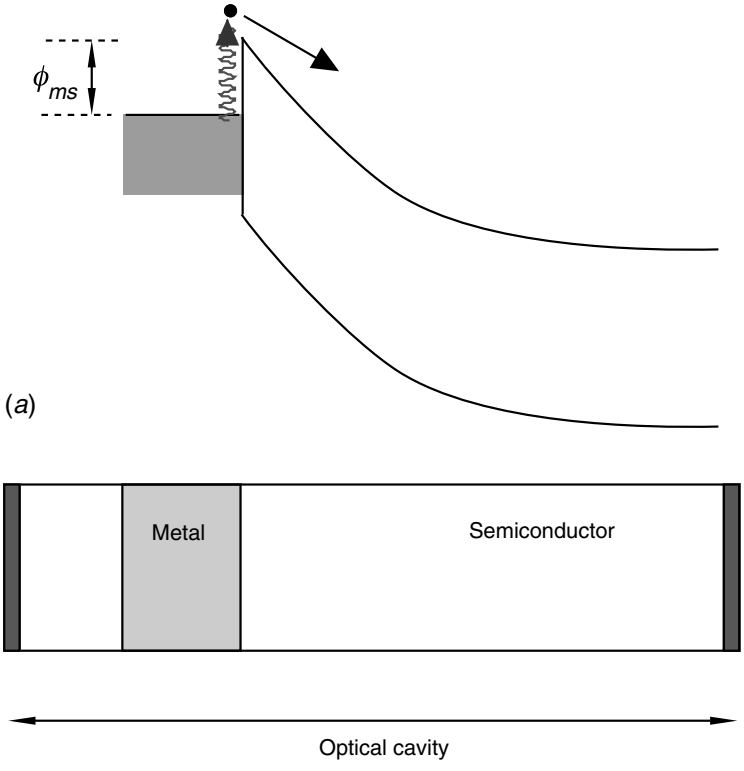


Fig. 11.14. Operation of an internal emission Schottky photodiode.

momenta are oriented at an angle θ from the normal direction of the interface possess a momentum component of $P\cos\theta$ and therefore an energy of $E\cos^2\theta$ in the appropriate direction (the Oz direction). They are a priori equally distributed with respect to θ. Over a solid angle of 4π, only a subset $P(E)$ will therefore be able to surmount the potential barrier (see Fig. 11.15):

$$P(E) = \frac{2\pi(1-\cos\theta)}{4\pi} = \frac{1-(\phi_{ms}/E)^{1/2}}{2} \tag{11.53}$$

We will then suppose that the density of states in the metal ρ_m (in $\mathrm{cm^{-3}\,eV^{-1}}$) is constant. The electronic flux effectively transferred to the semiconductor is then:

$$\frac{i_{\mathrm{ph}}}{q} = \int_{\phi_{ms}}^{hv} \rho_m P(E)\mathrm{d}E = \frac{1}{2}\rho_m hv\left[1-\left(\frac{\phi_{ms}}{hv}\right)^{1/2}\right]^2 \tag{11.54}$$

The *internal quantum efficiency* is given by the ratio between the flux i_{ph}/q and the absorbed photon flux $\rho_m hv$, i.e:

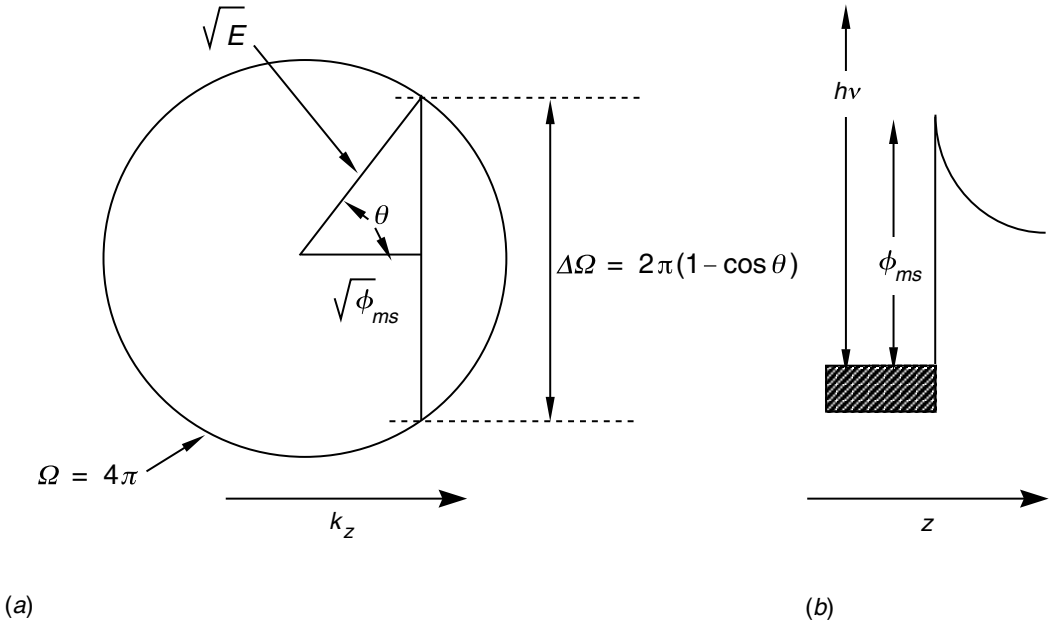


Fig. 11.15. (*a*) Solid angle restriction for electron recuperation in a Schottky photodiode. (*b*) Only electrons possessing energies between ϕ_{ms} and $h\nu$ are collected.

$$\eta_{\text{int}} = \frac{1}{2}\left[1 - \left(\frac{\phi_{ms}}{h\nu}\right)^{1/2}\right]^2 \tag{11.55}$$

which leads to a total quantum efficiency η of:

$$\eta = \eta_{\text{ext}}\eta_{\text{int}} \approx \frac{1}{2}(1 - \mathrm{e}^{-\alpha L_B})\left[1 - \left(\frac{\phi_{ms}}{h\nu}\right)^{1/2}\right]^2 \tag{11.56}$$

We have to note that (11.56) – derived using this approach – is extremely simple and is not reproduced by experiment. Given that the density of states in a three-dimensional metallic medium goes as $\rho_m \propto \sqrt{E}$ (see for example Eq. (5.20)) and taking into account the Pauli exclusion principle in the metal, we could in fact find:

$$\eta_{\text{int}} = \frac{1}{8E_F h\nu}(h\nu - \phi_{ms})^2$$

where E_F is the Fermi energy in the metal. This last expression is confirmed by experiment and is known as *Fowler's law*. Figure 11.16 shows the efficiency of a Si/IrSi internal emission diode which obeys Fowler's law. Expression (11.55) still leads to reasonable results for $h\nu \approx \phi_{ms}$.

Equation (11.56) explains why Schottky detectors are not intrinsically good

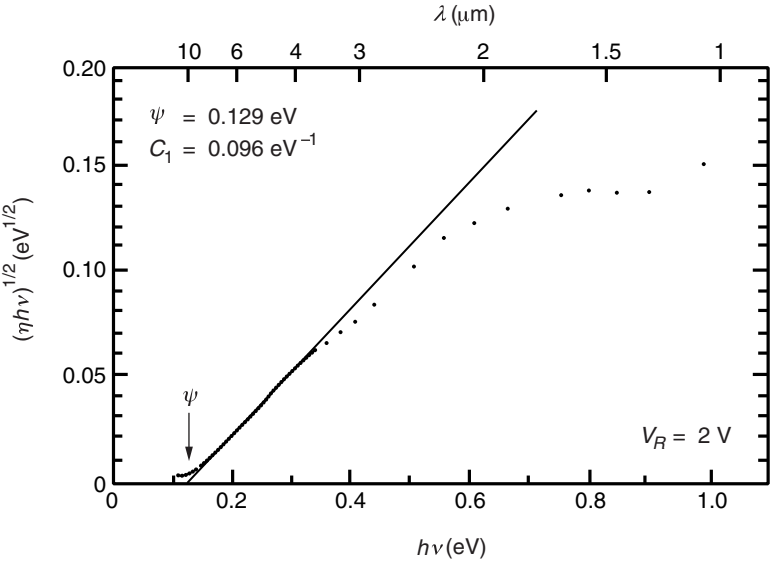


Fig. 11.16. Fowler plot of the yield V for a p-type Si/IrSi internal emission Schottky photodiode (after B. Y. Tsaur, M. M. Weeks, R. Trubiano, P. W. Pelegrini, and T. R. Yew, *IEEE Electron. Dev. Lett*. **EDL-9**, 650 (© 1988 IEEE)).

detectors. First, given the small ballistic mean free path for hot electrons in metals, the external quantum efficiencies are low, as we indicated at the beginning of this section (i.e. ∼10%). Additionally, 10% above the cut-off, i.e. for $h\nu \approx 1.1\phi_{ms}$, the internal quantum efficiency is 10^{-3}, which leads to a global quantum efficiency of 10^{-4} – far from spectacular to say the least!

Nonetheless, these detectors have seen a fair level of success in the infrared imaging industry; as in these devices, the range of detected wavelengths is not imposed by the semiconductor, but by the height of the barrier between the metal and the semiconductor. Silicon (semiconductor *extraordinaire* in microelectronics!) can then be used to fabricate integrated detector arrays alongside their measurement circuits. In this case, the only challenge is to find a good Si/metal pair. For example, p-type Si/PtSi possesses a cut-off wavelength of 5.6 μm. Detector arrays consisting of 1024 × 1024 pixels have been fabricated using such detectors, allowing the acquisition of high-resolution infrared images (with low detectivity).

11.6 Quantum well photodetectors (QWIPs)

In Section 8.6, we saw how the quantization of electron motion perpendicular to the interfaces in quantum wells based on, for instance, the GaAs/$Al_xGa_{1-x}As$ system lead to the emergence of energy subbands. If such wells are doped n type

(with an effective two-dimensional carrier concentration ρ_s in cm^{-2}), additional electrons populate the subbands according to Fermi–Dirac statistics, beginning with the fundamental subband e_1. If photons are incident upon the quantum well, they will be in a privileged position to promote electronic transitions between the conduction subbands.

Two kinds of optical transition can exist:

- *Transitions between two bound states* $\text{e}_1 \rightarrow \text{e}_2$ (bound-to-bound *transitions*). We saw in Section 8.6 that such transitions lead to (nearly) monoenergetic absorption features, i.e. only photons with energy within a spectrally broadened distribution $\hbar\Gamma$ about $h\nu = \varepsilon_2 - \varepsilon_1 = e_{12}$ are absorbed.

  The absorption coefficient within the well is given by (8.87b):

$$\alpha = \frac{\pi q^2 z_{12}^2}{\hbar} \rho_s \frac{Z_0}{n_{\text{sc}}} h\nu \frac{\hbar\Gamma/\pi}{(e_{12} - h\nu)^2 + (\hbar\Gamma)^2} f(\theta) \tag{11.57a}$$

  where we recall that Z_0 is the vacuum impedance (377 Ω), n_{sc} is the semiconductor's optical index, z_{12} is the dipolar matrix element (an approximation for it being given in (3.D.24) – a value typically on the order of 1 nm), and θ is the angle between the illumination axis and the normal vector of the quantum well. As seen in Section 8.7.2, the function $f(\theta)$ varies as $\sin^2\theta/\cos\theta$ as long as the cross-section of the light beam by the detector surface resides within the detector area and $\sin^2\theta$ otherwise. The $\sin^2\theta$ dependence emerges from the selection rules for $\mathbf{k}_{\|}$ as explained in Fig. 8.13.

- *Bound-to-free transitions.* Dominate when there is only one subband within the quantum well (see (3.D.27)). In this case, the transitions occur between localized states within the quantum well (energy difference and the bottom of the barrier conduction band $e_1 = E_c - \varepsilon_1$).

  The absorption coefficient is then given by integral (11.57a) evaluated over the continuum states $|k_z\rangle$ in the barrier conduction band. The states $|k_z\rangle$ have an energy $E_f = \hbar^2 k_z^2/2m_c$ above the bottom of the barrier conduction band, where m_c is the effective mass in the conduction band. We saw in Complement 1.A all the difficulties relating to normalization of the continuum states and the necessity of introducing a fictitious box of width L_{fict} to proceed with the normalization. To calculate the bound-to-free absorption in quantum wells, we must introduce the following integral:

$$\alpha = \frac{\pi q^2}{\hbar} \rho_s \frac{Z_0}{n_{\text{sc}}} \hbar\omega \int_0^\infty \frac{L_{\text{fict}}}{\pi} |\langle k_z|z|1\rangle|^2 \frac{\hbar\Gamma/\pi}{\left[\frac{\hbar^2 k_z^2}{2m_c} - (h\nu - e_1)\right]^2 + (\hbar\Gamma)^2} \text{d}k_z f(\theta) \tag{11.57b}$$

  where we recall that L_{fict}/π is the one-dimensional momentum density of states in the continuum (see Complement 1.A). Expression (11.57b) can be placed into

a form better suited to calculation by introducing the dipole volume $z(E)^3$ which emerges as the limit of $L_{\text{fict}}|\langle k_z|z|1\rangle|^2$ as $L_{\text{fict}} \to \infty$ and by integrating over the final state energies E_f, so that:

$$\alpha = \int_0^\infty \alpha_D(h\nu, E_f)\mathrm{d}E_f f(\theta) \tag{11.58a}$$

where α_D is the differential absorption defined as:

$$\alpha_D(h\nu, E_f) = \frac{q^2\sqrt{2m_c}}{2\hbar^2}\rho_s\frac{Z_0}{n_{\text{sc}}}h\nu\frac{z(E_f)^3}{\sqrt{E_f}}\frac{\hbar\Gamma/\pi}{[E_f - (h\nu - e_1)]^2 + (\hbar\Gamma)^2} \tag{11.58b}$$

The fine curves in Fig. 11.17 represent the energies $\varepsilon_2 - \varepsilon_1$ for *bound-to-bound intersubband transitions* in a GaAs/Al_xGa_{1-x}As quantum well as a function of the Al fraction x and the well thickness.

The thick curve represents the locus of the concentration–thickness values which lead to a *quasi-resonance* situation, i.e. where the energy of the first excited state e_2 sits just at the top of the well (and at the bottom of the barrier). This is the configuration that is used in quantum well detectors.

Indeed, in this case, the structure possesses a double advantage: the oscillator strength for such transitions is still important (being a close cousin of bound-to-bound transitions), while the promoted electron is free to propagate within the excited state (paralleling the behaviour of a bound-to-free transition). Figure 11.17 shows that we may therefore adjust the parameters of a quantum well to obtain the quasi-resonance condition over a wavelength range of 5 to 20 μm, i.e. spanning a portion of the far-infrared spectrum. The mid-infrared region can be accessed in a similar fashion by using InGaAs/AlGaAs quantum wells.

Electrons excited into the barrier continuum are swept out by an applied electric field leading to a photocurrent. The electric field E is produced by the external application of a suitable bias voltage across (for example) two highly doped GaAs contact layers. This multiple quantum well (or *multi-quantum well*) detector structure appears in Fig. 11.18. We shall see that increasing the number of quantum wells leads trivially to an increase in the total light absorption levels, but also to an increase in the overall quantum efficiency and detectivity of the device.

At sufficiently low temperatures, the electrons are trapped within the fundamental quantum well subband and, ideally, the system is insulating. Under the influence of a photon flux, electrons are photoionized out of the quantum wells and into the barrier regions leading to a photocurrent which can be measured by an ammeter. Figure 11.19 shows the calculated (using (11.58)) and measured absorption spectra for a multi-quantum well detector. It is worth noting, however, that

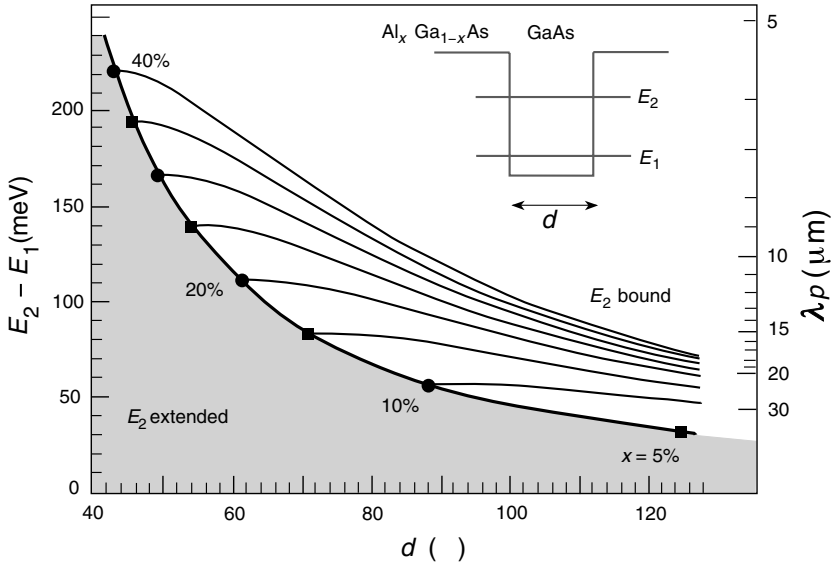


Fig. 11.17. Energies for bound-to-bound (narrow curves), bound-to-free (shaded region), and bound-to-quasi-resonant (thick curve) transitions in GaAs/Al$_x$Ga$_{1-x}$As quantum wells.

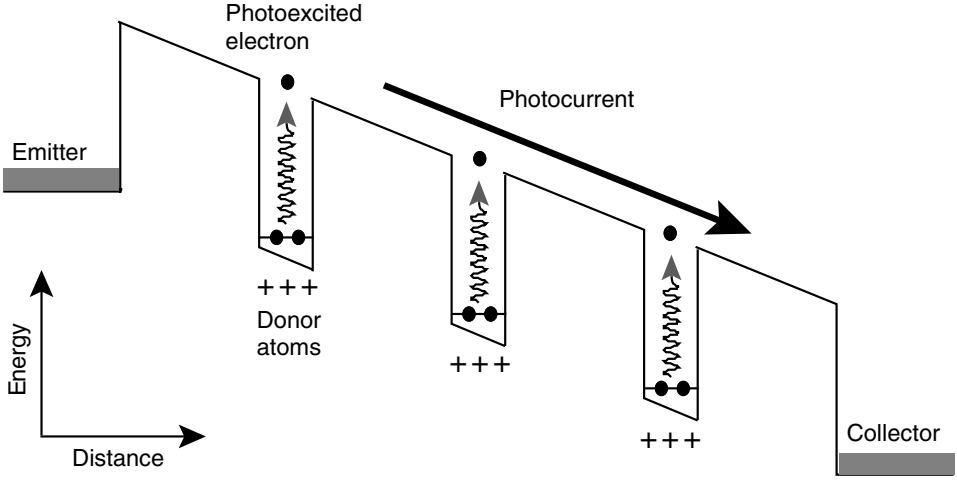


Fig. 11.18. Operation of a multiple quantum well detector.

relative to other semiconductor-based detector types, quantum well detectors are narrow bandwidth detectors.

Two models can be used to describe the operation of a quantum well photodetector.

- *The photoconductor model (Fig. 11.20*a). The generation rate per square cm in a single quantum well with an absorption coefficient α_{1QW} subject to a photon flux Φ_0 is:

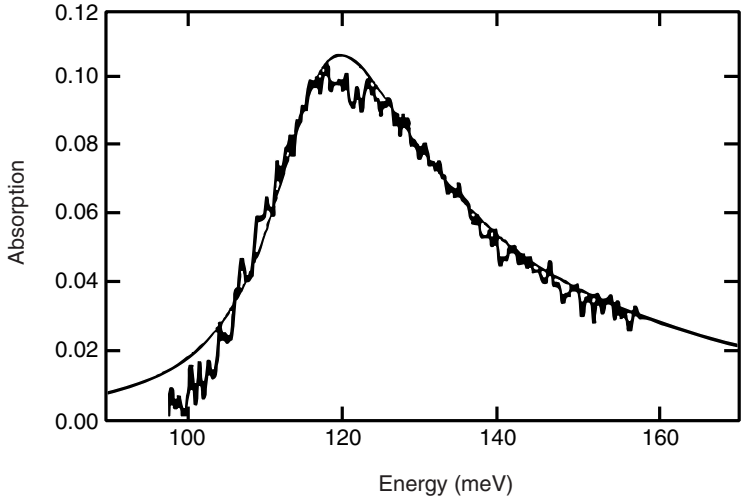


Fig. 11.19. Experimental and calculated absorption spectra for a multi-quantum well detector in a planar waveguide geometry ($\theta = 90°$). (Courtesy of F. Luc@THALES.)

$$G_{\text{op}} = p_e \alpha_{1\text{QW}} \Phi_0 \tag{11.59}$$

where p_e is the probability that a photoexcited electron will be successfully emitted into the barrier continuum and swept out by the electric field. At a stationary state, this rate must be compensated by the capture rate which keeps filling the quantum wells with electrons. The recombination rate is then given by:

$$R = \frac{\rho_{2\text{D}}}{\tau_c} \tag{11.60}$$

where τ_c is the capture time of an electron in the barrier by the quantum well and $\rho_{2\text{D}}$ is the two-dimensional density of photoexcited carriers. This *capture time* τ_c is dominated by Fröhlich interactions (see Complement 6.B) and is typically between 1 and 10 ps. Setting (11.59) and (11.60) equal to each other determines the two-dimensional density $\rho_{2\text{D}}$. The three-dimensional carrier density $\rho_{3\text{D}}$ can be obtained by assuming the carriers are uniformly distributed throughout the barrier material of thickness L, such that:

$$\rho_{3\text{D}} = \frac{\rho_{2\text{D}}}{L} = \frac{p_e \alpha_{1\text{QW}} \tau_c \Phi_0}{L} \tag{11.61}$$

Following the same reasoning as in Section 11.2, we find the *photoresponsivity* for a multi-quantum well detector to be given by:

$$\mathscr{R} = \eta_{1\mathrm{QW}} G_{1\mathrm{QW}} \frac{1}{hv/q}$$
$$\eta_{1\mathrm{QW}} = p_e \alpha_{1\mathrm{QW}} \text{ and } G_{1\mathrm{QW}} = \frac{\tau_c}{\tau_{\mathrm{tr}}} = \frac{\tau_c}{L/v_d} \tag{11.62}$$

where $\eta_{1\mathrm{QW}}$ is the *quantum efficiency* for a single quantum well, τ_{tr} is the transit time for an electron in a barrier of length L with a velocity $v_d = \mu E$, and $G_{1\mathrm{QW}}$ is the single period *photoconductive gain*. In this last expression, we note that the responsivity of a photodetector does not depend on the number of quantum wells. This rather surprising result is a consequence of the conservation of electrical current. On the other hand, we will see that the detectivity of an N well device is proportional to $\sqrt{N}$.

- *The photoemissive model (Fig. 11.20*b*)*. Using this model, the current is determined by the condition that, at stationary state, the carrier flux captured by the quantum wells $p_c J_{\mathrm{ph}}/q$ exactly balances out the flux of photoionized electrons outside the well $p_e \alpha_{1\mathrm{QW}} \Phi_0$. The coefficient p_c is the *capture probability* for an electron crossing the quantum well. Setting both fluxes equal, we immediately obtain for the detector photoresponsivity:

$$\mathscr{R} = \eta_{1\mathrm{QW}} G_{1\mathrm{QW}} \frac{1}{hv/q}$$
$$\eta_{1\mathrm{QW}} = p_e \alpha_{1\mathrm{QW}} \text{ and } G_{1\mathrm{QW}} = \frac{1}{p_c} \tag{11.63}$$

We see that the expressions for the responsivity of a quantum well detector given by each model are identical if:

$$\frac{1}{p_c} = v_d \frac{\tau_c}{L} \tag{11.64}$$

This last equation relates the quantum well capture time to the capture probability for an electron crossing a quantum well.

Example

We consider a quantum well structure subject to an electric field produced by 1 V across a distance of 1 μm, or $10^4\ \mathrm{V\,cm^{-1}}$. The carrier mobility in the barrier is $10^2\ \mathrm{cm^2\,V^{-1}\,s^{-1}}$, the barrier thickness between successive wells L is 250 Å, and the capture time τ_c is 10 ps. The velocity v_d is then $10^6\ \mathrm{cm\,s^{-1}}$ and the capture probability p_c is 0.25. The photoconductive gain of a quantum well $G_{1\mathrm{QW}}$ is then 4 assuming an ionization probability p_e of 1.

Using a grating coupler, the absorption coefficient for a typical quantum well is

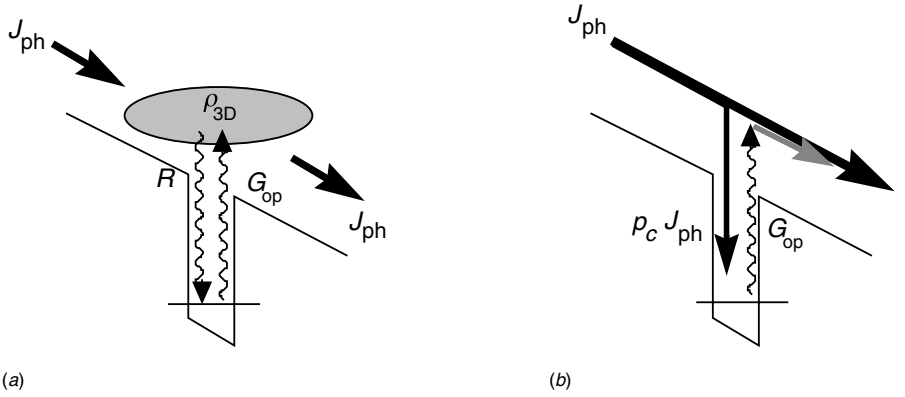


Fig. 11.20. Two models for electron transport across a multiple quantum well detector: (*a*) photoconduction model and (*b*) photoemission model.

$\alpha_{1QW} = 1\%$ for $\lambda = 10.6\,\mu m$ (117 meV). The responsivity for such a quantum well detector is then $\mathscr{R} = 4 \times 10^{-2}/0.117 = 0.34\,A\,W^{-1}$.

Addressing the detectivity aspects, we have the noise in an N well detector given by (11.A.42):

$$i_{GR}^2 = 2qI_0(2 - p_c)\frac{1}{Np_c} \tag{11.65}$$

The *dark current* I_0 can be obtained by first calculating the density $\rho_{2D,th}$ of thermally activated carriers in the barrier conduction band. We saw in (8.43) that this density is given by:

$$\rho_{3D,th} \approx \frac{1}{L}\frac{m_c}{\pi\hbar^2}\int_{E_c}^{\infty} e^{-(E-E_F)/kT}dE = \frac{1}{L}\frac{m_c kT}{\pi\hbar^2}e^{-(E_c-E_F)/kT} = \frac{n_c}{L}e^{-(E_c-E_F)/kT} \tag{11.66}$$

where n_c is the *critical density* (8.43) in the conduction band, and the Fermi level E_F is obtained by writing that the quantum wells are doped to a concentration of ρ_s, such that (see (8.45)):

$$E_F - \varepsilon_1 \approx \frac{\rho_s}{m_c/\tau\hbar^2} \tag{11.67a}$$

and $E_c - E_F$ is the difference between the barrier level and the Fermi level. For a quasi-resonant detector at $\hbar v$, we must have:

$$E_c - E_F = \hbar v - (E_F - \varepsilon_1) = \hbar v - \frac{\rho_s}{m_c/\pi\hbar^2} \tag{11.67b}$$

The dark current is obtained by setting $I_0 = AJ_0 = Aq\rho_{3D,th}v_d$, and the *detectivity*

(assuming $p_e \approx 1$) is then given by:

$$D^* \approx \frac{\mathscr{R}}{\sqrt{2(2-p_c)qJ_0/Np_c}} \tag{11.68}$$

that is:

$$D^* \approx \frac{\eta_{1QW}}{h\nu} \frac{N^{1/2}}{\left[2(2-p_c)p_c v_d \frac{m_c kT}{\pi\hbar^2 L} e^{-((E_c-E_F)/kT)}\right]^{1/2}} \tag{11.69}$$

Detectivity for a multi-quantum well structure

from which we obtain the aforementioned dependence on $\sqrt{N}$. This last equation also indicates that if we increase the doping concentration in the wells, we can expect to improve the quantum efficiency by increasing α_{1QW} (see (11.58)). The thermal barrier height $E_c - E_F$, however, decreases according to (11.67b). There exists, therefore, a compromise which will optimize the detectivity (11.69).

Figure 11.21 shows the temperature dependence of the detectivity for an 8.5 μm quasi-resonant 50 well detector. We observe an $e^{h\nu/2kT}$ dependence common to all quantum detectors.

Widespread interest in quantum well detectors derives from two principal advantages:

- The detection wavelength can be adjusted at will. To do so, one need only specify a certain quantum well thickness and Al fraction in the barrier at the time of growth (this aspect of design being referred to as *quantum engineering*).
- Standardized microfabrication methods are particularly well suited to the production of large-scale detector arrays, which in contrast to Schottky detector technology, offer good detectivity.

We note, finally, that since quantum well detectors cannot detect light at normal incidence, a method must be found to allow the photons to couple efficiently to intersubband transitions. This can be achieved by etching a grating coupler on the detector surface (see Fig. 11.22).

Example

We will calculate the detectivity for a quantum well structure at 77 K assuming the following physical parameters:

$\eta_{1QW} = 1\%$
$\hbar\nu = 0.155\,\text{eV}$ ($\lambda = 8\,\mu\text{m}$)
$N = 50$
$p_c = 0.25$
$v_d = 10^6\,\text{cm s}^{-1}$
$m_c/\pi\hbar^2 = 2.78 \times 10^{13}\,\text{eV}^{-1}\,\text{cm}^{-2}$ and $n_c = 7.2 \times 10^{11}\,(T/300)\,\text{cm}^{-2}$

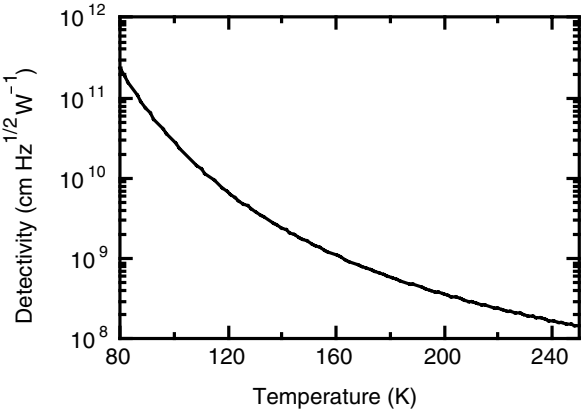


Fig. 11.21. Temperature dependence of the detectivity for a multi-quantum well structure ($N = 50$) assuming a quantum efficiency per well of 1%.

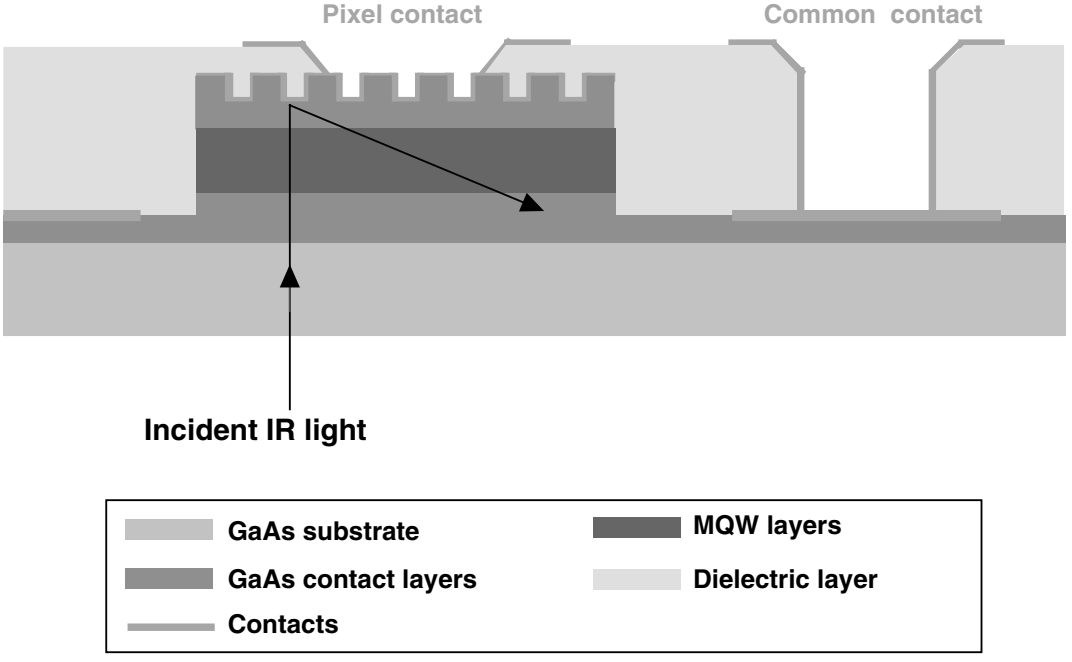


Fig. 11.22. Single pixel in a multiple quantum well detector array. A grating coupler redirects normally incident light to increase overlap with allowed intersubband transitions.

$L = 250\,\text{Å}$

$\rho_s = 5 \times 10^{11}\,\text{cm}^{-2}$

The Fermi level in the well is then (11.67b):

$$E_c - E_F = 0.155\,\text{eV} - 5 \times 10^{11}/2.78 \times 10^{13}\,\text{eV} = 0.155 - 0.018 = 137\,\text{meV}$$

and the detectivity is given by (11.69):

$$D^* \approx \frac{10^{-2}}{0.138 \times 1.6 \times 10^{-19}\,\mathrm{J}} \frac{\sqrt{50}}{\sqrt{4 \times 0.25 \times 10^{6}\,\mathrm{cm\,s^{-1}} \times \dfrac{7.2 \times 10^{11}\,(T/300)\,\mathrm{cm^{-2}}}{250 \times 10^{-8}\,\mathrm{cm}}\, \mathrm{e}^{-(0.138/8.63\times 10^{-5}\,T)}}}$$

or $3 \times 10^{11}\,\mathrm{cm\,Hz^{1/2}\,W^{-1}}$ at 77 K.

11.7 Avalanche photodetectors

In an avalanche photodiode the response is increased by using electron–hole pair multiplication to amplify the number of photogenerated carriers. *Impact ionization* was described in Complement 6.C, where the multiplication coefficient α_n was defined as being the number of secondary electron–hole pairs $\alpha_n \mathrm{d}x$ generated by an electron travelling a distance $\mathrm{d}x$ in the presence of a strong electric field. The coefficient α_p similarly describes the number of pairs produced by a hole. These coefficients depend drastically upon the field strength while being negligible at low fields (see Fig. 6.C.3). In an avalanche diode, we can obtain a strong field by applying a considerable reverse bias across a *p–i–n* structure (see Fig. 11.23).

There are therefore three possible sources for a current crossing such a diode (of cross-sectional area A): primary generation (either thermally created, or resulting from photon absorption) $AqG\mathrm{d}x$ in a layer of thickness $\mathrm{d}x$; impact generation due to electrons $\alpha_n I_n \mathrm{d}x$; and impact generation due to holes $\alpha_p I_p \mathrm{d}x$. These mechanisms lead to the following equations for the steady state current:

$$\begin{aligned} -\frac{\mathrm{d}I_n}{\mathrm{d}x} &= \alpha_n I_n + \alpha_p I_p + AqG \\ \frac{\mathrm{d}I_p}{\mathrm{d}x} &= \alpha_n I_n + \alpha_p I_p + AqG \end{aligned} \tag{11.70}$$

The continuity of the total current $I = I_n + I_p$ is assured by these equations, and by replacing I_n in the second equation, we obtain:

$$\frac{\mathrm{d}I_p}{\mathrm{d}x} = (\alpha_p - \alpha_n) I_p + \alpha_n I + AqG \tag{11.71}$$

which possesses the general solution:

$$I_p(x) = C \exp[-(\alpha_n - \alpha_p)x] + \frac{\alpha_n I + AqG}{\alpha_n - \alpha_p} \tag{11.72}$$

The constant C depends upon the boundary conditions at the contacts: n at $x = 0$, and p at $x = L$. Ideally, the n contact cannot inject any holes $I_p(0) = 0$, and the p

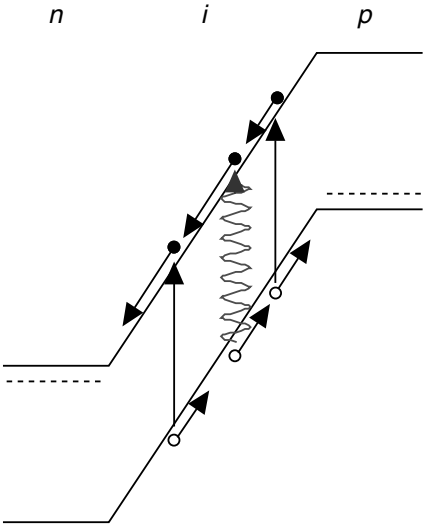


Fig. 11.23. In an avalanche diode, a photogenerated electron–hole pair creates secondary pairs by impact ionization.

contact cannot inject any electrons $I_n(L) = 0$. From the conservation of total current, it follows that $I = I_n(0) = I_p(L)$, so that:

$$I = AqG\frac{\exp[(\alpha_n - \alpha_p)L] - 1}{\alpha_n - \alpha_p\exp[(\alpha_n - \alpha_p)L]} \tag{11.73}$$

In this last equation, we see that the current flux I/Aq is equal to the generation current G, multiplied by a factor M defined as:

$$M = \frac{I}{AqGL} = \frac{1}{L}\frac{\exp[(\alpha_n - \alpha_p)L] - 1}{\alpha_n - \alpha_p\exp[(\alpha_n - \alpha_p)L]} \tag{11.74}$$

Generation current multiplication factor

For the special case where $\alpha_n = \alpha_p = \alpha$, the *multiplication factor* simplifies to:

$$M = \frac{1}{1 - \alpha L} \tag{11.75}$$

For a given field strength, (11.74) imposes a limit on the length of the avalanche region. If L increases to a value making the denominator equal to zero (Fig. 11.24), the multiplication factor tends towards infinity, and any infinitesimal carrier generation produces an avalanche response which short circuits the diode. This surge can actually destroy the diode unless the current is limited elsewhere in the circuit. Equation (11.74) also shows that an ideal material for obtaining a strong multiplication factor in intrinsic regions of minimal thickness would be one in

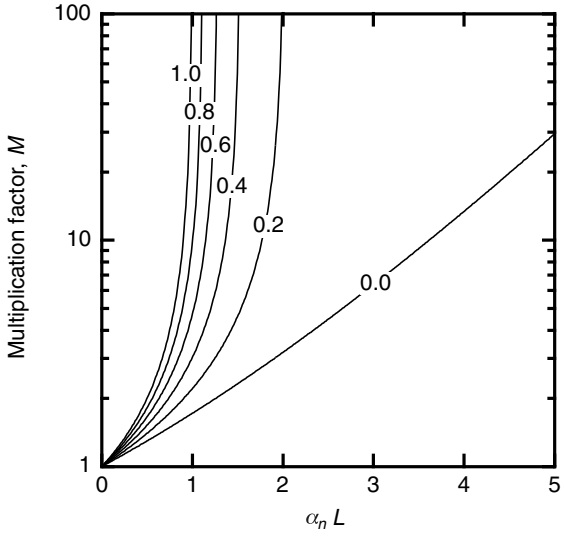


Fig. 11.24. Multiplication factor according to (11.74) for several different ratios of α_p/α_n.

which the avalanche coefficients for the electrons and holes (α_n and α_p) are as close as possible. However, we will see in Complement 11.A.5 that this is at the expense of a much more important increase in noise and, therefore, decrease in detectivity.

We therefore see that avalanche leads to a *photoconductivity gain*, but for fundamentally different reasons than those leading to gain in a photoconductor. We will see in Complement 11.A.5 that avalanche photodiodes possess additional sources of noise resulting from the avalanche process.

Several possible geometries can be used which can draw upon the avalanche effect. For instance, an intrinsic region can be used to absorb light *and* amplify the signal by avalanche. We can also absorb light in the weak field region (e.g. in one of the two contact layers) and only use the intrinsic region for avalanche amplification. In fact, if in the n and p contact layers, the photon absorption creates a non-equilibrium density of minority carriers, a portion of these carriers will diffuse towards the avalanche region, leading to an injection current (a hole current at $x = 0$ and an electron current at $x = L$). In *separate absorption–multiplication photodiodes*, a large gap semiconductor is used as an avalanche region and a small gap semiconductor is used as a light absorption region.

Let us take as an example, an electron current $I_n(L)$ injected from the p contact at $x = L$. In (11.70)–(11.72) we have $G = 0$, and the boundary conditions at $x = L$ have changed. The general solution (11.72) with the boundary condition $I_p(0) = 0$ then gives:

$$I_p(x) = \frac{\alpha_n I}{\alpha_n - \alpha_p}\{1 - \exp[-(\alpha_n - \alpha_p)x]\} \tag{11.76}$$

and at $x = L$, we have:

$$I = I_n(L) + I_p(L) = I_n(L) + \frac{\alpha_n I}{\alpha_n - \alpha_p}\{1 - \exp[-(\alpha_n - \alpha_p)L]\} \tag{11.77}$$

which gives us the *electron injection current multiplication factor*:

$$M_n = \frac{I}{I_n(L)} = \frac{(\alpha_n - \alpha_p)\exp[(\alpha_n - \alpha_p)L]}{\alpha_n - \alpha_p\exp[(\alpha_n - \alpha_p)L]} \tag{11.78}$$

Multiplication factor for injected electrons

Similarly, we obtain a *hole injection current multiplication factor* at $x = 0$:

$$M_p = \frac{I}{I_p(0)} = \frac{(\alpha_n - \alpha_p)}{\alpha_n - \alpha_p\exp[(\alpha_n - \alpha_p)L]} \tag{11.79}$$

Multiplication factor for injected holes

As for the volumetric generation current multiplication factor M, these multiplication factors for the injected carriers diverge if $\alpha_n L$ attains a certain threshold (see Fig. 11.24).

Avalanche photodiodes are widely used to detect small signals, with silicon avalanche detectors being used most commonly to detect wavelengths shorter than 1 μm. In telecommunications applications, InGaAs-based photodetectors are used to detect signals at wavelengths of 1.55 μm.

FURTHER READING

R. J. Keyes, ed., *Optical and Infrared detectors*, Topics in Physics Vol. 19, 2nd Edn, Springer Verlag, Berlin (1980).

H. Mathieu, *Physique des Semiconducteurs et des Composants Électroniques*, Masson, Paris (1987).

G. H. Rieke, *Detection of Light*: *From the Ultraviolet to the Submillimeter*, Cambridge University Press, Cambridge (1994).

A. Rose, *Concepts in Photoconductivity and Allied Problems*, Wiley Interscience, New York (1963).

S. M. Sze, *Physics of semiconductor devices*, Wiley Interscience, New York (1981).

. . . and on quantum well detectors:

B. Levine, *J. Appl. Phys.* **74**, R1–R81 (1993).

E. Rosencher, B. Vinter, F. Luc, L. Thibaudeau, P. Bois, and J. Nagle, *IEEE J. Quantum Electron.* **QE-30**, 2875 (1994).

Complement to Chapter 11

11.A Detector noise

In Complements 3.A and 4.D, we broached upon a few aspects of noise theory in order to take into account the statistical properties of light. As the notion of 'signal-to-noise ratio' forms the basis for light [and signal!] detection, we will further develop in this complement the mathematical formalism of noise theory. Also addressed here are the physical mechanisms responsible for noise generation in photodetectors.

Before engaging in the main goals of this complement, we wish to derive a rather simple (and far from rigorous) argument which will allow us to derive the formula for *generation noise* rapidly (the aim of this complement). This line of reasoning will be helpful as a conceptual guide (a lifeline) during our elaboration of a more precise theory.

We consider a photon flux Φ over a characteristic time T, incident upon a photodetector with a quantum efficiency η (see Fig. 11.A.1). This characteristic time is the integration time of the detector which is also $1/2\Delta v$ where Δv is the detection bandwidth. During this time interval, a number N of photoelectrons are created. N is a random number, with an average given by $\bar{N} = \eta\Phi T$, and if N fluctuates according to a Poisson distribution, its variance is given by $\sigma_N^2 = \bar{N}$. The photogenerated current can be estimated by the number of carriers created per unit time as $i = qN/T$. The average and the variance of the photocurrent are then:

$$\begin{aligned} I &= i = \frac{q\bar{N}}{T} \\ \overline{i_{\Delta v}^2} &= \left(\frac{q}{T}\right)^2 \sigma_N^2 = 2qI\Delta v \end{aligned} \tag{11.A.1}$$

We thereby obtain formula (11.A.38), which relates the average value of the current to its variance. We will now perform this derivation in a more rigorous fashion.

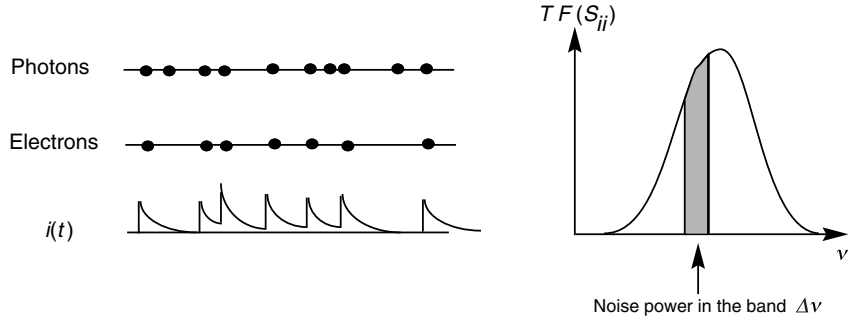


Fig. 11.A.1. (*a*) Noise in the detection process, and (*b*) associated noise spectrum in the photogenerated current.

11.A.1 Fluctuations

We consider a current $i(t)$ which circulates in a circuit. In a real situation, this current fluctuates about a mean value which corresponds to the signal level. For an observer devoid of any means of signal processing (e.g. signal integration, etc.), this noise will impede the detection of a signal smaller than the magnitude of the random *fluctuations*. We will now develop the physical and mathematical tools to allow us to deal with this simple, yet fundamental model for noise in a quantitative fashion.

The current $i(t)$ in this case is an example of a *random* (*or stochastic*) *process*. A random process is a function resulting from a random draw $i(t;\zeta)$, where ζ is a random variable (see Fig. 4.D.2). More precisely, such a process is described by an ensemble of functions $i(t;\zeta)$, $-\infty < t < \infty$, associated with a probability distribution $f(\zeta)$, where $f(\zeta)$ is the occurrence probability of the function $i(t;\zeta)$. As usual, we define the ensemble average $\langle i(t)\rangle = \int i(t;\zeta)f(\zeta)\mathrm{d}\zeta$, and $\langle i^2(t)\rangle = \int i^2(t;\zeta)f(\zeta)\mathrm{d}\zeta$. For stationary stochastic processes, which are the only ones we will consider here, this average does not depend on t. Furthermore, we redefine $i(t)$ by subtracting out the average $\langle i(t)\rangle$ (which in fact constitutes the signal) from the fluctuating current, so that the average $\langle i(t)\rangle$ is null. This is done for reasons of convenience as it lightens many of the ensuing expressions.

While the simple averages do not depend on time, the temporal aspects of these fluctuations do not disappear. The current at time t_1 can depend to varying degrees on the value at time t. This effect refers to the *memory* of the stochastic process. A measure of this memory is found using the *autocorrelation function* for i, which is the average of $i(t)i(t_1)$ over the ensemble:

$$S_{ii}(t,t_1) = \langle i(t)i(t_1)\rangle = \int i(t;\zeta)i(t_1;\zeta)f(\zeta)\mathrm{d}\zeta \tag{11.A.2}$$

The idea behind (11.A.2) is that, if the process is without a memory, then $i(t)$ and

$i(t_1)$ are independent random variables, and the average of the product is equal to the product of the averages, making (11.A.2) null. For a stationary process, the function S_{ii} depends only on the difference $\tau = t_1 - t$. Furthermore, we clearly have $S_{ii}(\tau) = S_{ii}(-\tau)$.

The Fourier transform of the autocorrelation function is referred to as the *power spectrum* of i:

$$\tilde{S}_{ii}(\omega) = \int S_{ii}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\mathrm{d}\tau \tag{11.A.3}$$

Power spectrum of $i(t)$

We can better understand the reason for calling this a power spectrum by calculating the square of the current $i(t)$ over a large time interval T. We can expand the current $i(t)$ sampled between $-T/2$ and $T/2$, in terms of a Fourier series:

$$\begin{aligned} i(t) &= \sum_{-\infty}^{\infty} \tilde{\imath}(\omega_j)\mathrm{e}^{\mathrm{i}\omega_j t}, \qquad \omega_j = \frac{2\pi j}{T} \\ \tilde{\imath}(\omega_j) &= \frac{1}{T}\int_{-T/2}^{T/2} i(t)\mathrm{e}^{-\mathrm{i}\omega_j t}\mathrm{d}t, \end{aligned} \tag{11.A.4}$$

We obtain for the square of the current:

$$i^2(t) = \sum_{j,k} \tilde{\imath}(\omega_j)\tilde{\imath}(\omega_k)^*\mathrm{e}^{\mathrm{i}(\omega_j-\omega_k)t} \tag{11.A.5}$$

giving for the time average of the square:

$$\frac{1}{T}\int_{-T/2}^{T/2} i^2(t)\mathrm{d}t = \sum_k |\tilde{\imath}(\omega_k)|^2 \tag{11.A.6}$$

which is simply Parseval's theorem. Inserting the expression for $i(\omega_k)$ into this last expression, we find:

$$\begin{aligned} \frac{1}{T}\int_{-T/2}^{T/2} i^2(t)\mathrm{d}t &= \sum_k \left|\frac{1}{T}\int_{-T/2}^{T/2} i(t)\mathrm{e}^{-\mathrm{i}\omega_k t}\right|^2 \mathrm{d}t \\ &= \sum_k \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{d}\tau\mathrm{e}^{-\mathrm{i}\omega_k\tau}\frac{1}{T}\int_{-T/2}^{T/2} i(t)i(t+\tau)\mathrm{d}t \end{aligned} \tag{11.A.7}$$

This last integral represents the time average of the variable $i(t)i(t+\tau)$. If the ensemble average of this integral in the limit $T \to \infty$ is equal to the autocorrelation

function $S_{ii}(\tau)$, then this process is referred to as an *ergodic process*. We will suppose in what follows that photodetection mechanisms are ergodic processes. It can be shown that this property is satisfied if the integral of $\tau S_{ii}(\tau)$ converges.

The consequence of ergodicity on (11.A.7) is therefore that:

$$\left\langle \frac{1}{T} \int_{-T/2}^{T/2} i^2(t)\mathrm{d}t \right\rangle = \sum_k \frac{1}{T} \int_{-T/2}^{T/2} S_{ii}(\tau)\mathrm{e}^{-\mathrm{i}\omega_k\tau}\mathrm{d}\tau \tag{11.A.8}$$

and in the $T \to \infty$ limit:

$$\left\langle \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} i^2(t)\mathrm{d}t \right\rangle = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \tilde{S}_{ii}(\omega) = \int_{-\infty}^{\infty} \tilde{S}_{ii}(v)\mathrm{d}v \tag{11.A.9}$$

where we have transformed the sum into an integral using the relation:

$$\frac{1}{T}\sum_k f(\omega_k) \to \int \frac{\mathrm{d}\omega}{2\pi} f(\omega) \tag{11.A.10}$$

as $\Delta\omega_k = 2\pi/T$. Equation (11.A.9) signifies that the total average of $i^2(t)$ is distributed over a frequency spectrum given by $\tilde{S}(v)$ and that the power over a band of positive frequencies Δv is:

$$\overline{i_v^2} = [\tilde{S}(-v) + \tilde{S}(v)]\Delta v = 2\tilde{S}(v)\Delta v \tag{11.A.11}$$

Average power in a frequency band Δv

since we do not make any difference between positive and negative frequencies (see Fig. 11.A.1). For calculations, we need to extract yet another important property from the noise power spectrum. We will consider a linear filter transfer function $h(\omega)$ which transforms a signal $i_{\mathrm{in}}(\omega)$ into an output signal $i_{\mathrm{out}}(\omega) = h(\omega)i_{\mathrm{in}}(\omega)$. According to the convolution theorem, we have that:

$$i_{\mathrm{out}}(t) = \int h(t')i_{\mathrm{in}}(t - t')\mathrm{d}t' = h(t)^* i_{\mathrm{in}}(t) \tag{11.A.12}$$

We now seek the correlation function between i_{out} and i_{in}:

$$\begin{aligned} S_{oi}(\tau) &= \langle i_{\mathrm{out}}(t)i_{\mathrm{in}}(t - \tau)\rangle \\ &= \int h(t')\langle i_{\mathrm{in}}(t - t')i_{\mathrm{in}}(t - \tau)\rangle \mathrm{d}t' \\ &= \int h(t')S_{ii}(\tau - t')\mathrm{d}t' = h(\tau)^* S_{ii}(\tau) \end{aligned} \tag{11.A.13}$$

The autocorrelation function for i_{out} is obtained by multiplying (11.A.12) by

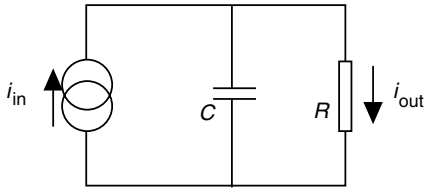


Fig. 11.A.2. An integrating circuit.

$i_{out}(t+\tau)$:

$$S_{oo}(\tau) = \langle i_{out}(t+\tau) i_{out}(t) \rangle$$

$$= \int h(t') \langle i_{out}(t+\tau) i_{in}(t-t') \rangle \mathrm{d}t' \tag{11.A.14}$$

$$= \int h(t') S_{oi}(\tau + t') \mathrm{d}t' = h(-\tau)^* S_{oi}(-\tau)$$

or again,

$$S_{oo}(\tau) = h(-\tau)^* h(\tau)^* S_{ii}(\tau) \tag{11.A.15}$$

which yields the Fourier transform:

$$\tilde{S}_{oo}(\omega) = |h(\omega)|^2 \tilde{S}_{ii}(\omega) \tag{11.A.16}$$

Transformation of the power spectrum by a filter

This important result allows one to recover (11.A.11) from the average noise power, as the filter ($|h(\omega)| = 1$, if $\omega_1 < \omega < \omega_1 + \Delta\omega$ and $h(\omega) = 0$ otherwise) transmits the power $\tilde{S}(\omega)\Delta\omega$. Additionally, this equation shows that the calculation for noise propagation in a linear circuit is completely analogous to the calculation for a normal signal, if we represent the noise by an amplitude source i_ν, whose square is given by (11.A.11). Armed with these general results, we can turn to the problem of identifying the physical origins of detector noise.

Example

For the simple *integrating circuit* in Fig. 11.A.2, we easily find:

$$h(\omega) = \frac{i_{out}(\omega)}{i_{in}(\omega)} = \frac{1}{1 + \mathrm{i}\omega\tau}, \ \tau = RC$$

and (11.A.16) shows us that if the power spectrum of a noise source at the entrance is $\tilde{S}_{ii}(\omega)$ then the output power spectrum is:

$$\tilde{S}_{oo}(\omega) = \frac{1}{1 + (\omega\tau)^2} \tilde{S}_{ii}(\omega)$$

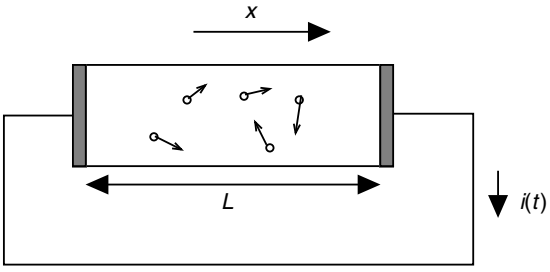


Fig. 11.A.3. In a semiconductor, the random motions of free carriers produce a fluctuation current in the circuit.

The filter therefore significantly reduces the noise by eliminating the portion of the power spectrum above $\omega = 1/\tau$. The greater the value of τ, the greater the extent of the filtering.

11.A.2 Physical origin of noise

In a shorted semiconductor sample, the instantaneous current which circulates is conserved everywhere in the circuit and is given by:

$$i(t) = \frac{A}{V}\sum_n - qv_{xn} = -q\frac{N}{L}\langle v_x \rangle_N \tag{11.A.17}$$

where A and V refer, respectively, to the area and the volume of the semiconductor sample, and the sum over n is performed for all N carriers in the sample. On average, this current is null. However, since the velocity and the number of carriers can fluctuate, $i(t)$ generally differs from being exactly zero at any given time. These deviations from zero are therefore responsible for a current noise (see Fig. 11.A.3). Two sources of noise coexist in this last equation:

- The velocity of each carrier can fluctuate due to scattering by thermal motions. This is the source of *thermal noise* (also referred to as *Johnson* or *Nyquist noise*).
- The number of carriers can fluctuate either due to generation–recombination processes (referred to as *generation–recombination noise*), or because of the fact that each contact can inject or capture the current one electron at a time with some necessary time elapsing between each event (referred to as *shot-noise*).

As these mechanisms are distinct from each other, they may be treated as independent non-correlated sources of noise.

11.A.3 Thermal noise

In order to calculate the thermal noise, we will first consider an electron in a semiconductor possessing neither field nor a carrier density gradient. Boltzmann's

equation (see Chapter 6) gives the evolution of the distribution function for such an electron:

$$\frac{\partial f(\mathbf{v},t)}{\partial t} = -\frac{f(\mathbf{v},t) - f_0(\mathbf{v})}{\tau_c(\mathbf{v})} \tag{11.A.18}$$

where $f(\mathbf{v},t)$ is the probability at time t that an electron possesses a velocity $\mathbf{v}$ (Eq. (6.7)). We have used the approximation for the relaxation time for the collision integral (Eq. (6.12)). $f_0(\mathbf{v})$ is the Maxwell distribution at thermodynamic equilibrium.

The solution to this equation is quite simple:

$$f(\mathbf{v},t) = f(\mathbf{v},0)\exp\left(-\frac{t}{\tau_c}\right) + f_0(\mathbf{v})\left[1 - \exp\left(-\frac{t}{\tau_c}\right)\right] \tag{11.A.19}$$

If we fix the velocity at time $t = 0$ to $\mathbf{v}_0$, we then have $f(\mathbf{v},0) = \delta(\mathbf{v} - \mathbf{v}_0)$, and (11.A.19) gives the *conditional distribution function*, i.e. the probability that the velocity at time t is $\mathbf{v}$ given that it was $\mathbf{v}_0$ at $t = 0$:

$$f(\mathbf{v},t|\mathbf{v}_0,0) = \delta(\mathbf{v} - \mathbf{v}_0)\exp\left[-\frac{t}{\tau_c(\mathbf{v}_0)}\right] + f_0(\mathbf{v})\left\{1 - \exp\left[-\frac{t}{\tau_c(\mathbf{v}_0)}\right]\right\} \tag{11.A.20}$$

The joint distribution function, i.e. the probability that the velocity be $\mathbf{v}$ at time t and $\mathbf{v}_0$ at time $t = 0$, is consequently:

$$\begin{aligned} &f(\mathbf{v},t;\mathbf{v}_0,0) \\ &= f_0(\mathbf{v}_0)\delta(\mathbf{v} - \mathbf{v}_0)\exp\left[-\frac{t}{\tau_c(\mathbf{v}_0)}\right] + f_0(\mathbf{v}_0)f_0(\mathbf{v})\left\{1 - \exp\left[-\frac{t}{\tau_c(\mathbf{v}_0)}\right]\right\} \end{aligned} \tag{11.A.21}$$

From this joint distribution, we obtain the autocorrelation function for v_x:

$$\begin{aligned} S_{xx}(t) &= \int \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{v}_0 v_x v_{0x} f(\mathbf{v},t;\mathbf{v}_0,0) \\ &= \int \mathrm{d}\mathbf{v}_0 v_{0x}^2 f_0(\mathbf{v}_0)\exp\left(-\frac{t}{\tau_c(\mathbf{v}_0)}\right) + 0 \end{aligned} \tag{11.A.22}$$

for $t \geq 0$. As for all other stationary processes, we have $S_{xx}(t) = S_{xx}(-t)$. Furthermore, if τ_c was simply a constant, integration of (11.A.22) would give:

$$S_{xx}(t) = \frac{kT}{m^*}\exp\left(-\frac{|t|}{\tau_c}\right) \tag{11.A.23}$$

The noise power spectrum is the Fourier transform:

$$\tilde{S}_{xx}(\omega) = \int_{-\infty}^{\infty} S_{xx}(t)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t \tag{11.A.24}$$

which yields:

$$\tilde{S}_{xx}(\omega) = \int \mathrm{d}\mathbf{v}_0 \frac{2\tau_c(\mathbf{v}_0)v_{0x}^2 f_0(\mathbf{v}_0)}{1 + \mathrm{i}\omega\tau_c(\mathbf{v}_0)} \tag{11.A.25}$$

For usual electronic frequencies, $\omega\tau_c \ll 1$ (τ_c is of the order of ps), and we find using expression (6.16) for the mobility:

$$\tilde{S}_{xx}(\omega) = 2\frac{kT}{q}\mu \tag{11.A.26}$$

Now, each electron contributes by $-qv_x/L$ to the current i, and since the motion of the electrons is uncorrelated, we have for the power spectrum of i:

$$\begin{aligned}\tilde{S}_{ii}(\omega) &= \frac{q^2}{L^2}\left\langle \sum_n v_{xn}(t)\sum_m v_{xm}(0)\right\rangle \\ &= \frac{q^2}{L^2}\sum_n \langle v_{xn}(t)v_{xn}(0)\rangle = \frac{q^2}{L^2}N\tilde{S}_{xx}(\omega)\end{aligned} \tag{11.A.27}$$

and in combination with (11.A.26):

$$\tilde{S}_{ii}(\omega) = 2kT\frac{N}{L^2}q\mu = 2kTG \tag{11.A.28}$$

Thermal noise power spectrum in electrical current

where $G = 1/R$ is the conductance of the semiconductor. The relationship between the fluctuations described by the current autocorrelation function S_{ii} and the conductance G is a particular case of the quite general *fluctuation–dissipation theorem*. The connection between the linear response G and the fluctuations in i show that *a dissipative process cannot exist without fluctuations or noise*. It is therefore impossible to measure the response of a system to a perturbation without the system being simultaneously a source of noise.

Applying (11.A.11) for the power in a bandwidth Δv due to thermal noise, we obtain:

$$\overline{i_v^2} = 4kTG\Delta v = \frac{4kT}{R}\Delta v \tag{11.A.29}$$

Thermal noise in a bandwidth Δv

This last expression can be interpreted as follows: the thermal power Ri_v^2 dissi-

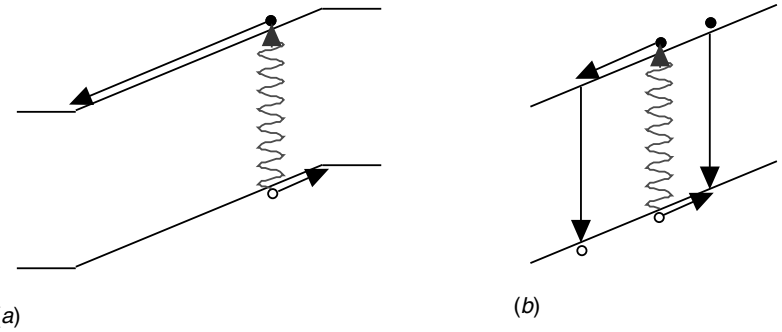


Fig. 11.A.4. (*a*) Following pair generation, an electron created in a diode generates a current by moving towards the *p* contact where it recombines with a hole, similarly a contribution is made by the hole as it moves towards the *n* contact where it recombines with an electron. In a photoconductor (*b*), the created electron recombines with another hole in the semiconductor.

pated in a resistor R over a bandwidth Δv is equivalent to an average thermal energy $4kT$ dissipated during a time Δv^{-1}.

Example
We seek the standard deviation of the current due to thermal noise at 300 K in a 50 Ω resistor by integrating the signal over 1 ms. The spectral bandwidth is 10^3 Hz and expression (11.A.29) gives $\sqrt{(4 \times 1.3 \times 10^{-23}\,\mathrm{J\,K^{-1}} \times 300\,\mathrm{K} \times 10^3\,\mathrm{Hz}\ 50^{-1}\Omega^{-1})}$ or 0.57 nA. This current is far from being negligible.

11.A.4 Generation–recombination noise

The electron–hole generation process does not create a perfectly constant current as electrons and holes carry discrete charge and the exact instant of their creation is random. Current fluctuations in this case originate from fluctuations in the density of free carriers in the semiconductor. From (11.A.17), we see that the average velocity of carriers must be other than zero to allow a density fluctuation to generate a current fluctuation. Contrary to thermal noise, *generation–recombination noise* is only present if the current is already non-zero. This then requires that an electric field or a density gradient be present in the semiconductor.

We will follow the fate of an electron created at time t_k, travelling with velocity v, until its tragic disappearance from the conduction band through recombination at a later time $t_k + \tau_k$ (see Fig. 11.A.4). The contribution of this event 'k' to the current in the circuit is $i_k(t - t_k)$, where:

$$i_k(t) = \begin{cases} -\dfrac{qv}{L}, & 0 < t < \tau_k \\ 0, & t < 0 \ \text{ or } \ \tau_k < t \end{cases} \tag{11.A.30}$$

The total current is therefore:

$$I(t) = \sum_k i_k(t - t_k) \tag{11.A.31}$$

and we can write the average current as:

$$\langle I \rangle = -\frac{qv}{L} \langle \tau \rangle \langle \dot{N} \rangle \tag{11.A.32}$$

where $\dot{N}$ is the generation rate and $\langle \tau \rangle$ is the average electron lifetime. This generation can have several physical origins: generation at an impurity site (as in the Shockley–Read mechanism studied in Section 6.5), the Auger effect (see Complement 6.D), band-to-band thermionic emission (see Complement 10.B), or photon absorption. In all cases, we can suppose that the distribution of creation times follows a Poisson distribution, i.e. that there is no temporal correlation between creation events.

According to our now well established procedure, the power spectrum for the generated current I is:

$$\begin{aligned} S_{II}(\omega) &= \int_{-\infty}^{\infty} \langle I(t) I(t+\tau) \rangle \mathrm{e}^{-\mathrm{i}\omega\tau} \mathrm{d}\tau \\ &\int_{-\infty}^{\infty} \left\langle \sum_k i_k(t - t_k) \sum_l i_l(t - t_l) \right\rangle \mathrm{e}^{-\mathrm{i}\omega\tau} \mathrm{d}\tau \end{aligned} \tag{11.A.33}$$

This expression can be calculated using the Fourier transform for the elementary contribution i_k:

$$\begin{aligned} i_k(t - t_k) &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} i_k(\omega) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{e}^{-\mathrm{i}\omega t_k} \\ i_k(\omega) &= \int_{-\infty}^{\infty} i_k(t) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t = -\frac{qv}{L} \frac{1 - \mathrm{e}^{-\mathrm{i}\omega\tau_k}}{\mathrm{i}\omega} \end{aligned} \tag{11.A.34}$$

which, once inserted into (11.A.30) gives:

$$\begin{aligned} S_{II}(\omega) &= \int \mathrm{d}\tau \int \frac{\mathrm{d}\omega_1}{2\pi} \int \frac{\mathrm{d}\omega_2}{2\pi} \mathrm{e}^{\mathrm{i}\omega_2\tau} \mathrm{e}^{-\mathrm{i}\omega\tau} \\ &\quad \times \left\langle \mathrm{e}^{\mathrm{i}(\omega_1+\omega_2)t} \sum_k i_k(\omega_1) \mathrm{e}^{-\mathrm{i}\omega_1 t_k} \sum_l i_l(\omega_2) \mathrm{e}^{-\mathrm{i}\omega_2 t_l} \right\rangle \\ &= \int \frac{\mathrm{d}\omega_1}{2\pi} \left\langle \mathrm{e}^{\mathrm{i}(\omega_1+\omega)t} \sum_k i_k(\omega_1) \mathrm{e}^{-\mathrm{i}\omega_1 t_k} \sum_l i_l(\omega) \mathrm{e}^{-\mathrm{i}\omega t_l} \right\rangle \end{aligned} \tag{11.A.35}$$

In this summation average, we have the two following distinct cases:

- $k \neq l$ *for which each sum becomes*:

$$\left\langle \sum_k i_k(\omega_1) \mathrm{e}^{-\mathrm{i}\omega_1 t_k} \right\rangle = 2\pi\delta(\omega_1)\langle i_k(0)\rangle \tag{11.A.36}$$

  as the creation times t_k and lifetimes τ_k are uncorrelated. In the integral (11.A.35) for S_{II} these averages give the square of the mean current.

- $k = l$, in which case the average of the double sum in (11.A.35) may be written:

$$\left\langle \sum_k i_k(\omega_1) i_k(\omega) \mathrm{e}^{-\mathrm{i}(\omega_1+\omega)t_k} \right\rangle = 2\pi\delta(\omega_1 + \omega) \left\langle \sum_k |i_k(\omega)|^2 \right\rangle \tag{11.A.37}$$

  which leads to the simple result for the *current fluctuation* autocorrelation function:

$$\begin{aligned} S_{\Delta I \Delta I}(\omega) &= \left\langle \sum_k |i_k(\omega)|^2 \right\rangle = \left(\frac{qv}{L}\right)^2 \left\langle \sum_k \frac{2 - 2\cos\omega\tau_k}{\omega^2} \right\rangle \\ &\simeq \left(\frac{qv}{L}\right)^2 \left\langle \sum_k \tau_k^2 \right\rangle = \langle \dot{N} \rangle \left(\frac{qv}{L}\right)^2 \langle \tau^2 \rangle \end{aligned} \tag{11.A.38}$$

By using Eq. (11.A.32) for the average current, we obtain the result:

$$\begin{aligned} &S_{\Delta I \Delta I}(\omega) = \langle I \rangle \frac{qv\langle\tau\rangle}{L} \frac{\langle\tau^2\rangle}{\langle\tau\rangle} = gq\langle I\rangle \frac{\langle\tau^2\rangle}{\langle\tau\rangle} \\ &\text{with the gain } g = \frac{v\langle\tau\rangle}{L} = \frac{\langle\tau\rangle}{\tau_{\mathrm{tr}}} \end{aligned} \tag{11.A.39}$$

where the *photoconductive gain* was introduced in Section 11.3, and τ_{tr} is the electron transit time across the creation–recombination region. We then have the following general result for the *generation–recombination noise*:

$$\overline{i_v^2} = 2qgI \frac{\langle\tau^2\rangle}{\langle\tau\rangle} \Delta v \tag{11.A.40}$$

Generation–recombination noise

Thus (as predicted), this noise only exists in the event of a non-zero average current. Three situations involving this expression are of particular interest:

1. *Photodiode without recombination.* The lifetime distribution is particularly simple. A created electron will have a lifetime τ_{tr} which will take it to the p contact, where it will instantaneously recombine upon arriving. We therefore

have $g = 1$ and $\langle \tau^2 \rangle = \langle \tau \rangle^2$ with no fluctuation in τ. The noise is then purely a generation noise:

$$\overline{i_v^2} = 2qI\Delta v \tag{11.A.41}$$

Generation noise

Example

Assuming that a 1 nA current crosses a photodiode, and a bandwidth of 1 kHz, (11.A.41) leads to a generation noise of $\sqrt{(2 \times 1.6 \times 10^{-19}\,\text{C} \times 10^{-9}\,\text{A} \times 10^3\,\text{Hz})}$ or 0.56 pA.

2. *Photoconductor with recombination.* We can suppose the radiative lifetime to be distributed according to a Poisson process:

$$P(\tau) = \frac{1}{\langle \tau \rangle} \exp\left(-\frac{\tau}{\langle \tau \rangle}\right) \tag{11.A.42}$$

which has the property (among many others) that $\langle \tau^2 \rangle = 2\langle \tau \rangle^2$. This leads to a generation–recombination noise:

$$\overline{i_v^2} = 4qgI\Delta v \tag{11.A.43}$$

Generation–recombination noise

Generation–recombination noise can be considered as a generation noise $2q(I/g)\Delta v$ amplified by the gain g, to which is added the recombination noise. In a photodiode, the recombination noise is null; however, in a photoconductor it is of the same type as the generation noise. This doubles the generation noise, and is responsible for the factor of 4 in Eq. (11.A.43).

3. *Generation–capture noise in a quantum well photoconductor.* As an electron traverses a multi-quantum well structure, we saw that it may be captured by a well with probability p_c. The probability that an electron will be captured after travelling n wells beyond the well from which it was initially photoexcited is $p_c(1 - p_c)^{n-1}$ and the average lifetimes are:

$$\langle \tau \rangle = \frac{L_w}{v p_c},\ \langle \tau^2 \rangle = \left(\frac{L_w}{v}\right)^2 \frac{2 - p_c}{p_c^2} \tag{11.A.44}$$

where L_w is the distance between successive wells. The noise according to (11.A.40) is then:

$$\overline{i_v^2} = 2(2 - p_c)qgI\Delta v \tag{11.A.45}$$

Generation–capture noise in a MQW structure

with the two limits

- $p_c \to 1$ (the photodiode limit, as capture of the electron by the nearest neighbouring well (with respect to electron's well of origin) is certain),
- $p_c \to 0$ (the photoconductor limit, as the capture probability is Poissonian as the electron crosses many wells before being captured).

We recall that $g = 1/p_c N$, where N is the number of wells.

11.A.5 Multiplication noise

In Section 11.7, we saw how, in avalanche diodes, the generation current is multiplied by impact ionization. As the generation current is also a source of noise, we may expect this noise to be amplified by avalanche multiplication. The impact ionization process, however, is also a random process, and thereby adds to the overall noise. To describe the noise in an avalanche detector, we need to generalize the equations in Section 11.7 to be able to find the current $I(x_0)$ due to a generation current source $qGA\Delta x\delta(x - x_0)$ located at $x = x_0, 0 \le x_0 \le L$. Equation (11.70) may then be written:

$$\begin{aligned} \frac{\mathrm{d}I_p}{\mathrm{d}x} &= \alpha_n I_n + \alpha_p I_p + qGA\Delta x\delta(x - x_0) \\ -\frac{\mathrm{d}I_n}{\mathrm{d}x} &= \alpha_n I_n + \alpha_p I_p + qGA\Delta x\delta(x - x_0) \end{aligned} \tag{11.A.46}$$

with the boundary conditions $I_p(0) = I_n(L) = 0$. For a total current $I(x_0)$, the solution for I_p employing these boundary conditions is:

$$I_p(x; x_0) = \begin{cases} \dfrac{\alpha_n I(x_0)}{\alpha_n - \alpha_p}(1 - \mathrm{e}^{-\Delta\alpha x}), & 0 < x < x_0 \\ \dfrac{\alpha_n I(x_0)}{\alpha_n - \alpha_p}(1 - k\mathrm{e}^{\Delta\alpha(L-x)}), & x_0 < x < L \end{cases} \tag{11.A.47}$$

where $\Delta\alpha = \alpha_n(1 - k) = \alpha_n - \alpha_p$, and $k = \alpha_p/\alpha_n$.

For $x = x_0$ we obtain by integration of (11.A.46) about $x = x_0$:

$$I_p(x_0^+; x_0) - I_p(x_0^-; x_0) = qGA\Delta x \tag{11.A.48}$$

which allows us to determine $I(x_0)$:

$$I(x_0) = qGA\Delta x \frac{\Delta\alpha \mathrm{e}^{\Delta\alpha x_0}}{\alpha_n - \alpha_p \mathrm{e}^{\Delta\alpha L}} = qGA\Delta x M(x_0) \tag{11.A.49}$$

where the *multiplication factor* M now depends on x_0:

$$M(x_0) = \frac{(1 - k)\mathrm{e}^{\Delta\alpha x_0}}{1 - k\mathrm{e}^{\Delta\alpha L}} \tag{11.A.50}$$

The conclusion to be drawn from (11.A.49) is that a generation source at x_0 yields through multiplication a current in the circuit which is $M(x_0)$ times the original generation current. We may therefore consider the multiplication region as a filter in the sense of Section 11.A.1, as it provides a transfer function $M(x_0)$. We also see that the multiplication factors deduced in Section 11.7 are the following particular cases of $M(x_0)$:

1. Injection at $x = 0$:

$$M_p = M(0) = \frac{1-k}{1-k\mathrm{e}^{\Delta\alpha L}} \tag{11.A.51}$$

2. Injection at $x = L$:

$$M_n = M(L) = \frac{(1-k)\mathrm{e}^{\Delta\alpha L}}{1-k\mathrm{e}^{\Delta\alpha L}} \tag{11.A.52}$$

3. Uniform bulk generation:

$$M = \frac{\int_0^L I(x_0)\mathrm{d}x_0}{\int_0^L qGA\mathrm{d}x_0} = \frac{\mathrm{e}^{\Delta\alpha L}-1}{\alpha_n L(1-k\mathrm{e}^{\Delta\alpha L})} \tag{11.A.53}$$

The fact that the avalanche region behaves as a linear filter is important in terms of noise calculation. We know given (11.A.16) that the noise spectrum at the output of the filter is $|M(x_0)|^2$ times the entrance noise. As the sources at different locations x_0 are independent, it is sufficient to calculate the total noise contributed from all sources and multiply the result by $M(x_0)^2$.

At each location, there are three generation sources having the associated noise:

$$\begin{aligned}
\overline{i_{vG}^2} &= 2q\Delta v qG(x_0)A\Delta x \\
\overline{i_{vn}^2} &= 2q\Delta v\alpha_n I_n(x_0)\Delta x \\
\overline{i_{vp}^2} &= 2q\Delta v\alpha_p I_p(x_0)\Delta x
\end{aligned} \tag{11.A.54}$$

where $G(x_0)\mathrm{A}\Delta x$ is the thermal and optical generation rate, $\alpha_n I_n(x_0)$ is the generation current produced by electron collisions, and $\alpha_p I_p(x_0)$ corresponds to that produced by holes. $I_n(x_0)$ and $I_p(x_0)$ are, respectively, the mean electron and hole currents. The case involving the injection of electrons or holes at the contacts is obtained by replacing $qG(x_0)A\Delta x$ by $I_n(L)\delta(x_0 - L)$ or $I_n(0)\delta(x_0)$ leading to the *injection noise*:

$$\overline{i_{vG}^2} = 2q\Delta v I_n(L)\delta(x_0 - L)$$
$$\text{or } \overline{i_{vG}^2} = 2q\Delta v I_p(0)\delta(x_0) \tag{11.A.55}$$

The total detector noise is then:

$$\overline{i_v^2} = \int_0^L M^2(x_0)(\overline{i_{vG}^2} + \overline{i_{vn}^2} + \overline{i_{vp}^2})\mathrm{d}x_0$$
$$= 2q\Delta v A \int_0^L M^2(x_0)[qG(x_0) + \alpha_n I_n(x_0) + \alpha_p I_p(x_0)]\mathrm{d}x_0 \tag{11.A.56}$$

We now consider the three following cases:

1. *Bulk generation, i.e.* $G(x_0) = G, 0 < x_0 < L$. We have according to (11.74) $I = qGALM$ and:

$$I_p(x) = qAG\frac{\alpha_n LM + 1}{\Delta\alpha}(1 - \mathrm{e}^{-\Delta\alpha x})$$
$$I_n(x) = I - I_p(x) \tag{11.A.57}$$

Inserting these equations into (11.A.56) and using the expression for M given in (11.74), we find:

$$\overline{i_v^2} = 2q\Delta v A q G(1 + \alpha_n LM)\int_0^L M^2(x)\mathrm{e}^{-\Delta\alpha x}\mathrm{d}x$$
$$= 2q\Delta v A q G L M(1 + \alpha_n LM)(1 + \alpha_p LM) \tag{11.A.58a}$$
$$= 2q\Delta v I M\frac{(1 + \alpha_n LM)(1 + \alpha_p LM)}{M}$$

which we may also write as:

$$\overline{i_v^2} = 2q\Delta v I M F(M)$$
$$F(M) = \frac{(1 + \alpha_n LM)(1 + \alpha_p LM)}{M} \tag{11.A.58b}$$

Consequently, the noise is composed of the generation noise $2q\Delta v I$ amplified by the avalanche factor M, which plays the same role as g in a photoconductor (see (11.A.43)), and a *multiplication noise factor* $F(M)$, which describes the noise added by the avalanche process. In Fig. 11.A.5, we show this factor as a function

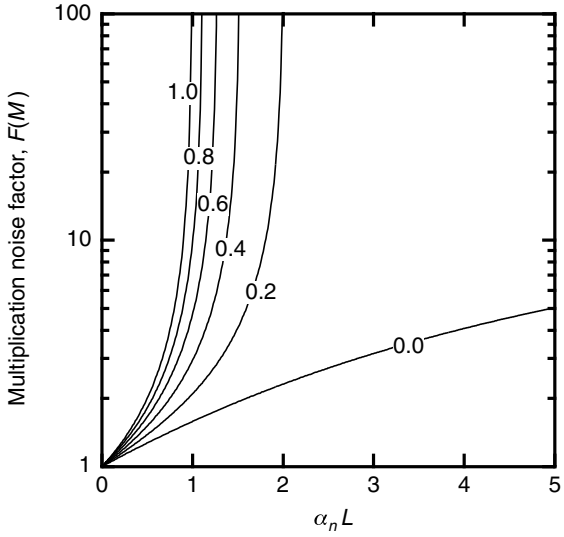


Fig. 11.A.5. Multiplication noise factor according to Eq. (11.A.54) for different ratios of α_p/α_n.

of the product $\alpha_n L$ for several values of the ratio $k = \alpha_p/\alpha_n$. We note that this factor is minimized when $k = 0$, or $1/k = 0$, but that it is asymptotically proportional to M when the two ionization coefficients have the same value. In order to obtain high performance detectors, it thus becomes necessary to find materials and polarization configurations in which one of the two coefficients dominates.

2. *Electron injection*, *i.e.* $\mathrm{qG(x_0) = I_n(L)\delta(x_0 - L)}$. We then have $I = M_n I_n(L)$, and according to Eqs. (11.76)–(11.78):

$$I_p(x) = \frac{\alpha_n M_n I_n(L)}{\Delta\alpha}(1 - \mathrm{e}^{-\Delta\alpha x}) = \frac{M_n I_n(L)}{1 - k}(1 - \mathrm{e}^{-\Delta\alpha x})$$
$$I_n(x) = I - I_p(x) \tag{11.A.59}$$

which gives, after introduction into Eq. (11.A.55) and simplification (note that $\mathrm{e}^{\Delta\alpha L} = M_n/(1 - k + kM_n)$):

$$\begin{aligned}\overline{i_v^2} &= 2q\Delta v I_n(L)\left[M_n^2 + \alpha_n M_n \int_0^L M^2(x)\mathrm{e}^{-\Delta\alpha x}\mathrm{d}x\right] \\ &= 2q\Delta v I_n(L)[M_n^2 + kM_n^3 + (1 - 2k)M_n^2 - (1 - k)M_n] \\ &= 2q\Delta v I M_n\left[kM_n + (1 - k)\left(2 - \frac{1}{M_n}\right)\right]\end{aligned} \tag{11.A.60}$$

The interpretation of this result is the same as in the preceding case. Furthermore, we clearly see that the *multiplication noise factor* is minimized for $k = 0$ and increases catastrophically as α_p approaches α_n.

3. *Hole injection*, *i.e.* $qG(x_0) = I_p(0)\delta(x_0)$. This case can be deduced from the preceding one by replacing $I_n(L)$ with $I_p(0)$, α_n with α_p and therefore k with $1/k$ yielding:

$$\overline{i_v^2} = 2q\Delta v I M_p F(M_p)$$
$$F(M_p) = \left[\frac{1}{k}M_p + \left(1 - \frac{1}{k}\right)\left(2 - \frac{1}{M_p}\right)\right] \tag{11.A.61}$$

This case is only of practical interest if $\alpha_p \gg \alpha_n$.

In concluding this complement, we see that there are several sources of noise whose behaviour we summarize in Fig. 11.A.6. Most notably, we found that:

- *photon noise* is transformed into photoelectron noise by the detection mechanism – we saw in this complement that its fate is the same as the signal, i.e. it suffers the same quantum losses ($\eta < 1$) and experiences the same gains (g in case of photoconductors, M in the case of avalanche photodiodes) as the detected signal;
- *gain noise* is specific to avalanche detectors with the factor $F(M)$ described by (11.A.58b);
- the *readout circuit noise* (e.g. amplifier, integrator, . . .), however, is independent of photoconductive gain.

It is therefore clear that one benefits from an internal amplification (photoconductive gain g or avalanche multiplication M) before resorting to electronic amplification.

More specifically, if the amplifier noise is σ_{ampli} (A $\text{Hz}^{1/2}$), the signal-to-noise ratio for the complete detection chain is given by (11.73) and (11.A.58b):

$$S/N = \frac{MAq\eta\Phi_0}{\sqrt{2Aq^2\Delta v\eta\Phi_0 M^2 F(M) + \sigma_{\text{ampli}}^2\Delta v}} \tag{11.A.62}$$

where A is the surface area, Φ_0 is the photon flux, and η is the quantum efficiency of the detector. This is to be compared with the ratio of the signal to noise obtained by using a photodiode in the same detection circuit, i.e:

$$S/N = \frac{qA\eta\Phi_0}{\sqrt{2Aq^2\Delta v\eta\Phi_0 + \sigma_{\text{ampli}}^2\Delta v}} \tag{11.A.63}$$

We will now compare the signal-to-noise ratios between a photodiode and a silicon avalanche diode with $k = \alpha_p/\alpha_n = 0.1$ and $\alpha_n L = 3$. In this case, $M = 100$ and $F(M) = 12$ (see Figs. 11.24 and 11.A.5). In Fig. 11.A.7, we compare (11.A.62)

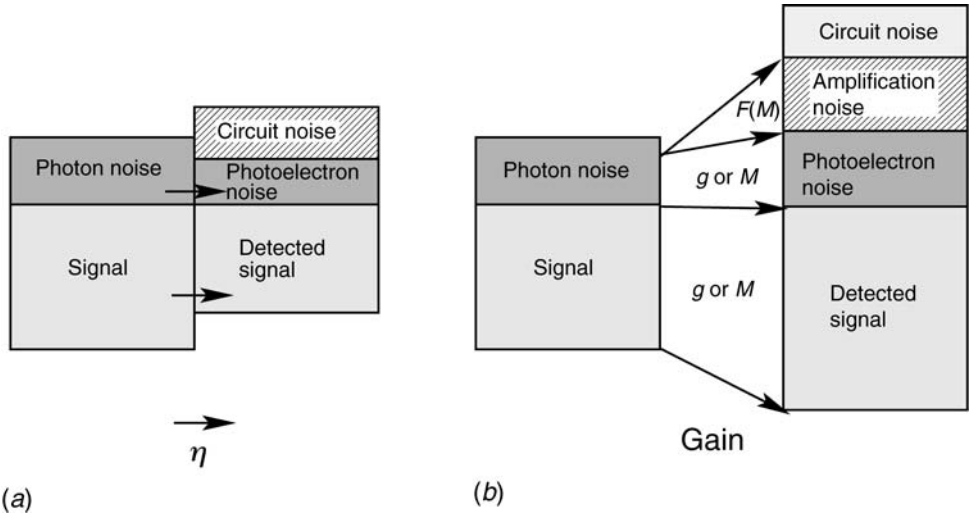


Fig. 11.A.6. Different sources of photodetection noise in a photodiode (*a*) and in a gain detector (photoconductor or avalanche) (*b*).

with (11.A.63). We notice that the use of an avalanche photodiode is advantageous for small signal levels when we want to extract the detection noise from the amplifier noise. More precisely this is the case when:

$$\Phi_0 < \frac{1 - 1/M^2}{F(M) - 1} \frac{\sigma^2_{\text{ampli}}}{2q^2 A\eta} \approx \frac{1}{F(M)} \frac{\sigma^2_{\text{ampli}}}{2q^2 A\eta} \tag{11.A.64}$$

Therefore, if we assume a light signal made up of 1 eV photons, detected using a 1 mm^2 detector with a quantum efficiency of 0.5, and an amplifier noise of 0.1 pA Hz$^{-1/2}$, this leads to a limit flux of 3.2×10^{12} photons cm^{-2} s^{-1}. So for an incident power less than 0.5 μW, an avalanche detector would be the detector of choice.

FURTHER READING

R. J. Keyes, Ed., *Optical and Infrared Detectors*, Topics in Physics Vol. 19, 2nd Edn, Springer Verlag, Berlin (1980).

R. J. McIntyre, *IEEE Trans. Electron Dev.* **ED-19**, 164 (1966).

A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York (1965).

11.B Detectivity limits: performance limits due to background (BLIP)

As we saw in Complement 11.A, all detectors contribute noise to the idealized signal. We are therefore not as interested in the detected signal (say as given by the

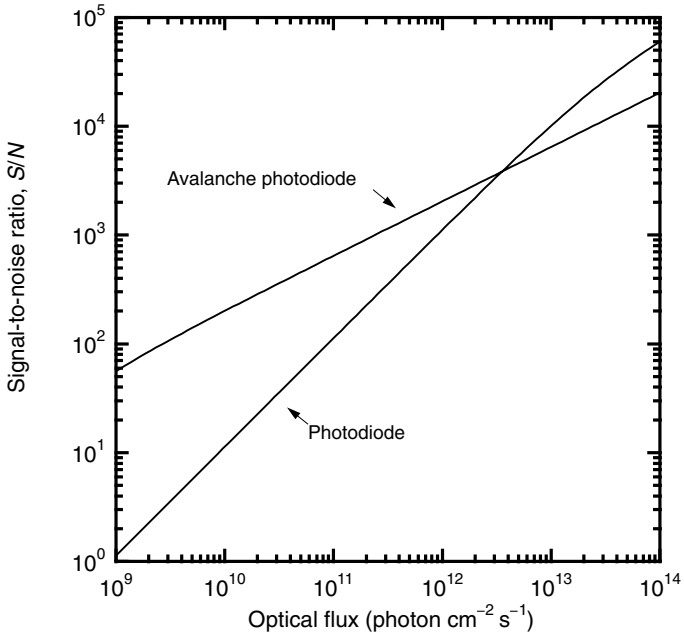


Fig. 11.A.7. Signal-to-noise comparison between 1 mm^2 diode detectors with a bandwidth $\Delta v = 50$ Hz.

photocurrent i_s) as we are in being able to extract this signal from a noise background (given by a noise current i_n). Detecting a signal, therefore consists in obtaining as large a S/N ratio as possible.

$$S/N = \frac{i_s}{i_n} \tag{11.B.1}$$

In a general sense, a signal i_s (A) is the product of a response $\mathscr{R}$ (A W^{-1}) to an incident power P_{inc} (W). The noise i_n takes the form $i_n = (quAJ_d\Delta v)^{1/2}$ (see Complement 11.A), where u is a constant (2 or 4 depending on whether a photovoltaic detector or a photoconductive detector possessing a gain of g is used); A is the area of the detector; J_d is the background current density; and Δv is the bandwidth of the measurement system, i.e. $2/t_{\text{int}}$, where t_{int} is the integration time of the signal measurement. The signal-to-noise ratio is then:

$$S/N = \frac{\mathscr{R}P_{\text{inc}}}{\sqrt{uqAJ_d\Delta v}} \tag{11.B.2}$$

The minimum detectable power is that which corresponds to a signal-to-noise ratio of 1. This quantity is referred to as the *noise equivalent power* (NEP) and is given by:

$$\mathrm{NEP} = \frac{\sqrt{uqAJ_d\Delta v}}{\mathscr{R}} \tag{11.B.3}$$

Noise equivalent power

This last expression shows that the performance of a detector in terms of NEP depends upon the bandwidth Δv and the detector surface area. These two dependencies result from nothing else than a manifestation of the law of large numbers $m/\sqrt{m}$, where m is the number of detection events in a certain spectral range. To allow comparisons between different detector technologies, we will introduce a more intrinsic measure of performance, the *detectivity* D^* defined as:

$$D^* = \frac{\sqrt{A\Delta v}}{\mathrm{NEP}} \tag{11.B.4}$$

or:

$$D^* = \frac{\mathscr{R}}{\sqrt{uqJ_d}} \tag{11.B.5}$$

Detectivity versus background current density J_d (cm $\mathrm{Hz}^{1/2}\,\mathrm{W}^{-1}$)

It is now time to define the background current density J_d. This current can have two origins:

- An intrinsic origin due to the fact that the detector picks up the blackbody radiation from the environment at T_B. This noise component is given by:

$$J_B = qg\int_{\lambda_1}^{\lambda_2}\eta(\lambda)\frac{\mathrm{d}\Phi_B(\lambda)}{\mathrm{d}\lambda}\mathrm{d}\lambda \tag{11.B.6a}$$

  where $\eta(\lambda)$ is the quantum efficiency of the detector for a photon with wavelength λ; λ_1 and λ_2 define the spectral detection range; g is the photoconductive gain ($g = 1$ for a photodiode); and $(\mathrm{d}\Phi_B(\lambda)/\mathrm{d}\lambda)\Delta\lambda$ is the photon flux over the wavelength range $\Delta\lambda$ ($\mathrm{d}\Phi_B/\mathrm{d}\lambda$ is in $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mu\mathrm{m}^{-1}$). This latter quantity can be expressed in terms of the blackbody emittance given in Complement 2.B as:

$$\frac{\mathrm{d}R}{\mathrm{d}\lambda} = \frac{hc}{\lambda}\frac{\mathrm{d}\Phi_B}{\mathrm{d}\lambda} \tag{11.B.6b}$$

  where c is the speed of light.
- An extrinsic origin due to the detector itself: this is the *dark current* which relates to conduction by free carriers in photoconductors (see Eq. (11.26)), or the *leakage current* J_{sat} in photodiodes (see Eq. (11.44)). To set these ideas straight, we will consider the noise in a reverse biased $V < 0$ photodiode (the factor

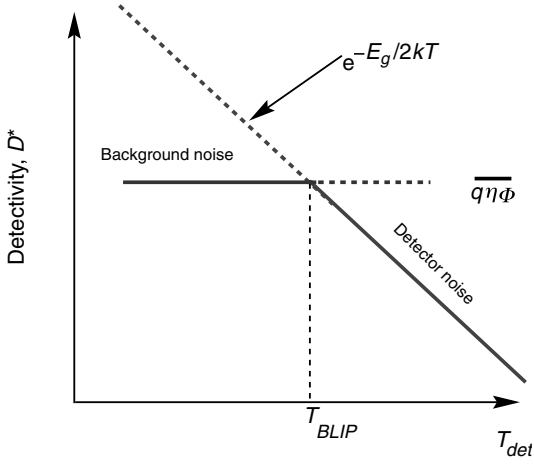


Fig. 11.B.1. Definition of the BLIP temperature.

$e^{qV/kT} = 0$ in (11.44) and $u = 2$). The global detectivity at a wavelength λ_0 and a background temperature T_B is then given by:

$$D^*(\lambda_0, T_B) = \frac{\eta(\lambda_0)}{1.24/\lambda_{\mu m}} \frac{1}{2^{1/2}\left[qJ_{sat}(T_{det}) + q^2 \int_{\lambda_1}^{\lambda_2} \eta(\lambda) \frac{d\Phi_B(\lambda)}{d\lambda} d\lambda \right]^{1/2}} \tag{11.B.7}$$

Global detectivity for a photodiode

where we have added the dependence of the current density J_{sat} to the detector temperature T_{det} (see Eq. (11.50)). As the detector temperature decreases, i_{sat} drops exponentially (as $e^{-E_g/kT_{det}}$). We see clearly in this last expression that there is an operation temperature below which it is useless to cool the detector further. The system is then dominated by background current, and we say that the detector is in the BLIP regime (for background limited infrared performance – see Fig. 11.B.1). The temperature required to reach the BLIP regime T_{BLIP} is given by:

$$J_{sat}(T_{BLIP}) = J_B = q \int_{\lambda_1}^{\lambda_2} \eta(\lambda) \frac{d\Phi_B(\lambda)}{d\lambda} d\lambda \tag{11.B.8}$$

BLIP temperature

Equation (11.B.7) permits one to calculate the detectivity limits for a photodiode, yielding:

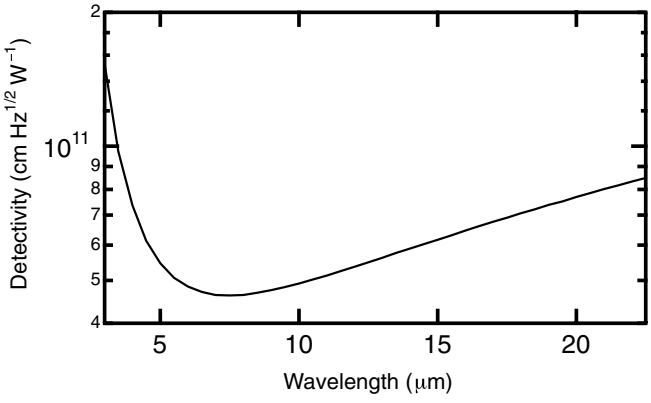


Fig. 11.B.2. Detectivity limit for an ideal photodiode ($\eta = 1$) as a function of detection wavelength.

$$D^*_{\mathrm{BLIP}}(\lambda_0, T_B) = \frac{\eta(\lambda_0)}{h\nu/q} \frac{1}{2^{1/2} q \left[\int_{\lambda_1}^{\lambda_2} \eta(\lambda) \frac{\mathrm{d}\Phi_B(\lambda)}{\mathrm{d}\lambda} \mathrm{d}\lambda \right]^{1/2}} \tag{11.B.9a}$$

To find the fundamental detectivity limit due to blackbody radiation at the cut-off wavelength λ_0 (where the detectivity is maximal), we set $\eta(\lambda) = 1$ for λ between $\lambda_1 = 0$ and $\lambda_2 = \lambda_0$, which leads to:

$$D^*_{\mathrm{BLIP,max}}(\lambda_0, T_B) = \frac{1}{h\nu} \frac{1}{2^{1/2} [\Phi_B(\lambda_0)]^{1/2}} \tag{11.B.9b}$$

BLIP detectivity

Figure 11.B.2 shows these fundamental limits as a function of detection wavelength λ for a background temperature T_B of 300 K. The detectivity limit at room temperature decreases considerably as the detection wavelength increases up to about 8 μm. This results from variation in the spectral emittance of blackbody radiation at 300 K, which exhibits a maximum at $\sim$10 μm (see Fig. 11.B.2). The linear increase in $D^*_{\mathrm{BLIP,max}}$ for $\lambda > 8$ μm occurs as the number of photons required to add up to 1 W of energy increases as λ becomes larger.

In an actual application, a detector observes a detail against a background through an objective (see Fig. 11.B.3). If the surfaces surrounding the detector are cooled, then the flux received from the background at temperature T_B decreases by $\sin^2 \phi$, where ϕ is the acceptance angle of the aperture.

The noise due to the background blackbody flux Φ_B therefore diminishes according to $\sin^2 \phi$. The detectivity in the BLIP regime for the system is then given by:

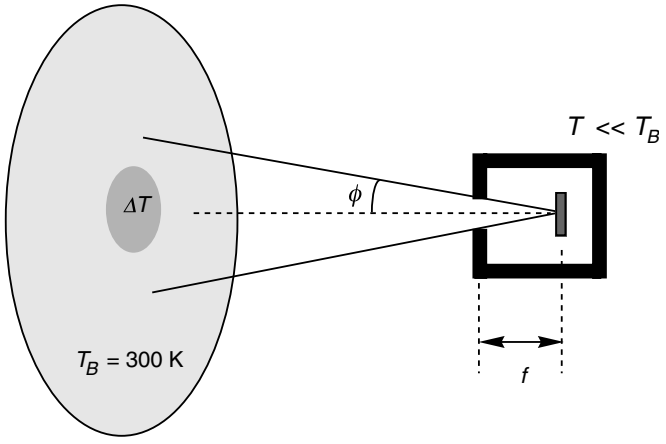


Fig. 11.B.3. In an infrared detection system, the blackbody flux is screened by cooling the detector enclosure. The detector observes the background with an acceptance angle of ϕ.

$$D^*_{\mathrm{BLIP}}(\lambda_0, T_B, \phi) = \frac{D^*_{\mathrm{BLIP}}(\lambda_0, T_B)}{\sin\phi} \tag{11.B.10}$$

We therefore see that the BLIP detectivity can be significantly increased by diminishing the acceptance angle of the detector. Generally, a lens with a diameter D is placed at the entrance of the detector assembly at a distance f (corresponding to the focal length of the lens) from the detector. The ratio of f/D then gives f-number (FN) of the system and $\tan\phi = 1/(2 \times \mathrm{FN})$. As $\sin^2\phi = 1/(1 + 1/\tan^2\phi)$, Eq. (11.B.10) may be written:

$$D^*_{\mathrm{BLIP}}(\lambda_0, T_B, \phi) = \sqrt{1 + 4\mathrm{FN}^2}\, D^*_{\mathrm{BLIP}}(\lambda_0, T_B) \tag{11.B.11}$$

Often, infrared detectors are used to resolve a thermal image immersed within a continuous background of thermal noise (see Fig. 11.B.3). A reasonable question which arises in such a case is what minimum thermal contrast can be resolved by the detector. The answer is in fact quite simple. This *noise equivalent temperature difference* (NETD) is given by the change in temperature which leads to a signal-to-noise ratio of 1, i.e. which leads to power emission equal to the NEP defined in (11.B.3). Figure 11.B.4 shows $D^*(\lambda_0, T_B = 300\,\mathrm{K})$ for several quantum detectors. From (2.B.12), a thermal variation ΔT for a blackbody leads to a variation $C_d(\Delta\lambda)\Delta T$ in the emittance $\mathrm{d}R/\mathrm{d}\lambda$ over the spectral range $\Delta\lambda$ (C_d is in $\mathrm{W\,cm^{-2}\,K^{-1}}$). The NETD may then be written as:

$$\mathrm{NEP} = \mathrm{NETD} \times \frac{\mathrm{d}}{\mathrm{d}T}\int_{\Delta\lambda}\frac{\mathrm{d}R}{\mathrm{d}\lambda}\mathrm{d}\lambda = C_d(\Delta\lambda)\mathrm{NETD} \tag{11.B.12}$$

giving for NETD as a function of the detectivity D^*:

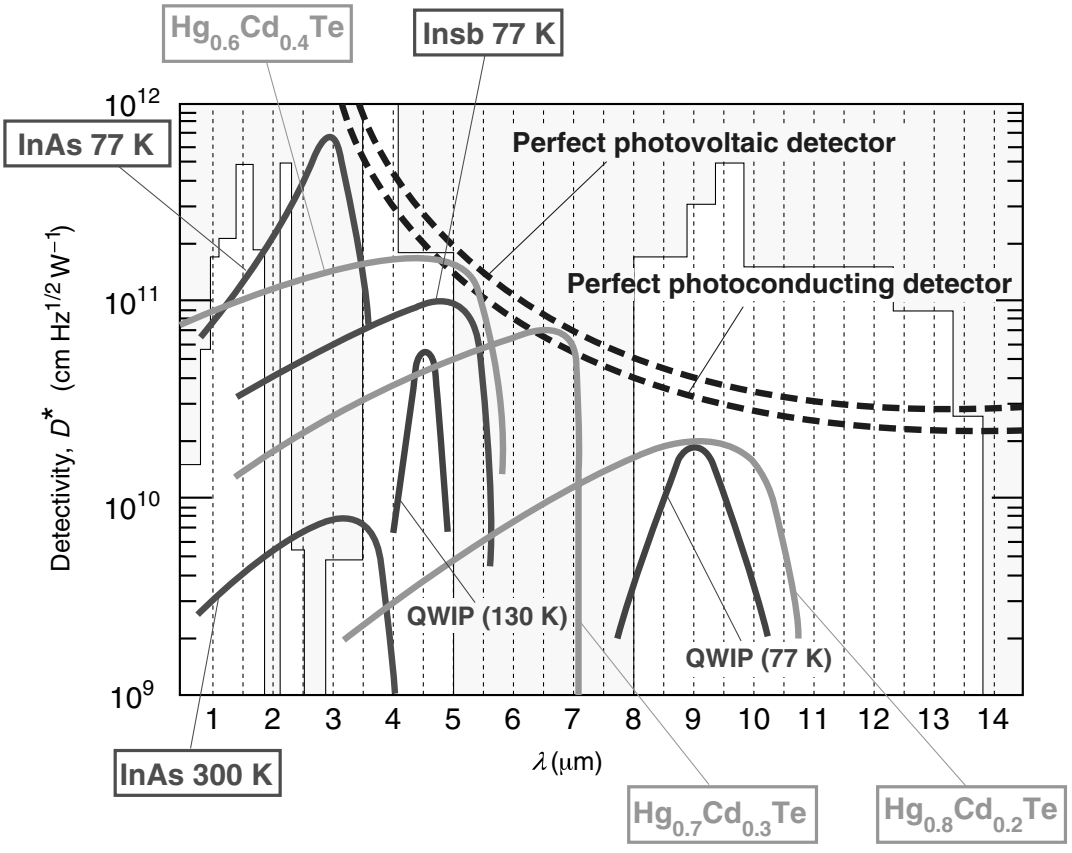


Fig. 11.B.4. Detectivity (D^*) plots for various quantum detectors.

$$\mathrm{NETD} = \frac{1}{C_d(\Delta\lambda)} \frac{\sqrt{A\Delta v}}{D^*} = \frac{1}{(\mathrm{d}/\mathrm{d}T) \int_{\Delta\lambda} (\mathrm{d}R/\mathrm{d}\lambda)\mathrm{d}\lambda} \frac{\sqrt{A\Delta v}}{D^*} \tag{11.B.13}$$

Figure 11.B.5 shows the correspondence between the noise equivalent temperature difference NETD and the detectivity for a $40 \times 40\,\mu m^2$ detector with a bandwidth of 50 Hz. We see that extremely small thermal contrasts can be resolved.

Example

MATHEMATICA program for calculating the noise equivalent temperature difference for a detector in the $3 \rightarrow 5\,\mu m$ band, with a detection area of $40 \times 40\,\mu m^2$, a bandpass of 50 Hz, and an f-number of $f/1$:

```
(*universal constants *)
c = 2.988 10^8 (* m/s*);
```

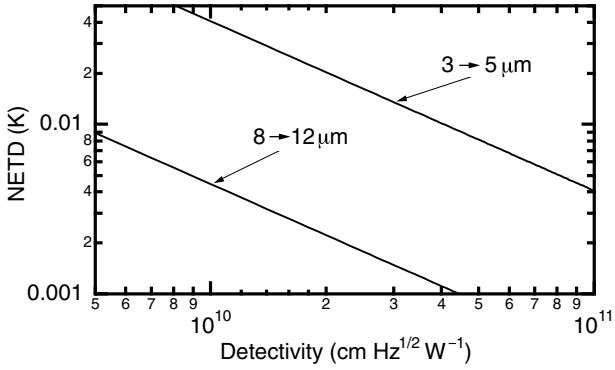


Fig. 11.B.5. Correspondence between the noise equivalent temperature difference NETD and the detectivity across two spectral bands for a $40 \times 40\,\mu m^2$ detector, with a bandpass of 50 Hz, and an FN of 1.

```
k = 1.38 10^-23 (*J/K*);
h = 6.625 10^-34 (*J.s*);
hb = h/(2*3.1416);
m0 = 0.91 10^-30 (*kg*);
(* Wavelength Band in μm *)
lambda1 = 3;lambda2 = 5;
(* Blackbody as a function of temperature*)
temp = .; lambda = .;
emm: = 2*3.1416*h*c^2*(lambda*10^-6)^-5/(Exp[h*c/((lambda*10^-6)*k*temp)]-1)
emm  = emm * 10^-6 (* W/m^2/μm*);
contr = D[emm,temp]; (*contrast used in the NETD calculation*)
temp = 300 (*Blackbody temperature*);
(*Differential contrast over the spectral band*)
cc = NIntegrate[contr,{lambda,lambda1,lambda2}] (*W/m^2/K*)
   *10 = -4 (* W/cm^2/K*);
df = 50;A = 16 10^-6 (*cm2*);dies = 1;
NETD = Sqrt[df/A]/Det/cc*(1]4*dies^2);
Plot[NETD,{Det,10^10,10^11}]
```

Results from this calculation appear in Fig. 11.B.5.

FURTHER READING

G. Gaussorgues and S. Chomet, *Infrared Thermography*, Kluwer, Boston (1993).

R. J. Keyes, Ed., *Optical and Infrared Detectors*, Topics in Physics Vol. 19, 2nd Edn, Springer Verlag, Berlin (1980).

12 Optical frequency conversion

12.1 Introduction

One of the most impressive accomplishments of wave optics is its success in providing a coherent and succinct description of the interactions between electromagnetic waves and matter (gases, solids, etc.). Maxwell's equations, which describe the propagation of light, and the Laplace–Lorentz equations, which describe the source terms of light, allow one to take into account the phenomena of refraction, diffusion, and diffraction of light by dense media. It is amazing – to say the least – that such a theory can account for the complex interactions of an electromagnetic wave with an immense ensemble of atoms (each approximated in terms of individual harmonic oscillators), by means of a simple optical index n_{op}. Such an achievement is reminiscent and, indeed, on par with the level of concision achieved by the concept of effective mass in representing the interaction of a conduction electron with a crystalline lattice.

In the description of all these effects, an electromagnetic wave with (angular) frequency ω forces free carriers into oscillatory motion at the same frequency, leading to radiative re-emission at this same frequency. This behaviour is a natural by-product of the linear equations we have employed up until now. In this chapter, we will see that non-linear media (i.e. materials whose response to external excitations contains non-linear terms), may by used to perform frequency conversion as evidenced, for example, by second harmonic light generation or optical parametric oscillations.

12.2 A mechanical description for second harmonic frequency generation

We already know that the mechanical model of an elastically bound electron successfully describes several optical effects, such as spectral dispersion (Complement 7.B). We now seek the behaviour of an electron subject to a confinement potential containing higher-order non-quadratic terms. To keep things simple, we shall suppose that the confinement potential $U(x)$ is one dimensional, and write it as (see Fig. 12.1):

$$U(x) = \frac{1}{2} m\omega_0^2 x^2 + \frac{1}{3} mDx^3 \tag{12.1}$$

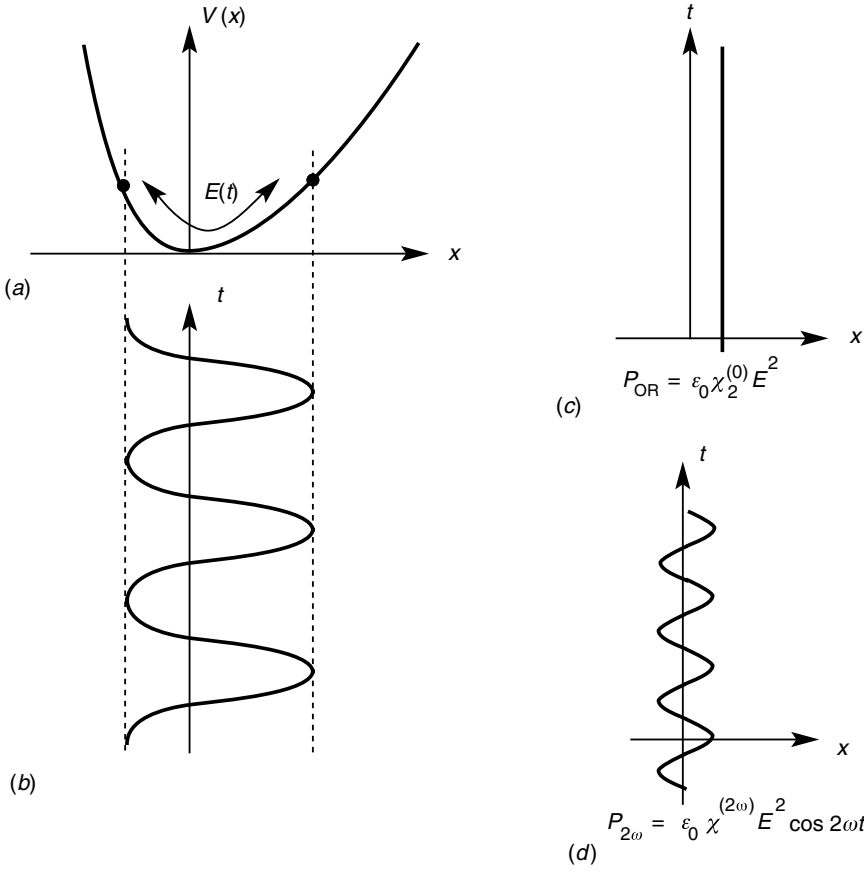


Fig. 12.1. An electron confined by an asymmetric potential and excited by an electromagnetic field $E\cos\omega t$ (*a*) experiences anharmonic oscillations about its equilibrium point (*b*). Its motion exhibits a continuous component (*c*) and a component at 2ω (*d*) which, respectively, give rise to a constant polarization (optical rectification) and the generation of a second harmonic signal.

where ω_0 is the natural oscillation frequency of the oscillator in the linear regime, m is the vacuum electron mass, and D is the non-linear coefficient (in $\mathrm{m}^{-1}\,\mathrm{s}^{-2}$). This system is then subjected to a dipolar electrical driving force:

$$F = qE\cos(\omega t) = \frac{qE}{2m}(\mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}) \tag{12.2}$$

where (as usual) c.c. designates the complex conjugate. The motion of the particle $x(t)$ therefore obeys the differential equation:

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x + Dx^2 = \frac{qE}{2m}(\mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}) \tag{12.3}$$

where γ is the friction coefficient of the particle and determines the width of the associated optical transition (see Complement 3.A). As the motion of the particle must clearly be periodic (with frequency ω and accompanying harmonics), we can perform a harmonic analysis of $x(t)$ by writing it as:

$$x(t) = \frac{1}{2}(x_0 + x_1 e^{i\omega t} + x_2 e^{i2\omega t} + \cdots + \text{c.c.}) \tag{12.4}$$

We will assume for the time being that $x_0 = 0$, i.e. that there is no constant polarization induced in the system. Later on we will see that such a term can exist, leading to the phenomenon of *optical rectification*. Substituting (12.4) into (12.3) we obtain:

$$\begin{aligned} &-\frac{\omega^2}{2}(x_1 e^{i\omega t} + \text{c.c.}) - 2\omega^2(x_2 e^{i2\omega t} + \text{c.c.}) + \frac{i\omega\gamma}{2}(x_1 e^{i\omega t} + \text{c.c.}) + i\omega\gamma(x_2 e^{i2\omega t} + \text{c.c.}) \\ &+ \frac{\omega_0^2}{2}(x_1 e^{i\omega t} + \text{c.c.}) + \frac{\omega_0^2}{2}(x_2 e^{i2\omega t} + \text{c.c.}) + \frac{D}{4}(x_1^2 e^{2i\omega t} + 2x_1 x_2^* e^{-i\omega t} \\ &+ x_1 x_1^* + x_2 x_2^* + 2x_1 x_2 e^{3i\omega t} + x_2^2 e^{4i\omega t} + \text{c.c.}) = \frac{qE}{2m}(e^{i\omega t} + \text{c.c.}) \end{aligned} \tag{12.5}$$

This expression may at first seem complicated, but, as is often the case in non-linear optics, it can be considerably simplified by grouping like terms.

We will first consider the linear response, i.e. the terms in $e^{i\omega t}$ and neglect those in D. We immediately find:

$$x_1 = \frac{qE}{m}\frac{1}{(\omega_0^2 - \omega^2) + i\omega\gamma} \approx \frac{qE}{2\omega m}\frac{1}{(\omega_0 - \omega) + i\gamma/2} \tag{12.6}$$

for $\omega \approx \omega_0$. The motion $x_1(t) = x_1 e^{i\omega t} + \text{c.c.}$ gives rise to a linear polarization in the medium:

$$P_1(t) = Nqx_1(t) = Nqx_1(e^{i\omega t} + \text{c.c.}) \tag{12.7}$$

where N is the volumetric density of the systems which interact with the wave. We can then identify term by term expression (12.7) with the definition for linear susceptibility given in (3.24):

$$P_1(t) = \frac{\varepsilon_0}{2}(\chi_1^{(\omega)} E e^{i\omega t} + \text{c.c.}) \tag{12.8}$$

which leads to:

$$\chi_1^{(\omega)} = \frac{Nq^2}{2\omega m\varepsilon_0}\frac{1}{(\omega_0 - \omega) + i\gamma/2} \tag{12.9}$$

This last result is equivalent to that provided by the classical Lorentz model investigated within the context of quantum mechanics in Chapters 2 and 3. Following an inductive line of reasoning, we define the non-linear second-order susceptibility as:

$$P_2(t) = \frac{\varepsilon_0}{2}(\chi_2^{(2\omega)} E^2 e^{i2\omega t} + \text{c.c.}) = \frac{Nq}{2}(x_2 e^{i2\omega t} + \text{c.c.}) \tag{12.10}$$

The term in x_2 is generated by the non-linear quadratic term Dx^2 in (12.3). An expression for it can be obtained through (12.5) by identifying terms in $e^{2i\omega t}$:

$$x_2(-4\omega^2 + 2i\omega\gamma + \omega_0^2) = -\frac{1}{2} D x_1^2 \tag{12.11a}$$

We see in this last expression that it is the term in x_1^2 which drives the 2ω motion of the electron. Employing (12.6), we find:

$$\begin{aligned} x_2 &= -\frac{q^2 D}{2m^2} \frac{1}{[(\omega_0^2 - \omega^2) + i\omega\gamma]^2[(\omega_0^2 - 4\omega^2) + 2i\omega\gamma]} E^2 \\ &\approx -\frac{q^2 D}{24m^2\omega^3} \frac{1}{[(\omega_0 - \omega) + i\gamma/2]^2[(\omega_0 - 2\omega) + (2/3)i\gamma]} E^2 \end{aligned} \tag{12.11b}$$

After substituting this last equation into (12.10), we are in a position to find the second-order non-linear susceptibility:

$$\chi_2^{(2\omega)} = -\frac{Nq^3 D}{24\varepsilon_0 m^2 \omega^3} \frac{1}{[(\omega_0 - \omega) + i\gamma/2]^2[(\omega_0 - 2\omega) + (2/3)i\gamma]} \tag{12.12}$$

Several interesting points are worthy of mention now. First, the system is doubly resonant, i.e. there is a resonance at $\omega = \omega_0$ and at $\omega = 2\omega_0$. Additionally, when comparing (12.12) and (12.9), we see that this model predicts the following relationship between linear and non-linear optical susceptibilities:

$$\frac{\chi_2^{(2\omega)}}{(\chi_1^{(\omega)})^2 \chi_1^{(2\omega)} \varepsilon_0^2} = \frac{mD}{2N^2 q^3} = \delta^{(2\omega)} \tag{12.13}$$

As shown in (12.13), the parameter $\delta^{(2\omega)}$ (referred to as the '*Miller parameter*') should be very similar for all materials, and it effectively is. Table 12.1 presents values for non-linear second-order susceptibilities in different semiconductors. Looking at this table, we conclude $\delta^{(2\omega)} \approx 3 - 8 \times 10^9$ SI.

We will now build an *extremely crude* model for a non-symmetric crystal, which will allow us to obtain a value for D starting from fundamental constants. To do so, we will assume that the electrons in such a crystal are subject to an attractive potential resulting from a nucleus with charge $2q$ and a second nucleus with charge

Table 12.1. *Optical indices* n_ω *at a fundamental wavelength of 10.6 μm, second harmonic optical indices* $n_{2\omega}$ *at 5.3 μm, quadratic susceptibilities* χ_2*, and Miller parameters for several semiconductors* ($n_{2\omega}$ *and* δ_{Miller} *are unknown for InSb*)

Material	n_ω	$n_{2\omega}$	χ_2 (pm V^{-1})	δ_{Miller} (10^9 SI)
InSb	3.95	?	1634	?
InAs	3.49	3.54	419	2.8
GaSb	3.8	3.82	628	3.2
GaAs	3.27	3.30	368	5.4
CdTe	2.69	2.71	168	6.4
ZnTe	2.69	2.70	90	4.7
ZnSe	2.42	2.43	78	8.6

q separated from the first by a distance a. This potential may be written as:

$$V(x) = \frac{-q^2}{4\pi\varepsilon_0}\left(\frac{1}{x} + \frac{2}{a-x}\right) \tag{12.14}$$

In the vicinity of the minimum this potential may be expanded as:

$$V(x) = \frac{-q^2}{4\pi\varepsilon_0 a}\left[5.83 + 24\left(\frac{x}{a}\right)^2 - 17\left(\frac{x}{a}\right)^3 + \cdots\right] \tag{12.15}$$

The non-linear coefficient D in (12.1) then follows through the identification of similar terms, yielding:

$$D = -51\frac{q^2}{4\pi\varepsilon_0 m a^4} \tag{12.16}$$

Assuming a typical interatomic distance a of 5 Å, we find $D = 2 \times 10^{41}\,\text{m}^{-1}\,\text{s}^{-2}$, which leads to a Miller parameter $\delta^{(2\omega)}$ of 6×10^9 SI for a density N of 6×10^{28} atoms m^3 (see Eq. (12.13)). Our simple model, which attributes the origin of optical non-linearity in a material to an asymmetry in the atomic potentials between its constituent atoms, leads to results that are compatible with experimentally determined values. We present in Complement 12.A a more rigorous quantum mechanical derivation using density matrix formalism.

Finally, it is obvious that the quantity χ_2 is fundamentally a tensor. If the fundamental wave has components (E_x, E_y, E_z), the second harmonic wave will have components (P_x, P_y, P_z) which in the most general case, will be determined by all quadratic combinations between the components E_x, E_y, and E_z, i.e:

$$\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \varepsilon_0 \underline{\underline{\chi_2^{(2\omega)}}} \begin{bmatrix} E_x^2 \\ E_y^2 \\ E_z^2 \\ E_y E_z \\ E_z E_x \\ E_x E_y \end{bmatrix} \tag{12.17}$$

The tensors $\underline{\underline{\chi_2^{(2\omega)}}}$ possess properties that depend strongly upon the symmetries of a given crystal. We will not enter here into a classification of these different symmetries, but rather content ourselves with describing the operation of various optoelectronic devices which employ the non-linear optical properties of these materials.

12.3 An electromagnetic description of quadratic non-linear optical interaction

The anharmonic displacement of electrons in a non-linear material subject to excitation by an electromagnetic wave leads to light emission at frequencies of $2\omega, 3\omega, \ldots$. The calculation of optical emission due to these non-linear source terms is, as is the case in linear optics, taken into account using the powerful formalism provided by Maxwell's equations. We recall here Maxwell's equations in the absence of free charges ($\rho = 0$) and conduction current ($\mathbf{j} = 0$):

$$\begin{aligned} &\nabla \cdot \mathbf{E} = 0 \\ &\nabla \cdot \mathbf{B} = 0 \\ &\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}\mathbf{B} \\ &\nabla \times \mathbf{B} = \mu_o \frac{\partial}{\partial t}\mathbf{D} = \mu_o \frac{\partial}{\partial t}(\varepsilon_o \mathbf{E} + \mathbf{P}) \end{aligned} \tag{12.18}$$

where $\mathbf{P}$ is the polarization vector and is the sum of the linear polarization $\mathbf{P}_l$ given by (12.8) or $\mathbf{P}_l = \varepsilon_0(n_{\text{op}}^2 - 1)\mathbf{E}$ (with $n_{\text{op}}^2 = 1 + \chi_L$) and the non-linear polarization $\mathbf{P}_{\text{nl}}$ as given for instance in (12.10). The ensemble of equations in (12.18) may be written in the more compact form:

$$\nabla^2 \mathbf{E} - \left(\frac{n_{\text{op}}}{c}\right)\frac{\partial^2}{\partial t^2}\mathbf{E} = \mu_o \frac{\partial^2}{\partial t^2}\mathbf{P}_{\text{nl}} \tag{12.19}$$

Propagation equation with a source term

In order to simplify the notation and concentrate on the concepts, we will assume

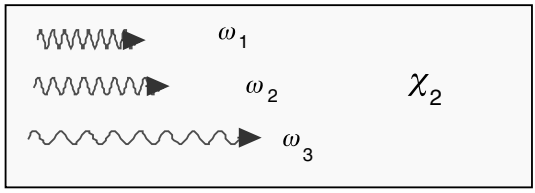


Fig. 12.2. In a medium characterized by a non-linear optical susceptibility χ_2, three waves ω_1, ω_2, and ω_3 can exchange energy during propagation by intermediary of the quadratic polarization terms.

the system is one dimensional and that all electromagnetic waves propagate along the Oz axis.

As illustrated in Fig. 12.2, we will consider three electromagnetic waves with frequencies ω_i ($i = 1$, 2, and 3 with $\omega_3 > \omega_2 > \omega_1$) which propagate within the non-linear crystal. These three waves may be described as *envelope* functions, i.e. they may be written in the form:

$$E^{\omega_i}(z,t) = \frac{1}{2}[E_i(z)\mathrm{e}^{\mathrm{i}(\omega_i t - k_i z)} + \text{c.c.}] \tag{12.20}$$

We recall that the envelope function approximation stipulates that the amplitude variations $E_i(z)$ are small in comparison to the scale of the involved wavelengths $\lambda_i = 2\pi/k_i$. This is to say:

$$\left|\frac{\mathrm{d}E_i}{\mathrm{d}z}k_i\right| \gg \frac{\mathrm{d}^2E_i}{\mathrm{d}z^2} \tag{12.21}$$

The envelope function approximation

Furthermore, the solutions to Maxwell's linear equations (without *source terms* as in (12.19)) lead to the dispersion relations:

$$\omega_i = k_i\frac{c}{n_i} \tag{12.22}$$

These three waves interact in the crystal through the effect of non-linear polarization. We will suppose that the non-linear polarization vector is oriented along Oy and we will only consider its norm P_{nl}. The non-linear source term possesses terms in $E^{\omega_1}(z,t)E^{\omega_2}(z,t)$ which will generate terms of angular frequency $\omega_3 = \omega_1 + \omega_2$. Similarly, source terms in $E^{\omega_3}(z,t)E^{\omega_2}(z,t)$ (or $E^{\omega_3}(z,t)E^{\omega_1}(z,t)$) will generate waves of frequency $\omega_1 = \omega_3 - \omega_2$ (or $\omega_2 = \omega_3 - \omega_1$). We therefore expect the following three mechanisms:

mechanism A: $\omega_3 = \omega_1 + \omega_2$ sum frequency
mechanism B: $\omega_1 = \omega_3 - \omega_2$ difference frequency
mechanism C: $\omega_2 = \omega_3 - \omega_1$ difference frequency

The non-linear source term corresponding to mechanism *B*, for example, is given by immediate generalization of (12.10):

$$P_{\mathrm{nl}}(z,t) = \frac{\varepsilon_0 \chi_2}{2}\left[E_2(z)^* E_3(z) \mathrm{e}^{\mathrm{i}((\omega_3-\omega_2)t-(k_3-k_2)z)} + \mathrm{c.c.}\right] \tag{12.23}$$

It is important to note at this stage that the presence of $\mathrm{e}^{+\mathrm{i}\omega_i t}$ or $\mathrm{e}^{-\mathrm{i}\omega_i t}$ in the time dependence will be accompanied, respectively, by $E_i(z)$ or $E_i^*(z)$ in the envelope function. Moreover, it is clear that the non-linear susceptibility $\chi^{(\omega_3-\omega_2)}$ has no reason to equal the second harmonic generation susceptibility $\chi_2^{(2\omega)}$. Nonetheless, to keep our notation simple, we will write this susceptibility under the general form of χ_2.

We need now only substitute the expression for the non-linear polarization (12.23) as the source term in Maxwell's equations, (12.19). We will calculate each of the terms in this equation. The first may be written as:

$$\begin{aligned}\frac{\partial^2}{\partial z^2}E^{\omega_1}(z,t) &= \frac{1}{2}\left\{\frac{d^2}{\mathrm{d}z^2}\left[E_1(z)\mathrm{e}^{\mathrm{i}(\omega_1 t-k_1 z)}\right] + \mathrm{c.c.}\right\} \\ &= \frac{1}{2}\left[\left(\frac{\mathrm{d}^2}{\mathrm{d}z^2}E_1 - 2\mathrm{i}k_1\frac{\mathrm{d}}{\mathrm{d}z}E_1 - k_1^2 E_1\right)\mathrm{e}^{\mathrm{i}(\omega_1 t-k_1 z)} + \mathrm{c.c.}\right] \\ &\approx \frac{1}{2}\left[\left(-2\mathrm{i}k_1\frac{\mathrm{d}}{\mathrm{d}z}E_1 - k_1^2 E_1\right)\mathrm{e}^{\mathrm{i}(\omega_1 t-k_1 z)} + \mathrm{c.c.}\right]\end{aligned} \tag{12.24}$$

where we have made use of the envelope function approximation (12.21). The propagation equation, (12.19), may therefore be written as:

$$\begin{aligned}&-\frac{1}{2}\left(2\mathrm{i}k_1\frac{\mathrm{d}}{\mathrm{d}z}E_1 + k_1^2 E_1\right)\mathrm{e}^{\mathrm{i}(\omega_1 t-k_1 z)} + \mathrm{c.c.} \\ &= -\left(\frac{\omega_1 n_1}{c}\right)^2\left(\frac{E_1}{2}\mathrm{e}^{\mathrm{i}(\omega_1 t-k_1 z)} + \mathrm{c.c.}\right) - \frac{\varepsilon_0\mu_0\omega_1^2}{2}(\chi_2 E_3 E_2^* \mathrm{e}^{\mathrm{i}(\omega_1 t+(k_3-k_2)z)})\end{aligned} \tag{12.25}$$

As $\omega_1 = k_1 c/n_1$ and $\omega_1 = \omega_3 - \omega_2$, this equation can be simplified and rewritten more compactly as:

$$\mathrm{i}k_1\frac{\mathrm{d}}{\mathrm{d}z}E_1\mathrm{e}^{-\mathrm{i}k_1 z} + \mathrm{c.c.} = \frac{\varepsilon_0\mu_0\omega_1^2}{2}\left[\chi_2 E_3 E_2^* \mathrm{e}^{-\mathrm{i}(k_3-k_2)z} + \mathrm{c.c.}\right] \tag{12.26}$$

We are now in a position to account for the variation of the complex amplitude $E_1(z)$ in the material (as a function of the evolution of the amplitudes of the two source-waves $E_2(z)$ and $E_3(z)$), which describes the difference frequency mechanism $\omega_1 = \omega_3 - \omega_2$. Clearly, the two other mechanisms $\omega_3 = \omega_1 + \omega_2$ and $\omega_2 = \omega_3 - \omega_1$ occur concurrently. The amplitudes $E_1(z)$, $E_2(z)$, and $E_3(z)$ are related by

the following series of three differential equations:

$$\omega_1 = \omega_3 - \omega_2 \leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z}E_1 = -\mathrm{i}\frac{\omega_1}{2n_1 c}\chi_2 E_3 E_2^* \mathrm{e}^{-\mathrm{i}\Delta k z}$$

$$\omega_2 = \omega_3 - \omega_1 \leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z}E_2 = -\mathrm{i}\frac{\omega_2}{2n_2 c}\chi_2 E_1^* E_3 \mathrm{e}^{-\mathrm{i}\Delta k z} \tag{12.27}$$

$$\omega_3 = \omega_1 + \omega_2 \leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z}E_3 = -\mathrm{i}\frac{\omega_3}{2n_3 c}\chi_2 E_1 E_2 \mathrm{e}^{\mathrm{i}\Delta k z}$$

Second-order parametric interactions

We have left the reader to derive the last two equations in (12.27). We note that the added frequencies $+\omega_i$ lead to an E_i term in (12.27), whereas the subtracted frequencies $-\omega_i$ yield an E_i^* term on the right-hand side of the equation. Δk is the *phase mismatch* given by:

$$\Delta k = k_3 - k_1 - k_2 \tag{12.28}$$

Phase mismatch

12.4 Optical second harmonic generation

We describe here the particular case where $\omega = \omega_1 = \omega_2$ and $\omega_3 = 2\omega$. This situation corresponds to *optical second harmonic generation*. Equation (12.27) then becomes:

$$\frac{\mathrm{d}}{\mathrm{d}z}E_\omega = -\mathrm{i}\frac{\omega}{2n_\omega c}\chi_2 E_{2\omega} E_\omega^* \mathrm{e}^{-\mathrm{i}\Delta k z}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}E_{2\omega} = -\mathrm{i}\frac{\omega}{n_{2\omega} c}\chi_2 E_\omega^2 \mathrm{e}^{\mathrm{i}\Delta k z} \tag{12.29}$$

with new notation for E_ω being evident. The origin of the Δk term is that the source field $P_{2\omega}$ is synchronous with the field E_ω (which generates it), which has a propagation speed given by c/n_ω, whereas the field $E_{2\omega}$ possesses its own propagation velocity as imposed by $n_{2\omega}$ (see Fig. 12.3).

We will suppose for the time being that the non-linear conversion efficiency is weak and consequently that $E_\omega(z)$ remains fairly constant over the interaction volume, i.e. that $E_\omega(z) = E_0$. The complete calculation is provided in Complement 12.C. The second differential equation in (12.29) can be easily integrated along the interaction path (the non-linear crystal length from 0 to L) to find:

$$E_{2\omega}(L) = -\mathrm{i}\frac{\omega}{n_{2\omega} c}\chi_2 E_0^2 \frac{\mathrm{e}^{\mathrm{i}\Delta k L} - 1}{\mathrm{i}\Delta k} \tag{12.30}$$

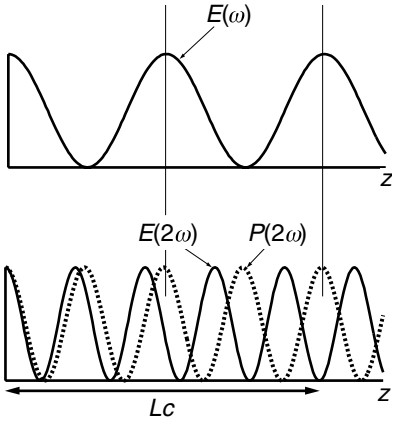


Fig. 12.3. Origin of phase mismatch between the non-linear polarization field $P(2\omega)$ and the second harmonic field $E(2\omega)$.

We can use the technique employed earlier in (1.77) for time-dependent perturbations, which involves multiplying the numerator and denominator in (12.30) by $e^{-i\Delta kL/2}$, leading to:

$$|E_{2\omega}(L)| = \frac{\omega}{n_{2\omega}c}\chi_2 E_0^2 L \,\mathrm{sinc}\left(\frac{\Delta kL}{2}\right) \tag{12.31}$$

In general, we are more interested in the optical power converted into second harmonic radiation, which relates to the amplitude $E_{2\omega}$ according to:

$$P_{2\omega} = \frac{1}{2Z_0} n_{2\omega} |E_{2\omega}(L)|^2 \tag{12.32}$$

where Z_0 is the vacuum impedance ($Z_0 = (\mu_0/\varepsilon_0)^{1/2} = 377\,\Omega$). The second harmonic frequency conversion efficiency is then:

$$\frac{P_{2\omega}}{P_\omega} = 2\frac{Z_0^3}{n_{2\omega} n_\omega^2}(\omega\varepsilon_0\chi_2 L)^2 \mathrm{sinc}^2\left(\frac{\Delta kL}{2}\right) P_\omega \tag{12.33}$$

Second harmonic generation yield

This last equation brings to light the role played by the phase mismatch Δk. If this term is null, the conversion efficiency increases quadratically with the interaction length L ($\mathrm{sinc}(0) = 1$), and there is a permanent exchange of energy between the fundamental wave and the second harmonic wave across the interaction trajectory, until the pump power is depleted of course (see Complement 12.C). On the other hand, if Δk is not zero, the efficiency varies as $\sin^2(\Delta kL/2)$, which oscillates periodically along the interaction path (see Fig. 12.4). In this case, the energy periodically cycles between both waves during propagation in the crystal. The

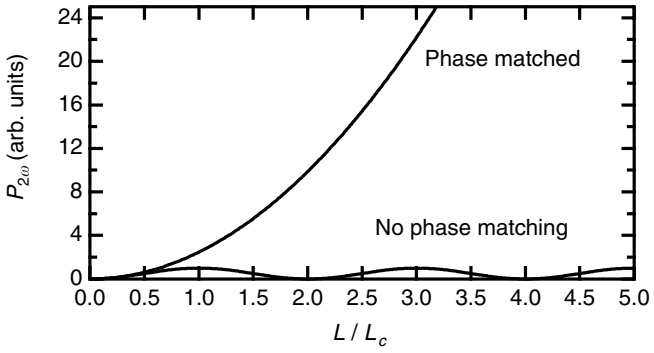


Fig. 12.4. Variation in power of the second harmonic signal $P_{2\omega}$ as a function of propagation distance in the non-linear medium for phase matched and phase mismatched conditions.

distance L_c at the end of which the energy transfer is at a maximum is given by $\Delta k L_c = \pi$, i.e:

$$L_c = \frac{\lambda_0}{4(n_{2\omega} - n_\omega)} \tag{12.34}$$

Phase mismatch length

where λ_0 is the vacuum wavelength of the fundamental electromagnetic wave. This quantity, L_c, is often inappropriately designated in the literature as a *coherence length*. We prefer the term phase *mismatch length*. The existence of this maximum conversion length clearly relates to the natural dispersion occurring in non-linear material (except near resonances where dispersion may be anomalous), where $n(2\omega) > n(\omega)$. Examination of (12.29) and (12.30) allow one to understand the oscillatory nature of the energy exchange between fundamental and second harmonic waves. It can be shown that for $z < L_c$, the work performed by the wave ω on the dipoles at 2ω is positive, whereas the inverse situation occurs over $L_c < z < 2L_c$.

Example

We wish to use a GaAs crystal to convert a beam of 10.6 μm radiation into light at 5.3 μm. GaAs is a material with a very strong non-linear susceptibility at 10.6 μm. We will suppose in this experimental configuration that $\chi_2 = 100\,\text{pm V}^{-1}$. Other required constants are:

$n(5.3\,\mu\text{m}) - n(10.6\,\mu\text{m}) = 2.5 \times 10^{-2}$ (see the Sellmeier relation in Complement 7.B)
$n(5.3\,\mu\text{m}) \approx n(10.6\,\mu\text{m}) \approx 3$

The optimal length of GaAs to obtain the greatest conversion efficiency is then

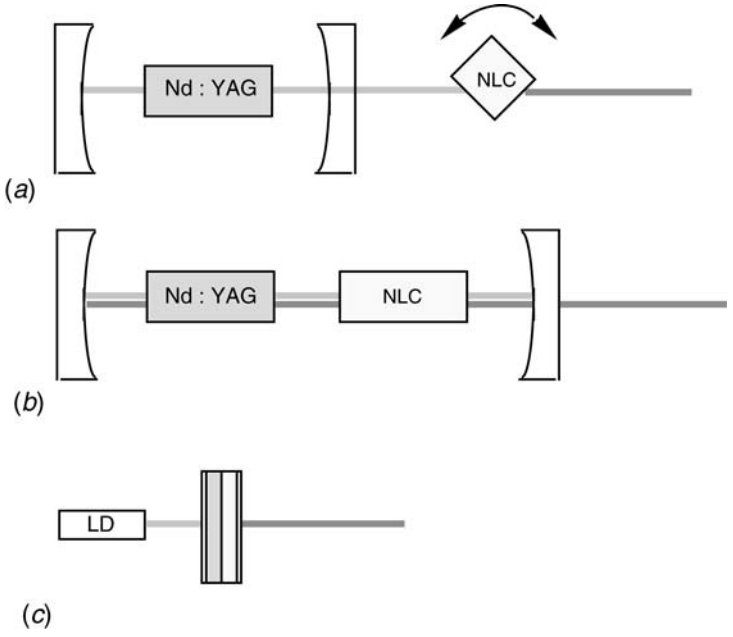


Fig. 12.5. Different configurations for second harmonic generation at 530 nm using a Nd: YAG laser at 1.06 μm. (*a*) Non-linear crystal (NLC) placed outside the optical cavity oriented to allow angular phase matching, (*b*) intracavity conversion using a NLC oriented for angular phase matching, and (*c*) microchip conversion with mirrors deposited on either side of the Nd: YAG and NLC layers.

given by $L_c = 10.6\,\mu\text{m}/(4 \times 2.5 \times 10^{-2}) = 106\,\mu\text{m}$. The accessible efficiency is then:

$$\frac{P_{2\omega}}{P_\omega} = 2 \times \left(\frac{377}{3}\Omega\right)^3 \times [1.8 \times 10^{14}\,\text{s}^{-1} \times 10^{-4}\,\text{m} \times 8.85 \times 10^{-12}\,\text{Fd}\,\text{m}^{-1} \times 10^{-10}\,\text{m}\,\text{V}^{-1}]^2 P_\omega$$

i.e:

$$\frac{P_{2\omega}}{P_\omega} = 10^{-11} P_\omega(\text{W}\,\text{cm}^{-2})$$

Assuming $P_\omega = 1\,\text{MW}\,\text{cm}^{-2}$, the conversion efficiency is only 10^{-5}, which is disappointing. The cause of this low conversion efficiency is the phase mismatch which limits the effective interaction length to 100 μm.

Figure 12.5 describes various experimental configurations for achieving optical frequency doubling: (*a*) the doubling crystal can be placed outside the laser cavity, (*b*) the doubling crystal can be placed inside the optical cavity to benefit from optical feedback from the cavity mirrors (see Complement 12.E), and (*c*) a very

compact (*microchip*) configuration involving the deposition of mirrors on either side of adjacent YAG and doubling crystal layers – the structure can be pumped using a laser diode.

12.5 Manley–Rowe relations

Examination of (12.27) reveals a lack of aesthetic symmetry between the equations (presence of different ω_i and n_i in each equation). To make these equations symmetrical, Manley and Rowe suggested the introduction of the following quantities:

$$A_i = \sqrt{\frac{n_i}{\omega_i}}\,\mathrm{E}_i \tag{12.35}$$

The interest in using these modified amplitudes is that they relate directly to the flux $\Phi_{\omega i}$ of photons (having energy $\hbar\omega_i$). In fact, the light power is given by:

$$P_{\omega i} = \frac{1}{2Z_0} n_i |E_i|^2 = \frac{1}{2Z_0} \omega_i |A_i|^2 \tag{12.36}$$

so that:

$$\Phi_{\omega i} = \frac{P_{\omega i}}{\hbar\omega_i} = \frac{1}{2\hbar Z_0} |A_i|^2 \tag{12.37}$$

in which neither ω_i nor n_i appear. Substituting this newly defined amplitude into the conversion equations, (12.27), we obtain the more symmetric set:

$$\begin{aligned}
\omega_1 = \omega_3 - \omega_2 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_1 = -\mathrm{i}\kappa A_3 A_2^* \mathrm{e}^{-\mathrm{i}\Delta k z} \\
\omega_2 = \omega_3 - \omega_1 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_2 = -\mathrm{i}\kappa A_1^* A_3 \mathrm{e}^{-\mathrm{i}\Delta k z} \\
\omega_3 = \omega_1 + \omega_2 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_3 = -\mathrm{i}\kappa A_1 A_2 \mathrm{e}^{\mathrm{i}\Delta k z}
\end{aligned} \tag{12.38}$$

Non-linear coupling equations between photon fluxes

where κ is the non-linear coupling coefficient:

$$\kappa = \frac{1}{2} \frac{\chi_2}{c} \sqrt{\frac{\omega_1 \omega_2 \omega_3}{n_1 n_2 n_3}} \tag{12.39}$$

Equation (12.38) is particularly powerful if we can assume that the material is non-dispersive (i.e. $\Delta k = 0$). This condition can be achieved, as we will see later,

using artificial phase matching techniques. We will now illustrate the superiority of (12.38) over (12.27). To do so, we multiply each equation i in (12.38) by A_i^* and add to it its complex conjugate. We then obtain:

$$\begin{aligned} A_1^* \frac{\mathrm{d}}{\mathrm{d}z} A_1 + A_1 \frac{\mathrm{d}}{\mathrm{d}z} A_1^* &= -\mathrm{i}\kappa A_1^* A_2^* A_3 + \mathrm{i}\kappa A_1 A_2 A_3^* \\ A_2^* \frac{\mathrm{d}}{\mathrm{d}z} A_2 + A_2 \frac{\mathrm{d}}{\mathrm{d}z} A_1^* &= -\mathrm{i}\kappa A_1^* A_2^* A_3 + \mathrm{i}\kappa A_1 A_2 A_3^* \\ A_3^* \frac{\mathrm{d}}{\mathrm{d}z} A_3 + A_3 \frac{\mathrm{d}}{\mathrm{d}z} A_3^* &= -\mathrm{i}\kappa A_1 A_2 A_3^* + \mathrm{i}\kappa A_1^* A_2^* A_3 \end{aligned} \tag{12.40}$$

which leads to the Manley–Rowe equations:

$$\frac{\mathrm{d}}{\mathrm{d}z}(|A_1|^2) = \frac{\mathrm{d}}{\mathrm{d}z}(|A_2|^2) = -\frac{\mathrm{d}}{\mathrm{d}z}(|A_3|^2) \tag{12.41}$$

The Manley–Rowe equations express the fact that the photon flux is conserved. Therefore, in a sum frequency experiment, photons with energy $\hbar\omega_3$ can only be produced at the detriment of the numbers of constituent $\hbar\omega_1$ and $\hbar\omega_2$ photons. Alternately, the Manley–Rowe relations reveal an asymmetry in the energy exchanges. In a difference frequency mechanism, ($\omega_1 = \omega_3 - \omega_2$), we see that photons $\hbar\omega_3$ are consumed, while $\hbar\omega_2$ photons are created even if they are already present. We will return to this aspect later on (see Section 12.8).

We therefore see how optical frequency conversion processes can be interpreted in corpuscular terms. The photons $\hbar\omega_1$, $\hbar\omega_2$, and $\hbar\omega_3$ cannot exchange their energies unless both energy and momentum conservation between the particles is conserved by the interaction (see Fig. 12.6):

$$\begin{aligned} \hbar\omega_1 + \hbar\omega_2 &= \hbar\omega_3 \\ \mathbf{k}_1 + \mathbf{k}_2 &= \mathbf{k}_3 \end{aligned} \tag{12.42}$$

Energy and momentum conservation laws for photons

12.6 Parametric amplification

Equation (12.38) shows that it is possible to amplify a signal with frequency ω_1 (*signal beam*) through non-linear interaction with a secondary wave at ω_3 (*pump beam*) (see Fig. 12.7). As expected from energy conservation considerations (12.42), a third beam is produced at a frequency ω_2 (referred to as an *idler beam*).

To simplify these calculations, we will make the same assumption as in the case

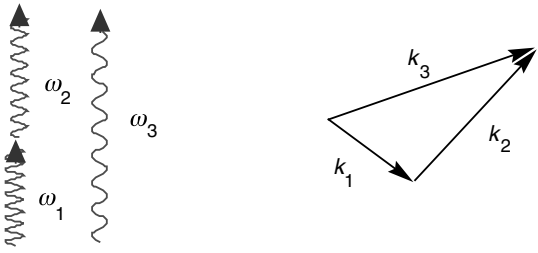


Fig. 12.6. Energy transfer conditions between three waves ω_1, ω_2, and ω_3 during interaction in a non-linear medium may be interpreted according to energy and momentum conservation laws which govern the constituent photons.

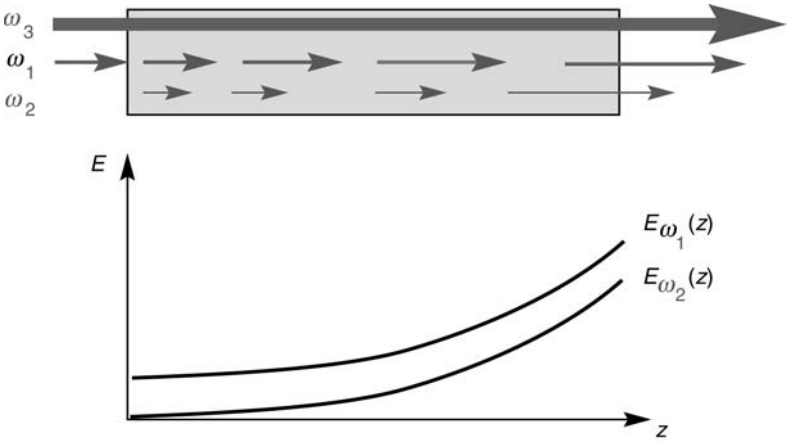


Fig. 12.7. In parametric amplification, an incident signal at frequency ω_1 is amplified while it propagates in a non-linear medium via quadratic non-linear interaction with a pump beam at ω_3. Conservation of photon energy requires the creation of an *idler* beam at ω_2.

of second harmonic generation, and suppose that the amplitude of the pump beam $A_3(z)$ is not affected by the non-linear interaction. In other words, the conversion efficiency is presumed to be weak (substantiated earlier in the example of Section 12.3 and later on in this section). In such a situation, we speak of an *undepleted pump beam*. The complete calculation is given in Complement 12.C.

Having made this approximation, (12.38) takes the simple form:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} A_1 &= -\mathrm{i}gA_2^* \\ \frac{\mathrm{d}}{\mathrm{d}z} A_2^* &= \mathrm{i}gA_1 \end{aligned} \tag{12.43}$$

where the *parametric gain* g (cm^{-1}) is given by:

$$g = \frac{1}{2}\frac{\chi_2}{c}\sqrt{\frac{\omega_1\omega_2\omega_3}{n_1n_2n_3}}A_3 = \frac{1}{2}\frac{\chi_2}{c}\sqrt{\frac{\omega_1\omega_2}{n_1n_2}}E_3 \tag{12.44}$$

and A_3 and E_3 are the constant amplitudes of the pump beam. The solution to this system of differential equations, (12.43), is trivial. We find[1]:

$$\begin{aligned} A_1(z) &= A_1(0)\mathrm{ch}(gz) - \mathrm{i}A_2(0)^*\mathrm{sh}(gz) \\ A_2(z) &= A_2(0)\mathrm{ch}(gz) - \mathrm{i}A_1(0)^*\mathrm{sh}(gz) \end{aligned} \tag{12.45}$$

We will suppose that no idler beam has been injected at $z = 0$, i.e. that $A_2(0) = 0$. The evolution of the signal and idler beams may then be written as:

$$\begin{aligned} A_1(z) &= A_1(0)\mathrm{ch}(gz) \\ A_2(z) &= -\mathrm{i}A_1(0)^*\mathrm{sh}(gz) \end{aligned} \tag{12.46}$$

These two waves experience exponential amplification as a function of propagation distance in the material (as in the case of laser amplification) – this situation is referred to as *parametric amplification*. Figure 12.7 shows the variation in the power levels E_ω for different waves in the material.

Example

We wish to amplify *a* $\lambda = 10.6\,\mu\mathrm{m}$ wave in a GaAs crystal which we will assumed to be phase matched (i.e. $\Delta k = 0$ – we will see later on how this can be achieved in practice). We will suppose for the particular configuration used in this experiment that $\chi_2 = 100\,\mathrm{pm\,V^{-1}}$. We will calculate the parametric gain for the crystal assuming a pump power $P_{\omega 3}$ of $5\,\mathrm{MW\,cm^{-2}}$ at $5.3\,\mu\mathrm{m}$.

The corresponding electric field amplitude is:

$$E_3 = (2P_{\omega 3}Z_0/n_{\mathrm{op}})^{1/2} = (2 \times 5 \times 10^{10}\,\mathrm{W\,m^{-2}} \times 377\,\Omega/3)^{1/2} = 3.5 \times 10^6\,\mathrm{V\,m^{-1}}$$

The parametric gain is therefore $g = \chi_2/2c \times \omega/n_{\mathrm{op}} \times E_3$ or:

$$g = (10^{-10}\,\mathrm{mV^{-1}})/(6 \times 10^8\,\mathrm{m\,s^{-1}}) \times 1.8 \times 10^{14}\,\mathrm{s^{-1}} \times 3.5 \times 10^6\,\mathrm{V\,m^{-1}}/3 \\ = 0.35\,\mathrm{cm^{-1}}$$

which is quite low considering the rather high pump power.

It is worth remarking that Eq. (12.38) states that *it is not possible to amplify a signal with frequency* ω_3 *using pump beams* ω_1 *or* ω_2 *of lesser frequency*. In fact, assuming an undepleted pump beam $A_1(z) = A_1$ a similar calculation to the preceding one shows that if $\omega_3 < \omega_1$, then the variations in the signal amplitude $A_3(z)$ will be sinusoidal (in $\cos(gz)$) and not exponential. Mathematically, this has its origin in the presence of E_i or E_i^* in the non-linear coupling equations depending on whether we increase or decrease the frequency (see Eq. (12.38)) – a point insisted upon earlier. In more physical terms, this comes from the fact that, according to Manley–Rowe, for each photon ω_1 annihilated to produce a photon

[1] ch = cosh, sh = sinh

ω_3, a photon ω_2 should also be consumed. And such photons are not present in the system! We will return to this aspect later in Section 12.8.

12.7 Optical parametric oscillators (OPOs)

As demonstrated in the preceding example, the optical gain associated with parametric amplification is quite low. We therefore cannot expect reasonable conversion efficiencies from an experimental set-up such as the one depicted in Fig. 12.7. This situation is completely analogous to that encountered in Section 4.4 in the course of our study of laser amplification. We will therefore make use of the same stratagem which allowed us to obtain laser oscillations from even weakly amplifying gain media – *optical feedback*.

In this case, we place a non-linear crystal in the centre of an optically resonant Fabry–Pérot cavity (see Complement 9.D) consisting of two mirrors with complex reflection coefficients r_e and r_s with $r_i = (R_i)^{1/2} e^{i\phi_i}$ (where $i = e$ or s for entrance or exit, respectively). R_i are the mirror reflectances and ϕ_i their associated phase shifts ($\phi_i = \pi$ for metallic mirrors).

We will study two distinct cases (see Fig. 12.8).

1. *A simply resonant oscillator*. The mirrors possess reflectances R_i close to unity for the signal wave, but virtually null for the idler wave.
2. *A doubly resonant oscillator*. The mirrors possess reflectances R_i close to unity for both the signal and idler waves.

In either case, mirrors fulfilling these requirement can be obtained through the deposition of dielectric multilayers (see Complement 9.D). In the following development the pump power is supposed to be undepleted. The complete case is treated in Complement 12.F.

12.7.1 Simply resonant optical parametric oscillators (SROPOs)

This case bears a complete resemblance to the laser cavity. To keep from repeating ourselves needlessly, we will assume as in Section 4.4 that the active crystal fills the entire space between the two mirrors (i.e. that L is simultaneously the length of the active crystal and the optical cavity). As the cavity is not resonant for the complementary wave, the amplitude $A_2(z)$ remains weak and the amplification coefficient for the wave during a single pass through the medium is given by (12.46) to be $\cosh(gL)$, even after multiple passes.

We consider a signal beam with a weak amplitude a_1 (starting at the entrance mirror M_e and directed towards the exit mirror M_s) subject to parametric amplification using a pump beam with frequency ω_3 (see Fig. 12.8*a*). Once at the exit mirror M_s, the complex amplitude of the signal wave is then $a_1 \cosh(gL) e^{-ik_1 L}$. The

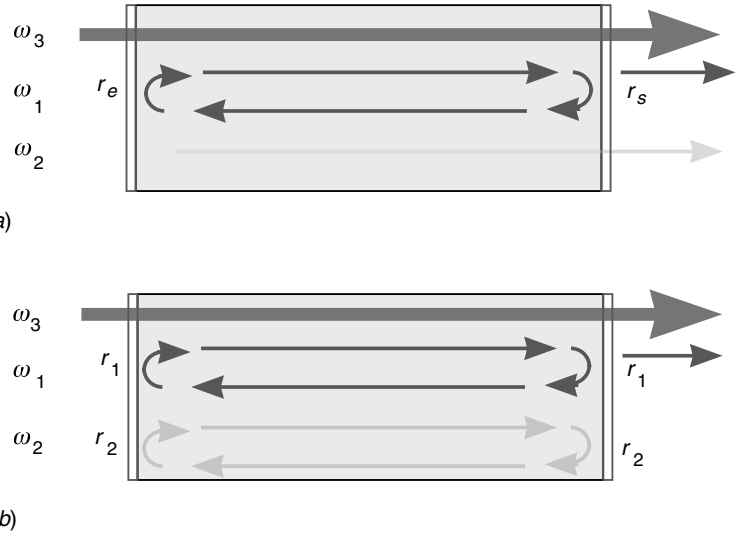


Fig. 12.8. Principle of optical parametric oscillation (OPO). The non-linear crystal is placed in a cavity between two mirrors M_e and M_s. When the parametric gain that the signal wave ω_1 extracts from the pump beam ω_3 exceeds the cavity losses, the system begins to oscillate at frequency ω_1. We can either make the signal wave ω_1 oscillate (using a simply resonant cavity) (*a*), or (*b*) make both ω_1 and the complementary wave ω_2 oscillate (using a doubly resonant cavity).

wave is then reflected, its amplitude becoming $a_1 r_s \cosh(gL) e^{-ik_1 L}$, and goes on to travel through the crystal in the opposite direction. The waves ω_1 and ω_3 then propagate in opposite directions and the phase matching condition cannot be met, with the result being that no amplification takes place over this leg of the trip. Its amplitude is then $a_1 r_s \cosh(gL) e^{-2ik_1 L}$, once it returns to the entrance mirror M_e. Finally, the signal wave experiences a second reflection at the entrance mirror, and its amplitude becomes $a_1 r_e r_s \cosh(gL) e^{-2ik_1 L}$, and so on.

The signal wave in the resonator at $z = 0$ then results from the sum of the contributions made by these multiple passes:

$$A_1(0) = a_1\{1 + r_e r_s \mathrm{ch}(gL) e^{-2ik_1 L} + [r_e r_s \mathrm{ch}(gL) e^{-2ik_1 L}]^2 + \cdots\} \tag{12.47}$$

or:

$$A_1(0) = \frac{a_1}{1 - r_e r_s \mathrm{ch}(gL) e^{-2ik_1 L}} \tag{12.48}$$

The system begins to oscillate spontaneously once the denominator in (12.48) goes to zero, i.e. given the two conditions:

$$\sqrt{R_e R_s}\,\mathrm{ch}(g_{\text{threshold}}L) = 1 \tag{12.49a}$$

$$\frac{\phi_e + \phi_s}{2} - k_1 L = m\pi,\ m = 0, 1, 2, \ldots \tag{12.49b}$$

The first condition indicates that the system possesses a threshold gain $g_{\text{threshold}}$ above which it begins to oscillate *spontaneously* at frequency ω_1. The optical parametric oscillator therefore 'fragments' the photon at ω_3 into two photons (ω_1, ω_2), such that $\omega_1 + \omega_2 = \omega_3$. Clearly, as was the case with laser oscillations, parametric oscillations are built up from quantum noise within the cavity. As the gain is generally weak, $\cosh(gL) \approx 1 + (gL)^2/2$ and condition (12.49a) becomes for $R_e = R_s = R \approx 1$:

$$\frac{g_{\text{threshold}}L}{\sqrt{2}} \approx \sqrt{1 - R} \tag{12.50}$$

SROPO threshold

This last condition indicates that the system begins to oscillate as soon as the cavity gain $g_{\text{threshold}}$ exceeds the cavity losses $(1 - R)^{1/2}$ due to light transmission by the mirrors.

The second condition is the phase restriction, which fixes the allowed modes in the cavity. The OPO can only oscillate over these allowed modes. Nonetheless, we may well ask at what wavelength the OPO will begin to oscillate if it is pumped at a certain frequency ω_3. In fact, contrary to the case of laser oscillations, the oscillation frequency is not determined by an atomic transition. All pairs (ω_1, ω_2) which satisfy $\omega_1 + \omega_2 = \omega_3$, are permitted a priori. Also, the phase conditions, (12.49a), are only slightly discriminating for macroscopic crystal sizes (>1 mm) and so do not allow one to select a particular (ω_1, ω_2) pair. We will see in Complement 12.B that it is the phase matching condition $(\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3)$ which determines a particular ω_1 frequency. This condition is generally obtained by rotating the crystal into some particular crystallographic orientation which allows the phase condition $(\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3)$ to be satisfied for a desired ω_1.

Example

We wish to induce oscillations in a cavity possessing crystalline GaAs phase matched for a wavelength of 9 μm and pumped by a 5 μm pump beam. The GaAs crystal is 5 mm long and possesses a non-linear susceptibility of 100 pm V^{-1}. The crystal cavity is coated at both ends by mirrors with reflectances of 98%. We wish to know the threshold power level for the pump beam to initiate optical parametric oscillations in the structure.

The wavelength of the complementary wave is $1/(1/5 - 1/9)$ μm or 11.9 μm. The frequencies of the signal and complementary waves are then 2.1×10^{14} and $1.6 \times 10^{14}\ s^{-1}$, respectively.

The parametric threshold gain for the system is given by (12.50) to be:

$$g_{\text{threshold}} = \sqrt{2} \times (1 - 0.98)^{1/2}/0.5\,\text{cm} = 0.4\,\text{cm}^{-1}$$

This parametric threshold gain corresponds to the pump electric field E_3 given by (12.44):

$$\begin{aligned} g_{\text{threshold}} &= 10^{-10}\,\text{m V}^{-1}/(6 \times 10^{8}\,\text{m s}^{-1}) \times (2.1 \times 10^{14}\,\text{s}^{-1} \\ &\quad \times 1.6 \times 10^{14}\,\text{s}^{-1})^{1/2} \times (E_{3,\text{threshold}}\,\text{V m}^{-1})/3 \\ &= 40\,\text{m}^{-1}, \text{ i.e. } E_{3,\text{threshold}} = 4 \times 10^{6}\,\text{V m}^{-1} \end{aligned}$$

The pump power at oscillation threshold is therefore:

$$P_{3,\text{threshold}} = 1/(2Z_0) n_{\text{op}} E^2_{3,\text{threshold}} = 6\,\text{MW cm}^{-2}$$

which is considerable.

12.7.2 Doubly resonant optical parametric oscillator (DROPO)

In this case, the complementary wave ω_2 also experiences optical feedback from the mirrors and is no longer negligible in comparison with the signal wave ω_1. We are therefore led to the formalism presented in Complement 9.D and used to describe optical wave propagation in a Fabry–Pérot cavity. We label as r_1 and $r_2 (r_i = (R_i)^{1/2} \mathrm{e}^{\mathrm{i}\phi_i}$ for $i = 1, 2)$ the reflectivity coefficients for the two mirrors (assumed to be the same for the entrance and exit mirrors) for the signal and complementary waves, respectively. We introduce the complex amplitude vector $\mathbf{A}(z)$:

$$\mathbf{A}(z) = \begin{bmatrix} A_1(z)\mathrm{e}^{-\mathrm{i}k_1 z} \\ A_2^*(z)\mathrm{e}^{\mathrm{i}k_2 z} \end{bmatrix} \tag{12.51}$$

Propagation of the complementary and signal waves between the cavity entrance and exit mirrors, and described by the system of equations in (12.45), can be placed into matrix form as:

$$\mathbf{A}_R(L) = \mathbf{M}\mathbf{A}(0) \tag{12.52}$$

where $\mathbf{M}$ is the matrix:

$$\mathbf{M} = \begin{bmatrix} \text{ch}(gL)\mathrm{e}^{-\mathrm{i}k_1 L} & -\mathrm{i}\,\text{sh}(gL)\mathrm{e}^{-\mathrm{i}k_1 L} \\ \mathrm{i}\,\text{sh}(gL)\mathrm{e}^{\mathrm{i}k_2 L} & \text{ch}(gL)\mathrm{e}^{\mathrm{i}k_2 L} \end{bmatrix} \tag{12.53}$$

and $\mathbf{A}_R$ designates the amplitude vector propagating towards the right. Once the wave is at the exit mirror, a portion of the light is reflected and is given by the vector:

$$\mathbf{A}_L(L) = \mathbf{RMA}(0), \text{ with } \mathbf{R} = \begin{bmatrix} r_1 & 0 \\ 0 & r_2^* \end{bmatrix} \tag{12.54}$$

where $\mathbf{A}_L$ corresponds to the complex amplitude directed towards the left. As in the preceding case, the signal and complementary waves cannot interact with the pump wave in this propagation direction as they cannot satisfy the phase matching requirement (12.42). The amplitude vector $\mathbf{A}_L$ at the entrance mirror prior to reflection is then:

$$\mathbf{A}_L(0) = \mathbf{LRMA}(0), \text{ with } \mathbf{L} = \begin{bmatrix} \mathrm{e}^{-\mathrm{i}k_1 L} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}k_2 L} \end{bmatrix} \tag{12.55}$$

Finally, the complementary and signal waves are reflected at the entrance mirror and the amplitude vector becomes:

$$\mathbf{A}_R(0) = \mathbf{RLRMA}(0) \tag{12.56}$$

The oscillation condition can be obtained by demanding that the sum of all these contributions diverge, as in the preceding case. We could also require that these waves be reproduced identically at $z = 0$ (a condition equivalent to the first), i.e. that $\mathbf{A}_R(0) = \mathbf{A}(0)$. This imposes the condition:

$$\det(\mathbf{RLRM} - \mathrm{Id}) = 0 \tag{12.57}$$

where Id is the 2×2 identity matrix. This last condition is obtained by performing the matrix multiplications leading to:

$$\det\begin{bmatrix} (r_1^2 \mathrm{ch}(gL)\mathrm{e}^{-2\mathrm{i}k_1 L} - 1) & -\mathrm{i}r_1^2 \mathrm{sh}(gL)\mathrm{e}^{-2\mathrm{i}k_1 L} \\ \mathrm{i}r_2^{*2} \mathrm{sh}(gL)\mathrm{e}^{2\mathrm{i}k_2 L} & (r_2^{*2} \mathrm{ch}(gL)\mathrm{e}^{2\mathrm{i}k_2 L} - 1) \end{bmatrix} = 0 \tag{12.58}$$

The evaluation of this determinant poses no particular problem and yields the condition:

$$\mathrm{ch}(gL) = \frac{R_1 \mathrm{e}^{-\mathrm{i}\Delta\phi_1} R_2 \mathrm{e}^{\mathrm{i}\Delta\phi_2} + 1}{R_1 \mathrm{e}^{-\mathrm{i}\Delta\phi_1} + R_2 \mathrm{e}^{\mathrm{i}\Delta\phi_2}} \tag{12.59}$$

where the dephasing terms $\Delta\phi_i$ are given by:

$$\Delta\phi_i = 2(k_i L - \phi_i), \; i = 1, 2 \tag{12.60}$$

Equation (12.59) yields a non-trivial solution only if the following three conditions are met:

$$k_1 L - \phi_1 = m\pi$$
$$k_2 L - \phi_2 = n\pi \tag{12.61}$$
$$\text{ch}(gL) = \frac{1 + R_1 R_2}{R_1 + R_2}$$

The first two conditions in (12.61) are nothing else than the stationary wave conditions in the cavity for the signal and idler waves. The third condition indicates that the cavity establishes parametric oscillations for the two waves once the parametric gain g exceeds the mirror losses R_1 and R_2. As in the case of a simple resonator, we have $\cosh(gL) \approx 1 + (g_{\text{threshold}}L)^2/2$ and $R_1 \approx R_2 \approx 1$ where $g_{\text{threshold}}$ is the parametric gain at oscillation threshold. This threshold gain is given by:

$$g_{\text{threshold}} L \approx \sqrt{(1 - R_1)(1 - R_2)} \tag{12.62}$$

DROPO threshold

This last expression is to be compared with condition (12.50) for the simply resonant cavity case. The doubly resonant configuration decreases the oscillation threshold by a factor of:

$$F = \frac{g_{\text{threshold simple, resonance}}}{g_{\text{threshold double, resonance}}} \approx \sqrt{\frac{2}{1 - R_2}} \tag{12.63}$$

Example

We will now calculate the oscillation threshold for the OPO described in the example in Section 12.6.1, i.e. for a 5 mm long GaAs crystal phase matched for a 9 μm wave, pumped using a 5 μm beam. Both ends of the crystal are coated this time with mirrors having 98 and 99.8% reflectances for the signal and idler waves, respectively. We seek the threshold power for the pump beam to initiate optical parametric oscillations in the structure.

The oscillation threshold is reduced by a factor F given in (12.63):

$F = [2/(2 \times 10^{-3})]^{1/2}$ or a factor of 31

The optical pump power at oscillation threshold is therefore:

$P_{3,\text{threshold}} = 6/31\,\text{MW cm}^{-2}$ or $190\,\text{kW cm}^{-2}$

We now summarize some of the particularities associated with the phenomenon of parametric oscillation:

- All frequency pairs (ω_1, ω_2) satisfying $\omega_1 + \omega_2 = \omega_3$, where ω_3 is the oscillation frequency of the pump beam, can be generated by an OPO. The explicit pair of frequencies which will undergo OPO is determined by the stationary phase

conditions (12.49b) and (12.61) (to begin with, the cavity mirrors must be properly selected) and, especially, by the requirement for momentum conservation between the involved photons (12.42). We can alter the pair (ω_1, ω_2) and thereby the wavelengths generated by the OPO by changing the phase matching conditions (for instance, by altering the orientation of the non-linear crystal). In practice, an OPO acts as a tuneable light source (see Complement 12.B).

- As was the case for lasers, parametric oscillations are built up from noise in the cavity. This noise does not result from spontaneous emission but from quantum noise within the cavity (see Section 2.5).

12.8 Sum frequency, difference frequency, and parametric oscillation

We saw in the beginning of Section 12.2 that interaction between three waves with frequencies ω_1, ω_2, and ω_3 in a crystal possessing a non-linear second-order susceptibility will lead to the generation of sum and difference frequencies. We have also just seen that parametric oscillations may also result. It is now time to draw a distinction between these different regimes. We therefore consider two waves of angular frequencies ω_2 (pump) and ω_1 (signal) incident upon a non-linear crystal. We will assume $\omega_2 > \omega_1$. Equation (12.38) lists the three processes we can expect:

Difference frequency generation (DFG) 4 $\omega_3' = \omega_2 - \omega_1$

We will assume that the crystal orientation has been chosen in order to satisfy the DFG phase matching condition, i.e. $\Delta\mathbf{k} = \mathbf{k}_2 - \mathbf{k}_1 - \mathbf{k}_3' = \mathbf{0}$. After a small change in notation, the two relevant equations from (12.38) may be written:

$$\begin{aligned} \omega_3' = \omega_2 - \omega_1 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_1 = -\mathrm{i} g_2 A_3^* \\ \omega_1 = \omega_2 - \omega_3' &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_3 = -\mathrm{i} g_2^* A_1 \end{aligned} \tag{12.64}$$

where $g_2 = \kappa A_2$ and may be supposed real. This system can be trivially solved to obtain the conversion efficiency corresponding to the ratio of the converted power $P_3(L)$ to the power of the incident signal $P_1(0)$:

$$\frac{P_3(L)}{P_1(0)} = \frac{\omega_3'}{\omega_1} \sin^2(g_2 L) \tag{12.65}$$

Energy is transferred from the signal wave ω_1 to the difference wave ω_3' and the pump wave ω_2. This translates in terms of 'photon number' notation (see Chapter 2) as $|n_1, n_2, n_3'\rangle \rightarrow |n_1 - 1, n_2 + 1, n_3' + 1\rangle$. At the end of a distance $L = \pi/2g_2$,

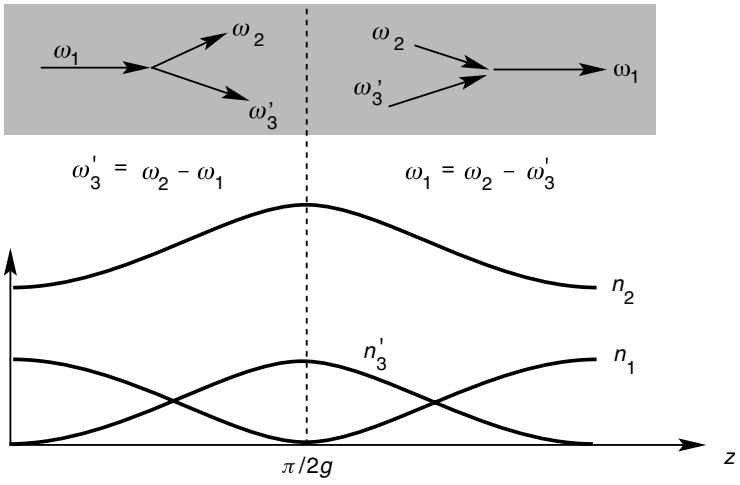


Fig. 12.9. Various energy transfer processes involved in difference frequency generation in a non-linear medium.

the energy transfer process changes direction, and photons from the ω'_3 and ω_2 waves recombine to yield signal photons with ω_1, which translates into $|n_1, n_2, n'_3\rangle \rightarrow |n_1 + 1, n_2 - 1, n'_3 - 1\rangle$ (see Fig. 12.9). Consequently, photon splitting and recombination may simultaneously be present in parametric interactions. We will see in Complement 12.F that the prevailing mechanism depends on the relative phases between the different waves.

We note that, even at the *conversion efficiency* maximum (i.e. for $g_2 L = \pi/2$), the conversion ratio P_3/P_1 cannot exceed unity:

$$\left(\frac{P_3}{P_1}\right)_{\max} = \frac{\omega'_3}{\omega_1} < 1 \tag{12.66}$$

This last expression is reminiscent of the fact that parametric conversion of a photon with energy $\hbar\omega_1$ into a photon of lesser energy $\hbar\omega'_3$ cannot proceed with an efficiency greater than $\hbar\omega'_3/\hbar\omega_1$. For weak efficiencies in comparison with unity, (12.65) can be expressed more simply by noting that:

$$\frac{P_3(L)}{P_1(0)} \approx \frac{\omega_3}{\omega_1}(g_2 L)^2 \tag{12.67}$$

which gives, using (12.44) to express g_2 as a function of the electric field E_2, and (12.32) to express E_2 as a function of the power per unit surface area P_2:

$$\frac{P_3(L)}{P_1(0)} = \frac{Z_0^3}{2n_1 n_2 n_3}(\varepsilon_0 \chi_2 \omega_3 L)^2 P_2 \tag{12.68}$$

Conversion efficiency for optical frequencies

We thus find an expression similar to Eq. (12.33) obtained within the context of second harmonic generation.

Sum frequency generation (SFG) $\omega_3 = \omega_2 + \omega_1$

Here, we again suppose that the crystal orientation has been chosen to satisfy the SFG phase matching condition $\Delta\mathbf{k} = \mathbf{k}_2 + \mathbf{k}_1 - \mathbf{k}_3 = \mathbf{0}$. The two relevant equations from (12.38) are in this case:

$$\begin{aligned} \omega_3 = \omega_1 + \omega_2 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_3 = -\mathrm{i} g_2 A_1 \\ \omega_1 = \omega_3 - \omega_2 &\leftrightarrow \frac{\mathrm{d}}{\mathrm{d}z} A_1 = -\mathrm{i} g_2^* A_3 \end{aligned} \tag{12.69}$$

As before, this system of equations can be solved trivially, with the result revealing a behaviour identical to that given earlier in (12.65). This time, however, energy is transferred from the signal and pump waves, to the sum wave ω_3. In 'photon number' notation, this translates to: $|n_1, n_2, n_3\rangle \rightarrow |n_1 - 1, n_2 - 1, n_3 + 1\rangle$. At the end of the distance $L = \pi/2g_2$, the energy transfer changes direction and the photons from the sum wave ω_3 'fragment' into ω_1 and ω_2. This process may be written as $|n_1, n_2, n_3\rangle \rightarrow |n_1 + 1, n_2 + 1, n_3 - 1\rangle$ (see Fig. 12.10).

Equation (12.69) is quite similar to (12.64), obtained during discussion of difference frequency generation. Expression (12.68) for frequency conversion efficiency therefore remains identical in the case of optical sum frequency generation. Additionally, we now have a maximum efficiency given by:

$$\left(\frac{P_3}{P_1}\right)_{\max} = \frac{\omega_3}{\omega_1} > 1 \tag{12.70}$$

This efficiency can therefore exceed unity. There is no mystery in this observation as the increase in the power of the beam resulting from the conversion of photons with energy $\hbar\omega_1$ into photons of greater energy $\hbar\omega_3$ proceeds at the expense of the pump beam photons $\hbar\omega_2$.

Sum or difference frequency generation then proceeds in the non-linear crystal according to which of the associated phase matching conditions are met (i.e. either $\mathbf{k}_3 = \mathbf{k}_1 + \mathbf{k}_2$, or $\mathbf{k}'_3 = \mathbf{k}_2 - \mathbf{k}_1$, respectively).

Parametric oscillation

This case differs from the two others. In parametric oscillation, the pump energy photon $\hbar\omega_2$ 'fragments' *spontaneously* into two lesser energy photons $\hbar\omega_1$ and $\hbar\omega'_3$ (see Fig. 12.11). 'Spontaneously' here is meant to imply that no external effect has been introduced from the outside to seed the $\hbar\omega_1$ or $\hbar\omega'_3$ photons.

In analogy with (4.28a,b), from Section 4.6 on laser oscillations, and given the

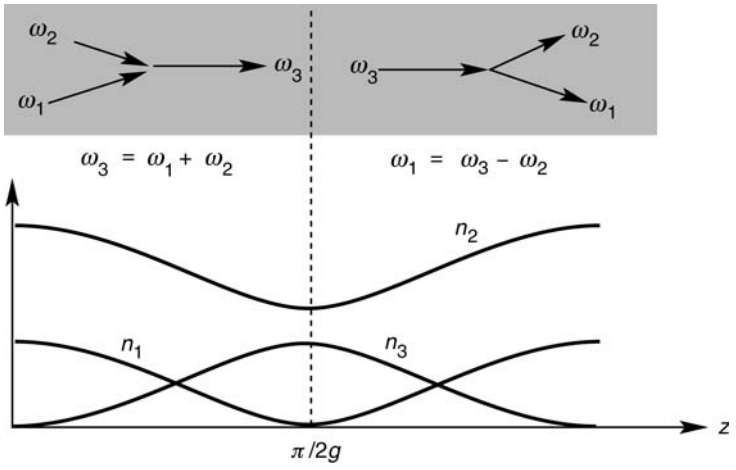


Fig. 12.10. Various energy transfer processes involved in sum frequency generation in a non-linear medium.

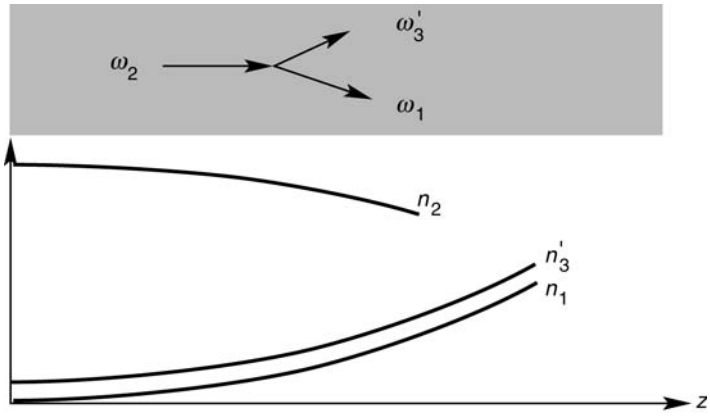


Fig. 12.11. Energy transfer process involved in parametric oscillation.

Manley–Rowe equations, we might jump to the conclusion that the output power for a parametric oscillator is given by:

$$\frac{P_1}{\omega_1} = \frac{P_3}{\omega_3'} = \frac{P_{2,\text{threshold}}}{\omega_2}\left(\frac{P_2}{P_{2,\text{threshold}}} - 1\right) \tag{12.71}$$

WRONG!

where $P_{2,\text{threshold}}$ is imposed by the oscillation conditions (12.50) or (12.62). This equation is quoted by many authors, but is never valid. In fact, the situation is far more complex and discussed in Complement 12.F, see Eqs. (12.F.20) and (12.F.33).

The spontaneous production of photons $\hbar\omega_1$ or $\hbar\omega_3'$ – which are necessary to

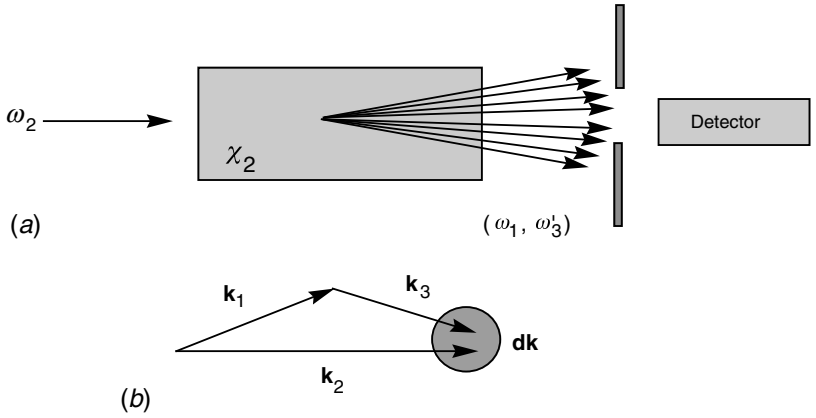


Fig. 12.12. Principle behind parametric fluorescence. (*a*) Photons with energy $\hbar\omega_2$ fragment spontaneously within the non-linear crystal into two photons with energies $\hbar\omega_1$ and $\hbar\omega_3'$. (*b*) Only photon pairs oriented towards the detector and satisfying momentum conservation to within **dk** are detected.

seed OPO oscillations – originate from *parametric fluorescence*. This physical phenomenon cannot be described within the classical framework developed up to this point (and is similar to the impasse reached in our discussion of laser oscillation and the necessity for introducing spontaneous emission into our model – see Chapter 4). Figure 12.12 depicts the conditions under which this effect can be observed. A pump beam of photons at $\hbar\omega_2$ with wavevector $\mathbf{k}_2$ is incident upon a non-linear crystal. Aligned with this beam, we have placed a detector sensitive only to photons with energy $\hbar\omega_1$ and $\hbar\omega_3'$. The angular acceptance of the experimental set-up is such that only photons originating from the non-linear crystal with wavevector $\mathbf{k}_2$ to within **dk** can be detected (see Fig. 12.12).

Spontaneously, an ensemble of photon pairs $(\hbar\omega_1, \hbar\omega_3')$ with $\hbar\omega_1 + \hbar\omega_3' = \hbar\omega_2$ will be created through parametric interaction. As this frequency conversion process must also respect momentum conservation, only photon pairs having total momenta $\mathbf{k}_1 + \mathbf{k}_3'$ that 'fall' within the acceptance window of the apparatus will be detected. This parametric fluorescence is responsible for triggering the parametric oscillations.

FURTHER READING

R. W. Boyd, *Non Linear Optics*, Academic Press, Boston (1992).
P. Butcher and D. Cotter, *The Elements of Nonlinear Optics*, Cambridge University Press, Cambridge (1990).
Y. S. Shen, *The Principles of Non Linear Optics*, Wiley, New York (1984).
R. L. Sutherland, *Handbook of Nonlinear Optics*, Marcel Dekker Inc., New York (1996).
A. Yariv, *Quantum Electronics*, 3rd Edn, Wiley, New York (1989).

Complement to Chapter 12

12.A A quantum model for quadratic non-linear susceptibility

In Section 12.2, within the mechanical model framework for non-linear optics, we were able to see how second-order non-linear susceptibility arises from asymmetries in the mechanical potentials (12.1) which confine the motions of electrons agitated by electromagnetic waves. While such a model is intuitively fruitful, it is in strong disagreement with the quantum mechanical nature of electrons in matter. We offer here such a quantum derivation. The very powerful and predictive formalism that we shall develop is founded on the density matrix approach presented in Section 1.7.

We consider a quantum system described by a Hamiltonian H_0 and possessing a spectrum of discrete non-degenerate states $|i\rangle$ with energy E_i, such that $H_0|i\rangle = E_i|i\rangle$. Examples of such systems are shown in Fig. 12.A.1. At thermal equilibrium, the associated density matrix ρ_0 is a diagonal matrix with diagonal elements $\rho_{ii}^{(0)} = n_i$ equal to the populations of the levels E_i as given by Fermi–Dirac statistics. The population densities are given in units of 'cm^{-2}' or 'cm^{-3}' depending on whether the system is two or three dimensional. The system is then excited by an electromagnetic wave, possessing an electric field given by:

$$E(t) = \tilde{E}\mathrm{e}^{\mathrm{i}\omega t} + \tilde{E}\mathrm{e}^{-\mathrm{i}\omega t} \tag{12.A.1}$$

In the $\mathbf{D}\cdot\mathbf{E}$ dipole approximation (see Chapter 3), interaction between the quantum system and the electromagnetic field is described by the Hamiltonian $H(t) = -q\hat{z}E(t)$. Under the influence of this time-dependent perturbation, evolution of the elements ρ_{ij} of the density matrix is given by the Schrödinger equation:

$$\frac{\partial \rho_{ij}}{\partial t} = \frac{1}{\mathrm{i}\hbar}[H_0 - q\hat{z}E(t), \rho]_{ij} - \Gamma_{ij}(\rho - \rho^{(0)})_{ij} \tag{12.A.2}$$

To keep things simple, we will suppose that only two relaxation rates Γ_{ij} are required to describe the system: $\Gamma_1 = 1/T_1$, for $i = j$, is the *population relaxation*

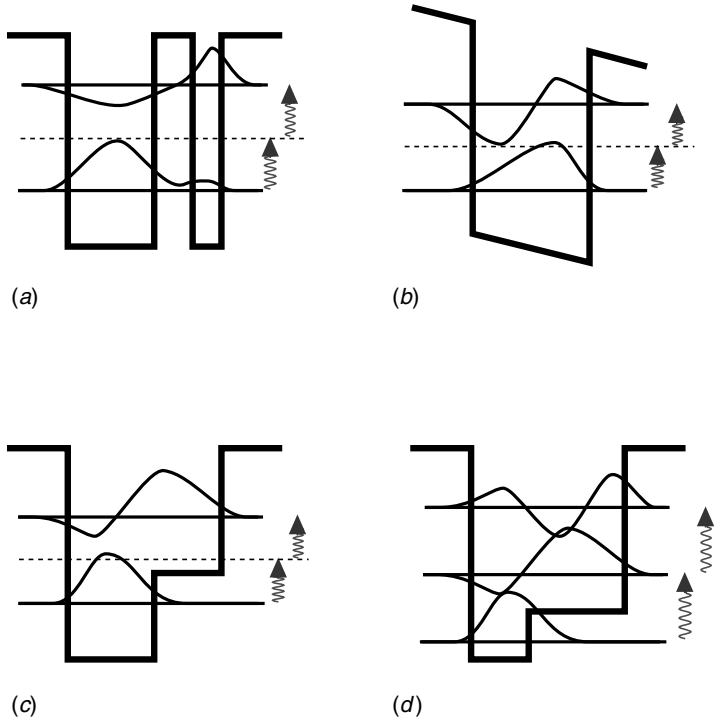


Fig. 12.A.1. A few examples of asymmetric quantum systems based on quantum wells.

rate (inelastic mechanism), i.e. the average de-excitation rate for an electron in an excited state, and $\Gamma_2 = 1/T_2$ is the *off diagonal relaxation rate* or *dephasing constant* corresponding to elastic scattering of electronic phases over various states. The density matrix can be expressed in an increasing series in the electric field $E(t)$ in the following iterative way:

$$\rho(t) = \sum_n \rho^{(n)}(t) \tag{12.A.3}$$

with:

$$\frac{\partial \rho_{ij}^{(n+1)}}{\partial t} = \frac{1}{\mathrm{i}\hbar}\{[H_0, \rho^{(n+1)}]_{ij} - \mathrm{i}\hbar\Gamma_{ij}\rho_{ij}^{(n+1)}\} - \frac{1}{\mathrm{i}\hbar}[q\hat{z}E(t), \rho^{(n)}]_{ij} \tag{12.A.4}$$

The term of order n ($\rho^{(n)}$) acts as a source term in the differential equation in $\rho^{(n+1)}$. Through recurrence, we see in this last equation that the contribution of the electric field $E(t)$ to the dipolar matrix element $\rho_{ij}^{(n)}$ is a polynomial of order n. The electrical polarization in the Oz direction is then given by the average value of the $q\hat{z}$ operator obtained (as with any other observable – see Section 1.7), by evaluating the trace of $\rho q\hat{z}$:

$$P(t) = \mathrm{Tr}(q\hat{z}\rho) = q\sum_i (\hat{z}\rho)_{ii} = q\sum_i\sum_k z_{ik}\rho_{ki} \tag{12.A.5}$$

The contribution of the nth order density matrix to electronic polarization is therefore a polarization of order n:

$$P^{(n)}(t) = \mathrm{Tr}(\rho^{(n)}q\hat{z}) \tag{12.A.6}$$

Now, the first two elements in the expansion for electronic polarization $P(t)$ of the system follow from expansion of the polynomial:

$$P(t) = \varepsilon_0\chi^{(1)}\tilde{E}\mathrm{e}^{\mathrm{i}\omega t} + \varepsilon_0\chi^{(2)}_{2\omega}\tilde{E}^2\mathrm{e}^{2\mathrm{i}\omega t} + \text{c.c.} + \varepsilon_0\chi^{(2)}_0|\tilde{E}^2| \tag{12.A.7}$$

where $\chi^{(1)}$, $\chi^{(2)}_{2\omega}$, and $\chi^{(2)}_0$ are the optical linear and quadratic susceptibilities (for second harmonic generation and optical rectification), respectively. The term in 2ω is due to the product of a term in $\mathrm{e}^{\mathrm{i}\omega t}$ with another in $\mathrm{e}^{\mathrm{i}\omega t}$, whereas the constant term results from the product of $\mathrm{e}^{\mathrm{i}\omega t}$ with $\mathrm{e}^{-\mathrm{i}\omega t}$. The term in $\chi^{(2)}_0$ gives rise to a constant electric field under the effect of an electromagnetic wave – this corresponds to the phenomenon of *optical rectification* and will not be dealt with further in this book. We are only interested in the second harmonic generation term in 2ω (or more generally, frequency $n\omega$), i.e. in expressions for $\rho^{(n)}$ of the form:

$$\rho^{(n)}(t) = \tilde{\rho}^{(n)}(\omega)\mathrm{e}^{\mathrm{i}n\omega t} + \tilde{\rho}^{(n)}(-\omega)\mathrm{e}^{-\mathrm{i}n\omega t} \tag{12.A.8}$$

In (12.A.4), we can replace the term $\mathrm{d}\rho^{(n+1)}_{ij}/\mathrm{d}t$ by $\mathrm{i}(n+1)\omega\rho^{(n+1)}_{ij}$. Furthermore, as the Hamiltonian H_0 is diagonal in the basis formed by its stationary states $|i\rangle$, we have:

$$\begin{aligned}[H_0, \rho^{(n+1)}]_{ij} &= \sum_k (H_0)_{ik}(\rho^{(n+1)})_{kj} - \sum_{k'}(\rho^{(n+1)})_{ik'}(H_0)_{k'j} \\ &= (E_i - E_j)(\rho^{(n+1)})_{ij}\end{aligned} \tag{12.A.9}$$

which leads to a simplified version of the recurrence relation in (12.A.4):

$$\tilde{\rho}^{(n+1)}_{ij}(\omega) = \frac{q[z, \tilde{\rho}^{(n)}]_{ij}}{\hbar[(n+1)\omega + \omega_{ij} - \mathrm{i}\Gamma_{ij}]}\tilde{E} \tag{12.A.10}$$

where $\omega_{ij} = (E_i - E_j)/\hbar$ is the Bohr frequency. We begin by applying the recurrence relation (12.A.10) for $n = 0$. This immediately yields:

$$\rho^{(1)}(t) = \tilde{\rho}^{(1)}(\omega)\mathrm{e}^{\mathrm{i}\omega t} + \tilde{\rho}^{(1)}(-\omega)\mathrm{e}^{-\mathrm{i}\omega t} \tag{12.A.11}$$

with:

$$\tilde{\rho}^{(1)}_{ij} = \frac{qz_{ij}(n_j - n_i)}{\hbar[(\omega + \omega_{ij}) - \mathrm{i}\Gamma_{ij}]}\tilde{E} \tag{12.A.12}$$

where z_{ij} is the matrix element $z_{ij} = \langle i|\hat{z}|j\rangle$. Substituting this expression into

(12.A.5) and (12.A.6), we obtain the linear polarization:

$$\tilde{P}^{(1)} = \frac{q^2}{\hbar}\sum_{i,k}\frac{|z_{ki}|^2(n_i - n_k)}{[(\omega + \omega_{ki}) - \mathrm{i}\Gamma_{ki}]}\tilde{E} \tag{12.A.13}$$

Under weak illumination, and at reasonable temperatures ($kT < E_n - E_1$), only level $|1\rangle$ is populated, and therefore only terms with i or $k = 1$ contribute to the linear susceptibility. Comparing it alongside Eq. (12.A.7), we see that the linear susceptibility is given by:

$$\chi^{(1)} = \frac{q^2 n_1}{\varepsilon_0 \hbar}\sum_{k\neq 1}\frac{|z_{k1}|^2}{[(\omega - \omega_{k1}) - \mathrm{i}\Gamma_2]} \tag{12.A.14}$$

which allows us to obtain the absorption α through $\alpha = \omega \mathrm{Im}\chi^{(1)}/n_{\mathrm{op}}c$, i.e:

$$\alpha = \frac{q^2 n_1 \omega}{n_{\mathrm{op}}\varepsilon_0 c\hbar}\sum_{k\neq 1}\frac{\Gamma_2|z_{k1}|^2}{[(\omega - \omega_{k1})^2 + \Gamma_2^2]} \tag{12.A.15}$$

or again:

$$\alpha = \frac{q^2 \pi n_1}{2mn_{\mathrm{op}}\varepsilon_0 c}\sum_{k\neq 1} f_{k1}\frac{\omega}{4\omega_{k1}}\frac{\Gamma_2/\pi}{[(\omega - \omega_{k1})^2 + \Gamma_2^2]} \tag{12.A.16}$$

where $f_{k1} = 2m(\mathrm{E}_k - \mathrm{E}_1)z_{1k}/\hbar^2$ is the oscillator strength for the transition $|1\rangle \rightarrow |k\rangle$, reproducing the result obtained in Chapter 3, i.e. Eq. (3.41).

The next iteration ($n = 1$) yields the quadratic susceptibility. We therefore start with the second-order element in the density matrix:

$$\rho^{(2)}(t) = \tilde{\rho}^{(2)}_{2\omega}(\omega)\mathrm{e}^{2\mathrm{i}\omega t} + \tilde{\rho}^{(2)}_{2\omega}(-\omega)\mathrm{e}^{-2\mathrm{i}\omega t} \tag{12.A.17}$$

given by (12.A.10):

$$[\tilde{\rho}^{(2)}_{2\omega}]_{ij} = \frac{1}{\hbar(2\omega + \omega_{ij} - \mathrm{i}\Gamma_{ij})}[qz, \tilde{\rho}^{(1)}]_{ij}\tilde{E} \tag{12.A.18}$$

We carry out the iteration by substituting (12.A.12) for $\rho^{(1)}$ into this last expression to find:

$$\begin{aligned}[\tilde{\rho}^{(2)}_{2\omega}]_{ij} = \frac{q^2}{\hbar^2}\frac{1}{(2\omega + \omega_{ij} - \mathrm{i}\Gamma_{ij})}\Bigg\{&\sum_l \frac{z_{il}z_{lj}(n_j - n_l)}{[(\omega + \omega_{lj}) - \mathrm{i}\Gamma_{lj}]} \\ &- \sum_l \frac{z_{il}z_{lj}(n_l - n_i)}{[(\omega + \omega_{il}) - \mathrm{i}\Gamma_{il}]}\Bigg\}\tilde{E}^2\end{aligned} \tag{12.A.19}$$

We then make use of the trace of the operator $\rho^{(2)}q\hat{z}$ to find the second-order electronic polarization $P^{(2)}(t)$ (see (12.A.6)). We need only use Eq. (12.A.17) for

definition of the quadratic susceptibility $\chi^{(2)}_{2\omega}$ to obtain (after performing a few acrobatics with the indices!):

$$\chi^{(2)}_{2\omega} = \frac{q^3}{\varepsilon_0 \hbar^2} \sum_i \sum_k \frac{1}{(2\omega + \omega_{ki}) - \mathrm{i}\Gamma_{ki}} \times \sum_l \mu_{ik}\mu_{kl}\mu_{li} \left[\frac{\rho_i - \rho_l}{(\omega + \omega_{li}) - \mathrm{i}\Gamma_{li}} - \frac{\rho_l - \rho_k}{(\omega + \omega_{kl}) - \mathrm{i}\Gamma_{kl}} \right] \tag{12.A.20}$$

This formula (in spite of being a little complicated) allows one to calculate the quadratic optical susceptibility for any quantum system, with the following two cases being of particular interest to us:

- *A two-level system.* After neglecting the antiresonant terms (terms in $\omega + \omega_{ij}$, with $\omega_{ij} > 0$) in (12.A.20), only two terms remain in the summation: i.e. those for which $k = 1, i = 2$, and $l = 1$ or 2. This yields for the quadratic susceptibility $\chi^{(2)}_{2\omega}$:

$$\chi^{(2)}_{2\omega} = \frac{q^3(n_1 - n_2)}{\varepsilon_0 \hbar^2} \frac{z_{12}^2 \delta_{12}}{(\omega - \omega_{21} - \mathrm{i}\Gamma_2)(2\omega - \omega_{21} - \mathrm{i}\Gamma_2)} \tag{12.A.21}$$

  where $\delta_{12} = \langle 2|\hat{z}|2\rangle - \langle 1|\hat{z}|1\rangle = z_{22} - z_{11}$ is the average electron displacement resulting from the transition from level $|1\rangle$ to $|2\rangle$. We see immediately that if the system possesses inversion symmetry (or more precisely, if electrons in the two levels possess the same average positions $z_{11} = z_{22}$) then the quadratic susceptibility goes to zero. A second point is that the quadratic susceptibility is maximized when the product $z_{12}^2 \delta_{12}$ is at a maximum. To maximize the quadratic susceptibility, the mean positions for wavefunctions describing states $|1\rangle$ and $|2\rangle$ must be as distant from one another as possible (δ_{12} maximal) while keeping the overlap between the wavefunctions (z_{12}) as large as possible. A satisfactory compromise between these mutually conflicting requirements can be reached through the *quantum engineering* of molecules and quantum wells for specific applications in non-linear optics. Far from resonance ($\omega_{12} \gg \omega$), (12.A.21) can be written:

$$\chi^{(2)}_{2\omega} = \frac{q}{\varepsilon_0} \frac{(n_1 - n_2)(z_{12}^2 \delta_{12})}{(\hbar\omega_{21}/q)^2} \tag{12.A.22}$$

Quadratic susceptibility for a two-level system far from resonance

  The term in the numerator is the quantity of charge contained by the asymmetric volume $z_{12}^2 \delta_{12}$. From this formula, we see that the quadratic susceptibility is expressed in units of m V^{-1} (and in pm V^{-1} in practice). While this last expression has not been derived within the framework of semiconductor physics, it nonetheless gives useful insight into the behaviour of these materials as one may consider the valence and conduction bands to be represented by two such distinct energy levels separated in energy by E_g. We see, assuming constant

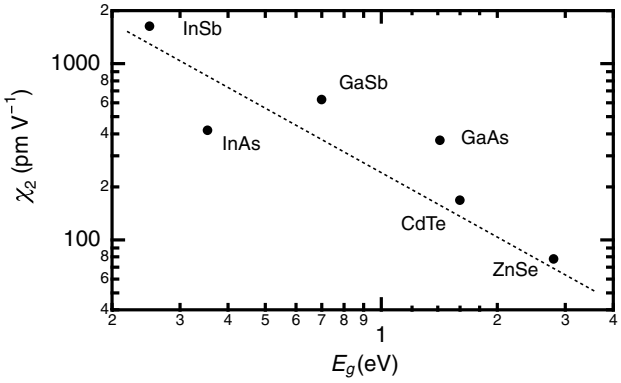


Fig. 12.A.2. Optical quadratic susceptibility at 10.6 μm for various semiconductors as a function of bandgap.

matrix elements z_{ij} (for the purpose of argument), that the non-linear susceptibility for semiconductors may be expected to vary according to $1/E_g^2$. Figure 12.A.2 shows the quadratic optical susceptibility χ_2 for different semiconductors at 10.6 μm as a function of their bandgaps E_g. The behaviour predicted by (12.A.22) is surprisingly well respected in spite of the simplicity of the model we used, which does not take into account the varying ionic tendencies of the chemical bonds, etc.

- *A three-level system.* For a three-level system with constant energy spacing ($E_2 - E_1 = E_3 - E_2 = \hbar\Omega$), (12.A.20) possesses a distinct maximum corresponding to the condition for double resonance $\omega \approx \omega_{12} \approx \omega_{23} \approx \Omega$. One term dominates the sum (12.A.2) in this case:

$$\chi^{(2)}_{2\omega} = \frac{q^3(n_1 - n_2)}{\varepsilon_0 \hbar^2} \frac{z_{12}z_{23}z_{31}}{(\omega - \Omega - i\Gamma_2)(2\omega - 2\Omega - i\Gamma_2)} \tag{12.A.23}$$

Here again, it is clear that the product $z_{12}z_{23}z_{31}$ is null if the system is symmetric. Near resonance ($\omega \approx \Omega$), the quadratic susceptibility becomes:

$$\chi^{(2)}_{2\omega,\max} = \frac{q}{\varepsilon_0} \frac{(n_1 - n_2)(z_{12}z_{23}z_{31})}{(\hbar\Gamma_2/q)^2} \tag{12.A.24}$$

Expression (12.A.22) for non-resonant systems and (12.A.24) for resonant systems lead to fairly reliable predictions for the optical quadratic non-linearities in quantum systems such as molecules or asymmetric quantum wells.

Example

1. *Asymmetric molecules.* Expression (12.A.22) allows one to calculate the order of magnitude of the non-linear susceptibility in a medium constituted of molecules possessing a resonance $\hbar\omega_{12}$ of about 2 eV, a density of 10^{28} molecules m^{-3}, and

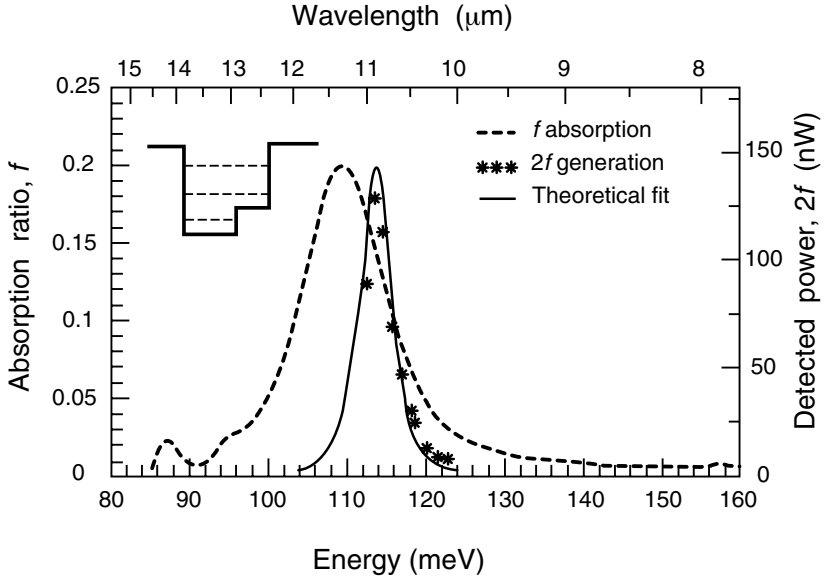


Fig. 12.A.3. Absorption and generation of second harmonic light in a doubly resonant asymmetric quantum well (from E. Rosencher and P. Bois, *Phys. Rev. B* **44**, 11 315 (1991); and P.Boucaud, *et al. Appl. Phys. Lett.* **57**, 215 (1990)).

matrix elements $z_{12} = 1$ Å and $\delta_{12} = 0.1$ Å. The ratio q/ε_0 is 1.8×10^{-8} V m. The average number of molecules in $z_{12}^2 \delta_{12}$ is therefore $10^{28} \times (10^{-10})^2 \times 10^{-11}$ or 10^{-3}. This leads to a quadratic susceptibility of:

$$\chi^{(2)} = 1.8 \times 10^{-8}\,\mathrm{V\,m} \times 10^{-3}/4\,\mathrm{V}^2 = 4.5\,\mathrm{pm\,V}^{-1}$$

which is effectively what is found for non-optimized asymmetric molecules.

2. *Asymmetric quantum wells.* We saw in Chapter 8 that optical intersubband transitions in a quantum well can by suitably represented as transitions between discrete energy levels. In this case the electrons possess the effective mass associated with their subband of origin. During growth, it is possible to introduce asymmetric compositional gradients in the quantum wells to produce wells with asymmetric confinement potentials (see Fig. 12.A.1).

Let us take as an example asymmetric quantum wells made of 6 nm of GaAs and 4.5 nm of $Al_{0.1}Ga_{0.9}As$ separated by 30 nm thick $Al_{0.4}Ga_{0.6}As$ barriers. The wells are doped $2 \times 10^{17}\,\mathrm{cm}^{-3}$ and the filling factor is of the order of $\frac{1}{4}$. A very simple MATHEMATICA program can be used to calculate the discrete energy levels E_1, E_2, and E_3, obtained as solutions to the one-dimensional Schrödinger equation $(p_z^2/2m^* + V(z))|i\rangle = E_i|i\rangle$ as well as the matrix elements z_{12}, z_{23}, and z_{31}. This system has been engineered in such a fashion as to provide identical $E_2 - E_1$ and $E_3 - E_2$ transitions energies equivalent to 10.6 μm. The matrix elements z_{12}, z_{23},

and z_{31} are, respectively, 2.1, 3.0, and 0.38 nm. We note that these matrix elements are larger than in molecules. This results from the small effective mass of electrons in GaAs. As the system is doubly resonant, (12.A.24) can be used to calculate the susceptibility at resonance. We find:

$$\chi^{(2)}_{2\omega,\max} = \frac{1.6 \times 10^{-19}\,\mathrm{C}}{8.85 \times 10^{-12}\,\mathrm{Fd\,m^{-1}}} \frac{2 \times 10^{23}\,\mathrm{m^{-3}}/4 \times (2.4 \times 10^{-27}\,\mathrm{m^3})}{(10^{-2}\,\mathrm{V})^2}$$
$$= 22\,000\,\mathrm{pm\,V^{-1}}$$

The conclusion is that significant non-linear optical effects can be obtained in doubly resonant systems. The experimental data in Fig. 12.A.3 confirm the resonant nature of this non-linear optical effect.

FURTHER READING

P. Butcher and D. Cotter, *The Elements of Nonlinear Optics*, Cambridge University Press, Cambridge (1990).
J. L. Oudar and J. Zyss, *Phys. Rev. A* **26**, 2016 (1982).
E. Rosencher and P. Bois, *Phys. Rev. B* **44**, 11 315 (1991).
Y. S. Shen, *The Principles of Non Linear Optics*, Wiley, New York (1984).

12.B Methods for achieving phase matching in semiconductors

Throughout the course of this chapter, we saw that very low frequency conversion efficiencies resulted when phase matching requirements could not be met (see, for example, (12.33)). This criterion is usually impossible to fulfil in any given material as it demands the absence of optical dispersion over the entire range of frequencies involved. For second harmonic generation, the phase matching condition can be written as:

$$n(2\omega) = n(\omega) \tag{12.B.1}$$

If this condition can be satisfied, the maximum conversion length L_c (12.34) becomes infinite. Of course, all materials (and semiconductors most particularly) exhibit dispersion. The fact that the smaller bandgap semiconductors (which possess increasingly large non-linear second-order susceptibilities – see (12.A.22)) also demonstrate increasingly dispersive behaviour (see Complement 7.C) is unfortunate to say the least!

Luckily, there are several means available to solve this problem. We will describe two commonly used techniques:

- birefringent phase matching
- quasi-phase matching

12.B.1 Birefringent phase matching

Many crystalline materials (and among them certain semiconductors) possess anisotropic optical properties due to their crystal structure. In this section, we will limit our discussion to *uniaxial crystals* – crystals that possess a single symmetry axis about which they can be rotated by any elementary symmetry operation (e.g. rotation by $\pi/3$, $\pi/4$, etc.) and remain unchanged. We will take this axis (referred to as the *optical axis*) to lie in the Oz direction. From elementary optical theory, such a crystal is known to be *birefringent*, i.e. the optical index of the medium is not constant for all propagation directions. More precisely, an electromagnetic wave propagating along the optical axis Oz encounters an index of refraction n_o (subscript 'o' for *ordinary*), whereas a wave propagating in any other direction in the Oxy plane experiences a different index of refraction n_e ('e' for *extraordinary*).

Furthermore, optics tells us that, in any electromagnetic propagation direction, only two principal polarization directions are possible, i.e. a wave polarized along one of these principal directions will remain linearly polarized in this direction while it propagates through the crystal. These two directions are geometrically determined as shown in Fig. 12.B.1. For this we consider the *index ellipsoid* defined by the ensemble of points (x, y, z) such that:

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = 1 \tag{12.B.2}$$

and an electromagnetic wave propagating with wavevector $\mathbf{k}$. Without loss of generality we may suppose the wavevector to be in the Oyz plane. The intersection of the plane perpendicular to $\mathbf{k}$ with the index ellipsoid produces an ellipse (see Fig. 12.B.1). The two principal axes (with one of them always being the Ox axis) are these two principal polarization directions, and the magnitude of each of the two principal axes of this ellipse gives the corresponding index of refraction for each polarization. It is clear from Fig. 12.B.1 that the index of refraction along the Ox axis is independent of the angle θ and is always equal to n_o: this is the *ordinary polarization*. On the other hand, the second allowed polarization direction (the *extraordinary polarization* corresponding to this vector $\mathbf{k}$) possesses an index of refraction that depends on the angle θ and is simply given by:

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \tag{12.B.3}$$

If the ordinary index of refraction n_o is less than the extraordinary index n_e, the

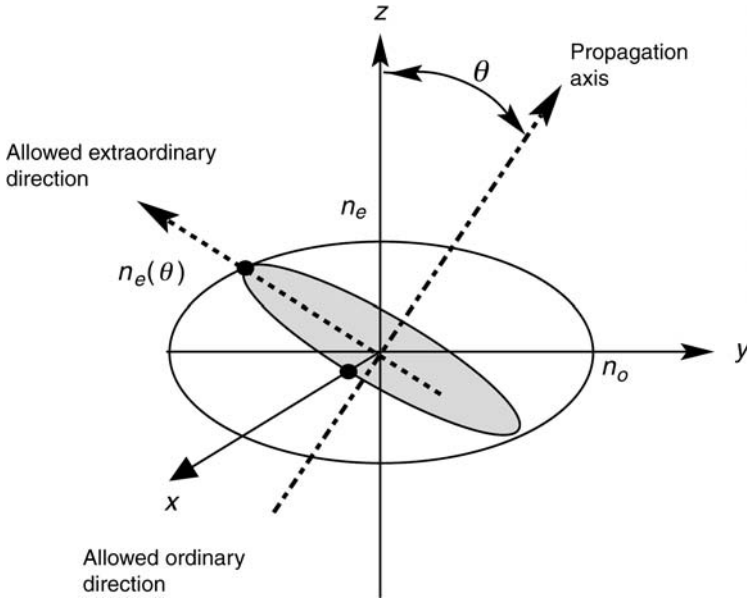


Fig. 12.B.1. Construction for the two allowed polarization directions in a birefringent medium. One of the two indices of refraction (n_e) depends on the angle θ between the wavevector and the optical axis. This figure corresponds to the situation for a negative uniaxial material (i.e. $n_o > n_e$).

ellipsoid will be narrow at the equator, and the crystal is a *positive uniaxial* material. Alternatively, if $n_o > n_e$ (as in Fig. 12.B.1), the crystal is a *negative uniaxial* material.

Birefringent phase matching makes use of this anisotropy. Under certain conditions, it is possible to satisfy the equality condition between the index of refraction $n(\omega)$ in one principal direction and $n(2\omega)$ in the other. More precisely, although in general we have $n(2\omega, \theta) > n(\omega, \theta)$, it may be possible using a negative uniaxial crystal to obtain:

$$n_e(2\omega, \theta_s) = n_o(\omega) \tag{12.B.4}$$

(In a positive uniaxial crystal we should use $n_o(2\omega) = n_e(\omega, \theta_s)$.) To solve this equation, we place ourselves in the Oyz plane and seek the angle θ_s which satisfies this equation. We trace the *normal ellipse for the extraordinary indices*, which is given by the ensemble of points M such that $OM(\theta) = n_e(\omega, \theta)$ and the sphere of ordinary indices given by points M for which $OM(\theta) = n_o(\omega, \theta)$. Figure 12.B.2 shows that this normal index ellipse can be obtained from the index ellipse by a $\pi/2$ rotation. We add to this figure the normal index ellipsoids corresponding to a frequency of 2ω. The intersection between the sphere $n_o(\omega)$ and the ellipse $n_e(2\omega, \theta)$ defines a cone about the optical axis along which phase matching between the wave at ω and its second harmonic 2ω occurs (see also Fig. 12.B.3).

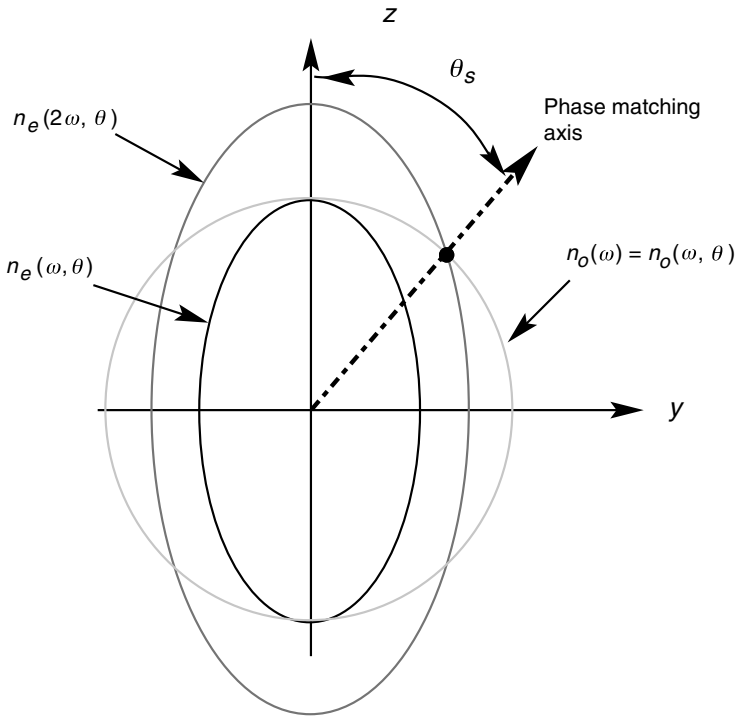


Fig. 12.B.2. Graphical method for obtaining phase matching conditions between an ordinary wave with frequency ω and a second harmonic (with frequency 2ω) extraordinary wave in a negative uniaxial crystal.

Clearly, a solution only exists if $n_o(\omega) > n_e(2\omega)(> n_e(\omega))$, i.e. provided the crystal is sufficiently birefringent. Using equalities (12.B.3) and (12.B.4), we see that the angle θ_s is given by the condition:

$$\frac{1}{n_o(\omega)^2} = \frac{\cos^2\theta_s}{n_o(2\omega)^2} + \frac{\sin^2\theta_s}{n_e(2\omega)^2} \tag{12.B.5}$$

or again:

$$\sin^2\theta_s = \frac{n_o(\omega)^{-2} - n_o(2\omega)^{-2}}{n_e(2\omega)^{-2} - n_o(2\omega)^{-2}} \tag{12.B.6}$$

Birefringent phase matching

This last equation allows one to calculate the phase matching angle directly once the dispersion relations are known for each of the polarizations. Most often these may be obtained through the Sellmeier (Eq. (7.B.2)) or Afromowitz (Eq. (7.B.10)) relations for large and small gap semiconductors, respectively.

Once the phase matching angle is known, we still need to calculate the value for the quadratic optical susceptibility starting from Eq. (12.17). The result will clearly

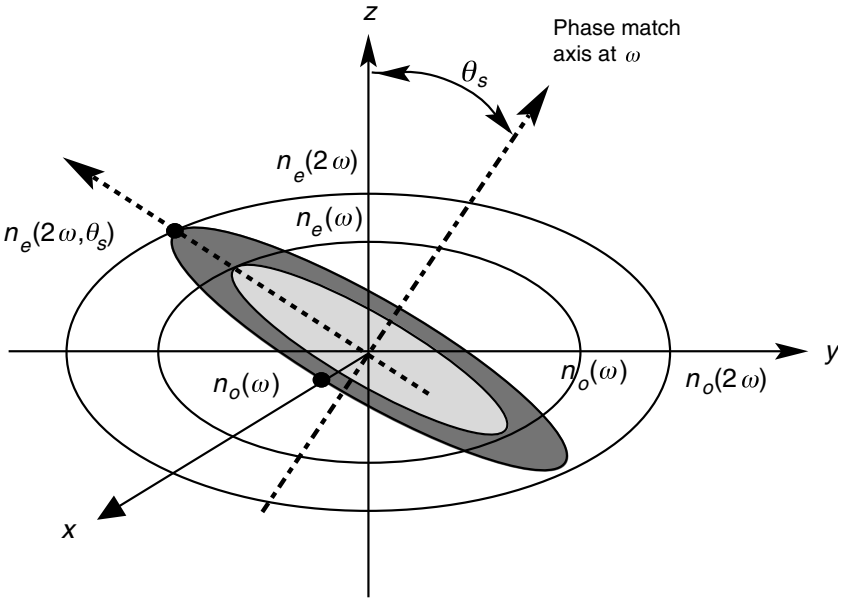


Fig. 12.B.3. The technique of birefringent phase matching represented by index ellipsoids at ω and 2ω for a negative uniaxial medium ($n_e < n_o$).

depend on the various non-zero terms in the tensor $\chi_2^{(2\omega)}$ and therefore upon the crystal symmetry. The reader can find descriptions for all relevant tensors in a non-linear optics reference text such as Sutherland (1996).

Example

Lithium niobate is one of the most popular non-linear optical materials. It is not a semiconductor, but rather a *metallic oxide* type insulator which possesses a strong birefringence and is therefore well suited to phase matching. Its ordinary and extraordinary indices of refraction are given by the Sellmeier-type formulas:

$$n^2 = A - \frac{B}{C - \lambda^2} - D\lambda^2$$

with:

	A	B	C	D
n_e	4.5820	0.099 169	0.044 432	0.021 950
n_o	4.9048	0.117 68	0.047 50	0.027 169

Figure 12.B.4 shows the resulting dispersion curves for this material.

We wish to phase match a lithium niobate crystal to frequency double light from a Nd : YAG laser at 1.3 μm to obtain red light with a wavelength of 0.65 μm. We

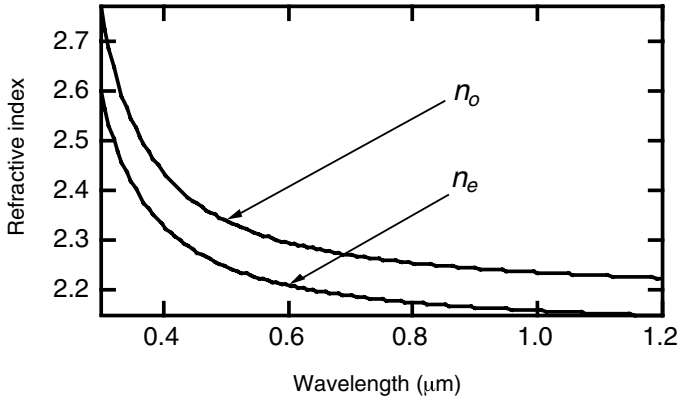


Fig. 12.B.4. Dispersion curves for lithium niobate.

therefore need to solve Eq. (12.B.6). The MATHEMATICA program listed below can be used to solve this type of problem.

```
ae=4.5820;be=0.099169;ce=0.044432;de=0.021950;
ao=4.9048;bo=0.11768;co=0.04750;do=0.027169;
ne2[l__]:= ae-be/(ce-l^2)- de*l^2
ne[l__]:= Sqrt[ne2[l]]
no2[l__]:= ao-bo/(co-l^2)- do*l^2
no[l__]:=Sqrt[no2[l]]
Funct=(1./no2[l]-1./no2[l/2])/(1./ne2[l/2]-1./no2[l/2]);
theta=ArcSin[Funct]*180/Pi;
N[theta]/.l->1.3
ParametricPlot[{theta,l},{l,1.1,3.5}]
```

The phase matching angle is found to be 45°.

Birefringent phase matching of parametric oscillations can be achieved following the same principles. In this case, however, we need to satisfy both the energy and momentum conservation requirements for the photons. Using a pump beam of frequency ω_3 and signal and idler beams with frequencies ω_1 and ω_2, the conditions given in (12.42) may be written:

$$\begin{aligned} \omega_1 + \omega_2 &= \omega_3 \\ n_1\omega_1 + n_2\omega_2 &= n_3\omega_3 \end{aligned} \qquad \text{(12.B.7a)}$$

As n_1/n_3 and n_2/n_3 are less than 1, optical dispersion will again prevent phase matching unless some artifice (such as birefringence) can be used to compensate. In a negative uniaxial material ($n_e < n_o$), it is natural to choose the signal and complementary waves as having an ordinary polarization, and a pump beam with an extraordinary polarization so that the ratios $n_o(\omega_1)/n_e(\omega_3)$ and $n_o(\omega_2)/n_e(\omega_3)$ render the two equations in (12.B.7a) compatible. This is known as the *eoo* configuration and corresponds to:

$$[n_o(\omega_2) - n_e(\omega_3)]\omega_2 = [n_e(\omega_3) - n_o(\omega_1)]\omega_1 \tag{12.B.7b}$$

with the two terms now being positive. The phase matching angle θ_s is given by the equation:

$$n_o(\omega_1)\omega_1 + n_o(\omega_2)\omega_2 = n_e(\omega_3, \theta_s)\omega_3 \tag{12.B.8}$$

where the variation of n_e as a function of θ is given by (12.B.3). Thus, any frequency ω_1 can be obtained through parametric interaction (or 'fragmentation') of the pump beam ω_3 as long as the following expression admits a solution for $\theta_s(\omega_1)$:

$$\omega_1 n_o(\omega_1) + (\omega_3 - \omega_1)n_o(\omega_3 - \omega_1) = \omega_3 \left\{ \left[\frac{\cos\theta_s}{n_o(\omega_3)} \right]^2 + \left[\frac{\sin\theta_s}{n_e(\omega_3)} \right]^2 \right\}^{-1/2} \tag{12.B.9}$$

Figure 12.B.5 shows the variation in wavelength of the signal and idler waves as a function of the phase matching angle θ for a pump beam of 1.06 μm in a lithium niobate crystal $LiNbO_3$. Other phase matching configurations are possible (e.g. *eoe*, *oeo*, . . .) and the interested reader is encouraged to consult more specialized texts for details on this.

Example

The phase matching diagram for a non-linear material represents the ensemble of wavelength pairs (for the signal and complementary beams) that can be generated at a given pump wavelength as a function of a parameter such as the angle θ between the wave and the optical axis of the crystal. The MATHEMATICA program below can be used to solve Eq. (12.B.9) (the dispersion curves are the same as those in the preceding example).

```
ae=4.5820;be=0.099169;ce=0.044432;de=0.021950;
ao=4.9048;bo=0.11768;co=0.04750;do=0.027169;
ne2[l__]:= ae-be/(ce-l^2)-de*l^2;ne[l__]:=Sqrt[ne2[l]];
no2[l__]:= ao-bo/(co-l^2)-do*l^2;no[l__]:=Sqrt[no2[l]];
lp=1.06;l2=1./(1./lp - 1./l1);
f1= ((lp/l1)*no[l1]+(lp/l2)*no[l2])^2;
f2=1/f1 -1/no2[lp];
f3= 1/ne2[lp]-1/no2[lp];
fctopo= Sqrt[f2/f3];theta= ArcSin[fctopo]*N[180/Pi];
ParametricPlot[{theta,l1},{l1,1.5,4.}]
```

With a means of achieving birefringent phase matching, the implementation of an optical parametric oscillator (OPO) as a (wavelength) tuneable source of coherent light is relatively straightforward (see Fig. 12.B.6). A laser cavity provides a powerful pump beam at ω_3. This beam then undergoes parametric interaction with a non-linear crystal situated within a singly or doubly resonant cavity. The mirrors which define the OPO cavity are designed to maintain an elevated reflectivity throughout the desired range of tuning wavelengths for the signal and

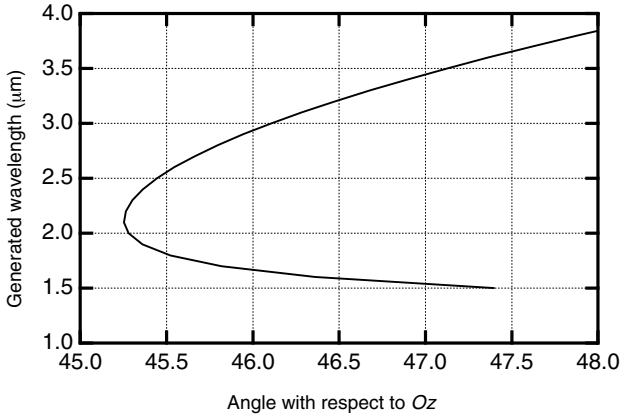


Fig. 12.B.5. Angular phase matching curve for lithium niobate at a pump wavelength of 1.064 μm. (Courtesy of T. Debuisschert@THALES.)

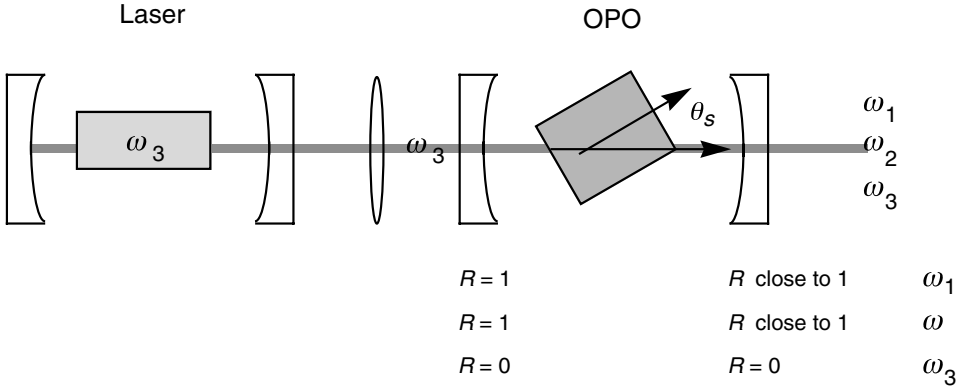


Fig. 12.B.6. Schematic for a wavelength tuneable doubly resonant OPO.

idler beams. Wavelength tuning of the OPO is then achieved by rotating the crystal to obtain the phase matching angle corresponding to the desired signal frequency ω_1 (Eq. (12.B.9)).

12.B.2 Quasi-phase matching

The idea behind this approach is to find a means of spatially modulating the optical non-linearity with some period (Λ) to supply a quasi-wavevector $K = 2\pi/\Lambda$ to satisfy the wavevector conservation requirement (as in diffraction theory). The quasi-phase matching condition consists in providing a quasi-wavevector such that $k_{2\omega} - 2k_{\omega} = K$.

This concept can be easily understood by returning to Eq. (12.29), which describes the evolution of the second harmonic field $E_{2\omega}(z)$ over the course of its propagation through the non-linear medium in the undepleted pump beam

approximation ($E_\omega = \text{constant}$). We add to this equation, however, a spatial variation for $\chi_2(z)$, which we write:

$$\chi_2(z) = \chi_2 f(z) \tag{12.B.10}$$

where $f(z)$ is a periodic function which oscillates between -1 and $+1$. We then obtain:

$$\frac{\mathrm{d}}{\mathrm{d}z} E_{2\omega} = -\mathrm{i}\frac{\omega\chi_2}{n_{2\omega}c} E_\omega^2 f(z)\mathrm{e}^{\mathrm{i}\Delta kz} \tag{12.B.11}$$

The converted field strength $E_{2\omega}$ at the end of a propagation distance L is then:

$$E_{2\omega}(L) = -\mathrm{i}\frac{\omega\chi_2}{n_{2\omega}c} E_\omega^2 \int_0^L f(z)\mathrm{e}^{\mathrm{i}\Delta kz}\mathrm{d}z \tag{12.B.12}$$

Expanding $f(z)$ as a Fourier series with period Λ:

$$f(z) = \sum_n f_n \mathrm{e}^{-\mathrm{i}n(2\pi/\Lambda)z} \tag{12.B.13}$$

Eq. (12.B.12) then gives:

$$E_{2\omega}(L) = -\mathrm{i}\frac{\omega\chi_2}{n_{2\omega}c} E_\omega^2 \sum_n f_n \int_0^L \mathrm{e}^{\mathrm{i}(\Delta kz - 2n\pi/\Lambda)}\mathrm{d}z \tag{12.B.14}$$

where in this last equation, we clearly see the role played by the periodic modulation of susceptibility. If an integer n exists such that:

$$k_{2\omega} - 2k_\omega = \frac{2n\pi}{\Lambda} \tag{12.B.15}$$

only a single term in (12.B.14) is non-zero for large L, leading to:

$$E_{2\omega}(L) = -\mathrm{i}\frac{\omega(f_n\chi_2)L}{n_{2\omega}c} E_\omega^2 \tag{12.B.16}$$

Comparing (12.B.15) with the expression for the maximum conversion length L_c (12.34), we see that the modulation periods are multiples of L_c given by:

$$\Lambda = 2nL_c \tag{12.B.17}$$

Expression (12.B.16) shows that the medium behaves as though it is phase matched, but with an effective non-linear susceptibility given by:

$$\chi_2^{\mathrm{eff}} = |f_n|\chi_2 \tag{12.B.18}$$

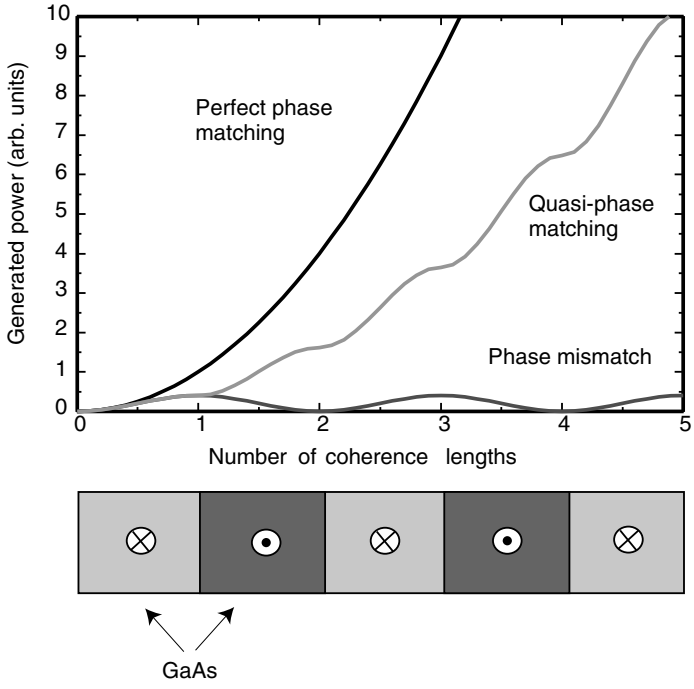


Fig. 12.B.7. Quasi-phase matching using a stack of alternately oriented GaAs layers. Each of the layers has a thickness equal to the phase mismatch length at the desired frequency.

where f_n is the Fourier series term for the periodic function $f(z)$ given by:

$$f_n = \frac{1}{\Lambda}\int_0^{\Lambda} f(z)\mathrm{e}^{\mathrm{i}n(2\pi/\Lambda)z}\mathrm{d}z \tag{12.B.19}$$

If $f(z)$ is a sinusoidal function ($f(z) = \sin 2\pi z/\Lambda$), then $f_n = f_1 = 1/2$ and $\chi_2^{\mathrm{eff}} = \chi_2/2$. A more realistic approach involves modulating the non-linear susceptibility by periodically reversing the direction of the non-linearity so that $f(z) = +1$ between 0 and $\Lambda/2$, and -1 between $\Lambda/2$ and Λ (see Fig. 12.B.7). Integration of (12.B.19) is then straightforward yielding:

$$\chi_2^{\mathrm{eff}} = \frac{2}{\pi}\chi_2 \tag{12.B.20}$$

Figure 12.B.7 shows the mechanism behind quasi-phase matching. The direction of the non-linearity is inverted before the cyclic energy transfer begins from the second harmonic wave to the fundamental wave. Several physical systems allow for this type of quasi-phase matching. In the mid-infrared, it is possible to use successive GaAs layers with alternating crystallographic orientations

(i.e. alternating $\langle 1\,1\,0\rangle$ and $\langle \bar{1}, \bar{1}, 0\rangle$ layers) at every phase mismatch length L_c (every 106 μm to frequency double 10.6 μm light into 5.3 μm light). In the near-infrared (to produce blue light by frequency doubling GaAs laser diode emission at 805 nm), it is possible to use the periodic reversal of domains in ferroelectric materials.

FURTHER READING

Optical frequency conversion techniques:
A. Yariv, *Quantum Electronics*, 3rd Edn, Wiley, New York (1989).
R. L. Sutherland, *Handbook of Nonlinear Optics*, Marcel Dekker Inc., New York (1996).

Birefringence:
M. Born and E. Wolf, *Principles of Optics*, 6th Edn, Pergamon Press, Oxford (1980).
J. M. Perez, Optique, 4eme édition, Masson, Paris (1997).

12.C Pump depletion in parametric interactions

During our investigation of parametric interaction in Chapter 12, the conversion efficiency between the pump beam and the signal and idler beams was assumed to be low enough to allow us to neglect any attenuation in the pump beam intensity (i.e. setting $E_\omega(z) = E_0$ in Section 12.3). In fact, efficiencies in excess of 50% are currently available in certain OPO materials. As energy must be conserved over the course of the parametric interaction, creation of non-linear waves must take place at the expense of the energy contained in the pump beam. An explicit description of the depletion of the pump beam intensity, however, requires a more complete account of the parametric interaction. Most notably, as we progress through this complement, we will uncover the primordial role played by the relative phases of the different waves in parametric interaction. In Chapter 12, this aspect of the problem was rendered inaccessible a priori by our earlier resolve to simplify the problem through application of the constant pump beam approximation.

We begin by treating the case of phase matched ($\Delta k = 0$) second harmonic generation ($\omega + \omega = 2\omega$):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} A_\omega &= -\mathrm{i}\kappa A_\omega^* A_{2\omega} \\ \frac{\mathrm{d}}{\mathrm{d}z} A_{2\omega} &= -\mathrm{i}\kappa A_\omega^2 \end{aligned} \tag{12.C.1a}$$

where we recall that the non-linear coupling parameter κ is given by:

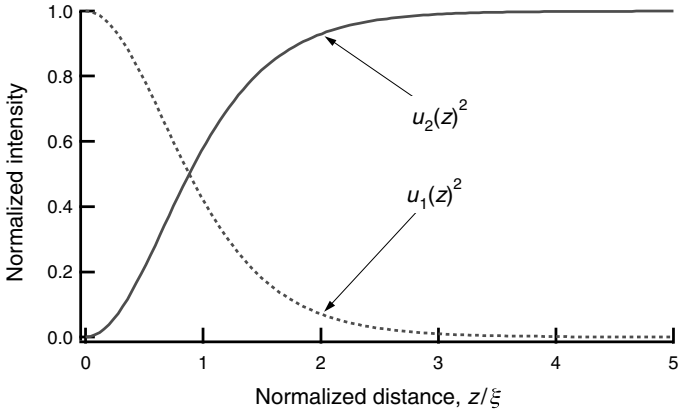


Fig. 12.C.1. Evolution of normalized intensities for the pump and second harmonic waves as a function of propagation distance normalized by the conversion length in the non-linear medium.

$$\kappa = \frac{1}{2}\frac{\chi_2}{c}\left(\frac{2\omega^3}{n_\omega^2 n_{2\omega}}\right)^{1/2} \tag{12.C.1b}$$

In order to take phase effects into account, we introduce the variables:

$$\begin{aligned} A_\omega &= u_1 e^{i\theta_1} \\ A_{2\omega} &= u_2 e^{i\theta_2} \end{aligned} \tag{12.C.2}$$

which, after introduction into (12.C.1) and identification of like terms, leads to the set of coupled differential equations:

$$\begin{aligned} \frac{d}{dz}u_1 &= \kappa u_1 u_2 \sin(\theta_2 - 2\theta_1) \\ \frac{d}{dz}u_2 &= \kappa u_1^2 \sin(2\theta_1 - \theta_2) \\ \frac{d}{dz}\theta_1 &= -\kappa u_2 \cos(\theta_2 - 2\theta_1) \\ \frac{d}{dz}\theta_2 &= -\kappa \frac{u_1^2}{u_2}\cos(2\theta_1 - \theta_2) \end{aligned} \tag{12.C.3}$$

These equations can be considerably simplified by seeking their *invariants* (or *constants of motion*). The first invariant is obtained by combining the first two equations in (12.C.3) leading to:

$$u_1(z)^2 + u_2(z)^2 = u_0^2 = \text{constant} \tag{12.C.4}$$

which is nothing other than a statement of energy conservation. The second

invariant is obtained by dividing the equations in du_i/dz by those in $d\theta_i/dz$ and by introducing the relative phase term $\theta = 2\theta_1 - \theta_2$ giving:

$$\frac{d\theta_1}{du_1} = -\frac{1}{u_1}\cot\theta$$
$$\frac{d\theta_2}{du_2} = -\frac{1}{u_2}\cot\theta \tag{12.C.5}$$

These two equations (12.C.5) can be combined to yield the single equation:

$$u_1^2(z)u_2(z)\cos\theta(z) = u_1^2(0)u_2(0)\cos\theta(0) = \Gamma \tag{12.C.6}$$

where Γ is a constant determined by the boundary conditions. This last equation shows how the phase difference between two waves evolves over the course of their non-linear interaction within the crystal. If we assume a zero starting amplitude for the second harmonic wave at the entrance of the non-linear crystal (i.e. assuming an absence of optical feedback), then $u_2(0) = 0$ and the constant Γ is zero. For $z > 0$, (12.C.6) is satisfied by a non-zero incident wave if and only if $\theta(z) = \text{constant} = (2m + 1)\pi/2$. In this case, u_0 in the energy conservation equation (12.C.4) is the amplitude of the pump wave at the entrance of the non-linear medium, and the equation for the amplitude $u_2(z)$ becomes:

$$\frac{d}{dz}u_2 = \kappa(u_0^2 - u_2^2) \tag{12.C.7}$$

Equation (12.C.7) can be readily integrated to obtain an expression for the evolution of amplitude in the non-linear medium:

$$u_2(z) = u_0\tanh\left(\frac{z}{\xi}\right) \tag{12.C.8a}$$

where ξ is the *frequency conversion distance* given by:

$$\xi = \frac{1}{\kappa u_0} = \frac{c}{\chi_2\omega}\frac{n^{3/2}}{\sqrt{Z_0 I}} \tag{12.C.8b}$$

Frequency conversion distance

where $n \approx n_\omega \approx n_{2\omega}$, I is the incident power, and Z_0 is the vacuum impedance (377 Ω). Figure 12.C.1 shows the variation in normalized intensities $(u_1/u_0)^2$ and $(u_2/u_0)^2$ as a function of normalized distance z/ξ.

We note that after propagating a distance roughly equal to the conversion length, 50% of the pump wave will have been converted into second harmonic radiation.

Example

We consider a crystal with a non-linear susceptibility of 10 pm V^{-1} and an index of

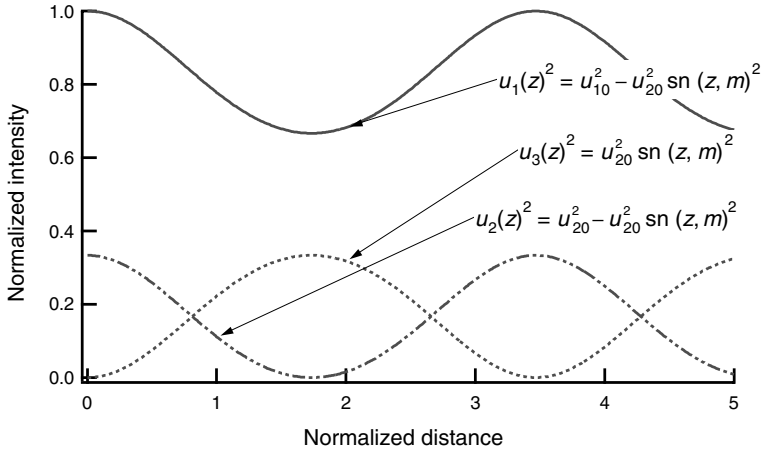


Fig. 12.C.2. Variation of the pump (ω_1 and ω_2) and the sum wave (ω_3) intensities as a function of normalized distance z/ξ_1 for an intensity ratio $m = (u_{20}/u_{10})^2$ of 1/3.

refraction of 1.5. The intensity of the incident wave (with $\lambda = 1\,\mu\mathrm{m}$) is 10 kW over an area of 1 mm^2. Equation (12.C.8b) then leads to:

$$\xi_{\mathrm{m}} = 8.2 \times 10^3 \frac{\lambda_{\mu\mathrm{m}}}{(\chi_2)_{\mathrm{pm\,V^{-1}}}} \frac{n^{3/2}}{\sqrt{I_{\mathrm{W\,m^{-2}}}}} \tag{12.C.8c}$$

or 1.5 cm.

We now interest ourselves with the more general case involving three wave parametric interaction $\omega_3 = \omega_1 + \omega_2$. Starting from (12.38) and generalizing the former change of variables (i.e. $A_k = u_k e^{i\theta_k}$) we obtain the following system of non-linear coupled equations:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}z}u_1 &= \kappa u_2 u_3 \sin\theta \\
\frac{\mathrm{d}}{\mathrm{d}z}u_2 &= \kappa u_1 u_3 \sin\theta \\
\frac{\mathrm{d}}{\mathrm{d}z}u_3 &= -\kappa u_1 u_2 \sin\theta \\
\frac{\mathrm{d}}{\mathrm{d}z}\theta_1 &= -\kappa \frac{u_2 u_3}{u_1}\cos\theta \\
\frac{\mathrm{d}}{\mathrm{d}z}\theta_2 &= -\kappa \frac{u_1 u_3}{u_2}\cos\theta \\
\frac{\mathrm{d}}{\mathrm{d}z}\theta_3 &= -\kappa \frac{u_1 u_2}{u_3}\cos\theta
\end{aligned} \tag{12.C.9}$$

where the phase θ is given by $\theta = \theta_3 - \theta_2 - \theta_1$. Seeking the invariants as before, we readily find:

$$\begin{aligned} u_2(z)^2 + u_3(z)^2 &= u_{23}^2 \\ u_1(z)^2 + u_3(z)^2 &= u_{13}^2 \\ u_1(z)^2 + u_2(z)^2 &= u_{12}^2 \end{aligned} \tag{12.C.10}$$

as a statement of energy conservation and:

$$u_1(z)u_2(z)u_3(z)\cos\theta(z) = \Gamma \tag{12.C.11}$$

for phase conservation.

We are now in a position to discuss an issue which was set aside in Chapter 12. If three waves ω_1, ω_2, and ω_3 are present at the entrance of a non-linear crystal, *how does the system decide which of the following conversions it will perform*: $\omega_1 + \omega_2 \rightarrow \omega_3$, $\omega_3 - \omega_1 \rightarrow \omega_2$, or $\omega_3 - \omega_2 \rightarrow \omega_1$? The answer is provided by the respective phases of the different waves. More particularly, the various mechanisms (sum, difference, and parametric amplification) are now differentiated according to the boundary conditions imposed on the various conservation laws (12.C.10) and (12.C.11).

Let us take as an example *sum frequency generation* (SFG) for which the two boundary conditions are:

$$u_1(0) = u_{10} = u_{13}, u_2(0) = u_{20} = u_{23}, u_3(0) = 0 \text{ and thus } \cos\theta(z) = 0 \tag{12.C.12}$$

The differential equation in u_3 is then:

$$\mathrm{d}Z = \frac{\mathrm{d}U_3}{\sqrt{1 - U_3^2}\sqrt{1 - mU_3^2}} \tag{12.C.13}$$

where Z is the normalized distance $Z = z/\xi_1$, ξ_1 is the conversion length for the wave ω_1, U_3 is the normalized amplitude $U_3 = u_3/u_{20}$, and m is the ratio of the initial amplitudes $m = (u_{20}/u_{10})^2$. The solution to the differential equation in (12.C.13) can be expressed with the help of a (little known!) function from numerical analysis: the *Jacobian elliptic function*, $sn(z)$ (see also Complement 12.F). As a discussion of the mathematical origin and significance of this function is beyond the scope of this text, it suffices for our purposes to note that it is tabulated in and available to most modern mathematical software packages. The integral of (12.C.13) is then given by:

$$z = \xi_1 \int_0^{u_3} \frac{\mathrm{d}x}{\sqrt{1 - x^2}\sqrt{1 - mx^2}} \tag{12.C.14}$$

or:

$$u_3(z) = u_{20}\mathrm{sn}\left(\frac{z}{\xi_1}, \frac{u_{20}^2}{u_{10}^2}\right) \tag{12.C.15}$$

The solutions yielding the intensities of the three waves are then:

$$\begin{aligned} u_1(z)^2 &= u_{10}^2 - u_{20}^2\mathrm{sn}^2\left(\frac{z}{\xi_1}, \frac{u_{20}^2}{u_{10}^2}\right) \\ u_2(z)^2 &= u_{20}^2 - u_{20}^2\mathrm{sn}^2\left(\frac{z}{\xi_1}, \frac{u_{20}^2}{u_{10}^2}\right) \\ u_3(z)^2 &= u_{20}^2\mathrm{sn}^2\left(\frac{z}{\xi_1}, \frac{u_{20}^2}{u_{10}^2}\right) \end{aligned} \tag{12.C.16}$$

Figure 12.C.2 shows the evolution of the normalized intensities for the various waves u_1/u_{10}, u_2/u_{10}, and u_3/u_{10} assuming an intensity ratio m of 1/3. We see that the Jacobian function sn(z) is periodic. The physical origin for this modulation is the periodic exchange of energy between the pump waves and the SFG one. Once the sum wave ω_3 has entirely depleted the pump wave ω_2, the inverse process occurs, and the pump wave ω_2 begins to regain its intensity at the expense of the sum wave.

This formalism describes in a precise manner the energy exchange mechanism between the various waves and the role played by their relative phases. Figure 12.C.2 shows how the energy contained in the pump waves ω_1 and ω_2 combine to form the sum wave ω_3 until complete depletion of the minority wave ω_2, at which time, the inverse process occurs, reconstituting ω_2 waves from ω_1 and ω_3 waves, etc.

FURTHER READING

M. Abramowitz and I. Stegun, *Handbook of Mathematical Tables*, Dover Publications, New York (1970).

J. A. Armstrong, N. Bloembergen, J. Ducuing, and P. S. Pershan, *Phys. Rev.* **127**, 1918 (1962).

Y. S. Shen, *The Physics of Nonlinear Optics*, Wiley, New York (1984).

12.D Spectral and temporal characteristics of optical parametric oscillators

Optical parametric oscillation (OPO) shares a great similarity with laser oscillation. In both cases, optical feedback provided by mirrors induce the oscillations, and set the conditions on the magnitude of the gain relative to the cavity losses and

on the resonant cavity wavelengths. We will concentrate here on two aspects which are unique to OPOs: their gain spectra and their dynamic response.

Parametric gain is imposed by geometrical considerations on the phase matching condition, in contrast to the situation with lasers where the gain is mediated by quantum transitions which occur at discrete atoms or within the band structure of a crystal. It is because of this fact that OPOs are extremely useful as tuneable coherent light sources for spectroscopic investigation of the atmosphere, biology, etc. Equation (12.33) indicates that the conversion efficiency for a parametric process such as $\omega_3 \rightarrow \omega_1 + \omega_2$ depends on the phase mismatch $\Delta k = k_3 - k_1 - k_2$ and is proportional to $\mathrm{sinc}^2(\Delta kL/2)$, where L is the length of the non-linear material. The full width at half maximum for the gain curve (i.e. the *spectral acceptance*) is therefore given by the condition:

$$\Delta kL = \pm\pi \tag{12.D.1}$$

At exact phase matching, photons in a collinear geometry are seen to satisfy the dual requirements for energy and momentum conservation:

$$\begin{aligned} &\frac{1}{\lambda_3} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \\ &\frac{\Delta k}{2\pi} = \frac{n_3(\lambda_3)}{\lambda_3} - \frac{n_1(\lambda_1)}{\lambda_1} - \frac{n_2(\lambda_2)}{\lambda_2} = 0 \end{aligned} \tag{12.D.2}$$

In this last equation, we did not specify the type of phase matching used. It could be of type $e \rightarrow o + o$ in which case $n_3 = n_e(\lambda_3)$, $n_1 = n_o(\lambda_1)$ and $n_2 = n_o(\lambda_2)$, or of any other type. Furthermore, in the case of quasi-phase matching with period Λ, the wavevector $2\pi/\Lambda$ must be added to the phase matching equation. We therefore seek repercussions in terms of the vector Δk of a variation $\Delta\lambda_1$. The pump wavelength λ_3 being fixed, the variations in the signal and idler waves are related via the derivative of the first equation in (12.D.2):

$$\frac{\Delta\lambda_1}{\lambda_1^2} = -\frac{\Delta\lambda_2}{\lambda_2^2} \tag{12.D.3}$$

The effect on phase mismatch is obtained by differentiating the second equation in (12.D.2):

$$\frac{\Delta k}{2\pi} = \frac{n_1}{\lambda_1^2}\Delta\lambda_1 + \frac{n_2}{\lambda_2^2}\Delta\lambda_2 - \frac{\mathrm{d}n_1}{\mathrm{d}\lambda_1}\frac{\Delta\lambda_1}{\lambda_1} - \frac{\mathrm{d}n_2}{\mathrm{d}\lambda_2}\frac{\Delta\lambda_2}{\lambda_2} \tag{12.D.4}$$

Taking (12.D.1) and (12.D.3) into account, this last equation leads to a full width at half maximum for the gain curve of:

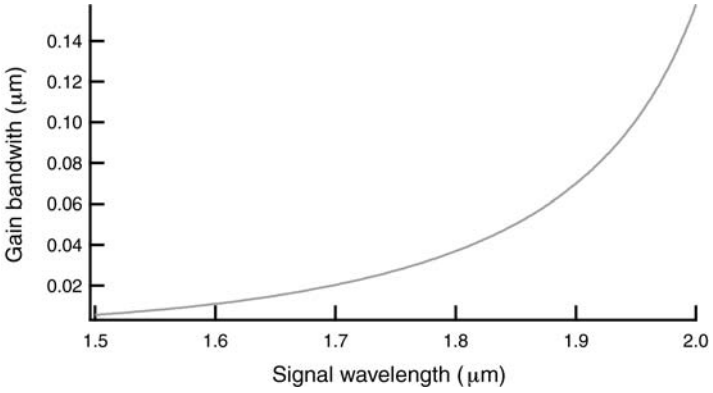


Fig. 12.D.1. Gain bandwidth for $e \rightarrow o + o$ type parametric oscillations in lithium niobate as a function of signal wavelength using a pump wavelength of 1.064 μm.

$$\Delta\lambda_1 = \frac{\lambda_1^2}{L}\left(n_1 - n_2 + \frac{dn_2}{d\lambda_2}\lambda_2 - \frac{dn_1}{d\lambda_1}\lambda_1\right)^{-1} \tag{12.D.5}$$

Spectral acceptance for an OPO

We note that the closer we approach the *degeneracy* condition (i.e. $\omega_1 \approx \omega_2$), the broader the gain curve. This behaviour is illustrated in the following example and in Fig. 12.D.1. In certain situations, it is desirable to obtain as narrow a gain curve as possible (e.g. in the case of single mode OPOs). It then becomes necessary to utilize configurations in which the polarizations for the signal and idler waves are different (e.g. configurations with $e \rightarrow e + o$, referred to as *type II* parametric interactions). If on the other hand, the polarizations of the signal and idler waves are identical, the interaction is of *type I*.

Example

Taking the Sellmeier curves for lithium niobate given in Complement 12.B, the program below calculates the dependence of gain bandwidth on signal wavelength using a pump wavelength of 1.064 μm.

```
ae=4.5820;be=0.099169;ce=0.044432;de=0.021950;
ao=4.9048;bo=0.11768;co=0.04750;do=0.027169;
ne2[l__]:= ae-be/(ce-l^2)-de*l^2;ne[l__]:=Sqrt[ne2[l]];
no2[l__]:= ao-bo/(co-l^2)-do*l^2;no[l__]:=Sqrt[no2[l]];
lp=1.06;l2=1./(1./lp - 1./l1);
L= 10^4;
dn[l__] = ∂_l no[l];
```

$$\Delta\lambda = \text{Abs}\left[\frac{l1^2}{L}(\text{no[l1]} = \text{no[l2]} + \text{dn[l2]}\ l2 - \text{dn[l1]}\ l1)^{-1}\right];$$

```
Plot [Δλ , {l1, 1.5, 2}]
```

Thus, all allowed cavity modes in the OPO which fall within the envelope of the gain spectrum for a non-linear crystal will experience amplification. The response of the OPO will therefore strictly depend upon the type of resonance within the cavity.

The first case involves an OPO in which only the signal wave ω_1 is resonant (i.e. the singly resonant OPO or SROPO case described in Fig. 12.D.2). All the allowed frequencies $\nu_{1m} = mc/2n_1L_{\text{cav}}$ separated by the cavity's free spectral range ($\Delta\nu_{\text{FSR}} = c/2n_1L_{\text{cav}}$) are amplified, corresponding to the number of modes $N = 2(\Delta\lambda_1/\lambda_1)(L_{\text{cav}}/\lambda_1)$ that are susceptible to oscillation. Taking $\Delta\lambda_1 = 0.01\,\mu\text{m}$, $\lambda_1 = 1.6\,\mu\text{m}$, and $L_{\text{cav}} = 4\,\text{cm}$, 312 allowed modes result. Under continuous-wave operation conditions (cw OPO), mode competition (see Complements 13.E and 13.I) will allow only the single mode closest to the peak of the gain curve to oscillate. Frequency tuning can then be achieved by changing the cavity length. The oscillation frequency varies linearly and continuously with the displacement of the cavity modes until an adjacent mode becomes closer to the gain curve maximum. When this happens, the oscillations 'hop' to this adjacent mode (a feature typical of cw SROPOs (see Fig. 12.D.2)). The effective *fine tuning range* is therefore equal to the free spectral range.

In a doubly resonant OPO (DROPO), the modes are now determined by the two conditions:

- $\omega_3 = \omega_1 + \omega_2$
- the eigenmodes ω_{1m} and $\omega_{2m'}$ both overlap within the OPO's gain curve.

The two conditions can be represented on a *Giordmaine–Miller diagram* (see Fig. 12.D.3). The signal and idler frequency axes are oriented in opposite directions and positioned in a manner so that any pair of vertically aligned frequencies ω_{1m} and $\omega_{2m'}$ add together to yield the pump frequency ω_3. Therefore, only those mode pairs ω_{1m} and $\omega_{2m'}$ that overlap to within their respective linewidths ($\Delta\omega_i = \Delta\omega_{\text{FSR}}/\mathscr{F}_i$ where $i = 1, 2$, and $\mathscr{F}_i$ is the finesse of the cavity at frequency ω_i) will be able to oscillate. Given the dispersion in the non-linear medium, the separation between modes ω_{1m} and $\omega_{2m'}$ will change as a function of frequency, and only a few modes will be able to overlap under the gain curve (even under pulsed conditions – see Fig. 12.D.3). DROPOs therefore intrinsically favour single mode operation to a greater extent than do SROPOs. On the other hand, they tend to be less stable. One of the reasons for this instability is the phenomenon of *mode clustering*. If the gain curve is sufficiently broad (often the case!), a pair of non-adjacent cavity modes can become closer than two adjacent modes (see Fig. 12.D.3). The oscillation frequencies will then fluctuate between these two mode clusters, leading to instabilities in the output characteristics of the OPO.

Parametric gain in a non-linear crystal is generally weak and increases linearly with pump power. It is therefore highly advantageous to operate OPOs in pulsed

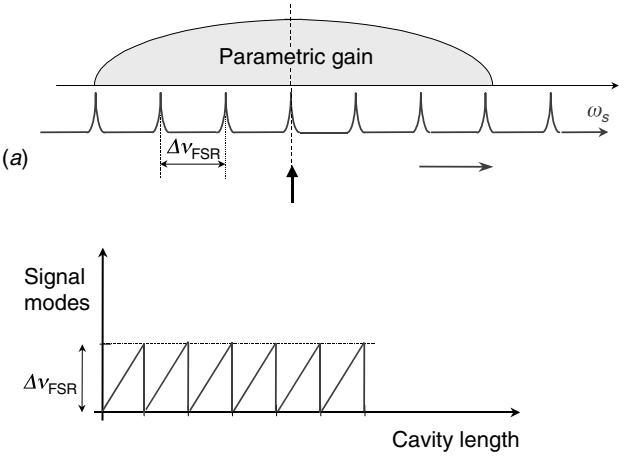


Fig. 12.D.2. Wavelength tuning for a SROPO: (*a*) all modes which lie within the gain curve envelope are susceptible to oscillation. In continuous-wave (cw) operation, only the mode closest to the gain curve maximum oscillates. (*b*) Variation of the cavity length allows frequency tuning over the free spectral range.

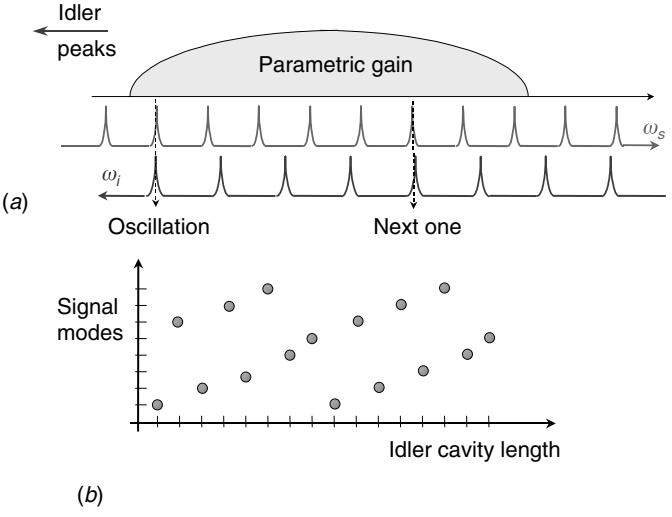


Fig. 12.D.3. Giordmaine–Miller diagram for a DROPO. (*a*) Only those modes which overlap under the parametric gain curve can oscillate. (*b*) In general, the modes tend to cluster together and the oscillation frequencies of the OPO fluctuate between these different values. As a result, DROPOs tend to be unstable.

mode where pulses with high peak intensities (produced say by a Q-switched laser) are used to pump a non-linear crystal. We are thus going to describe the transient behaviour of an OPO pulsed in the nanosecond regime, i.e. the pump pulse duration corresponds to few photon cavity lifetimes.

For that purpose, we are going to describe the evolution of different wave amplitudes during a photon round-trip in the cavity (between t_n and t_{n+1}). Moreover, we make the following assumptions:

- The relative phases of the waves are such that the constant of motion $\Gamma = 0$, i.e $\cos\theta = 0$ throughout the structures. This means that all the different phases θ_i stay constant in a non-linear crystal (see (12.C.9)).
- The parametric gain is supposed to be sufficiently small, so that (12.C.9) may be linearized.

With these two assumptions, (12.C.19) may then be written as:

$$\begin{aligned} u_1^n(L) &= u_1^n\left(\frac{L}{2}\right) + \frac{\kappa L}{2} u_2^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right) \\ u_2^n(L) &= u_2^n\left(\frac{L}{2}\right) + \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right) \\ u_3^n(L) &= u_3^n\left(\frac{L}{2}\right) - \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_2^n\left(\frac{L}{2}\right) \end{aligned} \tag{12.D.6}$$

In an OPO, the ω_i waves are fed back into the cavity input by mirrors of overall reflectivities r_i (supposed real for the time being) so that the new inputs at time t_{n+1} are:

$$\begin{aligned} u_1^{n+1}(0) &= r_1 u_1^n(L) = r_1\left[u_1^n\left(\frac{L}{2}\right) + \frac{\kappa L}{2} u_2^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right)\right] \\ u_2^{n+1}(0) &= r_2\left[u_2^n\left(\frac{L}{2}\right) + \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right)\right] \\ u_3^{n+1}(0) &= f(t_n) - r_3\left[u_3^n\left(\frac{L}{2}\right) + \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_2^n\left(\frac{L}{2}\right)\right] \end{aligned} \tag{12.D.7}$$

where $f(t)$ is the pump electric field at the OPO input. At time t_{n+1}, the wave amplitudes in the middle of the non-linear crystal are now:

$$u_1^{n+1}\left(\frac{L}{2}\right) = u_1^{n+1}(0) + \frac{\kappa L}{2} u_2^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right) = r_1 u_1^n\left(\frac{L}{2}\right) + (1+r_1)\frac{\kappa L}{2} u_2^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right)$$

$$\begin{aligned} u_2^{n+1}\left(\frac{L}{2}\right) &= u_2^{n+1}(0) + \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right) \\ &= r_2 u_2^n\left(\frac{L}{2}\right) + (1+r_2)\frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_3^n\left(\frac{L}{2}\right) \end{aligned} \tag{12.D.8}$$

$$\begin{aligned} u_3^{n+1}\left(\frac{L}{2}\right) &= u_3^{n+1}(0) - \frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_2^n\left(\frac{L}{2}\right) \\ &= f(t_n) + r_3 u_3^n\left(\frac{L}{2}\right) - (1+r_2)\frac{\kappa L}{2} u_1^n\left(\frac{L}{2}\right) u_2^n\left(\frac{L}{2}\right) \end{aligned}$$

We now introduce the new notation $a_i(t_n) = u_i^n\left(\frac{L}{2}\right)$. During a round-trip in the cavity, the different wave amplitude increments are then:

$$\frac{\mathrm{d}}{\mathrm{d}t} a_i(t_n) T_{\mathrm{RT}} = u_i^{n+1}\left(\frac{L}{2}\right) - u_i^n\left(\frac{L}{2}\right) \tag{12.D.9}$$

where T_{RT} is the round-trip time in the cavity, i.e. $T_{\mathrm{RT}} = 2L/c$. Equation (12.D.6) may now be written:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} a_1(t) &= -\frac{a_1}{\tau_1} + \gamma_1 a_2 a_3 \\ \frac{\mathrm{d}}{\mathrm{d}t} a_2(t) &= -\frac{a_2}{\tau_2} + \gamma_2 a_1 a_3 \\ \frac{\mathrm{d}}{\mathrm{d}t} a_3(t) &= f(t) - \frac{a_3}{\tau_3} - \gamma_3 a_1 a_2 \end{aligned} \tag{12.D.10}$$

OPO dynamic equations

These equations describe the temporal behaviour of an OPO. The different quantities appearing in (12.D.10) are given by:

$$\begin{aligned} \frac{1}{\tau_i} &= \frac{1-r_i}{T_{\mathrm{RT}}} \\ \gamma_i &= (1+r_i)\frac{\kappa c}{4} \end{aligned} \tag{12.D.11}$$

The quantities a_i are the real wave amplitudes at the middle of the non-linear crystal, related to the photon fluxes through (12.37), i.e. $\Phi_i = a_i^2/2\hbar Z_0$. These quantities may, however, be negative (corresponding to $\theta = -\pi/2$ in (12.C.9)), as

they describe the oscillatory transfer of energy between the various waves as they interact within the non-linear crystal. Here, $f(t)$ is the pump source term of the OPO (in amplitude), γ_i are the different non-linear coupling terms, and τ_i are the lifetimes of the different frequency photons in the cavity.

In order to describe the singly resonant OPO (SROPO) specifically, we must make the assumption that there is no feedback on the pump and idler waves so that $r_2 = r_3 = 0$. In (12.D.10), the idler cavity lifetime τ_2 may then be considered as negligible compared with the signal lifetime ($\tau_1 = T_{RT}/(1 - r_1)$) and the pump duration (typically a few nanoseconds). One may thus consider that the idler wave adiabatically follows the pump and signal variations. This is similar to considering that $\mathrm{d}a_2/\mathrm{d}t = 0$, so that:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}a_1(t) = -\frac{a_1}{\tau_1} + \gamma_1 a_2 a_3 \\ &a_2 = g_2 a_1 a_3 \\ &\frac{\mathrm{d}}{\mathrm{d}t}a_3(t) = f(t) - \frac{a_3}{\tau_3} - \gamma_3 a_1 a_2 \end{aligned} \qquad \text{(12.D.12)}$$

SROPO dynamic equations

with $g_2 = \kappa L/2$.

It is now interesting to solve (12.D.10) in the general case (SROPO, DROPO) under stationary state conditions, i.e. $f(t) = f =$ constant and $\mathrm{d}a_i/\mathrm{d}t = 0$. Equation (12.D.10) leads to:

$$\begin{aligned} &a_1(1 - \gamma_1\tau_1\gamma_2\tau_2 a_3^2) = 0 \\ &f = \frac{a_3}{\tau_3} + \gamma_3 a_1 a_2 \end{aligned} \qquad \text{(12.D.13)}$$

As was the case with the laser equations in (4.34) and (4.35), this equation admits two types of solutions:

- *Below threshold*. The signal and idler photon fluxes are null ($a_1 = a_2 = 0$) and the pump photon amplitude increases linearly within the cavity as a function of the pump amplitude ($a_3 = f\tau_3$).
- *Above threshold*. The pump photon flux is *clamped* to its threshold value given by $A_{3,\text{threshold}} = a_{3,\text{threshold}}^2 = 1/(\gamma_1\tau_1\gamma_2\tau_2)$. The signal photon flux is given by the second equation (12.D.13):

$$A_1 = a_1^2 = \frac{1}{\tau_3\gamma_3\tau_2\gamma_2}\left(\frac{f}{f_{\text{threshold}}} - 1\right) \qquad \text{(12.D.14)}$$

where the threshold is given by $f_{\text{threshold}} = a_{3,\text{threshold}}/\tau_3$. This means that the

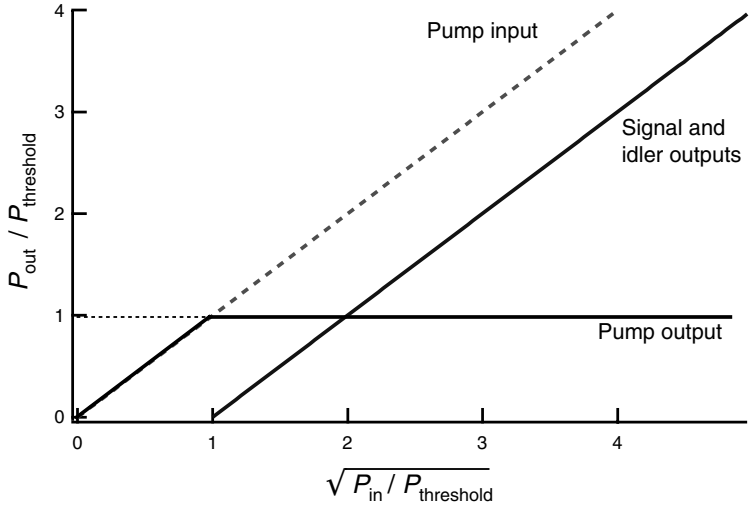


Fig. 12.D.4. OPO response as a function of the square root of pump power.

DROPO pump power threshold is proportional to $1/T_1T_2$, with T_i mirror transmission, which is effectively what we found in (12.62).

Equation (12.D.14) also suggests that the output signal power P_1 is proportional to $\sqrt{P_3/P_{3,\text{threshold}}} - 1$, as indicated in Fig. 12.D.4. This is markedly different from the linear $P_{in} - P_{out}$ behaviour of lasers. This point is generally missed by many authors and will be discussed in more detail in Complement 12.F.

The following example now illustrates the temporal evolution of a DROPO as described by (12.D.10). Figure 12.D.5 shows the temporal variation for the pump and signal waves at the output of an OPO cavity starting from these equations. The signal grows exponentially starting from a parametric fluorescence noise, while the signal and idler waves rapidly deplete the power in the pump beam.

Example

We study the temporal response of a DROPO pumped using Gaussian (shaped) pulses. The equations are normalized and solved using MATHEMATICA. We see that the signal cannot grow without being seeded by a minute quantity of *source* photons provided by parametric fluorescence.

```
g = 5; τ = 2; eq1 = ap'[t] == E^-(t/τ)^2 - ap[t] - g as[t] ac[t] ;
eq2 = as' [t] == - as[t] + g ac[t] ap[t] ;
eq3 = ac' [t] == -ac[t] + g as[t] ap[t];
sol = NDSolve[{eq1, eq2, eq3, ap[-5] == 0, as[-5] == 0.001, ac[-5] == 0.001},
   {ap[t], as[t], ac[t]}, {t, -5, 5}];
```

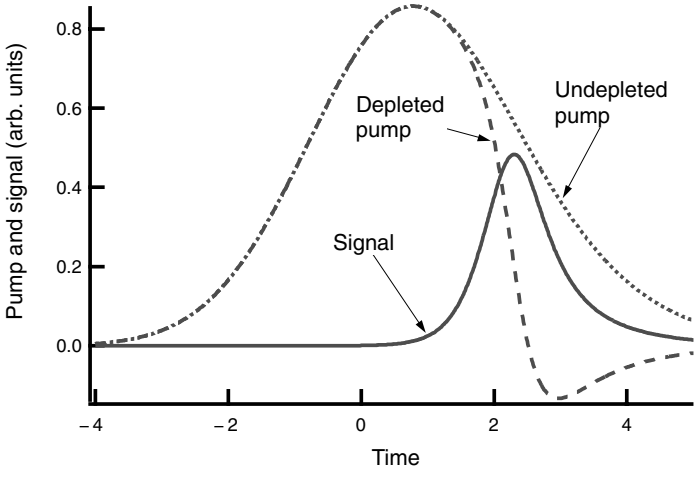


Fig. 12.D.5. Temporal profile of pump and signal pulsed amplitudes in an OPO obtained from the coupled mode equations (12.D.10).

```
P1 = Plot[Evaluate[ap[t] / . sol, {t, -5, 5}],
   PlotStyle→ RGBcolor[1, 0, 0], DisplayFunction→ Identity];
P3 = Plot[Evaluate[as[t] / . sol, {t, -5, 5}], PlotStyle→ RGBColor[0, 0, 1],
   DisplayFunction→ Identity];
```

FURTHER READING

J. A. Giordmaine and R. C. Miller, *Phys. Rev. Lett.* **14**, 973 (1965).
Y. S. Shen, *The Physics of Nonlinear Optics*, Wiley, New York (1984).

12.E Parametric interactions in laser cavities

As demonstrated by (12.33) or (12.68), the parametric conversion efficiency is proportional to pump beam intensity. It would therefore be advantageous if the parametric interaction could occur within the optical cavity of the pump laser itself as the pump beam intensity is $1/T$ (where T is mirror transmission) times greater than the pump output intensity (see Complement 9.D). We will therefore study the intracavity non-linear frequency conversion efficiency. In order to simplify our approach, and to emphasize the more important aspects of the problem, we will assume that:

1. Phase matching considerations have been taken into account by suitable design of the cavity, allowing us to focus only on the number of photons in the cavity.

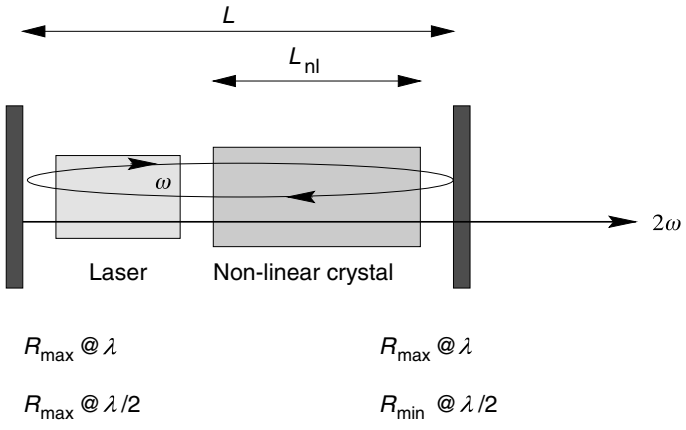


Fig. 12.E.1. Configuration for intracavity frequency doubling.

2. The wave amplitudes in the cavity are position independent: this approximation becomes reasonable as soon as the cavity finesses at ω and 2ω are greater than a few units.
3. The entire medium is characterized by a single optical index allowing us to neglect varying effects of refraction.

We will now develop the formalism for intracavity frequency doubling. The required set-up and relevant notations are illustrated in Fig. 12.E.1. We will assume that the parametric gain is weak enough to justify the undepleted pump beam approximation (see Section 12.3). During a time interval dt, the energies $\mathscr{P}_{\omega}\mathrm{d}t$ and $\mathscr{P}_{2\omega}\mathrm{d}t$ contained by the ω and 2ω waves leaving the non-linear medium are related in terms of photon number P_2 at 2ω and P_1 at ω in the cavity by:

$$\begin{aligned}\mathscr{P}_{\omega}\mathrm{d}t &= P_1\frac{c'\mathrm{d}t}{SL}\hbar\omega \\ \mathscr{P}_{2\omega}\mathrm{d}t &= P_2\frac{c'\mathrm{d}t}{SL}2\hbar\omega\end{aligned} \tag{12.E.1}$$

where c' is the speed of light in the medium, S is the normal cross-sectional area of the beam, and L is the total cavity length. The variation rates for the numbers of photons due to the non-linear interaction are given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P_2\bigg|_{\mathrm{NL}} &= \frac{1}{2}K_{\mathrm{nl}}P_1^2 \\ \frac{\mathrm{d}}{\mathrm{d}t}P_1\bigg|_{\mathrm{NL}} &= -K_{\mathrm{nl}}P_1^2\end{aligned} \tag{12.E.2}$$

where we remember that, two ω photons are required to produce a single 2ω

photon. To find an expression for K_{nl}, we recall that on the one hand, the power circulating in the cavity and the rate of change in the number of photons are related by:

$$\mathscr{P}_{2\omega} = \frac{1}{S}\frac{\mathrm{d}}{\mathrm{d}t}P_2 2\hbar\omega = \frac{K_{\mathrm{nl}}}{S}P_1^2\hbar\omega \tag{12.E.3}$$

while on the other hand, the intensities $\mathscr{P}_{\omega}$ and $\mathscr{P}_{2\omega}$ are given by (12.33):

$$\mathscr{P}_{2\omega} = \eta\mathscr{P}_{\omega}^2 \tag{12.E.4a}$$

where η is the conversion efficiency (in $\mathrm{m}^2\,\mathrm{W}^{-1}$):

$$\eta = 2\frac{Z_0^3}{n_{\omega}^2 n_{2\omega}}(\omega\varepsilon_0\chi_2 L_{\mathrm{nl}})^2 \tag{12.E.4b}$$

Neglecting pump beam depletion in the crystal, the coefficient K_{nl} follows from (12.E.3) and (12.E.4):

$$K_{\mathrm{nl}} = \eta\frac{c^2}{SL^2}\hbar\omega \tag{12.E.5}$$

We now turn our attention to the response of the intracavity laser medium.

The temporal evolution of the medium without the effect of frequency doubling is given by (4.34) and (4.35) while also taking into account the fact that the laser medium occupies only a fraction of the cavity.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}N &= \Gamma_2(N_0 - N) - K_l N P_1\\ \frac{\mathrm{d}}{\mathrm{d}t}P_1 &= K_l N P_1 - \Gamma_{\mathrm{cav}1}P_1\end{aligned} \tag{12.E.6a}$$

where N is the number of inverted atoms and K_l is the linear coupling coefficient:

$$K_l = \frac{c'\sigma_{\mathrm{op}}}{SL} \tag{12.E.6b}$$

In the absence of any parametric interaction, we immediately recover the stationary state behaviour. If the number of inverted atoms N is smaller than the threshold value $N_{\mathrm{threshold}}$, the number of photons P_1 will be zero and $N = N_0$ (proportional to the pump power). In the opposite situation, N will clamp to $N_{\mathrm{threshold}}$ as:

$$N_{\mathrm{threshold}} = \frac{\Gamma_{\mathrm{cav},1}}{K_l} = \frac{SL}{c'\sigma_{\mathrm{op}}\tau_{\mathrm{cav}1}} \tag{12.E.7a}$$

with the number of photons P_1 being given by:

$$P_1 = P_{\text{sat}}(r - 1) \tag{12.E.7b}$$

where $P_{\text{sat}} = \Gamma_2/K_l = SL/c\sigma_{\text{op}}\tau_2$ is the saturation photon number and r is the normalized pump rate $r = N_0/N_{\text{threshold}}$ (see Section 4.6). We now need only bring together the two mechanisms (amplification and frequency conversion) to describe the cavity response:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}N &= \Gamma_2(N_0 - N) - K_l N P_1 \\
\frac{\mathrm{d}}{\mathrm{d}t}P_1 &= K_l N P_1 - \Gamma_{\text{cav1}} P_1 - K_{\text{nl}} P_1^2 \\
\frac{\mathrm{d}}{\mathrm{d}t}P_2 &= \frac{1}{2}K_{\text{nl}} P_1^2 - \Gamma_{\text{cav2}} P_2
\end{aligned} \tag{12.E.8}$$

At stationary state, the second equation provides a means of relating the number of inverted atoms to the number of photons P_1:

$$N_\infty = N_{\text{threshold}} + \beta P_1 \tag{12.E.9a}$$

with

$$\beta = \frac{K_{\text{nl}}}{K_l} \tag{12.E.9b}$$

The inversion density is therefore no longer clamped! This results from the fact that the photons ω do not saturate the laser transition as they are eliminated within the cavity by turning into 2ω photons. Substituting (12.E.9) into the first equation in (12.E.8), the number of photons P_1 is given by the solution to the second-degree equation:

$$u[(r - 1) - \beta Y] = (1 + \beta Y)Y \tag{12.E.10}$$

where u is the ratio $u = \Gamma_2/\Gamma_{\text{cav1}}$ and $Y = P_1/N_{\text{threshold}}$ (see (4.36)). The number of photons leaving the cavity $\Gamma_{\text{cav2}}P_2$ is then given by the third equation in (12.E.8), i.e:

$$\Gamma_{\text{cav2}} P_2 = \frac{K_{\text{nl}}}{2} N_{\text{threshold}}^2 Y^2 \tag{12.E.11}$$

or:

$$\Gamma_{\text{cav2}} P_2 = \frac{\Gamma_2 \Gamma_{\text{cav1}}}{K_l} \frac{[\sqrt{(1 + u\beta)^2 + 4u\beta(r - 1)} - (1 + u\beta)]^2}{8u\beta} \tag{12.E.12}$$

This last quantity essentially depends on the parameter $u\beta = K_{\text{nl}}/(\Gamma_{\text{cav1}}K_l/\Gamma_2)$,

which is nothing else than the non-linear parameter normalized by the linear cavity parameters. An analysis of (12.E.12) shows that the non-linearity K_{nl} can be selected to yield an optimal efficiency, and this at any given pump rate r: it suffices that $u\beta = 1$, i.e. for a non-linear coupling coefficient given by:

$$K_{opt}P_{sat} = \Gamma_{cav1} \tag{12.E.13}$$

This last equation indicates that an optimal level of non-linear coupling occurs when the non-linear losses at saturation ($K_{opt}P_{sat}$) match the linear cavity losses. We see that there is an interest in minimizing the linear losses to keep the non-linear crystal length to a minimum. The optimal 2ω photon flux which can leave the cavity is:

$$\Gamma_{cav2}P_2 = \frac{1}{2}\Gamma_{cav1}P_{sat}(\sqrt{(r)} - 1)^2 \tag{12.E.14}$$

We note that this flux is proportional to $(\sqrt{(r)} - 1)^2$: beyond threshold, the bulk of the pump energy is converted into 2ω photons, with the 2ω photon flux becoming proportional to r, whereas the flux at the fundamental frequency is proportional to $\sqrt{(r)}$. The cavity therefore acts (almost) as a perfect frequency converter. Figure 12.E.2 shows the variations in the normalized inversion population $X = N/N_{threshold}$ and 2ω photons $Z = P_2/N_{threshold}$ as a function of the pump rate r. Declamping of the inversion population X and the $(\sqrt{(r)} - 1)^2$ dependence of the 2ω photon number Z are clearly apparent.

The time dependence of the signal is obtained with the help of (12.E.8), written (neglecting the spontaneous emission term $\Gamma_2(N_0 - N)$ during the optical pulse) as:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}X &\approx -XY \\ \frac{\mathrm{d}}{\mathrm{d}t}Y &= (X-1)Y - \beta Y^2 \\ \frac{\mathrm{d}}{\mathrm{d}t}Z &\approx \frac{\beta}{2}Y^2 \end{aligned} \tag{12.E.15}$$

Figure 12.E.3 shows the time dependence for the 2ω photon flux obtained by solving the differential equations in (12.E.15). As before (see Section 4.6.2), it was necessary to seed the medium with a few initial photons to simulate the effect of spontaneous emission (*parametric fluorescence*). Figure 12.E.3 also shows that for strong coupling coefficients ($\beta > 1$) the pulse lengths become significant but the overall integral (the pulse energy) is unchanged. This is good news: it means that small non-linearities or thin non-linear crystals are enough to provide good SHG

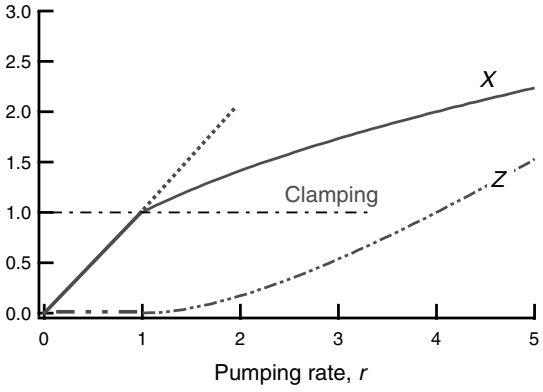


Fig. 12.E.2. Variation in the normalized number of gain medium atoms in the excited state X and the normalized 2ω photon flux Z as a function of pump rate r.

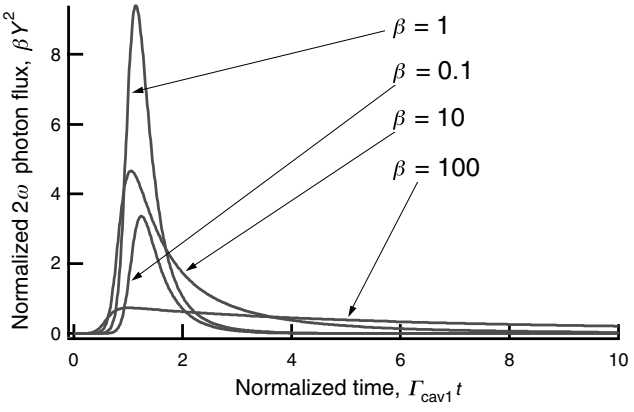


Fig. 12.E.3. Time dependence of second harmonic photon output for various β ratios between the non-linear and linear coefficients.

efficiency. In fact, intracavity SHG of 1.06 μm YAG laser has become the work-horse of today's high-power visible lasers, totally replacing ion gas lasers.

Example

The MATHEMATICA program below calculates the pulsed behaviour described by (12.E.15).

```
xi=10 ;β=.1 ;
eq1= y'[t]==y[t]*(x[t]-1)-β y[t]^2;
eq2= x'[t]== -x[t]*y[t] ;
sol=NDSolve[{eq1,eq2,x[0]==xi,y[0]==0.001},{x[t],y[t]},{t,0.,50}];
plot2=Plot[ Evaluate[β y[t]^2 /.sol,{t,0,50}],PlotStyle->
{RGBColor[.5,0,.5]},PlotRange->{0,1},DisplayFunction->Identity];
```

FURTHER READING

J. E. Murray, *J. Appl. Phys.* **41**, 609 (1970).
R. G. Smith, *IEEE J. Quantum Electron.* **6**, 215 (1970).

12.F Continuous wave optical parametric oscillator characteristics

Up to now, OPO characteristics have been analysed assuming a small frequency conversion efficiency. This allowed us to assume an undepleted pump amplitude in the crystal and to simplify the theoretical derivations in a very convenient way. In fact, as we will now see, this approximation is of very poor validity (conversion efficiencies higher than 93% have been reported!) and we are going to derive the OPO characteristics without this approximation.

For this purpose, we recall the propagation equations for the phases and amplitudes of the different fields (see (12.C.9)):

$$\frac{\mathrm{d}}{\mathrm{d}z}u_1(z) = \kappa u_2(z)u_3(z)\sin\phi(z) \tag{12.F.1a}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}u_2(z) = \kappa u_1(z)u_3(z)\sin\phi(z) \tag{12.F.1b}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}u_3(z) = -\kappa u_1(z)u_2(z)\sin\phi(z) \tag{12.F.1c}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\phi_1(z) = \kappa\frac{u_2(z)u_3(z)}{u_1(z)}\cos\phi(z) \tag{12.F.1d}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\phi_2(z) = \kappa\frac{u_1(z)u_3(z)}{u_2(z)}\cos\phi(z) \tag{12.F.1e}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\phi_3(z) = \kappa\frac{u_1(z)u_2(z)}{u_3(z)}\cos\phi(z) \tag{12.F.1f}$$

where $\phi = \phi_3 - \phi_1 - \phi_2$ is the non-linear relative phase shift and κ the non-linear coupling coefficient given by $\kappa = (\chi^{(2)}/2c)(\omega_1\omega_2\omega_3/n_1n_2n_3)^{1/2}$. From (12.F.1), one can derive the existence of four quantities that are constant in propagation (three of which are independent):

$$\Gamma = u_1(z)u_2(z)u_3(z)\cos\phi(z) \tag{12.F.2a}$$

$$m_1 = u_2(z)^2 + u_3(z)^2 \tag{12.F.2b}$$

$$m_2 = u_1(z)^2 + u_3(z)^2 \tag{12.F.2c}$$

$$m_3 = u_1(z)^2 - u_2(z)^2 \tag{12.F.2d}$$

These constants of motion are determined by the boundary conditions. Moreover the conservation of energy is easily written in the following way:

$$\alpha_1 u_1(z)^2 + \alpha_2 u_2(z)^2 + u_3(z)^2 = \frac{2Z_0 P_c}{\omega_3} = p_c \tag{12.F.3}$$

where $\alpha_i = \omega_i/\omega_3$ are the *OPO quantum defects* and P_c is the total optical power flowing inside the cavity.

The cavity induces a feedback effect for the different fields that are recirculated, implying the following self-consistency equations in the steady state regime for the field amplitudes E_i between the crystal entrance $z = 0$ and the crystal exit $z = L$:

$$E_i(0) = r_i e^{i[\theta_i + (\omega_i/c)L_i']} E_i(L) \tag{12.F.4}$$

where $r_i = \sqrt{R_i}$ and θ_i are the reflectivity coefficient and dephasing of the mirror for signal and idler waves, respectively; and L_i' the cavity length outside the crystal for mode *i*. On the other hand, (12.F.2) and (12.F.3) guarantee that the input and output powers are equal and are given by:

$$p_{\text{in}} = \frac{2Z_0 P_{\text{in}}}{\omega_3} = u_{30}^2 = u_{3L}^2 + (1 - R_1)\alpha_1 u_{1L}^2 + (1 - R_2)\alpha_2 u_{2L}^2 = p_{\text{out}} \tag{12.F.5}$$

where $u_{i0} = u_i(0)$ and $u_{iL} = u_i(L)$.

When the OPO is operating above threshold in the steady state regime, the oscillation frequencies of the signal and idler waves are determined by the energy conservation relation $\omega_1 = \omega_2 + \omega_3$ and by the round-trip equations, (12.F.4), for the different field amplitudes. As a result, the frequency tuning characteristics of the OPO strongly depend on the values and the resonant nature of the different fields in the cavity, as we will see below. We will now investigate the solutions of these coupled equations in the SROPO and DROPO configurations, respectively.

12.F.1 Singly resonant OPO

For the SROPO configuration, we thus suppose that the reflectivity of the idler is zero, i.e. $r_2 = 0$. The signal self-consistency equations read:

$$u_1(0) = r_1 u_1(L) \tag{12.F.6a}$$

$$u_2(0) = 0 \tag{12.F.6b}$$

$$\phi_1(0) - \phi_1(L) = \theta_1 + k_1 L + \frac{\omega_1 L'}{c} \tag{12.F.6c}$$

keeping in mind that the total phase of the electric field inside the non-linear

crystal is $\phi_1(z) + k_1 z$. Condition (12.F.6) implies that the constant of motion Γ is $\Gamma = 0$ in (12.F.2a), and therefore, since $u_1(L)u_2(L)u_3(L) \neq 0$ above threshold, that $\cos\phi(z) = 0$. *Threshold is thus obtained once phase* $\phi(z)$ *is identically* $(\pi/2) + p\pi$, where p is an integer. The sign of $\sin\phi$ (i.e. the parity of integer p) is of great importance in (12.F.1a–c): it tells whether the energy flows from the pump to the signal or *vice versa*.

One also deduces from (12.F.1d–f) that the three phases $\phi_i(z)$ $(i = 1, 2, 3)$ are separately constant and, in particular, that $\phi_1(L) = \phi_1(0)$. Equation (12.F.6c) then has a simple consequence:

$$k_1 L + \frac{\omega_1}{c} L' + \theta_1 = 2m\pi, \text{ where } m \text{ is an integer} \tag{12.F.7a}$$

One thus retrieves in the present general case the simple and intuitive condition that the SROPO oscillates only when *the signal field is resonant in the cavity*, which imposes a comb of possible signal frequencies ω_1, or longitudinal modes, like in a laser, given by the implicit equation:

$$\omega_1 = m\frac{2\pi c}{L' + n(\omega_1)L} \tag{12.F.7b}$$

Now, using (12.F.6a) and (12.F.6b), the flux conservation equations, (12.F.2), read:

$$m_1 = u_{30}^2 = p_{\text{in}} \tag{12.F.8a}$$

$$m_3 = u_{10}^2 = R_1 u_{1L}^2 = u_{1L}^2 - u_{2L}^2 \tag{12.F.8b}$$

$$u_{1L}^2 + u_{3L}^2 = R_1 u_{1L}^2 + p_{\text{in}} \tag{12.F.8c}$$

where both pump and signal amplitudes at the crystal output (respectively u_{3L} and u_{1L}) are to be determined as a function of the input power u_{30}^2. Assuming, for the time being, that the energy flows from the pump to the signal, (12.F.1b), describing the evolution of the idler wave, is now:

$$\frac{\mathrm{d}}{\mathrm{d}z} u_2 = \kappa\sqrt{(m_1 - u_2^2)(m_3 + u_2^2)} \tag{12.F.9}$$

which may be formally integrated as:

$$\kappa L = \int_0^{u_2(L)} \frac{\mathrm{d}u_2}{\sqrt{(m_1 - u_2^2)(m_3 + u_2^2)}} \tag{12.F.10}$$

This latter expression has a well known solution in terms of *Jacobian functions*:

$$u_2(L) = \mathrm{i}\sqrt{m_3}\,\mathrm{sn}\left(\mathrm{i}\sqrt{m_1}\,\kappa L \,\middle|\, -\frac{m_3}{m_1}\right) \tag{12.F.11}$$

In order to define the notation (which differs depending on authors), we have taken the following definition for the inverse of the Jacobian function sn:

$$a\int_0^x \frac{\mathrm{d}t}{\sqrt{(a^2-t^2)(b^2-t^2)}} = \mathrm{sn}^{-1}\left(\frac{x}{b}\,\middle|\,\frac{b^2}{a^2}\right) \tag{12.F.12}$$

Inverse Jacobian function

Now using the flux conservation condition for the pump over the crystal length (Eq. (12.F.8b)), (12.F.11) reads:

$$\mathrm{sn}^2\left(\mathrm{i}\kappa L\sqrt{p_{\mathrm{in}}}\,\middle|\, -R_1\frac{u_{1L}^2}{p_{\mathrm{in}}}\right) = 1 - \frac{1}{R_1} \tag{12.F.13}$$

Equation (12.F.13) is an implicit equation that yields the value of the signal amplitude u_{1L} at the crystal output as a function of input power p_{in}.

Let us first find the SROPO *oscillation threshold*: in this case the signal amplitude $u_{iL} \to 0$ with a non-zero pump amplitude $u_{30} \neq 0$. Since $\mathrm{sn}(\mathrm{i}x|m) \approx \mathrm{i}\sinh(x)$ when $m \to 0$, one derives from (12.F.13) the value of the normalized pump threshold p_s:

$$p_s = \left(\frac{\cosh^{-1}(1/\sqrt{R_1})}{\kappa L}\right)^2 \tag{12.F.14}$$

which is effectively the value obtained in the linearized theory for SROPO (Eq. (12.49)); a theory which is obviously valid when the signal has a vanishing value in the crystal. Introducing the normalized variables $X = p_{\mathrm{in}}/p_s$ and $Y = u_{1L}^2/p_s$, the implicit equation which yields the normalized signal power Y at the crystal output as a function of normalized input power X is now:

$$\mathrm{sn}^2\left(\mathrm{i}\cosh^{-1}\left(\frac{1}{\sqrt{R_1}}\right)\sqrt{X}\,\middle|\, -R_1\frac{Y}{X}\right) = 1 - \frac{1}{R_1} \tag{12.F.15}$$

The normalized output signal power Y' is the product of Y by the mirror transmission, i.e:

$$Y' = (1 - R_1)Y \tag{12.F.16}$$

Equation (12.F.15) holds for any value of pumping power and its solution depends on a single parameter: the signal reflectivity R_1. It is easily solved using a symbolic computation program. As an example, Fig. 12.F.1 shows the Y' versus X curves (i.e. normalized $P_{\mathrm{out}} - P_{\mathrm{in}}$ characteristics) for a value of signal reflectivity $R_1 = 90\%$.

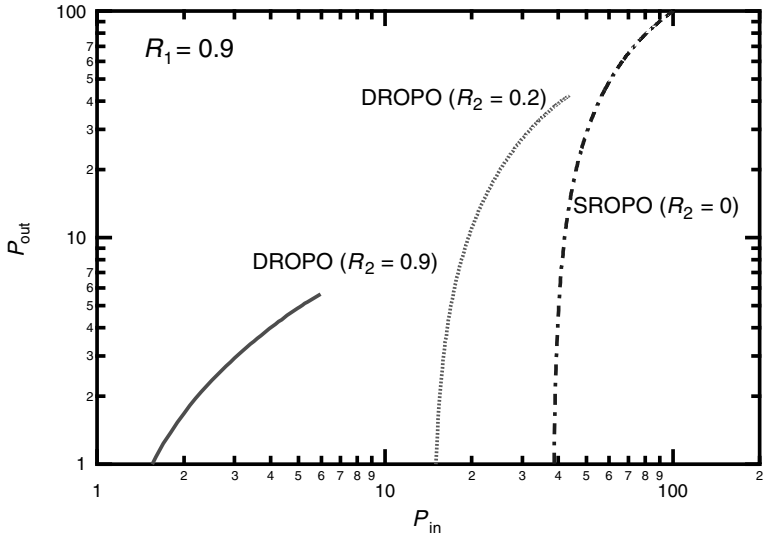


Fig. 12.F.1. Comparison between the $P_{in} - P_{out}$ characteristics of a balanced OPO ($R_1 = R_2 = 0.9$), a SROPO ($R_1 = 0.9, R_2 = 0$), and an unbalanced OPO ($R_1 = 0.9, R_2 = 0.2$). For comparison purposes, the intensities are normalized relative to the balanced DROPO threshold power.

For high-mirror reflectivities, i.e. when $\delta = 1 - R_1 \to 0$, an asymptotic expression may be derived from (12.F.15). Using the approximation $\cosh^{-1}(1/\sqrt{1-\delta}) \approx \sqrt{\delta}$, (12.F.15) may be written as:

$$\mathrm{sn}\left(\mathrm{i}\sqrt{\delta X}\middle| -\frac{Y'}{\delta X}\right) \approx \mathrm{i}\sqrt{\delta} \tag{12.F.17}$$

The asymptotic value of this latter expression is rather tricky to determine when $\delta \to 0$. It may be shown that, as soon as the reflectivity R_1 exceeds 80%, (12.F.17) is very well approximated by a *universal relation*, independent of the reflectivity:

$$X = \frac{Y'}{\sin^2\sqrt{Y'}} \tag{12.F.18}$$

Close to threshold ($X \approx 1$), one can derive a Taylor development from this expression:

$$Y \approx 6(\sqrt{X} - 1) \tag{12.F.19}$$

which, with output power at ω_1 given by $P_{\mathrm{out},1} = (1 - R_1)\omega_1 u_{1L}^2/2Z_0$, leads to the relation:

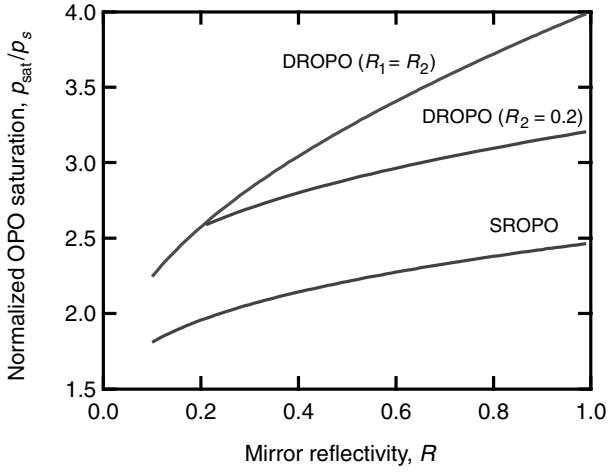


Fig. 12.F.2. Normalized saturation power of OPOs as a function of different mirror reflectivities. At these values, the OPO yield is 100%. For this input pump power, total pump depletion occurs.

$$\frac{P_{\text{out},1}}{\omega_1} = 6\frac{P_s}{\omega_3}\left(\sqrt{\frac{P_{\text{in}}}{P_s}} - 1\right) \tag{12.F.20}$$

SROPO cw $P_{out} - P_{in}$ characteristics

The value of the proportionality factor 6 in (12.F.20) is quite surprising, but can be derived directly by stating that the signal field E_1 is almost clamped in the cavity and that the pump (respectively idler) fields E_3 (respectively E_2) is then a cosine (respectively sine) function of position in the cavity.

Close examination of (12.F.15) also indicates that the output pump power can be completely depleted ($u_{3L} = 0$, or equivalently $X = Y'$) for an input power p_{sat} given by the relation:

$$\text{i}\kappa L\sqrt{p_{\text{sat}}} = \text{sn}^{-1}\left(i\sqrt{\frac{1}{R_1} - 1}\,\middle|\, -\frac{R_1}{1 - R_1}\right) \tag{12.F.21}$$

which, after some cumbersome elliptic function algebra, leads to:

$$X_{\text{sat}} = \frac{p_{\text{sat}}}{p_s} = \left[\frac{1}{\cosh^{-1}(1/\sqrt{R_1})}\right]^2 \frac{1 - R_1}{R_1} K^2\left(-\frac{1 - R_1}{R_1}\right) \tag{12.F.22}$$

where K is the elliptic integral $K(m) = \int_0^1 \text{d}t/[(1 - t^2)(1 - mt^2)]^{1/2}$. Figure 12.F.2 shows the variation of the saturation intensity X_{sat} as a function of reflectivity R_1. This numerical study shows that X_{sat} is weakly dependent on R (from 0.5 to 0.99) i.e. $X_{\text{sat}} \approx K^2(0) = (\pi/2)^2 \approx 2.4$. For input power higher than this value, part of the

signal and idler is converted back into pump power in the crystal. The input pump power p_{sat} is thus an optimum for which the OPO efficiency is 100%, i.e. all the input pump power is converted into signal and idler waves.

12.F.2 Doubly resonant OPO: the balanced DROPO

For the DROPO, the self-consistency equations are now:

$$u_1(0) = r_1 u_1(L) \tag{12.F.23a}$$

$$u_2(0) = r_2 u_2(L) \tag{12.F.23b}$$

$$\phi_1(0) - \phi_1(L) = \theta_1 + k_1 L + \frac{\omega_1 L_1'}{c} - 2m_1\pi \tag{12.F.23c}$$

$$\phi_2(0) - \phi_2(L) = \theta_2 + k_2 L + \frac{\omega_2 L_2'}{c} - 2m_2\pi \tag{12.F.23d}$$

where we have left the possibility of having two different cavity lengths for the signal and idler waves (L_1', L_2'). Concerning the phase conditions, above threshold none of the signal, idler, and pump fields is vanishing at some point of the crystal, which implies that the constant of motion Γ can take any non-zero value: therefore there is no constraint on the value of the relative phase $\phi(z)$.

Concerning the conditions on the amplitude (12.F.23a) and (12.F.23b), the flux conservation equation, (12.F.2d) gives:

$$(1 - R_1)u_1(L)^2 = (1 - R_2)u_2(L)^2 \tag{12.F.24}$$

which is nothing else than the equality of the idler and signal photon fluxes at the output of the OPO.

We will first assume, for simplicity's sake, that both reflectivities are equal, i.e. $R_1 = R_2$ (for instance at degeneracy): this is designated as the *balanced DROPO*. In that case, $u_1(z) = u_2(z) = u(z)$ and $\phi_1(z) = \phi_2(z)$. Equation (12.F.23c) leads to the phase condition:

$$\theta_1 + k_1 L + \frac{\omega_1 L_1'}{c} - 2m_1\pi = \theta_2 + k_2 L + \frac{\omega_2 L_2'}{c} - 2m_2\pi = \delta\phi \tag{12.F.25}$$

The system may now oscillate in detuned configurations, as the two sides of (12.F.25) are not necessarily equal to zero. However, the signal and idler wave detunings $\delta\phi$ must be equal with respect to the cold cavity resonance ($\delta\phi = 0$). Condition (12.F.25) together with photon energy conservation $\omega_1 + \omega_2 = \omega_3$ form an implicit equation which imposes given values for the signal and idler frequencies to be chosen among a comb of possible values depending on the two integers m_1 and m_2. This comb is continuously translatable along the frequency axis while changing the detuning $\delta\phi$.

Equations (12.F.1a–c) show that the transfer of energy between the pump and parametric waves is maximum for $\sin\phi(z) = 0$, which means a zero value for the constant Γ. In that particular case, similar to the SROPO, the individual phases $\phi_i(z)$ are constant in the crystal and the detuning $\delta\phi$ is zero. The converse (i.e. a zero detuning leading to a zero constant Γ) can also be shown. This is the case of *exact double resonance for the signal and idler waves* in the cavity. It is obtained, in practice, by tuning cavity lengths L_1' and L_2' into resonance, such as in a dual cavity DROPO. It can be shown, however, that the DROPO can oscillate even when this condition is not met, but with an enhanced threshold compared with the exact resonance case. With this latter restriction in mind, let us suppose that we are in the exact double resonance case, so that $\Gamma = 0$.

Since $m_3 = 0$, the power conservation equation, (12.F.2d) now reads:

$$p_c = u(z)^2 + u_3(z)^2 \tag{12.F.26}$$

Evolution of the signal wave is now given by:

$$\kappa L = \int_{\sqrt{R}u_L}^{u_L} \frac{\mathrm{d}u}{u\sqrt{(p_c - u^2)}} \tag{12.F.27}$$

where $u_L = u(L)$. This integrates trivially to give:

$$\log\left(\frac{1}{\sqrt{R}}\right) + \log\left(\frac{\sqrt{p_c} + \sqrt{p_c - Ru_L^2}}{\sqrt{p_c} + \sqrt{p_c - u_L^2}}\right) = \kappa L\sqrt{p_c} \tag{12.F.28}$$

The threshold is once again obtained when the idler signal $u_L \to 0$ with $p_c \neq 0$. This yields the threshold power for this balanced DROPO:

$$\kappa L\sqrt{p_s} = \log\left(\frac{1}{\sqrt{R}}\right) \tag{12.F.29}$$

Now the intracavity flux p_c is related to p_{in} and u_L through the relation (12.F.26), i.e. $p_c = p_{\text{in}} + Ru_L^2$. Using the same notation as in the SROPO case ($X = p_{\text{in}}/p_s$ and $Y = u_L^2/p_s$), the $P_{\text{out}} - P_{\text{in}}$ relation of (12.F.28) may then be written in normalized form:

$$\log\left[\frac{\sqrt{X + RY} + \sqrt{X}}{\sqrt{X + RY} + \sqrt{X - (1 - R)Y}}\right] = \log\left(\frac{1}{\sqrt{R}}\right)(\sqrt{X + RY} - 1) \tag{12.F.30}$$

Equation (12.F.30) is only valid if the output signal flux Y' is smaller than the

pump flux, i.e. $Y' = (1 - R)Y < X$. The equality $Y' = X$ is obtained for a *saturation power* p_{sat} given by (12.F.30), i.e:

$$p_{\text{sat}} = p_s(1 - R)\left[1 + \frac{\log(1 + \sqrt{1 - R})}{\log(1/\sqrt{R})}\right]^2 \tag{12.F.31}$$

At this value, the pump output power u_{3L} is then zero. Variation of the normalized saturation power as a function of mirror reflectivities $R = R_1 = R_2$ is shown in Fig. 12.F.2 where it is compared with the SROPO values. Figure 12.F.2 shows that, for reflectivity R in the 0.9–1 range, p_{sat} is about $4\,p_s$. At this value, i.e. four times above threshold, the OPO is 100% efficient. For input power above this value, signal power is reconverted into the pump wave: in that case, $\phi(z)$ changes from $\pi/2$ to $-\pi/2$ in (12.F.18). The calculation may then be completed and is left as an exercise.

Figure 12.F.1 now shows the $P_{\text{out}} - P_{\text{in}}$ characteristics of balanced DROPOs for mirror reflectivity values of 90% for both signal and idler waves. One can expand (12.F.30) into a series to obtain the *universal* $P_{\text{out}} - P_{\text{in}}$ *relation for a degenerate DROPO*, which holds below saturation as soon as the reflectivity exceeds 90%:

$$Y' = 4(\sqrt{X} - 1) \tag{12.F.32}$$

or else

$$\frac{P_{\text{out}}}{\omega_1} = 4\frac{P_s}{\omega_3}\left(\sqrt{\frac{P_{\text{in}}}{P_s}} - 1\right) \tag{12.F.33}$$

Balanced DROPO cw $P_{\text{out}} - P_{\text{in}}$ characteristics

Clearly, one finds that the OPO efficiency equals 1 when $P_{\text{in}} = P_{\text{sat}} = 4P_s$, which is consistent with (12.F.31).

12.F.3 Doubly resonant OPO: the general case

Finally, we shall briefly give below the equations describing DROPO in the general regime ($R_1 \neq R_2$), which are useful for practical applications but rather cumbersome to derive. Assuming an exact double resonance ($\Gamma = 0$), combining (12.F.1b) with (12.F.2) yields:

$$\kappa L = \int_{\sqrt{R_2}u_{2L}}^{u_{2L}} \frac{\mathrm{d}u_2}{\sqrt{(m_3 + u_2^2)(m_1 - u_2^2)}} \tag{12.F.34}$$

where, taking into account the flux conservation relation (12.F.24), the constants m_1 and m_3 are now given by:

$$m_1 = p_{\text{in}} + R_2 u_{2L}^2 \tag{12.F.35a}$$

$$m_3 = R_1 u_{1L}^2 - R_2 u_{2L}^2 = \frac{R_1 - R_2}{1 - R_1} u_{2L}^2 \tag{12.F.35b}$$

Using the Jacobian function defined in (12.F.12), Eq. (12.F.34) now reads:

$$\begin{aligned} \mathrm{i}\kappa L\sqrt{p_{\text{in}} + R_2 u_{2L}^2} = \mathrm{sn}^{-1}\left(\mathrm{i}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\frac{u_{2L}^2}{p_{\text{in}}+R_2u_{2L}^2}\right) \\ - \mathrm{sn}^{-1}\left(\mathrm{i}\sqrt{R_2}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\frac{u_{2L}^2}{p_{\text{in}}+R_2u_{2L}^2}\right) \end{aligned} \tag{12.F.36}$$

This last equation effectively yields Eq. (12.F.13) in the SROPO case ($R_2 = 0$). The expression for the threshold power p_s is easily obtained with $u_{2L} = 0$, i.e:

$$\kappa L\sqrt{p_s} = \log\left(\frac{\sqrt{1-R_1}+\sqrt{1-R_2}}{\sqrt{(1-R_1)R_2}+\sqrt{(1-R_2)R_1}}\right) \tag{12.F.37}$$

which is somewhat different from the usual expression derived in Eq. (12.61). Once again, the OPO reaches its optimum operation condition (yield = 100%) when the condition $p_{\text{in}} = (1 - R_2)u_{2L}^2$ is met, that is when:

$$\begin{aligned} \mathrm{i}\kappa L\sqrt{p_{\text{sat}}} = \sqrt{1-R_2}\left[\mathrm{sn}^{-1}\left(\mathrm{i}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\right)\right. \\ \left. - \mathrm{sn}^{-1}\left(\mathrm{i}\sqrt{R_2}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\right)\right] \end{aligned} \tag{12.F.38}$$

which allows one to trace the normalized value of saturation pump power X_{sat} as a function of reflectivities R_1 and R_2, i.e. intermediate between the SROPO and degenerate DROPO cases (see Fig. 12.F.2). The normalized $P_{\text{out}} - P_{\text{in}}$ relation in normalized form is rather cumbersome and reads:

$$\begin{aligned} \mathrm{i}\sqrt{X + R_2 Y} = \log\left[\frac{\sqrt{1-R_1}+\sqrt{1-R_2}}{\sqrt{(1-R_1)R_2}+\sqrt{(1-R_2)R_1}}\right] \\ \left[\mathrm{sn}^{-1}\left(\mathrm{i}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\frac{Y}{X+R_2Y}\right)\right. \\ \left. -\mathrm{sn}^{-1}\left[\mathrm{i}\sqrt{R_2}\sqrt{\frac{1-R_1}{R_1-R_2}}\middle| -\frac{R_1-R_2}{1-R_1}\frac{Y}{X+R_2Y}\right)\right] \end{aligned} \tag{12.F.39}$$

From this latter expression, all the different experimental situations developed above can be obtained, with some algebraic difficulty, to rederive the balanced

DROPO case. Figure 12.F.1 shows the numerical results for $P_{out} - P_{in}$ characteristics obtained for $R_1 = 90\%$ with an idler reflectivity $R_2 = 20\%$.

FURTHER READING

M. Abramowitz and I. Stegun, *Handbook of Mathematical Tables*, Dover Publications, New York (1970).

E. Rosencher and C. Fabre, *JOSA B* (to be published 2001).

13 Light emitting diodes and laser diodes

13.1 Introduction

In Chapter 6, we saw that a semiconductor driven away from thermodynamic equilibrium can emit light (in addition to blackbody radiation) when excited carriers recombine from one band to another. We also derived the Bernard–Durrafourg condition, which the energy distributions of the carrier populations must satisfy before optical amplification can occur. We will now show how this light emission can be put to use in *electroluminescent diodes* (alternatively known as *light emitting diodes* or *LEDs*) and *laser diodes*. In doing so, we will draw upon the contents of no less than five of the previous chapters:

- Chapter 4, which describes the physics of laser oscillations;
- Chapter 7, which describes the various optical emission mechanisms in semiconductors;
- Chapter 8, which describes the physics of semiconductor heterostructures and quantum well structures;
- Chapter 9, which describes waveguiding in optical heterostructures;
- Chapter 10, which describes carrier injection mechanisms in p–n diodes.

This chapter is fairly complex (and exciting!) in that it brings into play many of the different physical concepts elaborated over the course of this book. While making frequent use of material developed in other chapters, we will take the time to recap many of the key concepts to allow the reader to progress through this chapter without breaking stride on too many occasions.

13.2 Electrical injection and non-equilibrium carrier densities

Light emission in a semiconductor usually proceeds from electron–hole recombination in regions where they are in excess in comparison with levels allowed by thermodynamic equilibrium. (The only exception to this case is the *unipolar quantum cascade laser*, which we will encounter later in Complement 13.H.) Electron–hole recombination naturally occurs in a forward biased p–n junction where we see that minority carriers can coexist with majority carriers over distances of the order of L_D (the diffusion length). This is the phenomenon referred to as *electrical injection* (Chapter 10). We will consider such a junction in which both

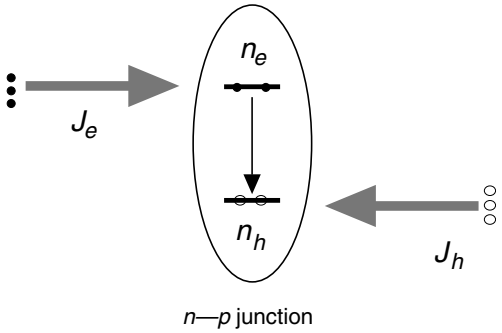


Fig. 13.1. Optical emission in a semiconductor results from the recombination of electron–hole pairs in regions where both carrier types coexist (e.g. as in a *p–n* junction).

electron J_e/q and hole J_h/q (in $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) fluxes converge (where J_e and J_h designate the electron and hole current densities, and q is the absolute value of the electronic charge). Given the requirement of charge neutrality, we have $J_e = J_h = J$ (see Fig. 13.1).

The problem now is to figure out the volumetric density of the non-equilibrium carriers n_e and n_h in the junction. This will allow us to calculate, with the help of quasi-Fermi levels (7.67), the stimulated (7.41) and spontaneous (7.47) emission rates in the structure. To do so, it is sufficient to write under stationary state conditions, that the total recombination rate R_{tot} (in $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$) in the volume V exactly equals the incident flux of carriers flowing through the surface S, i.e. $VR_{\mathrm{tot}} = SJ_e/q = SJ_h/q = SJ/q$. For no other reason than to simplify the required notation, we will assume that doping near the junction is negligible in comparison with the non-equilibrium carrier density. This guarantees that $n_e = n_h = n$.

The total recombination rate R_{tot} that compensates for the flux entering the junction region reflects contributions from at least four different mechanisms addressed in the preceding chapters:

- the non-radiative recombination rate $A_{\mathrm{nr}}n$ due to the presence of deep level defects, surface defects, . . . (see Section 6.5 and Complement 5.D);
- spontaneous radiative recombination Bn^2, where B is the bimolecular recombination coefficient given by (7.61) and (7.62);
- the Auger recombination term $C_{\mathrm{Aug}}n^3$ given in Complement 6.D;
- the stimulated recombination rate $R_{\mathrm{st}}N_{\mathrm{ph}}$, where $R_{\mathrm{st}}(=R_{\mathrm{abs}})$ is the stimulated recombination coefficient (7.41) and N_{ph} is the photon density.

To begin, we will focus on the behaviour of the device far below laser threshold, i.e. we will neglect simulated emission ($N_{\mathrm{ph}} = 0$). The carrier density is then given by the stationary state condition:

$$\frac{J_e}{qd} = \frac{J_h}{qd} = \frac{J}{qd} = A_{\mathrm{nr}}n + Bn^2 + C_{\mathrm{Aug}}n^3 \tag{13.1}$$

where $d = V/S$ is the effective thickness of the junction, i.e. the distance over which electron–hole recombination takes place. It is clear in (13.1) that the recombination terms are not equivalent. The first and last terms do not generate light (i.e. they are non-radiative), whereas the second term describes the spontaneous emission in the structure. We thus regroup these terms in the following manner:

$$\frac{1}{t_{\mathrm{nr}}} = A_{\mathrm{nr}} + C_{\mathrm{Aug}} n^2$$

$$\frac{1}{t_{\mathrm{rad}}} = Bn \tag{13.2}$$

$$\frac{1}{t_{\mathrm{tot}}} = \frac{1}{t_{\mathrm{nr}}} + \frac{1}{t_{\mathrm{rad}}}$$

where t_{rad}, t_{nr}, and t_{tot} are the *radiative*, *non-radiative*, and *total recombination times*, respectively. The non-equilibrium carrier density is then given by:

$$n = \frac{J t_{\mathrm{tot}}}{qd} \tag{13.3}$$

This very simple formula is one of the major results of this chapter. We note:

- The total recombination time t_{tot} depends on the carrier densities and, therefore, on the operation conditions. Often, this time is determined by the residual doping (in this case, we replace the carrier density n in (13.3) by the doping concentration N_d). In most instances, this quantity will be a constant determined by experiment.
- The Auger term $C_{\mathrm{Aug}} n^2$ depends explicitly on the carrier densities and becomes increasingly important as these concentrations increase.
- For a given current density the non-equilibrium carrier density is dramatically enhanced as the thickness of the active region decreases. For a homojunction, this thickness d is given by $L_{Dn} + L_{Dp}$ where the diffusion lengths of the minority carriers L_{Dn} (electrons in the p region) and L_{Dp} (holes in the n region) are given by (10.41) and typically equal anywhere from 1 to 10 μm (see Fig. 13.2). On the other hand, this thickness can be reduced to less than 100 nm in a heterostructure, and down to a few nanometers in a quantum well (see Fig. 13.2).

Once we have determined the non-equilibrium carrier density in (13.3), we can calculate the positions of the quasi-Fermi levels for each carrier type with the help of (7.67).

Having done so, one of three possible cases will emerge:

- spontaneous emission dominates over stimulated emission, and the diode is in the electroluminescent or LED regime;
- the medium is inverted (see the Bernard–Durrafourg criterion in (7.26b) and

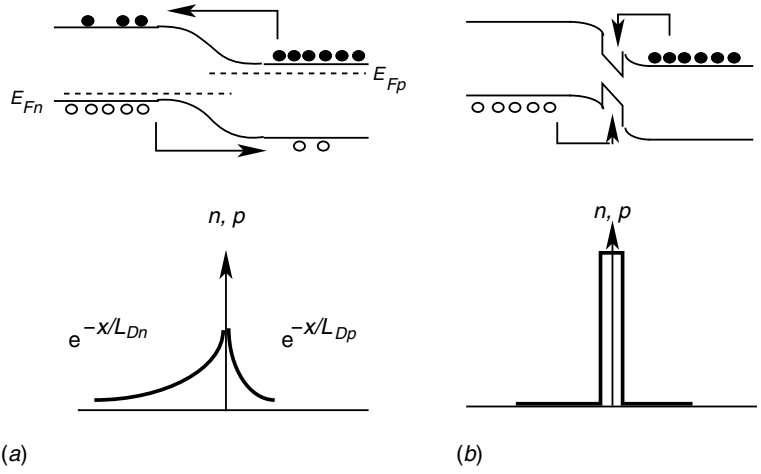


Fig. 13.2. For a given carrier flux, the density of electron–hole pairs is far greater in a heterojunction (*b*) than in a homojunction (*a*) where these carriers can diffuse more readily.

(7.69)) but the cavity losses are superior to the gain in the medium, and the diode is in the optical amplification regime (and is referred to as a *superluminescent* diode);

- stimulated emission dominates over spontaneous emission, and the cavity losses are exactly compensated by the medium gain placing the device in the laser oscillation regime (i.e. making it a laser diode).

We will interest ourselves for the time being with first case.

Example

An $Al_{0.25}Ga_{0.75}As/GaAs$ heterostructure with a thickness $d = 10\,\text{nm}$ possesses under certain conditions a total recombination time of 5 ns. A 1 mA current crossing a $10\,\mu\text{m} \times 100\,\mu\text{m}$ junction area results in a non-equilibrium carrier density (13.3) of:

$$n = 10^{-3}\,\text{A} \times 5 \times 10^{-9}\,\text{s}/(10^{-5}\,\text{cm}^2 \times 1.6 \times 10^{-19}\,\text{C} \times 10^{-6}\,\text{cm}) = 3.1 \times 10^{18}\,\text{cm}^{-3}$$

As this carrier density exceeds the effective density of states in the conduction band $N_c\,(=4.3 \times 10^{17}\,\text{cm}^{-3})$, the electron gas is degenerate and the quasi-Fermi level for the electrons (5.56b) is:

$$E_{Fn} - E_c = (1.05 \times 10^{-34}\,\text{J})^2/(2 \times 0.067 \times 0.9 \times 10^{-30}\,\text{kg})(3 \times \pi^2 \times 3.1 \times 10^{24}\,\text{m}^{-3})^{2/3}/1.6 \times 10^{-19}\,\text{C} = 116\,\text{meV}$$

Alternately, the hole density $p\,(=n)$ is less than the effective density of states in the valence band $N_v\,(=1.3 \times 10^{19}\,\text{cm}^{-3})$, and so the quasi-Fermi level is given by (5.47) as:

$$E_{Fp} - E_v = 0.0259\,\text{eV} \times \ln(1.3 \times 10^{19}/3 \times 10^{18}) = 38\,\text{meV}$$

As the energy spread between the electron and hole Fermi levels is greater than the bandgap, i.e:

$$(E_{Fn} - E_c) - (E_{Fp} - E_v) = 116 - 38\,\text{meV} > 0$$

the system is inverted according to the Bernard–Durrafourg criterion.

13.3 Electroluminescent diodes

13.3.1 Electroluminescence

As the excess carriers build up in the junction region, they distribute themselves in the conduction and valence bands, occupying increasingly elevated energy states (see Fig. 13.3). Using a single valence band model, the carrier distributions can be described in terms of the quasi-Fermi levels (see Section 7.6) E_{Fc} (for electrons in the conduction band) and E_{Fv} (for the holes in the valence band) as:

$$\begin{aligned} E_{Fc} &= E_c + kT\tilde{F}_{1/2}\left(\frac{n}{N_c}\right) \\ E_{Fv} &= E_v - kT\tilde{F}_{1/2}\left(\frac{n}{N_v}\right) \end{aligned} \tag{13.4}$$

where $\tilde{F}_{1/2}$ is the inverse of the Fermi–Dirac integral $F_{1/2}$, given by:

$$F_{1/2}(u) = \frac{1}{\Gamma(3/2)}\int_0^\infty \frac{x^{1/2}}{1 + \mathrm{e}^{(x-u)}}\,\mathrm{d}x \tag{13.5}$$

As discussed in Section 7.4, this displacement from equilibrium (as described by the quasi-Fermi levels for the electrons and holes) produces an increase in spontaneous emission with a spectral rate $R_{\text{spon}}(h\nu)$ per unit time, volume, and energy (for photons with energy $h\nu$) given by (7.47):

$$R_{\text{spon}}(h\nu) = \frac{1}{\tau_R}\rho_j(h\nu) f_c(h\nu)[1 - f_v(h\nu)] \tag{13.6}$$

where for quantities appearing in this last equation, we recall the definitions introduced in Chapter 7:

- τ_R is the *spontaneous radiative lifetime* in the emitting material, which can be assumed to be relatively independent of the energy of the emitted photons. This quantity is characteristic of a given semiconductor and depends upon its physi-

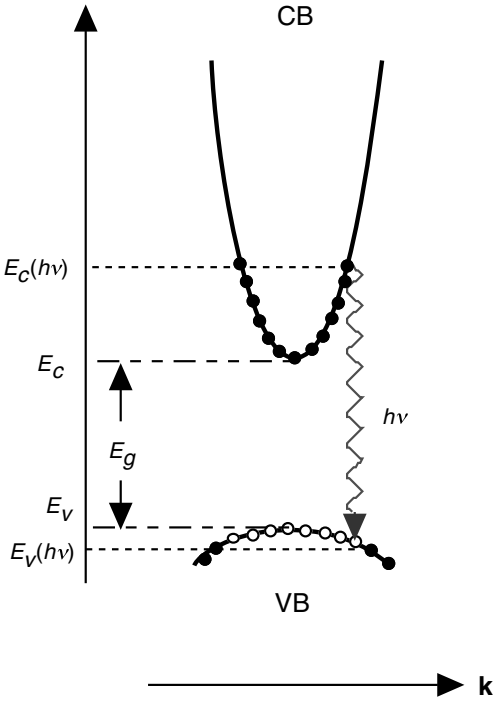


Fig. 13.3. Optical transitions in a semiconductor are *vertical* (i.e. **k** conserving) in energy. The states $E_c(h\nu)$ and $E_v(h\nu)$ are joined by optical transitions involving photons with energies $h\nu$: CB, conduction band; VB, valence band.

cal parameters (effective carrier masses, Kane matrix elements, etc.) as described by (7.37b).

- $\rho_j(h\nu)$ is the joint density of states given in (7.19). It counts the 'number' of states that satisfy the energy $h\nu = E_c(\mathbf{k}) - E_v(\mathbf{k})$ and momentum (constant **k**) conservation requirements for transitions between states with energy $E_v(\mathbf{k})$ in the valence band and $E_c(\mathbf{k})$ in the conduction band (see Fig. 13.3).
- $f_c(h\nu)$ and $f_v(h\nu)$ are the Fermi–Dirac functions describing the occupation of the states connected by the optical transitions, i.e. satisfying (7.25b,c):

$$f_c(h\nu) = \frac{1}{1 + \exp\left(\dfrac{E_c(h\nu) - E_{Fc}}{kT}\right)}$$

$$E_c(h\nu) = E_g + \frac{m_r}{m_c}(h\nu - E_g) \tag{13.7a}$$

and

$$f_v(h\nu) = \frac{1}{1 + \exp\left(\dfrac{E_v(h\nu) - E_{Fv}}{kT}\right)}$$
$$E_v(h\nu) = -\frac{m_r}{m_v}(h\nu - E_g) \tag{13.7b}$$

As a general rule, in the case of electroluminescent diodes, the excess carrier density is sufficiently weak to justify approximating the Fermi–Dirac functions by Boltzmann functions (i.e. the quasi-Fermi levels are far from the band extrema). As seen in Chapter 7, in this case the spectral spontaneous emission rate can be considerably simplified yielding (see (7.52a)):

$$R_{\text{spon}}(h\nu) = K_{\text{spon}}(h\nu - E_g)^{1/2}\exp\left(-\frac{h\nu - E_g}{kT}\right) \tag{13.8}$$

where K_{spon} is given by:

$$K_{\text{spon}} = \frac{(2m_r)^{3/2}}{\pi h^2 \tau_R}\exp\left(\frac{\Delta E_F - E_g}{kT}\right) \tag{13.9}$$

and ΔE_F is the difference between the quasi-Fermi levels $\Delta E_F = E_{Fc} - E_{Fv}$. Clearly, as the excess carrier density n increases, the factors $\exp(E_{Fc}/kT)$ and $\exp(-E_{Fv}/kT)$ rise as well, so that the spectral rate $R_{\text{spon}}(h\nu)$ is proportional to n^2, which is not surprising. The spontaneous emission lineshape is given by (13.8). It is trivial to show that the maximum intensity is emitted at $h\nu_{\text{peak}} = E_g + kT/2$ and that the full width at half maximum $\Delta h\nu \approx 1.8\,kT$ (see Fig. 13.4).

As λ_{peak} and $h\nu_{\text{peak}}$ are related by $\lambda_{\text{peak}}\,(\mu\text{m}) = 1.24/h\nu_{\text{peak}}\,(\text{eV})$, we immediately deduce that the full width at half maximum in terms of wavelength $\Delta\lambda$ is related to λ_{peak} by:

$$\Delta\lambda = 1.45\lambda_{\text{peak}}^2 kT \tag{13.10}$$

Figure 13.5 shows the spectral densities emitted by LEDs fabricated from various semiconductors. As predicted by (13.10), the emission spectrum becomes narrower in terms of wavelength as the emission energy moves from the infrared to the ultraviolet ends of the spectrum.

13.3.2 Internal and external efficiencies for LEDs

We will now shift our attention from the spectral distribution $R_{\text{spon}}(h\nu)$ of the emitted energy to the total intensity R_{spon} corresponding to the evaluation of the integral of (13.6) over the entire semiconductor band structure. This integration was carried out in Section 7.5. At that time, we were able to show that the radiative

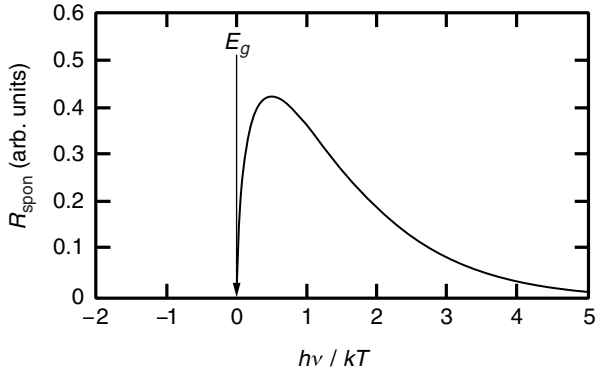


Fig. 13.4. Spectral distribution of the spontaneous emission rate. The width of the spectrum is of the order of $1.8\,kT$.

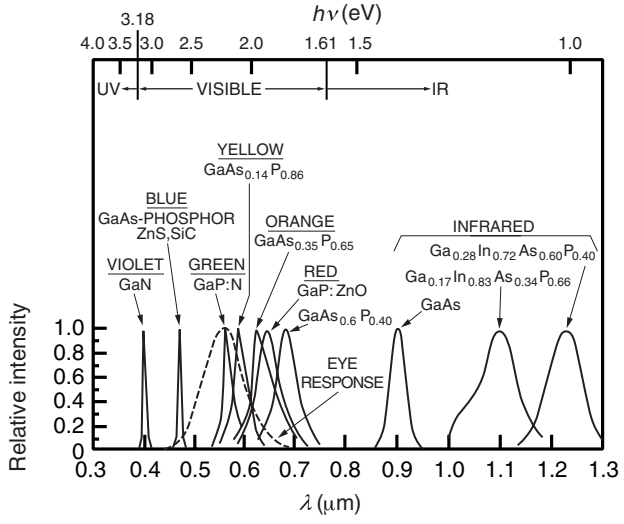


Fig. 13.5. Spectral distribution of LED emission for diodes fabricated from different semiconductors. The spectral width is proportional to λ^2_{peak} as predicted by (13.10). From S. M. Sze: *Physics of Semiconductor Devices*, copyright © Wiley Interscience, New York (1981), reprinted by permission of John Wiley & Sons, Inc.

recombination rate for the case where the carrier densities are far from being degenerate (as is the case with LEDs) is given (see (7.62) and (7.65)) by:

$$R_{\text{spon}} = Bn_e n_h = \frac{n}{t_{\text{rad}}} \tag{13.11}$$

where t_{rad} is the radiative recombination time and B is the bimolecular recombination coefficient which relates to the radiative lifetime τ_R according to (7.62). In this

situation, the total photon flux Φ released by the volume V of the junction takes the particularly simple form:

$$\Phi = R_{\text{spon}} \frac{V}{S} = \frac{n}{t_{\text{rad}}} d \tag{13.12}$$

We then introduce the coefficient η_i, which is given by the ratio:

$$\eta_i = \frac{t_{\text{tot}}}{t_{\text{rad}}} = \frac{1}{1 + t_{\text{rad}}/t_{\text{nr}}} \tag{13.13}$$

Internal quantum efficiency

This coefficient η_i is the *internal quantum efficiency* for the LED. It gives the percentage of electron–hole pairs that can successfully release their stored energy in the form of emitted photons immediately following recombination. Given (13.3), we see that the flux of emitted photons is given by:

$$\Phi = \eta_i \frac{J}{q} \tag{13.14}$$

This last equation clearly shows the role played by the internal quantum efficiency in terms of effectively transforming a carrier flux J/q into a photon flux Φ.

In spite of this, not all the light emitted by a p–n junction can escape the semiconductor. In fact, electroluminescent emission is not directional, it is Lambertian. Figure 13.6 shows all the optical mechanisms that contribute to diminishing the total efficiency of an LED:

- θ_1 rays (i.e. making small angles with the surface normal vector) travel across semiconductor material, which will absorb some portion of the emitted light. Clearly, this absorption can be minimized in a heterojunction by choosing constituent semiconductors for the n and p regions with bandgaps in excess of the photon energy of the light released by the structure. Nonetheless, a small portion of the small-angle light rays still experience partial reflection by the semiconductor–air interface. The fraction η_{diel} of normally incident light transmitted by the interface (i.e. the *dielectric efficiency*) is given by Fresnel's equation to be:

$$\eta_{\text{diel}} = 1 - \frac{(n_{\text{sc}} - 1)^2}{(n_{\text{sc}} + 1)^2} = \frac{4n_{\text{sc}}}{(n_{\text{sc}} + 1)^2} \tag{13.15}$$

  where n_{sc} is the optical index of the semiconductor material. For GaAs ($n_{\text{sc}} = 3.6$), the transmission efficiency η_{diel} is 0.7. This efficiency becomes greater as the semiconductor's optical index becomes smaller (i.e. as the diodes emit at shorter wavelengths). This is, for instance, well verified for blue-emitting GaN

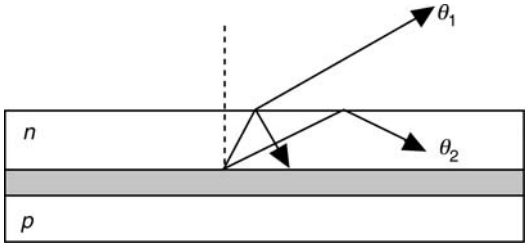


Fig. 13.6. The light rays that form an angle θ_1 with the normal surface vector of the LED experience partial Fresnel reflection, whereas those rays making an angle greater than θ_2 experience total internal reflection.

diodes for which $n_{sc} = 2.3$ and $\eta_{diel} = 0.85$.

- θ_2 rays emitted outside the cone defined by the critical angle θ_c are completely reflected. If the process ended there, the efficiencies for LEDs would be catastrophically low. In that case, the percentage of emitted light would only be 4%! In fact, LEDs are helped along by a second phenomenon known as *photon recycling*. A portion of the back-reflected photons are absorbed, forming electron–hole pairs; and, as such, through subsequent recombination events, stand a chance of being re-emitted ultimately to escape the semiconductor–air interface. All these processes combine to yield an effective transmission efficiency η_t which, when multiplied by the internal quantum efficiency, yields the fraction of electron–hole pairs converted into photons capable of escaping the LED. The external power P_{ext} is then given by the product of the external photon flux by he energy of the emitted photons:

$$P_{ext} = \eta_{ext} \frac{J}{q} hv \tag{13.16}$$

where η_{ext} is the *external quantum efficiency* given by $\eta_{ext} = \eta_t \eta_i$. It is customary to introduce the *response* $\mathscr{R}$ (W A^{-1}) for an LED as the ratio of the emitted light power density divided by the forward bias junction current density J, i.e. $\mathscr{R} = P_{ext}/J$ such that:

$$\mathscr{R} = \eta_{ext} \frac{hv}{q} = \eta_{ext} \frac{1.24}{\lambda_{peak}} \tag{13.17}$$

LED response (W A^{-1}) at a particular peak emission wavelength (μm)

We can now inquire about the energetic efficiency of an electroluminescent diode. If the series resistance for the diode is weak, the required applied voltage to obtain emission will be of the order of a semiconductor bandgap, i.e. hv/q. The electrical power invested in the diode is then $J \times hv/q$ (W cm^{-2}). Equation (13.16) gives us the *energy efficiency* of the diode $P_{ext}/(J \times hv/q)$ or $\sim\eta_{ext}$. The energy efficiency is therefore close to the external quantum efficiency η_{ext} (to within the series resis-

tance), i.e. of the order of 20%. This is to be compared with the energy efficiency of ~1% (in the visible spectrum) for an incandescent light bulb where the remaining ~99% of the electrical energy is dissipated as heat.

Examples

1. A GaAs/$Al_{0.25}Ga_{0.75}As$ LED emits at a peak wavelength of 0.87 μm. Its internal quantum efficiency is 50% and its dielectric efficiency η_{diel} is 20%. The response of this LED is therefore:

   $$\mathscr{R} = 0.5 \times 0.2 \times 1.24\,\mu m/0.87\,\mu m = 0.14\,W\,A^{-1}$$

   Assuming a pump current of 20 mA, the diode will emit 2.8 mW of optical power.
2. An InGaN/GaN diode emits at a peak wavelength of 0.4 μm. The internal efficiency is 50%, but its dielectric efficiency is 40% (the optical index of GaN is 2.3). The response of this diode is therefore:

   $$\mathscr{R} = 0.5 \times 0.4 \times 1.24\,\mu m/0.4\,\mu m = 0.62\,W/A^{-1}$$

   With a pump current of 20 mA, the diode will emit a total external power of 12.4 mW – clearly superior to that achieved by the GaAs/AlGaAs LED.
3. The bimolecular recombination coefficient B determined for GaAs is given in Table 7.1, Section 7.5 ($B = 7.2 \times 10^{-10}\,cm^3\,s^{-1}$). Assuming a heterostructure doping level $N_D = 2 \times 10^{17}\,cm^{-3}$, the radiative recombination time t_{rad} is then given by (7.65) to be $t_{rad} = 1/BN_D = 7$ ns. Assuming a non-radiative recombination time of 5 ns, the internal quantum efficiency is then 42%.

13.3.3 A few device issues

LED diodes have three principal uses: display panels, lighting, medium rate optical communications. They are fairly economical to deploy as they can be manufactured in parallel *en masse*. Figure 13.7 shows a typical device configuration for an electroluminescent diode. In this example, layers of $GaAs_xP_{1-x}$ are epitaxially deposited onto a GaP wafer. The composition x is then varied from $x = 0$ (GaP) to $x = 0.35$ ($E_g(GaAs_{0.35}P_{0.65}) = 0.62$ eV), allowing gradual lattice matching between the substrate and the electroluminescent layer. A layer of SiO_2 is then deposited over the entire wafer surface, and patterned using standard lithographic techniques to open holes into the dielectric layer. *P*-type dopant atoms are then diffused through the openings in the SiO_2 and a metallic layer is finally deposited on the top surface (making contact with the *p*-type material) and on the rear of the substrate. As the GaP is transparent to the emitted light, light travelling towards the substrate is reflected by the backside metallization towards the openings in the top metallization layer. This increases the external efficiency of

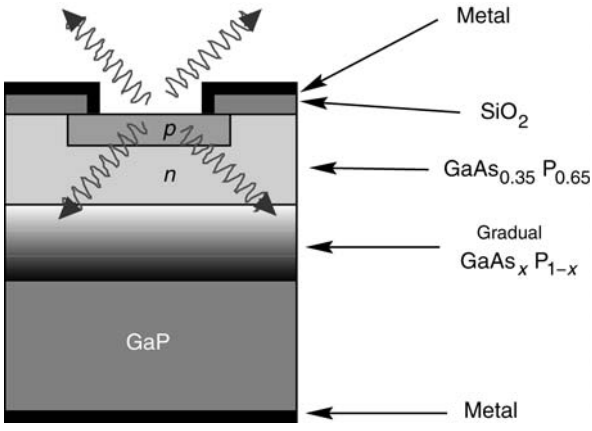


Fig. 13.7. Device structure for a LED with a transparent substrate.

the electroluminescent diode. All that is left to do is to dice the wafer into pieces to isolate the LEDs (or 'chips') and to mount them individually in windowed canisters. Assuming typical LED surface dimensions of $100 \times 100\,\mu m^2$, we can expect on the order of 200 000 discrete devices from a $20\,cm^2$ wafer (i.e. from a '2 inch wafer' in microelectronics' jargon).

As spontaneous emission in a semiconductor occurs isotropically, light emission in an LED is Lambertian (i.e. varies as $\cos\theta$ – see Fig. 13.8*a*). In order to increase the directionality of this emission (i.e. to increase the LED's brilliance), the component is generally encased in a parabolic lens fashioned out of epoxy resin (see Fig. 13.8).

Finally, the LED response times are equal to the total recombination time t_{tot}. As mentioned elsewhere, these times lie typically between 1 and 50 ns corresponding to modest bandpass frequencies of the order of a few hundred MHz.

13.4 Optical amplification in heterojunction diodes

The emphasis in Section 13.3 was on spontaneous emission. In Chapter 7, however, we saw that if a semiconductor is driven sufficiently far from thermodynamic equilibrium, it may become possible to create a population inversion between the electrons in the conduction band and the holes in the valence band. We were even able to establish the transparency condition, i.e. the required concentration of charge carriers n_{transp} in the bands to make the material effectively transparent to radiation, and beyond which the material begins to exhibit amplification (see (7.69)). Translated in terms of an electrical injection current using (13.3), the material becomes transparent when electrically driven with a *transparency current density* J_{transp} given by:

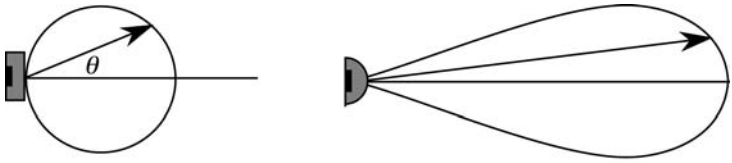


Fig. 13.8. (*a*) Electroluminescent emission is Lambertian. (*b*) To increase the brilliance of LEDs, the devices are encased in epoxy resin fashioned into parabolic lenses.

$$\tilde{F}_{1/2}\left(\frac{J_{\text{transp}}t_{\text{tot}}}{qdN_c}\right) + \tilde{F}_{1/2}\left(\frac{J_{\text{transp}}t_{\text{tot}}}{qdN_v}\right) = 0 \tag{13.18}$$

where N_c and N_v are the effective state densities in the conduction and valence bands, respectively. When driven beyond this transparency threshold, photons with energy hv that fulfil the Bernard–Durrafourg criterion:

$$E_g < hv < E_{Fc} - E_{Fv} \tag{13.19}$$

will experience optical amplification. The gain for an amplifying medium $\gamma(hv)$ was calculated in Section 7.3 within the context of optical susceptibility and in Section 7.4 during the discussion of rate equations. These results are sufficiently important as to warrant our repetition of them here. The gain is given by:

$$\gamma(hv) = \alpha_0(\omega)[f_c(hv) - f_v(hv)] \tag{13.20a}$$

where α_0 is the empty conduction band absorption (i.e. under zero current) given by:

$$\alpha_0(\omega) = \frac{q^2 x_{vc}^2}{\lambda_0 \varepsilon_0 h n_{\text{sc}}}\left(\frac{2m_r}{\hbar}\right)^{3/2}\sqrt{(\omega - E_g/\hbar)} \tag{13.20b}$$

We recall that f_c and f_v are the Fermi–Dirac functions given in (13.7), x_{vc} is the dipolar matrix element given by (7.11), m_r is the reduced effective mass (7.13), n_{sc} is the index of refraction for the semiconductor, and λ_0 is the vacuum wavelength for an electromagnetic wave of frequency ω.

Equation (13.20a) can also be written in a form which better resembles the equations for the atomic transition laser introduced in Chapter 4:

$$\gamma(hv) = \frac{\lambda^2}{8\pi\tau_R}\rho_j(hv)[f_c(hv) - f_v(hv)] \tag{13.20c}$$

It is clear that in (13.20c), the factor $\rho_j(hv)(f_c - f_v)$ plays the role of $g(hv)(N_2 - N_1)$ in the equations for the atomic transition laser.

Therefore, as the current density J passing through the diode becomes more important, the carrier density n given by (13.3) increases. This causes the

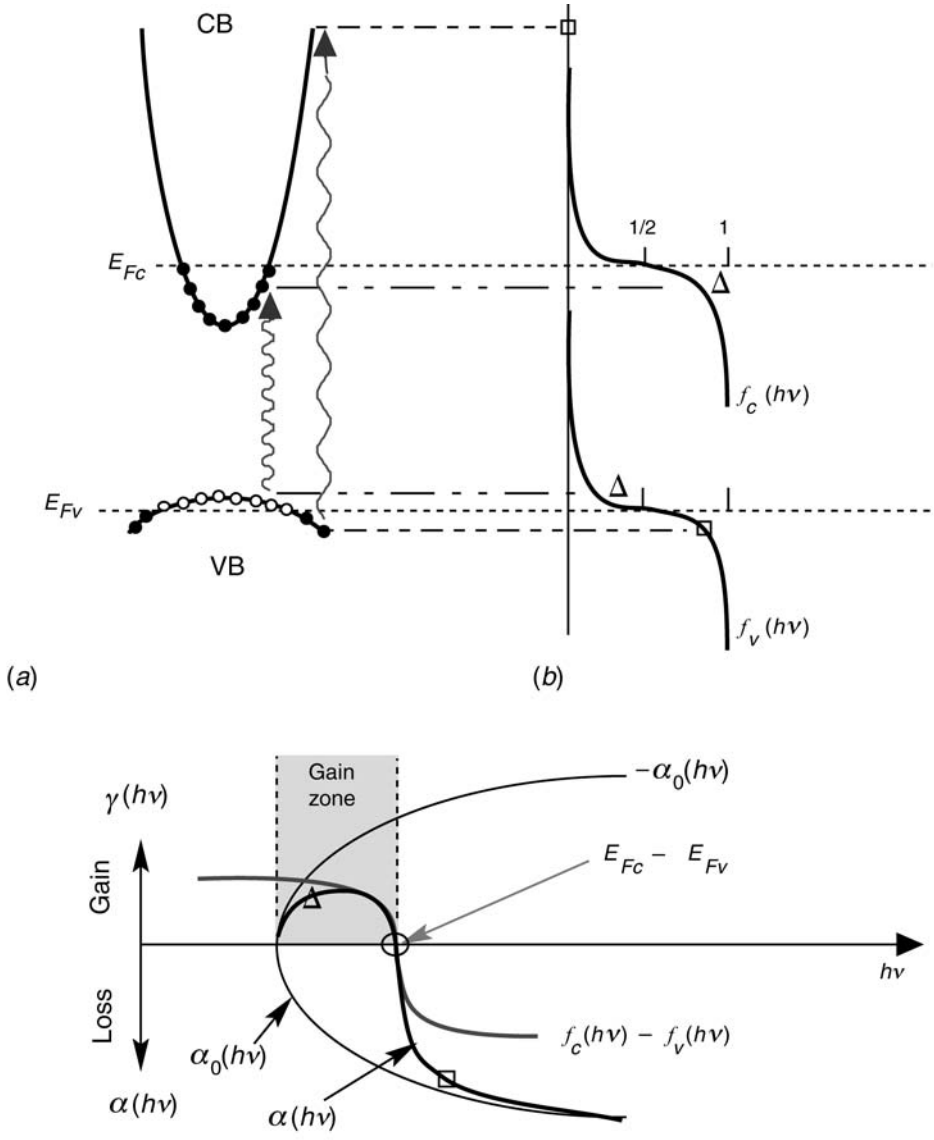


Fig. 13.9. Population inversion, and optical amplification in a semiconductor medium. This figure illustrates Eq. (13.20): (*a*) represents the two energy bands involved, (*b*) shows the associated Fermi–Dirac distributions for the quasi-Fermi levels in the conduction and valence bands, (*c*) shows the gain curve for the medium (thick dark curve) resulting from the product of the absorption $\alpha_0(h\nu)$ (thin curve) and the filling factor $f_c - f_v$ (thick grey curve). The symbols Δ and □ denote photons with energies satisfying and not satisfying the Bernard–Durrafourg condition, respectively.

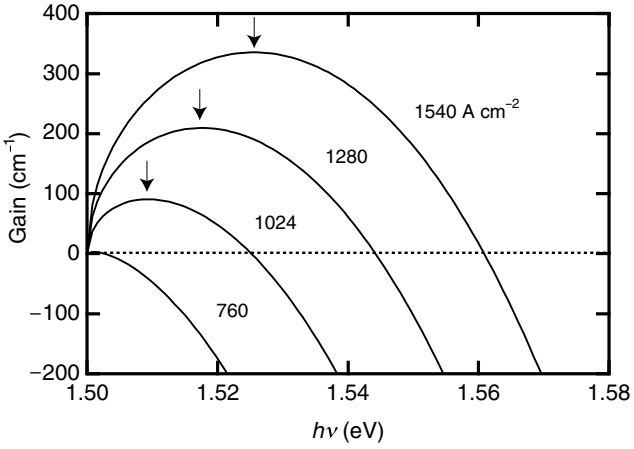


Fig. 13.10. Gain for a 100 nm GaAs/AlGaAs heterostructure (assuming a radiative lifetime of 5 ns, and an internal quantum efficiency of 0.5 giving $t_{tot} = 2.5$ ns) as a function of injection current density in the junction. The transparency current in this case sits around 760 A cm^{-2}.

quasi-Fermi levels given in (13.4) to penetrate further into the bands resulting in a broader and larger gain curve (13.20a) – see Fig. 13.9. Figure 13.10 shows the resulting gain curves for various injection current levels. As is already known from Section 7.6, and Fig. 7.7, the maximum gain γ_{max} (one of the most important points on these curves, as only those modes closest to γ_{max} will lase) increases nearly linearly with carrier density n above the transparency threshold n_{transp}. It is therefore customary to approximate the complex dependence of γ_{max} on n by a linear function:

$$\gamma_{max} = \gamma_0 \left(\frac{n}{n_{transp}} - 1 \right) \tag{13.21}$$

where γ_0 is obtained by a fitting procedure. For a density n between 0 and n_{transp}, the medium is absorbent. For $n \geq n_{transp}$, the medium becomes transparent and the dependence of γ_{max} on n is largely reproduced.

Equation (13.3) allows us to rewrite this expression in terms of the electrical injection current density J. The maximum gain γ_{max} is then given by:

$$\gamma_{max} = \gamma_0 \left(\frac{J}{J_{transp}} - 1 \right) \tag{13.22a}$$

where the transparency current is given by (13.18) and related to n_{transp} through:

$$J_{transp} = \frac{qd}{\eta_i t_{rad}} n_{transp} \tag{13.22b}$$

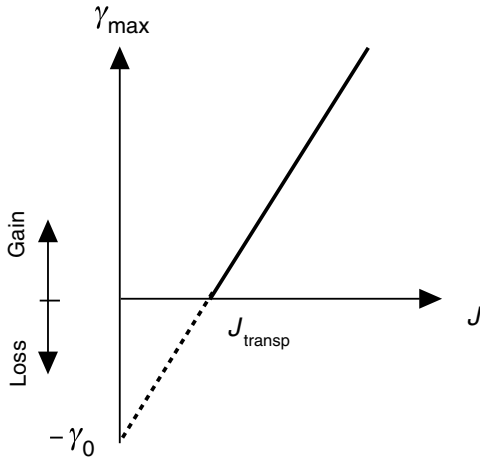


Fig. 13.11. Variation in the maximum gain for a semiconductor junction as a function of injection current density using the linear approximation model.

Example

We consider the amplification afforded by a double heterojunction GaAs/AlGaAs laser diode. Under experimental conditions, the radiative recombination time t_{rad} is found to be 5 ns. Figure 7.7 shows the transparency threshold carrier density n_{transp} to be $1.2 \times 10^{18}\,cm^{-3}$. We will further suppose for the diode: a quantum efficiency of 50%, a GaAs active-region thickness of d of 100 nm, a width of 10 μm, and a length of 300 μm. The transparency threshold current J_{transp} is then:

$$J_{transp} = 1.6 \times 10^{-19}\,C \times 10^{-5}\,cm \times 1.2 \times 10^{18}\,cm^{-3}/(0.5 \times 5 \times 10^{-9}\,s)$$
$$= 760\,A\,cm^{-2} \text{ or } 23\,mA \text{ in the diode}$$

For a carrier density n of $2.4 \times 10^{18}\,cm^{-3}$ (i.e. a level twice above threshold) and corresponding to an injection current of 46 mA, the maximum gain is $300\,cm^{-1}$. This leads to a gain G in the structure of $\exp(3 \times 10^{2}\,cm^{-1} \times 3 \times 10^{-2}\,cm)$ or 8100.

Semiconductor optical amplifiers (SOAs) are used as amplification stages at the outputs of laser diodes (referred to in this case as *monolithic optical amplifiers – MOPAs*). The challenge in this case is to keep the amplifiers from lasing. To this end, the exit windows are treated with antireflection coatings ($R < 10^{-6}$) to inhibit optical feedback. Furthermore, to maximize the energy that can be extracted from these devices, the amplifier cavity is *tapered*. This makes the cavity unstable, thereby impeding the formation of stationary optical modes at the amplification wavelength (see Fig. 13.12).

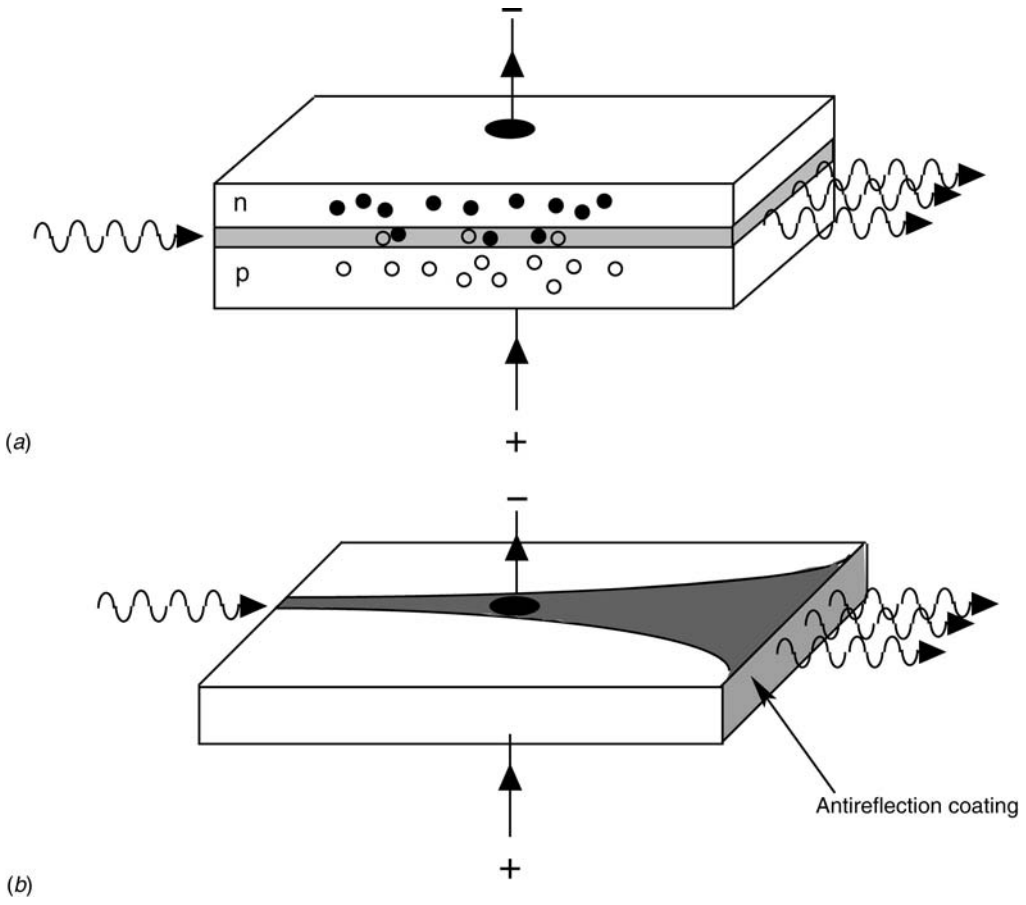


Fig. 13.12. Operation of a semiconductor optical amplifier. (*a*) The incident wave (provided, for instance, by the output of a laser diode) is amplified by the forward biased junction. One of the main challenges in designing an SOA is to produce a device which provides high gain without it lasing on its own. (*b*) Antireflection coatings and a tapered cavity geometry are used to impede the formation of stationary optical modes at the amplification wavelength.

13.5 Double heterojunction laser diodes

13.5.1 Laser threshold

A semiconductor *p–n* junction pumped electrically beyond its transparency threshold, will be able to amplify those electromagnetic modes (i.e. photons) that satisfy the Bernard–Durrafourg criterion. As described in Section 4.4, the medium will exhibit laser oscillations if it is subjected to *optical feedback*. Several methods exist that can provide the required feedback and are described in Complement 13.A. The simplest method involves using the natural dielectric mirror provided by

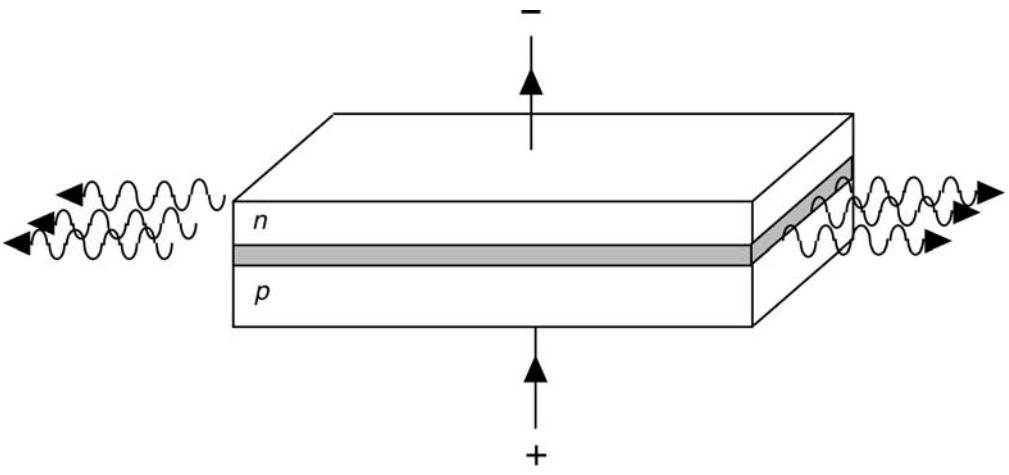


Fig. 13.13. Semiconductor injection laser with two cleaved mirrors.

the semiconductor–air interface. In this case, the semiconductor cavity is defined by cleaving the semiconductor crystal at two different locations along the same crystallographic plane resulting in perfectly parallel mirrors (see Fig. 13.13). The semiconductor–air interface yields a dielectric mirror with a reflectance R_m given by:

$$R_m = \frac{(n_{sc} - 1)^2}{(n_{sc} + 1)^2} \tag{13.23}$$

For GaAs, $n_{sc} = 3.6$ and the reflectance is $R_m = 0.32$. This value may seem low in comparison with mirrors used in other types of lasers (e.g. gas lasers, ion lasers, . . .). The gain made available by semiconductors (up to several thousand cm^{-1}), however, is considerable in comparison with the levels produced in alternate laser media. As a result, the deposition of high-reflectivity mirrors is not a requirement for observing laser oscillations in semiconductor devices. The physical origin of this large gain in semiconductors can be seen by comparing the expressions for gain in a dense medium $(\rho_j(h\nu)(f_c - f_v)\lambda^2/8\pi\tau_R)$ and in an atomic medium $(g(h\nu)(N_2 - N_1)\lambda^2/8\pi\tau_R)$:

- The density of available emission centres in semiconductors (some $10^{18}\,cm^{-3}$) greatly exceeds the levels available in gas media lasers ($\sim 10^{15}\,cm^{-3}$).
- The optical cross-sections are much larger for electrons in bands than for electrons in atomic orbitals. Therefore, at wavelengths of $\sim 1\,\mu m$, the radiative lifetime is typically 1 ms in ion or gas lasers, whereas it is only of the order of a few nanoseconds in semiconductors. This behaviour results from the delocalized nature of conduction electrons in condensed matter.

We saw in Section 4.4 that laser oscillations emerge when the gain in the amplifying medium tends to exceed the cavity losses in the optical resonator.

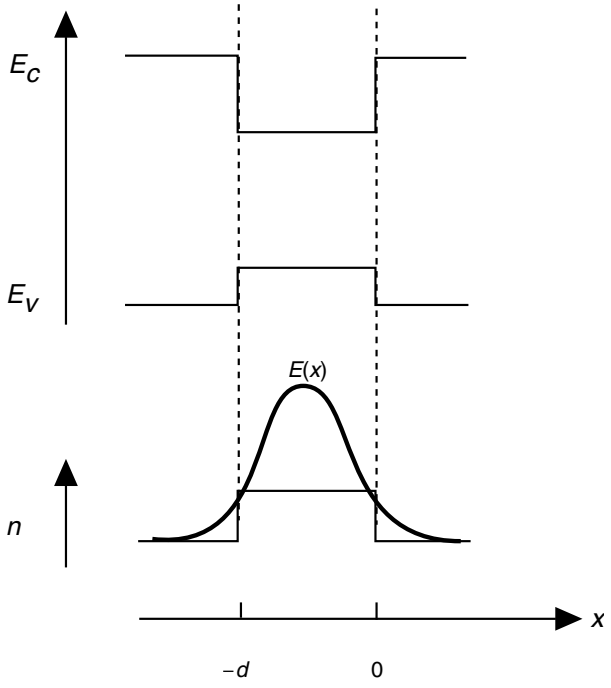


Fig. 13.14. Electron–hole recombination proceeds in the region with the smallest bandgap and hence the highest index of refraction. In addition to providing gain, the low bandgap semiconductor therefore also acts as an optical confinement region for the amplified optical modes.

There are two types of losses:

- Intrinsic losses α_m resulting from photon loss through the mirrors (necessary to some level as light must be able to escape the laser cavity!) – this was described in (4.21b) as:

$$\alpha_m = \frac{1}{2L} \ln \frac{1}{R_{m1} R_{m2}} \tag{13.24a}$$

  where we have allowed for the possibility of different reflectances (R_{m1} and R_{m2}) at each of the mirrors. By making one of the mirrors totally reflecting, the threshold for laser oscillation is reduced. L is the cavity length.
- Parasitic loss α_p resulting from free carrier absorption in the contact layers, roughness scattering, etc.

The condition for achieving laser threshold is then (see (4.21b)):

$$\gamma_{\text{threshold}}(h\nu) = \alpha_p + \frac{1}{2L} \ln \frac{1}{R_{m1} R_{m2}} \tag{13.24b}$$

This last expression does not take into account the fact that the amplified modes

(the *photons*) must also satisfy Maxwell's equations in the heterojunction. Correspondingly, this structure also acts as a *waveguide*. In a double heterostructure, the potential barriers (which have a larger bandgap than the quantum well material) possess a smaller index of refraction so that the electromagnetic waves remain trapped within the large index material (and, most importantly, within the gain region – see Fig. 13.14). We saw in Chapter 9, that the electromagnetic field is confined within the structure along the Ox axis normal to the interfaces and that the propagation coefficient β_z depends on the transverse mode indices (see, for example, (9.13)). The proportion of energy which is effectively present in the amplifying region is given by the confinement factor Γ (see (9.27)):

$$\Gamma = \frac{\int_{-d}^{0} |E(x)|^2 \mathrm{d}x}{\int_{-\infty}^{+\infty} |E(x)|^2 \mathrm{d}x} \tag{13.25}$$

Therefore, only a fraction Γ of the electromagnetic energy experiences amplification (see also Section 8.7.2) assuming both the barriers and wells are subject to the same losses. The condition for laser threshold is therefore modified, and becomes:

$$\gamma_{\text{threshold}}(h\nu) = \frac{1}{\Gamma}\left(\alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}}\right) = \frac{1}{\Gamma}(\alpha_p + \alpha_m) \tag{13.26a}$$

Threshold condition for a heterojunction laser

We will now focus our attention on the maximum gain given by (13.22a), this last equation can be put into a more explicit form in terms of the threshold current density $J_{\text{threshold}}$:

$$\gamma_0\left[\frac{\eta_i t_{\text{rad}}}{qdn_{\text{transp}}}J_{\text{threshold}} - 1\right] = \frac{1}{\Gamma}\left(\alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}}\right) \tag{13.26b}$$

As the confinement factor Γ decreases with the thickness of the double heterostructure d (see (9.30b)), we see that an optimum value for d exists. This situation is illustrated in Fig. 13.15 and contrasted against the corresponding variation in threshold current density for a homojunction laser diode. The large differences in the threshold current densities result on the one hand from the weak optical confinement in homojunction lasers and on the other from the reduced volume of material in the narrower heterostructures which needs to be inverted (see Fig. 13.2).

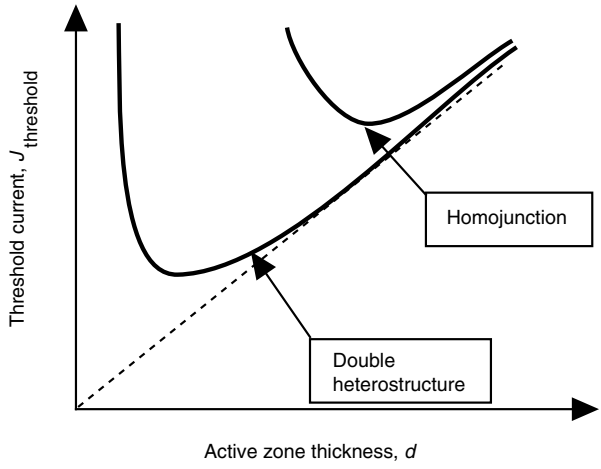


Fig. 13.15. Dependence of the threshold current density as a function of active region thickness d. Heterojunctions possess intrinsically lower threshold current densities due to improved overlap of the optical mode with the gain medium and increased carrier confinement.

Example

We consider a GaAs/$Al_xGa_{1-x}As$ heterojunction laser diode with a well thickness d of 100 nm. The radiative lifetime is 5 ns and the internal quantum efficiency at threshold is 50%. The diode is 500 μm long by 5 μm wide, and possesses a totally reflecting mirror. The Al fraction x in the barriers is sufficiently large to make the confinement factor $\Gamma = 1$. The loss per mirror α_m is then:

$$\alpha_m = \ln(1/0.32)/(2 \times 5 \times 10^{-2}\,\text{cm}) = 11.4\,\text{cm}^{-1}$$

We will assume the parasitic losses α_p to be $10\,\text{cm}^{-1}$, making $\alpha_{\text{tot}} = 21.4\,\text{cm}^{-1}$. The system will begin to lase once the maximum gain $\gamma_{\max}$ in the forward biased diode exceeds α_{tot}. Looking at Fig. 13.10, this corresponds to a threshold current density $J_{\text{threshold}}$ of $800\,\text{A cm}^{-2}$ (or given the surface area of the device, a threshold current $I_{\text{threshold}}$ of 20 mA).

The *photon lifetime* in a semiconductor cavity is given (see (4.23a,b)) by:

$$\tau_c = \frac{1}{\left(\alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}}\right)c'} \tag{13.27}$$

where $c' = c/n_{\text{op}}$ is the speed of light in the material. Assuming null parasitic losses, a typical device length L of 500 μm and mirror reflectances of 1 and 0.32, yields a photon lifetime of 5 ps. We therefore note that a consequence of weak optical feedback is a very short cavity lifetime for photons.

13.5.2 Output power

As long as the transparency threshold has not been reached, the diode behaves exactly as a LED, releasing spontaneous emission in all possible directions. Between the transparency threshold and the onset of laser oscillation, ($J_{\text{trans}} < J < J_{\text{threshold}}$), stimulated emission dominates over spontaneous emission. In this case radiative emission becomes directional as the gain $e^{\gamma z}$ promotes emission along directions close to the diode axis (the diode is referred to as being *superluminescent*). Once the maximum in the gain curve $\gamma_{\max}$ attains the threshold value $\gamma_{\text{threshold}}$, the electromagnetic modes (as we shall see later on) corresponding to this maximum will be amplified along the layer sandwiched between the double heterojunctions. Once this threshold has been surpassed, the carrier densities in the junction are *clamped* to their threshold values $n_{\text{threshold}}$ (see (4.22a,b)). In this case, condition (13.26) requiring equality between the gain and the cavity losses is satisfied. All additional carriers injected into the diode (in actuality, only a fraction of these as determined by the internal quantum efficiency η_i) will recombine immediately under the effect of stimulated emission. We will see, however, that as the current density rises above $J_{\text{threshold}}$ this quantum efficiency rapidly tends towards unity.

Experiment shows that, when driven beyond laser threshold, the semiconductor behaves as a homogeneous medium (see Section 4.6). The mode corresponding to the gain maximum sees its gain decrease until it exactly matches the cavity losses (see Fig. 4.6). The photon flux in the cavity can then be obtained by setting the saturated gain and the cavity losses equal to one another $\gamma(h\nu) = \gamma_0(h\nu)/(1 + \Phi/\Phi_{\text{sat}})$ (see (4.10)). This leads to (4.26a,b) or (4.28a,b). We will rederive this result, however, for the particular case of a semiconductor medium.

We return to the rate equations, (13.1), between the incident current in the junction J and the carrier density n, without this time forgetting to include the stimulated emission rate $R_{\text{st}}s$:

$$\frac{J}{qd} = A_{\text{nr}}n + Bn^2 + C_{\text{Aug}}n^3 + R_{\text{st}}s \tag{13.28}$$

where s is the volumetric density of the photons in the cavity. Above laser threshold, the gain in the medium is clamped to its threshold value $\gamma_{\text{threshold}}$, as are the carrier densities (i.e. $n = n_{\text{threshold}}$). Equation (13.28) can then be written:

$$\frac{J}{qd} = \frac{n_{\text{threshold}}}{\tau_{\text{threshold}}} + R_{\text{st}}s \tag{13.29a}$$

where $\tau_{\text{threshold}}$ is given by:

$$\tau_{\text{threshold}} = \frac{1}{A_{\text{nr}} + Bn_{\text{threshold}} + C_{\text{Aug}}n_{\text{threshold}}^2} \tag{13.29b}$$

We then introduce the threshold current density $J_{\text{threshold}}$:

$$J_{\text{threshold}} = \frac{qd}{\tau_{\text{threshold}}} n_{\text{threshold}} \tag{13.30}$$

which, once substituted into (13.29), leads to $R_{\text{st}}s = (J - J_{\text{threshold}})/qd$. To find the stimulated emission rate R_{st}, we recall that the population variation rate $\mathrm{d}n/\mathrm{d}t$ due to competition between absorption and stimulated emission is given by $\mathrm{d}n/\mathrm{d}t = c'\gamma s = R_{\text{st}}s$ (see (3.72a)). Once the gain becomes clamped, the stimulated emission rate becomes $R_{\text{st}} = \gamma_{\text{threshold}}c'$. Substituting this last expression into (13.29), we obtain the photon density in the cavity:

$$s = \frac{1}{qd\gamma_{\text{threshold}}c'}(J - J_{\text{threshold}}) \tag{13.31a}$$

This last expression indicates that all the current above threshold is consumed by stimulated emission. Complex mechanisms may lie at the origin of current leakage via non-radiative channels such as surface recombination (see Chapter 10), ballistic carriers which travel across the heterostructure without recombining, etc. To take these mechanisms into account, we introduce the *internal quantum efficiency* η_i, allowing us to rewrite (13.31a) as:

$$s = \eta_i \frac{1}{qd\gamma_{\text{threshold}}c'}(J - J_{\text{threshold}}) \tag{13.31b}$$

The laser output power P_{out} is then given either by (4.29a) or (4.30a). We recover this result by writing the output power as a product of: photon density in the cavity s, energy carried per photon hv, effective mode volume Lwd/Γ, and escape rate of photons from the cavity $c'\alpha_m$. Using this chapter's notation, the output power may be written:

$$\begin{aligned} P_{\text{out}} &= \left(\begin{matrix}\text{photon}\\ \text{energy}\end{matrix}\right)\left(\begin{matrix}\text{photon}\\ \text{density}\end{matrix}\right)\left(\begin{matrix}\text{effective mode}\\ \text{volume}\end{matrix}\right)\left(\begin{matrix}\text{photon escape}\\ \text{rate}\end{matrix}\right) \\ &= (hv) \quad (s) \quad \left(Lw\frac{d}{\Gamma}\right) \quad (c'\alpha_m) \end{aligned} \tag{13.32}$$

Given (13.26) and (13.31), we finally obtain:

$$P_{\text{out}} = \eta_i \frac{\alpha_m}{\alpha_m + \alpha_p} \frac{hv}{q}(I - I_{\text{threshold}}) \tag{13.33}$$

where I is the current $I = JLw$ so that:

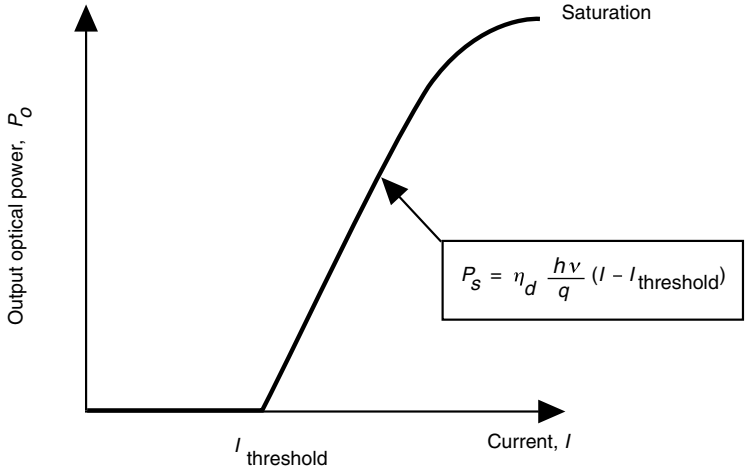


Fig. 13.16. *P–I* characteristics for a laser diode. The phenomenon of saturation is explained in Complement 13.F.

$$P_{\text{out}} = \eta_{\text{ext}} \frac{h\nu}{q} (I - I_{\text{threshold}}) \tag{13.34}$$

Output power from a laser diode as a function of pump current

In this last expression, the external quantum efficiency η_{ext} is given by:

$$\eta_{\text{ext}} = \eta_i \frac{\alpha_m}{\alpha_m + \alpha_p} = \eta_i \frac{\ln(1/R_m)}{\alpha_p L + \ln(1/R_m)} \tag{13.35}$$

We will now introduce a very important concept for semiconductor laser diodes: the *differential external quantum efficiency* η_d. This is the variation in the external photon flux $\mathrm{d}P_{\text{out}}/h\nu$ due to a variation in the diode current $\mathrm{d}J/q$:

$$\eta_d = \frac{\mathrm{d}P_{\text{out}}/h\nu}{\mathrm{d}J/q} \tag{13.36}$$

or, taking (13.34) and (13.35) into account:

$$\eta_d = \eta_i \frac{1}{1 + \alpha_p L/\ln(1/R_m)} \tag{13.37a}$$

In this simple approach, the quantum efficiencies η_{ext} and η_d are equivalent as $P_{\text{out}}(J)$ is linear. Clearly, this will not hold in general. It is common practice, in characterizing semiconductor lasers, to plot η_d^{-1} as a function of laser cavity length L (this is achieved by cleaving lasers of different length and measuring them individually). Equation (13.37a) then allows one to extract the internal quantum efficiency η_i and parasitic losses α_p for the specific devices (see Fig. 13.17).

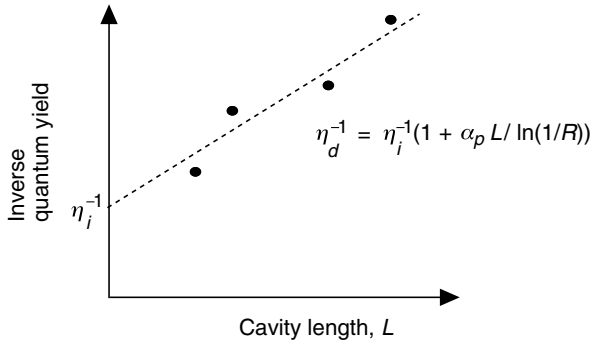


Fig. 13.17. Inverse differential quantum efficiency as a function of laser cavity length L.

Equation (13.34) determines the optical output power of the laser. It is interesting to compare this output power to the electrical power injected into the diode. Assuming the lack of any parasitic resistance, the voltage required to achieve the necessary diode injection current is of the order of the semiconductor bandgap E_g/q or $V \approx h\nu/q$. The electrical power is then $P_{\text{el}} = IV = Ih\nu/q$, so that:

$$\frac{P_{\text{out}}}{P_{\text{el}}} \approx \eta_{\text{ext}}\left(1 - \frac{I_{\text{threshold}}}{I}\right) \tag{13.37b}$$

This last ratio gives the '*wall-plug*' *efficiency*. As long as the structure possesses a good quantum efficiency and is pumped well above threshold, semiconductor heterojunction lasers are characterized by excellent electrical to optical power conversion efficiencies (especially when compared with gas lasers which have efficiencies usually less than 1%).

Example

Returning to the example in Section 13.5.1, we consider a $GaAs/Al_xGa_{1-x}As$ laser diode (with $d = 100\,\text{nm}$, $\Gamma = 1$, $t_{\text{rad}} = 5\,\text{ns}$, $\eta_i = 0.5$, $L = 500\,\mu\text{m}$, but this time possessing two mirrors with $R_m = 0.32$). We find that the losses due to the mirrors are $\alpha_m = \ln(1/0.32)/(5 \times 10^{-2}\,\text{cm}) = 22.8\,\text{cm}^{-1}$.

Assuming the parasitic losses α_p to be $10\,\text{cm}^{-1}$, the total losses α_{tot} then become $32.8\,\text{cm}^{-1}$. This results in a differential quantum efficiency at threshold η_e of $0.5 \times 22.8/32.8$ or 35%.

13.6 Quantum well laser diodes

13.6.1 Optical amplification in a quantum well structure: general case

We saw in the preceding section that the threshold current is to a large part proportional to the thickness of the active region in the laser diode (i.e. the width of

the potential well formed by the double heterojunction). To decrease the threshold current further, one is naturally drawn to decrease the size of the active region down to quantum length scales (i.e. to make use of potential wells that are so narrow as to confine and quantize the motions of carriers in directions perpendicular to the heterointerfaces). We spent some time exploring the physics of such *quantum wells* in Chapter 8. We recall here the important results from this chapter.

The electron and hole wavefunctions in quantum wells are given by the product of three terms: the periodic portion of the Bloch wavefunctions, which result from the fact that the carriers belong to the extrema of a particular semiconductor band ($u_c(\mathbf{r})$ for electrons, $u_v(\mathbf{r})$ for holes with $\mathbf{k} = 0$ for a model involving a single valence band with extrema at Γ); a portion that describes the free motion of carriers parallel to the well interfaces ($e^{i\mathbf{k}_{||}\mathbf{r}_{||}}$, where $\mathbf{r}_{||}$ and $\mathbf{k}_{||}$ are, respectively, the position and wavevectors of the wave in the plane of the quantum well); and a portion that describes the quantization of motion perpendicular to the heterointerfaces. Therefore, the wavefunctions for electron states in a subband n and heavy hole states in a subband m may be written as:

$$\begin{aligned} |e_n(\mathbf{k}_{||})\rangle &= u_c(\mathbf{r})e^{i\mathbf{k}_{||}\mathbf{r}_{||}}e_n(z) \\ |hh_m(\mathbf{k}_{||})\rangle &= u_v(\mathbf{r})e^{i\mathbf{k}_{||}\mathbf{r}_{||}}hh_m(z) \end{aligned} \tag{13.38}$$

The envelope functions $e_n(z)$ and $hh_m(z)$, are one-dimensional states with energies e_n and hh_m, and are solutions to the one-dimensional Schrödinger equations:

$$\begin{aligned} H_c|e_n\rangle &= \left[\frac{p_z^2}{2m_c} + V_e(z)\right]|e_n\rangle = e_n|e_n\rangle \\ H_v|hh_m\rangle &= \left[\frac{p_z^2}{2m_{hh}} + V_{hh}(z)\right]|hh_m\rangle = hh_m|hh_m\rangle \end{aligned} \tag{13.39}$$

where $V_e(z)$ and $V_{hh}(z)$ are the electron and heavy hole potentials (see Fig. 13.18). We recall that increasing energies in such diagrams are 'upwards' for electrons and 'downwards' for holes. All these concepts can be extrapolated to the light-hole and spin-orbit bands, but add considerable complexity to the required notation (as a result, we shall not treat these subbands in this chapter). Equations (13.38) and (13.39) show that the electrons and holes are distributed over energy subbands (additionally, these subbands run parallel to each other for a given carrier type). The extrema of these subbands (e_n and hh_m) are obtained from the Schrödinger equations, (13.39), and are:

$$\begin{aligned} e_n(\mathbf{k}_{||}) &= e_n + \frac{\hbar^2 k_{||}^2}{2m_c} \\ hh_m(\mathbf{k}_{||}) &= hh_m + \frac{\hbar^2 k_{||}^2}{2m_{hh}} \end{aligned} \tag{13.40}$$

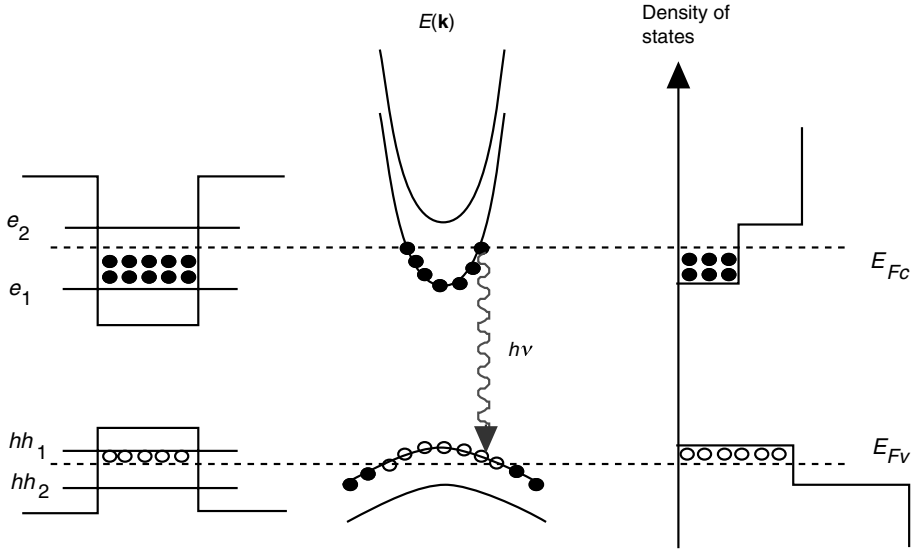


Fig. 13.18. Band structure, subbands $E(\mathbf{k}_{\|})$, and state densities in a quantum well.

(see Fig. 13.18). The density of states in each subband are constant, implying that the state densities for electrons and heavy holes are (see (8.41a,b)):

$$\rho_{2D,e}(E) = \rho_{2D,e} \sum_{n=1}^{\infty} \theta(E - e_n) = \frac{m_c}{\pi\hbar^2} \sum_{n=1}^{\infty} \theta(E - e_n)$$
$$\rho_{2D,hh}(E) = \rho_{2D,hh} \sum_{m=1}^{\infty} \theta(E - hh_m) = \frac{m_{hh}}{\pi\hbar^2} \sum_{m=1}^{\infty} \theta(E - hh_m) \tag{13.41}$$

where θ is the Heaviside function.

The optical *interband* transitions bridge states with identical $\mathbf{k}_{\|}$. A photon with energy $h\nu$ therefore couples states in the conduction subband n with states in the valence subband m related by:

$$h\nu = E_g + e_n + hh_m + \frac{\hbar^2 k_{\|}^2}{2m_r} \tag{13.42}$$

where m_r is the reduced effective mass given by $1/m_r = 1/m_c + 1/m_{hh}$ (see Fig. 13.18). The optical gain in this system for an electromagnetic wave with frequency ν was calculated in Section 8.7. It was obtained by summing the contributions due to optical transitions between the valence subbands m and conduction subbands n having a wavevector k between k and $k + \mathrm{d}k$. These result in a carrier inversion density of:

$$\mathrm{d}(N_c - N_v)^{nm} = \frac{\rho_{2D,e}(k)}{d} \{f_c^n[E_c(k)] - f_v^m[E_v(k)]\} \mathrm{d}k \tag{13.43}$$

The gain $\gamma(h\nu)$ can be obtained by integrating over k the contribution made by (13.43) to the optical susceptibility and by summing over all involved subbands. This result appears in (8.86a), but is rewritten here in a manner better suited to the present discussion:

$$\gamma(h\nu) = \alpha_{2D} \sum_{n,m=1}^{\infty} I_{n,m}[f_c^n(h\nu) - f_v^m(h\nu)]\theta(h\nu - E_g - e_n - hh_m) \tag{13.44}$$

where $I_{n,m}$ are the overlap integrals $|\langle n|m\rangle|^2$ and f_c^n and f_v^m are the Fermi distributions in subbands n and m:

$$f_c^n(h\nu) = \frac{1}{1 + \exp\{[E_c^n(h\nu) - E_{Fc}]/kT\}}$$
$$E_c^n(h\nu) = E_g + e_n + \frac{m_r}{m_c}(h\nu - E_g - e_n) \tag{13.45a}$$

and

$$f_v^m(h\nu) = \frac{1}{1 + \exp\{[E_v^m(h\nu) - E_{Fv}]/kT\}}$$
$$E_v^m(h\nu) = -\frac{m_r}{m_v}(h\nu - E_g - hh_m) \tag{13.45b}$$

The absorption coefficient α_{2D} (cm^{-1}) is the absorption for a quantum well with zero population (i.e. under null current conditions) and is given by:

$$\alpha_{2D} = \frac{2\pi q^2 x_{vc}^2 m_r}{\lambda_0 n_{sc} \varepsilon_0 \hbar^2 d} \tag{13.46a}$$

where λ_0 is the vacuum emission wavelength. As discussed at length in Section 8.7.2, (13.46a) represents the attenuation coefficient for light crossing a quantum layer of thickness d at normal incidence. The coefficient α_{2D} is therefore expressed in terms of cm^{-1}. This concept of an attenuation length in a quantum well is not particularly satisfying. As quantum well thicknesses are very thin in comparison to the attenuation length $1/\alpha_{2D}$, the attenuation given by $d\alpha_{2D}$ does not depend on the thickness of the quantum well (rigorously this only holds if there is a single confined level in the well). We therefore make use of the attenuation rate per well (a dimensionless quantity) $A_{2D} = d\alpha_{2D}$ given by (8.86b):

$$A_{2D} = \frac{2\pi q^2 x_{vc}^2 m_r}{\lambda_0 n_{sc} \varepsilon_0 \hbar^2} \tag{13.46b}$$

Attenuation coefficient for one quantum well

This last formula involves the manipulation of impressively large quantities. In

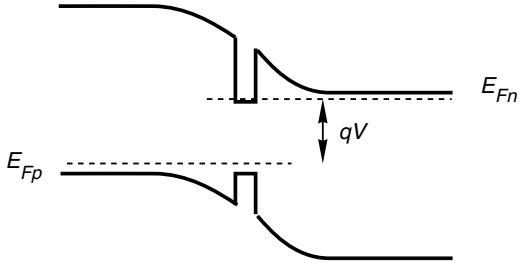


Fig. 13.19. A forward biased *p–n* diode containing a single quantum well.

order to express A_{2D} in terms of parameters with similar magnitudes, we introduce the two following energies: an interband energy $E_{vc} = \hbar^2 k_{vc}^2 / 2m_r$, where k_{vc} is the wavevector $k_{vc} = 2\pi / x_{vc}$, and an electrostatic energy $E_{\text{stat}} = q^2 / 4\pi\varepsilon_0\lambda_0$. The attenuation rate A_{2D} is then:

$$A_{2D} = \frac{16\pi^4}{n_{sc}} \frac{E_{\text{stat}}}{E_{vc}} \tag{13.47}$$

This last equation allows one to estimate the attenuation rates in quantum wells rapidly by drawing upon quantities with similar magnitudes. The following example illustrates the convenience afforded by this expression.

Example
We will calculate the attenuation rate for a GaAs quantum well. We recall the relevant parameters for GaAs:

$m_c = 0.067 m_e$; $m_v = 0.46 m_e$; $m_r = 0.058 m_e$
$x_{vc} = 6.14\,\text{Å}/\sqrt{2} = 4.3\,\text{Å}$ (see Tables 7.1 and 8.1)
$n_{sc} = 3.3$
$\lambda_0 = 0.8\,\mu\text{m}$

Equation (13.46b) yields an attenuation rate of:

$$A_{2D} = \frac{2\pi(1.6 \times 10^{-19}\,\text{C})^2 \times (4.3 \times 10^{-10}\,\text{m})^2 \times 0.058 \times 0.9 \times 10^{-30}\,\text{kg}}{0.8 \times 10^{-6}\,\text{m} \times 3.3 \times 8.85 \times 10^{-12}\,\text{Fd m}^{-1} \times (1.05 \times 10^{-34}\,\text{J s}^2)}$$
$$\approx 0.6\%$$

This calculation is cumbersome to say the least! What this really means is that MKSA units are not ideally suited to such a problem. This then is the motivation for rewriting (13.47) in terms of more easily manipulated quantities. Evaluating $E_{vc} = 3.99\,\text{eV}$ and $E_{\text{stat}} = 1.8\,\text{meV}$, we readily obtain the same result.

13.6.2 Transparency threshold

We consider a *p–n* diode containing a single quantum well (see Fig. 13.19). By forward biasing this diode, injected carriers accumulate in the well. The carrier

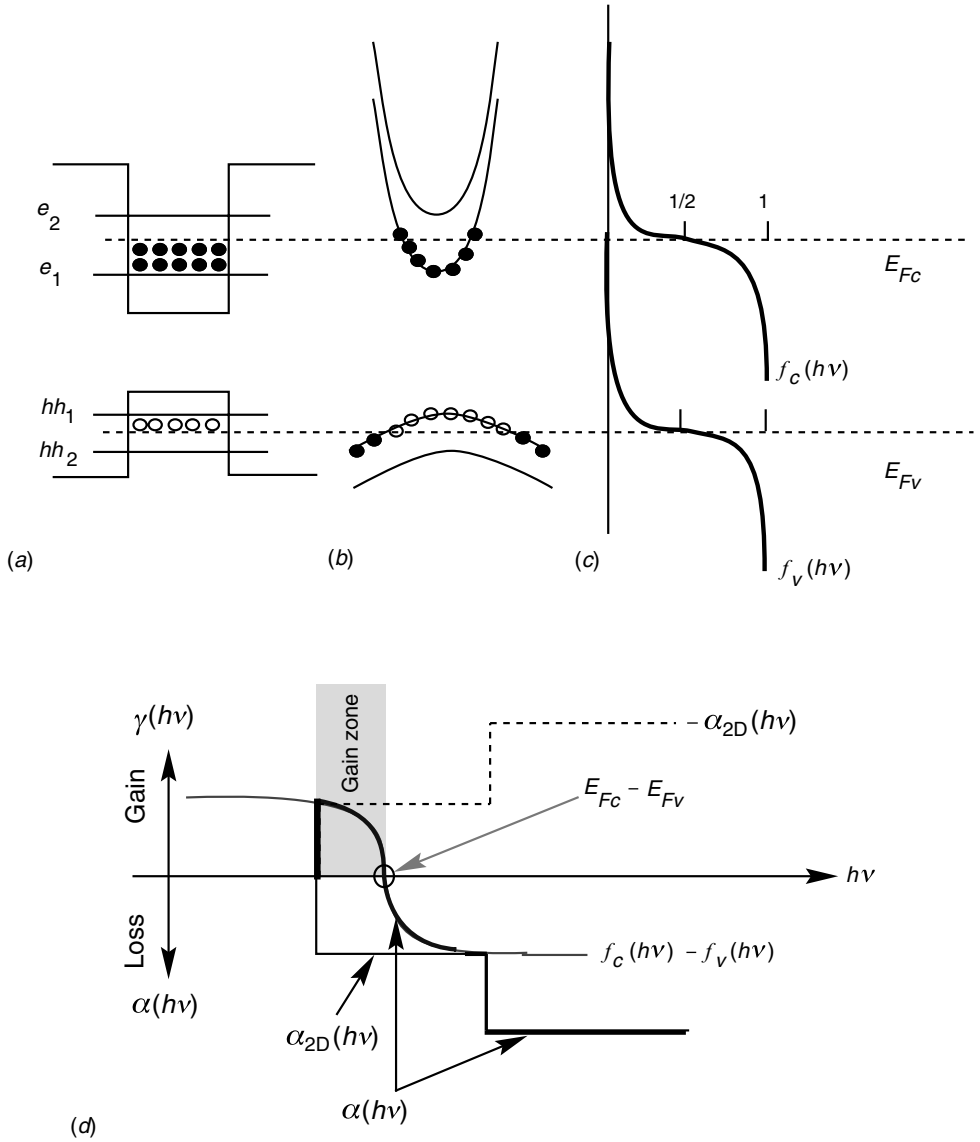


Fig. 13.20. As the quantum wells are filled with electron–hole pairs, the quasi-Fermi levels climb within the subbands. Once the energy separation between the two quasi-Fermi levels exceeds the bandgap, photons which satisfy the Bernard–Durrafourg condition are amplified: (*a*) band diagram, (*b*) subband structure, (*c*) associated Fermi–Dirac distributions, and (*d*) gain curve for the quantum well.

density per unit area n_s can be obtained in the same manner as in (13.3), i.e:

$$n_s = p_s = \frac{Jt_{\text{tot}}}{q} \tag{13.48}$$

With a sufficiently elevated current, we obtain a population inversion leading to

transparency in the quantum well (see Fig. 13.20). Close to transparency, the quasi-Fermi levels begin to penetrate the subbands. Only the $n = m = 1$ subbands are involved, and the gain may be written simply as:

$$\gamma(h\nu) = \alpha_{2D}[f_c^1(h\nu) - f_v^1(h\nu)]\theta(h\nu - E_g - e_1 - hh_1) \tag{13.49}$$

Optical gain due to the first e_1–hh_1 transition

Clearly, the spectral gain must be convoluted with a Lorentzian to take into account broadening mechanisms. The quasi-Fermi levels E_{Fc} and E_{Fv} are given by the conditions:

$$\begin{aligned} n_s &= \int_{-\infty}^{\infty} \rho_{2D,e}(E) f_c^1(E) \mathrm{d}E \\ n_s &= \int_{-\infty}^{\infty} \rho_{2D,hh}(E)[1 - f_v^1(E)] \mathrm{d}E \end{aligned} \tag{13.50}$$

These conditions can be calculated exactly by recalling that $\rho_{2D,e}$ and $\rho_{2D,hh}$ are constants given by (13.41). Therefore, for electrons, (13.50) is written:

$$n_s = \rho_{2D,e} \int_{E_g+e_1}^{+\infty} \frac{1}{1 + \exp[(E - E_{Fc})/kT]} \mathrm{d}E \tag{13.51}$$

where we have taken the energy at the top of the valence band to be zero. Setting $\exp((E - E_{Fc})/kT) = u$ and $\exp((E_g + e_1 - E_{Fc})/kT) = u_c$, this last expression takes the form:

$$n_s = \rho_{2D,e} kT \int_{u_c}^{+\infty} \frac{1}{u(1+u)} \mathrm{d}u = \rho_{2D,e} kT \ln\left(1 + \frac{1}{u_c}\right) \tag{13.52}$$

so that:

$$\begin{aligned} n_s &= n_c \ln\left[1 + \exp\left(\frac{E_{Fc} - E_g - e_1}{kT}\right)\right] \\ n_s &= n_v \ln\left[1 + \exp\left(\frac{hh_1 - E_{Fv}}{kT}\right)\right] \end{aligned} \tag{13.53}$$

where n_c and n_v are the *two-dimensional critical densities* given by:

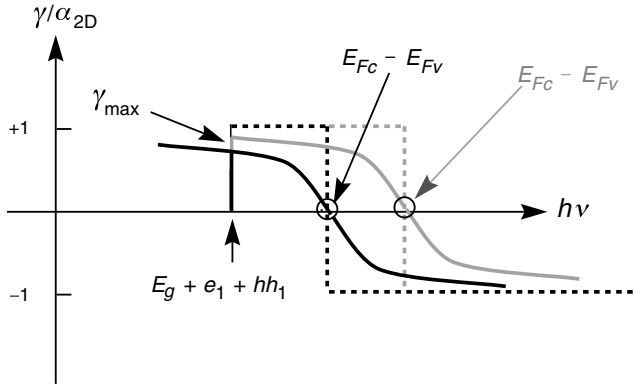


Fig. 13.21. Evolution of the gain curve for a quantum well laser for increasing pump currents. The dotted curves show the gain values at zero temperature. The black (grey) curves correspond to low (high) pump current conditions.

$$n_c = \rho_{2D,e} kT = \frac{m_e kT}{\pi \hbar^2}$$
$$n_v = \rho_{2D,hh} kT = \frac{m_{hh} kT}{\pi \hbar^2} \tag{13.54}$$

We will see that the transparency and threshold densities can be expressed as a product of these two-dimensional critical densities by a factor close to 1, typically between 1 and 5.

The method for calculating the gain is as follows: given a current density J, we find the carrier density given by (13.48), then the quasi-Fermi levels according to (13.53), and finally the gain with the help of (13.49) and the Fermi functions (13.45). Figure 13.21 shows the gain curve (13.48) as a function of photon energy for increasing carrier densities. We note an abrupt increase in the gain for $h\nu > E_g + e_1 + hh_1$. This is far more abrupt than is the case with a regular heterojunction and results from the staircase profile of the density of states in a two-dimensional structure (as opposed to $\sqrt{E}$ occurring in a three-dimensional structure). Therefore, the maximum gain $\gamma_{\max}$ is obtained when $h\nu = E_g + e_1 + hh_1$, i.e:

$$\gamma_{\max} = \alpha_{2D}[f_c^1(h\nu = E_g + e_1 + hh_1) - f_v^1(h\nu = E_g + e_1 + hh_1)] \tag{13.55}$$

Equations (13.53) and (13.45) allow one simply to relate the value of the Fermi function to the carrier density n_s:

$$1 - \mathrm{e}^{-n_s/n_c} = 1 - \frac{1}{1 + \exp[(E_{Fc} - E_g - e_1)/kT]} = f_c^1(E_g + e_1) \tag{13.56a}$$

Similarly:

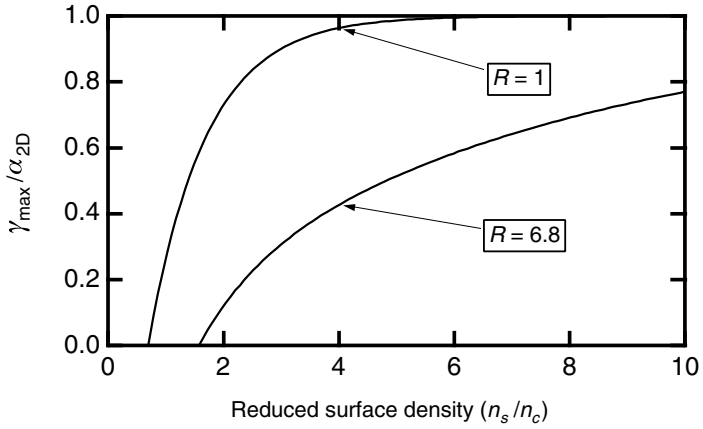


Fig. 13.22. Normalized gain (relative to the absorption α_{2D}) as a function of normalized carrier surface density (relative to the two-dimensional critical carrier density n_c).

$$e^{-p_s/n_v} = f_v^1(hh_1) \tag{13.56b}$$

The maximum gain (13.55) can then be written:

$$\gamma_{max} = \alpha_{2D}(1 - e^{-n_s/n_c} - e^{-n_s/R_{cv}n_c}) \tag{13.57}$$

Modal gain of a quantum well as a function of carrier density (cm^{-1})

where $R_{cv}(=m_{hh}/m_c)$ is the ratio of the effective masses for carriers in the conduction and valence bands. Figure 13.22 shows the variation in maximum gain as a function of reduced carrier surface density n_s/n_c with $R = 6.8$ (for GaAs) and $R = 1$. We see that the gain increases rapidly once the transparency condition has been reached, but saturates quickly. This results from the form of the two-dimensional density of states. From (13.57) the transparency threshold is reached once the maximum gain becomes positive, i.e. when the transparency threshold n_{tr} is reached:

$$e^{-n_{tr}/n_c} + e^{-n_{tr}/R_{cv}n_c} = 1 \tag{13.58}$$

For $R_{cv} = 1$, we have $n_{tr} = n_c\ln(2)$. The transparency current is always related to n_c by a numerical factor close to 1. This explains the importance of the concept of the *two-dimensional critical density* n_c. Figure 13.22 shows the transparency condition for different values of R_{cv}. We note that it is advantageous to have closely matched effective masses between the valence and conduction bands. This is in fact the motivation for growing strained laser structures (we will explore these structures in Complement 13.B).

It is worth noting that in the literature, the variation in maximum gain γ_{max} as a function of carrier surface density is often taken to be logarithmic:

$$\gamma_{\max} = \gamma_0 \ln\left(\frac{n}{n_{\mathrm{tr}}}\right) \tag{13.59}$$

where the constant γ_0 depends only on the effective mass ratio R_{cv} and is obtained by a fitting procedure (see below). This last formula leads to behaviours close to those predicted by (13.57) for values of n which approach the transparency threshold. This equation holds the advantage that it introduces the transparency threshold in a simple fashion into the gain formula. Note the difference from the similar, but linear, relation (13.21) for three-dimensional laser diodes.

Example

1. We will employ the following MATHEMATICA program to verify the equivalence of (13.57) and (13.59):

```
f=1-Exp[-x]-Exp[-x/R]
R=6.8
plot1=Plot[f,{x,.5,5}]
FindRoot[f==0,{x,1}]
x0=x/.%
g=Log[x/x0]
plot2=Plot[0.48*g,{x,.5,5}]
Show[plot1,plot2]
```

   Figure 13.23 compares the two expressions. By fitting, we additionally find that $\gamma_0 = 0.48\alpha_{2D}$.

2. For GaAs, the two-dimensional state densities are:

$$\rho_{2D,e} = m_c/\pi\hbar^2 = 2.8 \times 10^{13}\,\mathrm{cm}^{-2}\,\mathrm{eV}^{-1}$$
$$\rho_{2D,hh} = m_{hh}/\pi\hbar^2 = 1.9 \times 10^{14}\,\mathrm{cm}^{-2}\,\mathrm{eV}^{-1}$$

   The two-dimensional critical density in the conduction band is $n_c = \rho_{2D,e}kT$ or $2.8 \times 10^{13}\,\mathrm{cm}^{-2}\,\mathrm{eV}^{-1} \times 0.0259\,\mathrm{eV} = 7.25 \times 10^{11}\,\mathrm{cm}^{-2}$. Figure 13.22 shows that, in GaAs, $n_{\mathrm{tr}} = 1.6n_c = 1.16 \times 10^{12}\,\mathrm{cm}^{-2}$. In a 100 Å wide quantum well, this corresponds to a transparency threshold density of $1.16 \times 10^{18}\,\mathrm{cm}^{-3}$. This result is very close to that obtained for bulk material. In fact, the advantage of using quantum wells does not involve decreasing the threshold carrier densities, but rather in decreasing the transparency current densities and hence the threshold current densities.

The gain curve for a quantum well laser is in fact very complex. As the carrier densities increase in the wells, the electrons and holes populate higher energy states in the subbands and bring into play complex transitions: first the e_1–hh_1 transitions, then e_2–hh_2, etc. Figure 13.24 shows the results for a calculation which takes all of these different transitions into account.

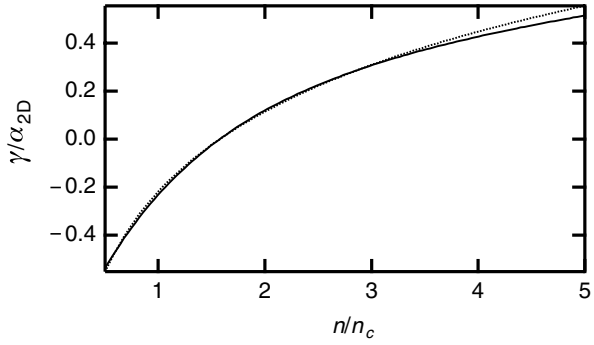


Fig. 13.23. Comparison between values for γ_{max} derived using the exact formula (13.57) (solid curve) and the approximation (13.59) (dotted curve).

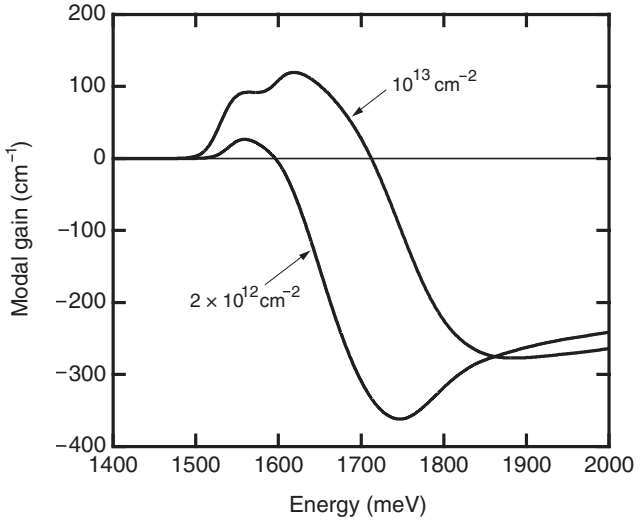


Fig. 13.24. Modal gain (i.e. the product of the gain by the confinement factor Γ) for a quantum well laser at two different carrier densities. Both the e_1–hh_1 and e_2–hh_2 transitions can be observed under the higher current injection conditions. (Courtesy of A. Fily@THALES.)

13.6.3 Laser threshold for a quantum well laser

Figure 13.25 shows the configuration for a quantum well laser with separate confinement layers for both carriers and photons (referred to as a *separate confinement heterostructure laser or SCH laser*). The $Al_xGa_{1-x}As/GaAs$ quantum well has been introduced into a lower index cavity (also composed of $Al_yGa_{1-y}As$ but with $y > x$) to confine the photons generated by the quantum well. The electromagnetic field is amplified along the quantum well layer. As described in Section 13.5.1 (see (13.26)), weak overlap between the guided wave and the

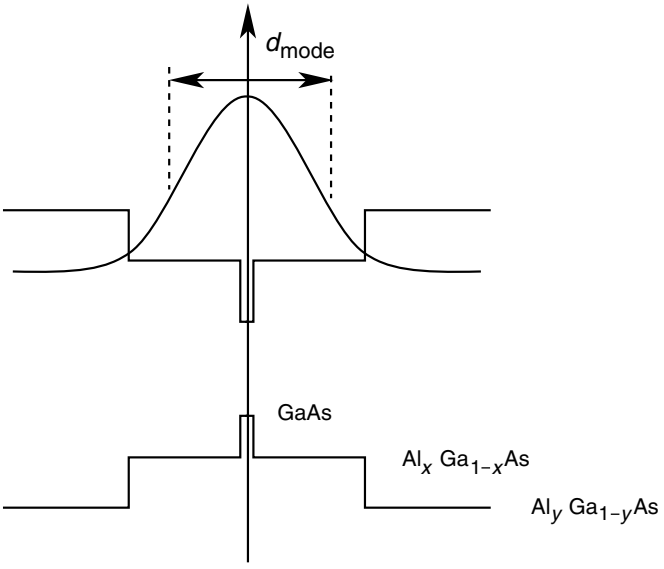


Fig. 13.25. Optical mode, composition, and band profiles for a separate confinement heterostructure quantum well laser.

quantum well leads to a reduction in the gain from γ to $\gamma\Gamma$, where Γ is the optical confinement factor given by (13.25). Given (13.59), the condition for achieving laser oscillations in a quantum well laser may be written:

$$\Gamma\gamma_0 \ln\left(\frac{n_{\text{threshold}}}{n_{\text{tr}}}\right) = \alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}} \tag{13.60}$$

where, in accordance with the usual notation, α_p is the parasitic absorption, L is the cavity length, and R_{m1} and R_{m2} are the mirror reflectances. The confinement factor is given by the ratio $\Gamma = d/d_{\text{mode}}$, where d_{mode} is a measure of the space occupied by the optical mode (see Fig. 13.25 and Section 9.4). The threshold current density $J_{\text{threshold}}$ is then given by (13.38). Assuming that the time t_{tot} depends only slightly on n (which is in fact a fairly crude approximation), the characteristic gain–current is then given by:

$$\gamma_{\max}(J) = \Gamma\gamma_0 \ln\left(\frac{J}{J_{\text{tr}}}\right) \tag{13.61a}$$

Logarithmic approximation for optical gain in a quantum well (cm^{-1})

where clearly the *transparency current density* J_{tr} (A cm^{-2}) is described by:

$$J_{\text{tr}} = q\frac{n_{\text{tr}}}{t_{\text{tot}}} \tag{13.61b}$$

Example

We will make use of the same physical system employed in the two previous sections, i.e. a 100 Å thick GaAs quantum well. The cavity is 500 μm in length and possesses a mirror with $R = 100\%$ and another with $R = 32\%$. The parasitic losses are $10\,\text{cm}^{-1}$. The threshold gain is therefore given by $\alpha_p - 1/2L\ln(R_s) = 21.4\,\text{cm}^{-1}$. The width of the mode d_{mode} is 5000 Å. The effective gain $\Gamma\alpha_{2D} = A_{2D}/d_{\text{mode}}$ is $0.55 \times 10^{-2}/5000$ Å or $110\,\text{cm}^{-1}$. To obtain the laser threshold current, we need to find the ratio for n/n_c such that $1 - \exp(-n/n_c) - \exp(-n/n_cR) = 21.4/110$, or $n = 1.3 \times 10^{12}\,\text{cm}^{-2}$ ($R = 6.8$ and $n_c = 7.25 \times 10^{11}\,\text{cm}^{-2}$). Assuming a lifetime of 5 ns and a surface area of 5×500 μm, we find a threshold current of 1 mA, i.e. a factor of 10 less than obtained for the heterostructure laser featured in Section 13.5.1.

13.6.4 Scaling rules for multi-quantum well lasers

We now compare the behaviour of heterojunction and single quantum well lasers. As we noted in previous sections, the required carrier densities at the transparency and laser thresholds are almost identical whether one considers a bulk semiconductor laser or a quantum well laser. On the other hand, the required current densities decrease in proportion to d, the width of the amplification region. Also, we noticed that the gain rises more rapidly as a function of drive current in a quantum well system owing to the two-dimensional density of states (see Figs. 13.10 and 13.22). On the down-side, the optical gain saturates more rapidly as a function of pump current for single quantum well lasers (at least until other quantum well transitions such as e_2–hh_2 come into play); while in bulk semiconductor heterojunction devices, the gain simply continues to rise. These characteristics are summarized in Fig. 13.26.

We are now in a position to ask what the optimal number of quantum wells N_{QW} would be to decrease the threshold current density to a minimum. Clearly, as the number N_{QW} increases (corresponding to the inclusion of multiple quantum wells in what is referred to as a *multi-quantum well* structure), the larger the gain and the losses occurring in the laser medium will be compensated for with greater ease. On the other hand, if the number of quantum wells is too large, the required threshold current density will increase in proportion to N_{QW}. There is therefore some optimal number of quantum wells which results from these opposing trends.

To determine this optimum number, we define the following quantities:

- G_1, the maximum gain for a single quantum well laser:

$$G_1 = \Gamma\gamma_0 \ln\left(\frac{J_1}{J_{\text{tr},1}}\right) \tag{13.62}$$

where $J_{\text{tr},1}$ is the transparency current density for a single quantum well and J_1

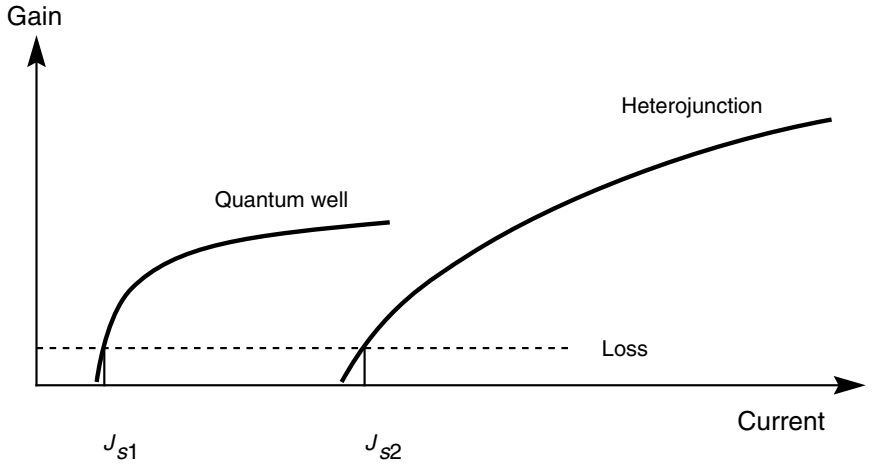


Fig. 13.26. Comparison of gain–current curves between bulk and single quantum well heterojunction lasers.

is the pump current density for the structure.

- G_N, the gain for a structure identical to the previous one, but possessing N quantum wells. Assuming coupling with the electromagnetic wave is identical for all wells, we have:

$$G_N = NG_1 = N\Gamma\gamma_0 \ln\left(\frac{J_1}{J_{\text{tr},1}}\right) \tag{13.63}$$

If J_N is the pump current density for the multi-quantum well structure, we must have:

$$J_N = q\frac{Nn_s}{t_{\text{tot}}} = NJ_1 \tag{13.64}$$

Substituting J_1 in (13.63) by its value given in (13.64), we obtain for the variation in optical gain as a function of the number of quantum wells N:

$$G_N = N\Gamma\gamma_0 \ln\left(\frac{J_N}{NJ_{\text{tr},1}}\right) \tag{13.65}$$

Figure 13.27 compares the gain–current curves (G_N) obtained for $N = 1, 2, 3$, and 4 wells. The transparency threshold is seen to increase in proportion to N.

The laser threshold current density is obtained by setting the gain (13.65) equal to the losses (13.60), i.e:

$$J_{\text{threshold},N} = \frac{NJ_{\text{tr},1}}{\eta}\exp\left[\frac{1}{N\Gamma\gamma_0}\left(\alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}}\right)\right] \tag{13.66}$$

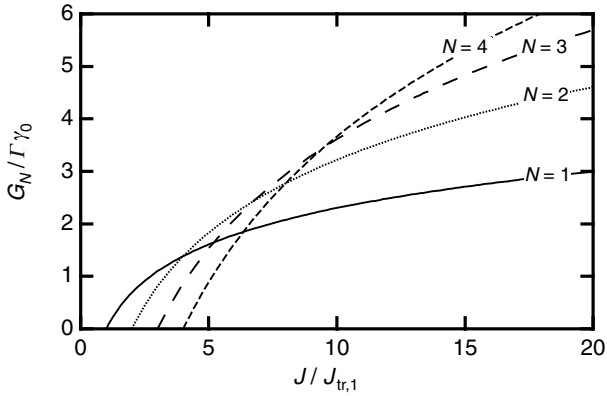


Fig. 13.27. Normalized gain for an N well multi-quantum well structure as a function of pump current density (normalized to the transparency current density for a single quantum well laser). The threshold current density increases in proportion to N.

where we have introduced the quantum efficiency η. This last expression allows most aspects of multi-quantum laser cavities to be optimized. The optimal number of wells can be obtained by setting $\mathrm{d}J_N/\mathrm{d}N = 0$, leading to:

$$N_{\text{opt}} = \text{Int}\left[\frac{1}{\Gamma\gamma_0}\left(\alpha_p + \frac{1}{2L}\ln\frac{1}{R_{m1}R_{m2}}\right)\right] + 1 = \frac{\text{cavity losses}}{\text{single well gain}} \tag{13.67}$$

where Int is the integer function. The interpretation of this last formula is straightforward. The current $I_{\text{threshold},N}$ is the product of (13.66) with Lw (where w is the width of the structure) and possesses a minimum as a function of the cavity length. The optimal length is obtained by taking the derivative of $I_{\text{threshold},N}$ with respect to L yielding:

$$L_{\text{opt}} = \frac{1}{2N\Gamma\gamma_0}\ln\frac{1}{R_{m1}R_{m2}} \tag{13.68}$$

Example

We consider a multi-quantum well laser with the following characteristics:

Single well parameters: $\Gamma = 0.1$, $\gamma_0 = 100\,\text{cm}^{-1}$
Cavity parameters: $\alpha_p = 10\,\text{cm}^{-1}$, $R_{m1} = 1, R_{m2} = 0.32$

Assuming a 500 μm long cavity, we obtain a loss of $21.4\,\text{cm}^{-1}$ and a gain per well of $10\,\text{cm}^{-1}$. The optimal number of quantum wells N_{opt} is 3.

In the case of a cavity possessing a single quantum well, the optimal length would be 570 μm.

13.7 Dynamic aspects of laser diodes

The temporal characteristics of laser diodes can be investigated following a similar approach to that employed in Section 4.7. Laser diodes, however, present a certain number of novel characteristics which must be included in the coupled dynamic equations (4.34) and (4.35). We rewrite these equations modified to include the particularities associated with laser diode operation.

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{J}{qd} - \frac{n}{t_{\mathrm{tot}}} - c'g(n)s(n - n_{\mathrm{tr}})$$
$$\frac{\mathrm{d}s}{\mathrm{d}t} = c'g(n)s(n - n_{\mathrm{tr}})\Gamma - \frac{s}{\tau_c} \tag{13.69}$$

Dynamic equations for a semiconductor laser cavity

The first two terms in the leading equation describe the supply of electron–hole pairs provided by the injection current, and the pair losses resulting from all recombination mechanisms with the exclusion of stimulated emission. The third term tallies the contribution due to stimulated emission and is equal to the photon density in the cavity (cm^{-3}) multiplied by the product $g(n)(n - n_{\mathrm{tr}})$, (where $(n - n_{\mathrm{tr}})$ corrects for the required transparency density for a semiconductor laser); $c' = c/n_{\mathrm{sc}}$ is the group velocity for photons in the semiconductor through which one can establish the relationship between photon flux and density (see Chapter 4). The term $g(n)$ ($\mathrm{cm}^{-1}\,\mathrm{cm}^{-3}$ or cm^2) describes the variation of *dynamic gain* as a function of carrier density:

$$g(n) = \frac{\mathrm{d}\gamma}{\mathrm{d}n} \tag{13.70a}$$

and is the characteristic slope of the $\gamma(n)$ curve (see Fig. 7.8 for heterojunction lasers and Fig. 13.22 for quantum well lasers). For heterojunction lasers, this gain is a constant given by (13.21):

$$\frac{\mathrm{d}\gamma}{\mathrm{d}n} = \frac{\gamma_0}{n_{\mathrm{transp}}} \tag{13.70b}$$

whereas for a quantum well laser, the dynamic gain is given by (13.59), or:

$$\frac{\mathrm{d}\gamma}{\mathrm{d}n} = \frac{\gamma_0}{n} \tag{13.70c}$$

where $n = n_s/d$. Comparing (13.69) and (4.1), we note that the dynamic gain is a

concept in close analogy to the *optical cross-section* σ_{op} in atomic systems. It would not be an exaggeration to consider the *dynamic gain* as an optical cross-section for each electron in the semiconductor medium. We will see that this optical cross-section for a semiconductor is considerable: a few $10^{-16}\,\text{cm}^2$ compared with $10^{-19}\,\text{cm}^2$ in ion-doped lasers (Cr, Nd, ...). It explains the very large amplification coefficients characteristic of semiconductor lasers.

In the second equation of (13.69), the first term reflects photon creation *into the waveguided mode* (hence the involvement of Γ). The second term represents the cavity losses (light scattering, mirror losses, etc.) through an effective photon lifetime τ_c which we recall equals (see (4.24) and (13.27)):

$$\tau_c = \frac{1}{c'[\alpha_p + (1/2L)\ln(1/R_{m1}R_{m2})]} \tag{13.71}$$

The calculation of dynamic response for a laser diode proceeds along classical lines (see Section 4.7). We will assume that the gain $g(n) = g$ is independent of n for small signals. We then designate as n_0 and s_0 the stationary state electron–hole pair and photon densities, respectively. These are given by the stationary solutions to (13.69):

$$\begin{aligned} n_0 - n_{tr} &= \frac{1}{c'g\Gamma\tau_c} \\ s_0 &= \Gamma\tau_c\left(\frac{J_0}{qd} - \frac{n_0}{t_{tot}}\right) \end{aligned} \tag{13.72}$$

Using these equations, we easily recover the results obtained in (13.34) for the laser output power as a function of drive current. We are in fact interested in the small signal dynamic response for a current modulation $J = J_0 + \delta j$ (with $\delta j \ll J_0$). We therefore write the carrier pair and photon densities in the form of $n = n_0 + \delta n$ and $s_0 + \delta s$. Neglecting second-order terms, we find the following system of differential equations in the small signal limit:

$$\begin{aligned} \frac{\mathrm{d}\delta n}{\mathrm{d}t} &= \frac{\delta j}{qd} - \delta n\left(\frac{1}{t_{tot}} + c'gs_0\right) - \frac{1}{\Gamma\tau_c}\delta s \\ \frac{\mathrm{d}\delta s}{\mathrm{d}t} &= \Gamma c'gs_0\delta n \end{aligned} \tag{13.73}$$

where we have made use of (13.72). We then study the harmonic response of the system by taking:

$$\delta j = \mathrm{Re}[\delta j(\omega)\mathrm{e}^{\mathrm{i}\omega t}];\ \delta n = \mathrm{Re}[\delta n(\omega)\mathrm{e}^{\mathrm{i}\omega t}];\ \delta s = \mathrm{Re}[\delta s(\omega)\mathrm{e}^{\mathrm{i}\omega t}] \tag{13.74}$$

The system of differential equations in (13.73) then leads to:

$$\frac{\delta s(\omega)}{s_0} = \frac{(\delta j/qd)\Gamma c' g}{-\omega^2 + [(1/t_{\text{tot}}) + c' g s_0]\text{i}\omega + c' g s_0/\tau_c} \tag{13.75}$$

The small signal frequency response is given by:

$$|\delta s(\omega)| = \frac{\delta s(0)}{(\tau_c/c' g s_0)\sqrt{[\omega^2 - (c' g s_0/\tau_c)]^2 + \omega^2[(1/t_{\text{tot}}) + c' g s_0]^2}} \tag{13.76}$$

This response is maximized at a *relaxation frequency* ω_R of:

$$\omega_R = 2\pi f_R = \sqrt{\frac{c' g s_0}{\tau_c} - \frac{1}{2}\left(\frac{1}{t_{\text{tot}}} + c' g s_0\right)^2} \approx \sqrt{\frac{c' g s_0}{\tau_c}} \tag{13.77a}$$

Relaxation frequency for a semiconductor laser

and a *damping coefficient* Γ_R given by:

$$\Gamma_R \approx \frac{c' g s_0}{2} = \frac{1}{2}\tau_c \omega_R^2 \tag{13.77b}$$

The physical origin for the oscillations and relaxation was discussed previously in Section 4.7.1. This corresponds to the oscillatory exchange of energy between the electron–hole pairs and the photon population, which are strongly coupled through stimulated emission. More precisely, as the electron–hole density increases, so does the gain, which triggers an increase in the photon density, which leads to an increased rate of stimulated emission, which decreases the carrier population (i.e. through recombination), . . . leading to oscillations in these quantities. Relaxation results from photon loss (by parasitic absorption or through mirror loss), and is described by τ_c in (13.77b). Figure 13.28 shows the temporal response of a semiconductor laser. The behaviour predicted by (13.76) is reproduced experimentally showing a rise in the frequency response towards a maximum value, with a subsequent decrease proportional to f^{-2}. Equation (13.77a) also predicts an increase in the relaxation frequency with pump current given the dependence of f_R on $s_0^{1/2}$. Finally, we note that the maximum utilization frequency (i.e. the frequency for which $|\delta s(\omega)| = |\delta s(0)|$ and therefore near f_R) increases with the slope $g(n)$ (or $g(J)$). Quantum well laser diodes therefore possess a significant advantage over bulk heterojunction lasers given their superior dynamic gain.

Example

We seek the relaxation frequency for a GaAs/AlGaAs heterojunction laser. The relevant parameters are:

Output power, $P_{\text{out}} = 10\,\text{mW}$
Photon energy, $h\nu = 1.4\,\text{eV}$
Output mirror reflectances, $R_{m1} = 1$; $R_{m2} = 0.32$

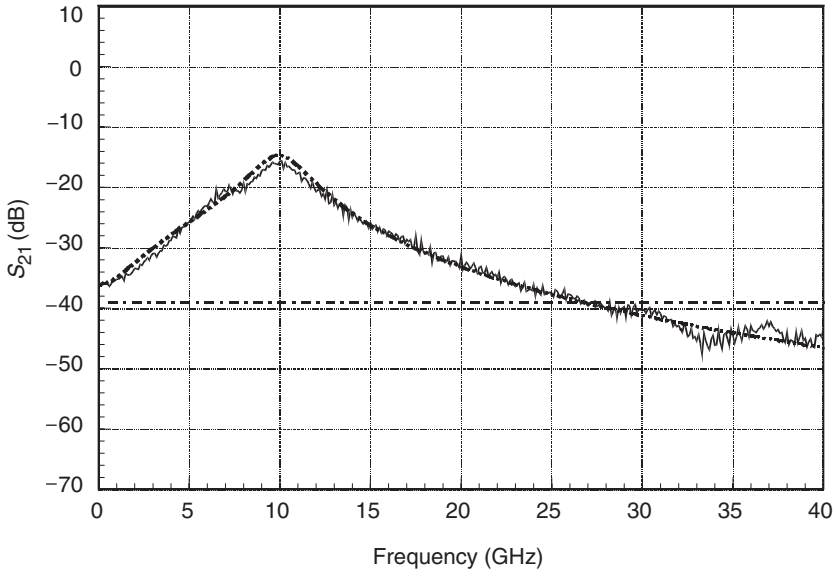


Fig. 13.28. Frequency response for an InGaAs/InP laser diode. (Courtesy of E. Goutain@THALES.)

Optical mode area $A = wd_{\text{mode}}$, where w (the cavity width) $= 2\,\mu\text{m}$, and d_{mode} (the effective mode thickness) $= 0.1\,\mu\text{m}$, giving $A = 2 \times 10^{-9}\,\text{cm}^2$
Cavity length, $L = 200\,\mu\text{m}$
Parasitic loss, $\alpha_p = 10\,\text{cm}^{-1}$
Optical index, $n_{\text{sc}} = 3.3$
Quantum efficiency, $\eta = 1$
Dynamic gain g, given in Fig. 7.8 is $3 \times 10^{-16}\,\text{cm}^2$

The photon density s_0 is given by (13.32), which we rewrite here as:

$$\frac{P_{\text{out}}}{A} = (1 - R)\frac{c}{n_{\text{sc}}}\hbar\omega s_0 \tag{13.78}$$

leading to a photon density s_0 of $3.6 \times 10^{15}\,\text{cm}^{-3}$. The photon lifetime in the cavity is $1/c'(\alpha_p - 1/2L\ln(R))$ or 2.8 ps. The relaxation frequency is then:

$$f_R = \frac{1}{2\pi}\sqrt{\frac{3 \times 10^{10}\,\text{cm s}^{-1}/3.3 \times 3 \times 10^{-16}\,\text{cm}^2 \times 3.6 \times 10^{15}\,\text{cm}^{-3}}{2.8 \times 10^{-12}\,\text{s}}} = 9.4\,\text{GHz}$$

13.8 Characteristics of laser diode emission

13.8.1 Spectral distribution

Figure 13.29 shows the spectral emission from a laser diode for different pump

currents. Below threshold (Fig. 13.29*a*), the spectral dependence of the emission follows the gain curve in the medium, and is modulated according to the allowed Fabry–Pérot modes in the cavity (see Fig. 4.9). As the laser cavity used in this experiment has a cavity length of 300 μm and a peak emission wavelength λ near 0.85 μm, the anticipated cavity mode spacing ($\Delta\lambda = \lambda^2/2n_{sc}L$ or 0.36 nm) is found to be in accordance with the experiment. In Fig. 13.29*b*, the cavity modes that experience gain in excess of the threshold requirement undergo laser oscillation. As *intraband* scattering mechanisms are extremely rapid (0.1 ps), semiconductors tend to act as homogeneous media and favour single mode emission. In spite of this, a number of concurrent causes can provoke multimode emission (e.g. *spectral hole burning*, *spatial hole burning* (see Chapter 4), filamentation, . . .).

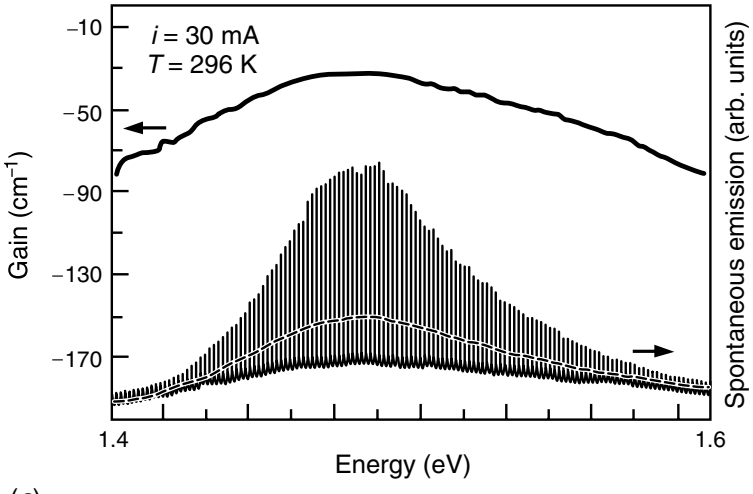
To obtain single mode operation, distributed Bragg mirrors are often incorporated into the device structure. These are known as *distributed feedback* (*DFB*) *lasers* and form the subject of Complement 13.A. Another possibility involves reducing the size of the cavity so that the cavity mode spacing exceeds the width of the gain curve. Given a typical gain bandwidth $\Delta\lambda$ of 100 nm, this requires a cavity length of $L = \lambda^2/2n_{op}\Delta\lambda$ or 1 μm. Such short cavities can be formed along the epitaxial deposition axis. Devices fabricated with this geometry are referred to as vertical cavity surface emission lasers (VCSELs) and are examined in Complement 13.C.

Finally, we recall that the light emitted by a laser diode below the transparency threshold is incoherent (see Chapter 4). In fact, edge emitting LEDs are laser diodes that possess gain levels below the laser oscillation threshold. In certain cases, the diode facets are even treated with antireflection coatings to suppress optical feedback and keep these devices from lasing under high injection current levels. Nevertheless, once the devices are driven beyond the transparency threshold, stimulated emission will dominate over spontaneous emission. In this regime, LEDs emit extremely brilliant (albeit incoherent) electroluminescence. Such devices are referred to as *superluminescent LEDs*.

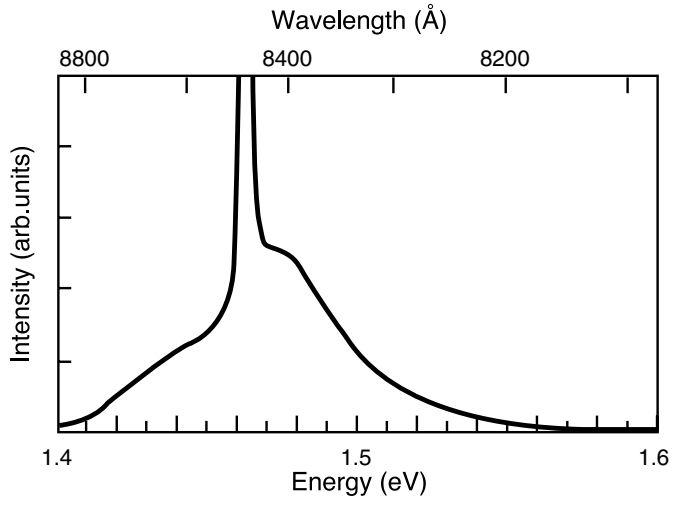
13.8.2 Spatial distribution

Calculation of the spatial distribution of the luminescent energy emitted by a laser diode is challenging, requiring the application of numerical techniques which we will outline below. Most notably, one makes use of the *far-field approximation*. We recall this approximation now.

A distribution of oscillating charges with frequency ω gives rise to an amplitude distribution for the electric field $\mathbf{E}_{nf}(\mathbf{r}')$ ($\mathbf{E}_{nf}$ is the near-field) across a small opening lying in the $z = 0$ plane, where $\mathbf{r}'$ runs along the major axis of the emission aperture (see Fig. 13.30). The far-field $\mathbf{E}_{ff}$ at a distant point $\mathbf{r}$ away from the opening is the Fourier transform of near-field, or more specifically:



(*a*)



(*b*)

Fig. 13.29. Emission spectrum from a GaAs/$Al_{0.18}Ga_{0.82}As$ quantum well laser diode (*a*) below and (*b*) above laser threshold. (Courtesy of J. Nagle@THALES.)

$$\mathbf{E}_{\mathrm{ff}}(\mathbf{r}) = \frac{-\mathrm{i}k\mathrm{e}^{\mathrm{i}kr}}{4\pi}\mathbf{u}_{\mathbf{r}} \times \int_{\text{opening}} \mathrm{d}\mathbf{r}'\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}'}[-2\mathbf{n} \times \mathbf{E}_{\mathrm{nf}}(\mathbf{r}')] \tag{13.79}$$

where $\mathbf{n}$ in the unit normal vector to the opening, $\mathbf{u}_{\mathbf{r}}$ is the unit normal vector in the direction of $\mathbf{r}$, and $\mathbf{k} = \omega/c\mathbf{u}_{\mathbf{r}}$ is the vacuum wavevector in the $\mathbf{r}$ direction (see Fig. 13.30).

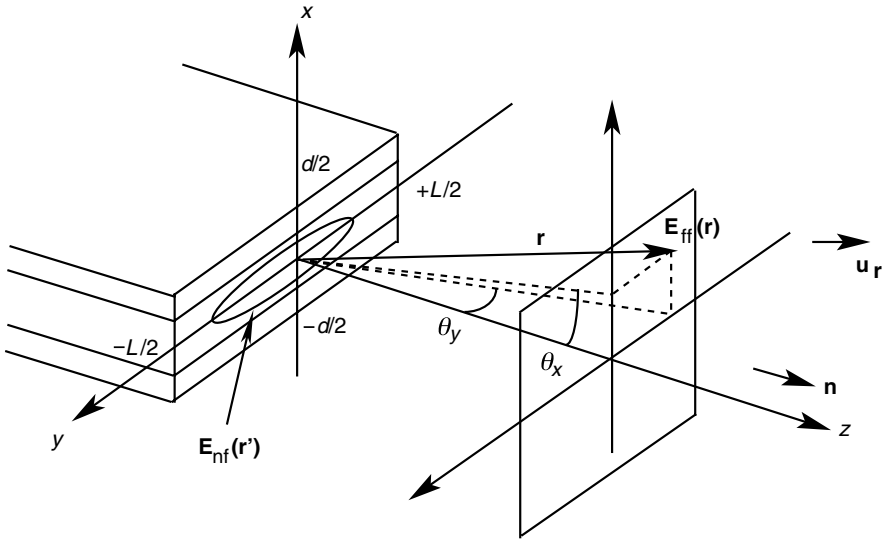


Fig. 13.30. Diagram for calculating the far-field power distribution of a semiconductor laser.

The emitted power is then given by $P(r) = (1/2)Z_0|E_{\mathrm{ff}}(r)|^2$, where Z_0 is the vacuum impedance ($Z_0 = 377\,\Omega$). The spatial distribution of the emitted intensity is of particular interest along two directions:

- *Along the growth axis.* In this case, the angular variation (as a function of $\theta_\perp$) is given by:

$$P_{\mathrm{ff}}(\mathbf{r}) \propto \cos^2\theta_\perp \left| \int_{-l/2}^{l/2} \int_{-d/2}^{d/2} \mathrm{e}^{-\mathrm{i}k\sin\theta_\perp x'} E_{\mathrm{nf}}(x', y')\mathrm{d}x'\mathrm{d}y' \right|^2 \tag{13.80}$$

  where l is the effective mode width (see Fig. 13.31). In a laser waveguide structure, the electromagnetic field is principally distributed between $-d_{\mathrm{mode}}/2$ and $+d_{\mathrm{mode}}/2$. Taking for the near-field $E_{\mathrm{nf}}(x, y)$ the function $C(x, d_{\mathrm{mode}}/2) \times C(y, l/2)$, where $C(x, a) = 1$ between $-a$ and $+a$ and 0 elsewhere, the Fourier transform (13.80) can be written:

$$P_{\mathrm{ff}}(\mathbf{r}) = P_{\max}\,\mathrm{sinc}^2\left(\frac{2\pi}{\lambda_0}\sin\theta_\perp\frac{d_{\mathrm{mode}}}{2}\right) \tag{13.81}$$

  Thus, we see that all the energy is confined between the angles $+\lambda_0/2d_{\mathrm{mode}}$ and $-\lambda_0/2d_{\mathrm{mode}}$. This result can be obtained by approximating the confined mode as a Gaussian beam with width $W_0 = d_{\mathrm{mode}}/2$ (see (4.E.5)). The divergence of the beam is then $\lambda_0/\pi W_0 = 2\lambda_0/\pi d_{\mathrm{mode}} \approx \lambda_0/2d_{\mathrm{mode}}$.
- *Direction parallel to the heterointerfaces.* The angular distribution is then:

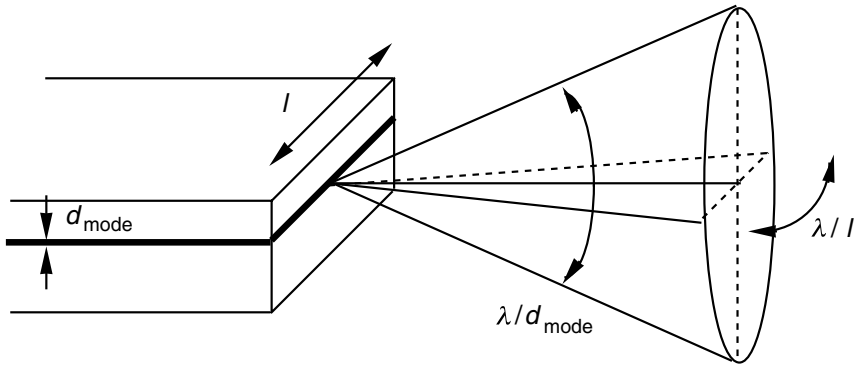


Fig. 13.31. Angular distribution for light emerging from the fundamental mode of a semiconductor waveguide.

$$P_{\mathrm{ff}}(\mathbf{r}) \propto \cos^2\theta_{\parallel} \left| \int_{-l/2}^{l/2} \int_{-d/2}^{d/2} \mathrm{e}^{-\mathrm{i}k\sin\theta_{\parallel} y'} E_{\mathrm{nf}}(x', y')\mathrm{d}x'\mathrm{d}y' \right|^2 \tag{13.82}$$

and the beam divergence is $\lambda_0/2l$.

Expressions (13.80) and (13.82) are extremely useful in allowing one to treat more general cases than the approximation dealt with above. Figure 13.31 summarizes the present results. For cases involving higher index modes, one obtains increased angular dispersions with correspondingly reduced intensities.

For a typical laser structure with a width of 5 μm and an effective mode thickness $d_{\mathrm{mode}} = 1\,\mu\mathrm{m}$, the angular divergence along the growth axis is of the order of (1 μm/1 μm) radians or $\sim 60^\circ$, and (1 μm/5 μm) radians or $\sim 10^\circ$ along the direction parallel to the heterointerfaces. The spatial distribution of emission from a semiconductor component is therefore generally elliptical in nature, with a preponderance of angular divergence along the growth axis. This implies the necessity of corrective optics in applications which require properly collimated light beams (e.g. *compact disk* lasers, efficient coupling into optical fibres, etc.).

FURTHER READING

H. C. Casey Jr and M. B. Panish, *Heterostructure Lasers*, *Part A*: *Fundamental Principles*, Academic Press, New York (1978).

S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley Interscience, New York (1995).

A. Yariv, *Quantum Electronics*, Wiley, New York (1989).

Complement to Chapter 13

13.A Distributed feedback (DFB) lasers

As discussed in Section 13.8, semiconductor lasers can oscillate simultaneously over multiple longitudinal modes. Such multimodal operation, however, is undesirable in a number of applications as it leads to: temporal broadening in optical fibres, collimation problems, etc. It is therefore advantageous in these instances to be able to select the single mode that will undergo laser amplification, by employing extremely wavelength selective mirrors. Such a mirror can be implemented using *distributed feedback*. This idea consists of incorporating within the laser cavity, a modulated waveguide structure with period Λ (see Fig. 13.A.1). This periodic perturbation, in a manner similar to a Bragg mirror (see Complement 9.D), acts as an extremely selective frequency filter. The formalism for describing an optical mode in a spatially modulated waveguide was presented in Complement 9.B. We recall the principal results here.

We will label as $A_l^+(z)$ and $\mathrm{A}_l^-(z)$, respectively, the amplitudes for the right and left propagating optical modes l, with a propagation constant β_l. The periodic modulation induces an energy transfer between the left and right propagating waves as described by the system of equations in (9.B.6):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} A_l^-(z) &= \mathrm{i}gA_l^+(z)\mathrm{e}^{-2\mathrm{i}\Delta\beta z} \\ \frac{\mathrm{d}}{\mathrm{d}z} A_l^+(z) &= -\mathrm{i}gA_l^-(z)\mathrm{e}^{2\mathrm{i}\Delta\beta z} \end{aligned} \tag{13.A.1}$$

where $\Delta\beta$ is the phase mismatch term given by:

$$\Delta\beta = \beta_l - \frac{\pi}{\Lambda} \tag{13.A.2}$$

Only those modes, l, with $\Delta\beta \approx 0$ can propagate within the structure. We designate as ω_B and β_B the frequencies and propagation constants which satisfy the Bragg condition $\Delta\beta = 0$. The coupling constant g (not to be confused with the

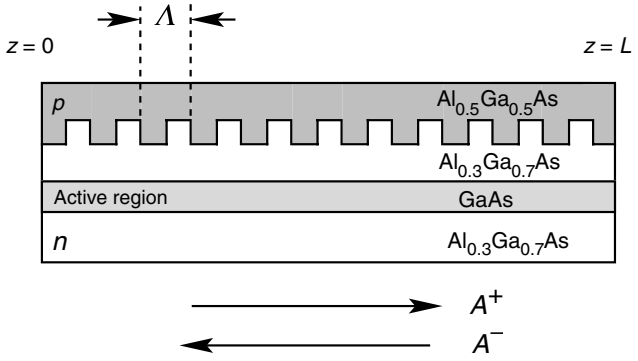


Fig. 13.A.1. Cross-section of a distributed feedback laser. The Al fractions in the diagram are merely suggestive of plausible layer compositions.

gain!) which describes the coupling efficiency between the left and right propagating waves may be expressed in the form of (9.B.6b) as:

$$g = \frac{\omega \varepsilon_0}{8 p_0} \int_{-\infty}^{+\infty} \varepsilon_M(x) E_l(x)^2 \mathrm{d}x \tag{13.A.3}$$

where ω is the frequency of the electromagnetic wave, $\varepsilon_M(x)$ is the (modulated) relative permittivity function in the grating region, $E_l(x)$ is the electric field distribution for the lth mode, and p_0 is the normalization constant necessary for having a homogeneous equation ($p_0 = 1\,\mathrm{W\,m^{-1}}$). An approximate expression for g is given in (9.B.16) for the case of a highly confining waveguide. To account for the presence of the gain medium (with modal gain γ), (13.A.1) takes the modified form:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} A_l^- &= \mathrm{i} g A_l^+ \mathrm{e}^{-2\mathrm{i}\Delta\beta z} - \gamma A_l^- \\ \frac{\mathrm{d}}{\mathrm{d}z} A_l^+ &= -\mathrm{i} g A_l^- \mathrm{e}^{2\mathrm{i}\Delta\beta z} + \gamma A_l^+ \end{aligned} \tag{13.A.4}$$

To solve this system of equations, we perform the variable transformation:

$$\begin{aligned} A_l^- &= a_l^- \mathrm{e}^{-\gamma z} \\ A_l^+ &= a_l^+ \mathrm{e}^{+\gamma z} \end{aligned} \tag{13.A.5}$$

which leads to the new system of equations:

$$\frac{\mathrm{d}}{\mathrm{d}z}a_l^- = \mathrm{i}g a_l^+ \mathrm{e}^{-2\mathrm{i}(\Delta\beta+\mathrm{i}\gamma)z}$$
$$\frac{\mathrm{d}}{\mathrm{d}z}a_l^+ = -\mathrm{i}g a_l^- \mathrm{e}^{-2\mathrm{i}(\Delta\beta+\mathrm{i}\gamma)z} \tag{13.A.6}$$

(13.A.6) is formally identical to the set of equations in (13.A.1) given the transformation $\Delta\beta \to \Delta\beta + \mathrm{i}\gamma$.

We can therefore employ the result from Complement 9.B. To do so, we need to specify the boundary conditions for the amplitudes in the guide. We will assume the wave to be incident from the left with amplitude $a_l^+(0) = a_0$ at the grating entrance ($z = 0$, see Fig. 13.A.1), and that no wave is incident from the right at the guide exit ($z = L$) so that $a_l^-(L) = 0$. Equation (9.B.10) may then be written[1]:

$$a_l^-(z) = a_0 \frac{g}{\mathrm{i}(\Delta\beta + \mathrm{i}\gamma)\mathrm{sh}(\delta L) + \delta \mathrm{ch}(\delta L)} \mathrm{e}^{-\mathrm{i}(\Delta\beta+\mathrm{i}\gamma)z} \mathrm{sh}[\delta(z-L)]$$
$$a_l^+(z) = a_0 \frac{\mathrm{e}^{+\mathrm{i}(\Delta\beta+\mathrm{i}\gamma)z}}{\mathrm{i}(\Delta\beta + \mathrm{i}\gamma)\mathrm{sh}(\delta L) + \delta \mathrm{ch}(\delta L)} \{\delta \mathrm{ch}[\delta(z-L)] - \mathrm{i}(\Delta\beta + \mathrm{i}\gamma)\mathrm{sh}[\delta(z-L)]\} \tag{13.A.7}$$

where this time, the coefficient δ is given (see (9.B.9b)) by:

$$\delta = \sqrt{g^2 + (\gamma - \mathrm{i}\Delta\beta)^2} \tag{13.A.8}$$

The presence of the $\pm\gamma z$ term in the exponential takes into account amplification within the waveguide. The denominators in (13.A.7) express the mirror-like role played by the periodic modulation. The denominators vanish if:

$$\delta \cosh \delta L = (\gamma - \mathrm{i}\Delta\beta)\sinh \delta L \tag{13.A.9a}$$

Threshold condition for a DFB laser

In fact, if this last expression is satisfied, the system will yield a non-zero output even for the situation involving zero input ($a_0 = 0$) – this is what characterizes laser oscillation. To obtain some understanding of the repercussions of this condition on the gain and phase matching requirements, we will analyse (13.A.9) for the Bragg frequency ω_B, i.e. for $\Delta\beta = 0$:

$$\sqrt{1 + \left(\frac{g}{\gamma}\right)^2} = \tanh(\sqrt{g^2 + \gamma^2}L) \tag{13.A.9b}$$

As the gain γ tends towards infinity, both sides of this last equation approach unity, but one from 1^+ and the other from 1^-. Clearly, (13.A.9b) cannot be satisfied and $\Delta\beta = 0$ no longer represents a propagating mode, i.e. the gain introduces new

[1] sh = sinh, ch = cosh.

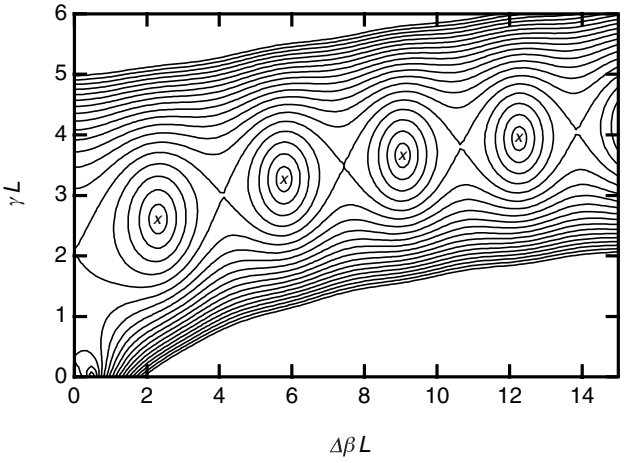


Fig. 13.A.2. Contour plot for (13.A.9) in the plane $\Delta\beta L - \gamma L$. Solutions of this equation are located at the centres of the closed curves.

conditions for phase matching. It is therefore necessary to solve (13.A.9) in the complex plane.

Figure 13.A.2 shows a contour plot for $1/|\delta\cosh\ \delta L - (\gamma - \mathrm{i}\Delta\beta)\sinh \delta L|^2$ for $gL = 0.5$. The (extremely useful) MATHEMATICA program used to obtain this plot is listed below:

```
gL=0.5;
deltaL= Sqrt[gL^2+(y-I*x)^2];
Den= 1/Abs[deltaL* Cosh[deltaL]-(y-I*x)*Sinh[deltaL]];
ContourPlot[Den, {x,0,15},{y,0,6}, PlotPoints->100,
ContourShading ->False,AspectRatio->.7]
```

We note that the required gain–length product γL to achieve threshold increases with the degree of excursion from the Bragg condition $\Delta\beta L = 0$. This means that distributed feedback favours only small deviations from the Bragg condition. To understand this behaviour better, we will solve (13.A.9) for the case involving strong gain, i.e. $\gamma \gg g$.

Equation (13.A.9a) may be written:

$$\mathrm{e}^{2\delta L} = \frac{1 + \delta/(\gamma - \mathrm{i}\Delta\beta)}{1 - \delta/(\gamma - \mathrm{i}\Delta\beta)} \tag{13.A.10a}$$

which can be developed into:

$$\mathrm{e}^{2L[\gamma - \mathrm{i}\Delta\beta + (1/2)(g^2/(\gamma - \mathrm{i}\Delta\beta))]} \approx -4\left(\frac{\gamma - \mathrm{i}\Delta\beta}{g}\right)^2 \tag{13.A.10b}$$

since from (13.A.8) we have:

$$\frac{\delta}{\gamma - \mathrm{i}\Delta\beta} = \sqrt{1 + \left(\frac{g}{\gamma - \mathrm{i}\Delta\beta}\right)^2} \approx 1 + \frac{1}{2}\left(\frac{g}{\gamma - \mathrm{i}\Delta\beta}\right)^2 \tag{13.A.10c}$$

We will begin by analysing the phase requirements for (13.A.10b) that will allow the lth mode to oscillate:

$$2\arctan\frac{\Delta\beta_l}{\gamma_l} + 2\Delta\beta_l L - \frac{g^2\Delta\beta_l L}{\gamma_l^2 + \Delta\beta_l^2} = (2m+1)\pi \tag{13.A.11}$$

Once the gain γ becomes large in comparison to the phase difference, (13.A.11) simplifies considerably giving:

$$\Delta\beta_l L \approx \left(m + \frac{1}{2}\right)\pi \tag{13.A.12}$$

The oscillation modes closest to the Bragg mode are those corresponding to $m = 0$, i.e. $\Delta\beta_0 L = \pi/2$ and $m = -1$, or $\Delta\beta_0 L = -\pi/2$, which is not very different from what we observe in Fig. 13.A.2 ($\Delta\beta_0 L \approx 2$). To express this condition in terms of allowed wavelengths, it is necessary to remember that $\beta_l = 2\pi n_{\mathrm{eff}}/\lambda$, where n_{eff} is the effective guide index for the lth mode (see Chapter 9) and that the *Bragg wavelength* $\lambda_B = 2n_{\mathrm{eff}}\Lambda$ is the resonant wavelength for the grating. Using (13.A.2), the two modes closest to the Bragg conditions are:

$$\begin{aligned} \frac{2n_{\mathrm{eff}}}{\lambda_0} - \frac{1}{\Lambda} &= \frac{1}{2L} \\ \frac{2n_{\mathrm{eff}}}{\lambda_{-1}} - \frac{1}{\Lambda} &= -\frac{1}{2L} \end{aligned} \tag{13.A.13a}$$

or, since $1/L \ll 1/\Lambda$:

$$\begin{aligned} \lambda_0 - \lambda_{\mathrm{B}} &\approx -n_{\mathrm{eff}}\frac{\Lambda^2}{L} \\ \lambda_{-1} - \lambda_{\mathrm{B}} &\approx n_{\mathrm{eff}}\frac{\Lambda^2}{L} \end{aligned} \tag{13.A.13b}$$

The wavelength spacing between the two allowed modes is:

$$\Delta\lambda = 2n_{\mathrm{eff}}\frac{\Lambda^2}{L} \tag{13.A.14}$$

which is the same result that would have been obtained for the case of two mirrors separated by a distance L.

Still following (13.A.10b), the condition on the gain at threshold is given by:

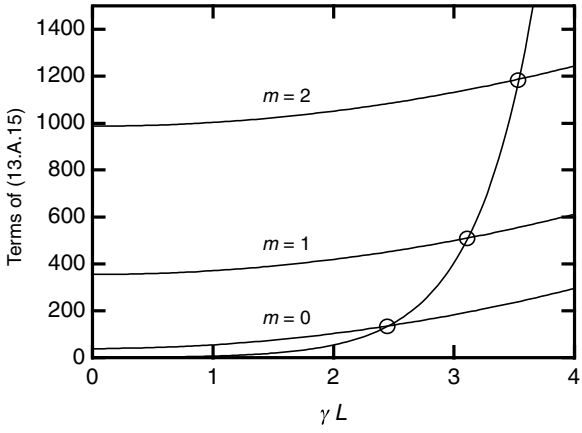


Fig. 13.A.3. Solutions to (13.A.15) with $gL = 0.5$ showing the necessary gain γL to achieve laser oscillation in the $m = 0$, 1, and 2 modes. Distributed feedback significantly favours single mode operation at $m = 0$ as it most closely approaches the Bragg condition.

$$e^{2\gamma_l L} \approx 4\frac{(\gamma_l L)^2 + (m + 1/2)^2\pi^2}{(gL)^2} \tag{13.A.15}$$

Equation (13.A.15) is an implicit function of γ_l, and Fig. 13.A.3 shows the points that satisfy (13.A.15) assuming $gL = 0.5$. It is clear that distributed feedback encourages the fundamental modes ($l = 0, -1$), with the result that distributed feedback lasers favour to a large extent *single longitudinal mode* laser oscillation.

Neglecting parasitic losses α_p, the condition for laser threshold is obtained for $\gamma_{\text{threshold}}L \approx 2$. Expression (13.26a) $\gamma_{\text{threshold}} = -1/2L\ln(R_{m1}R_{m2})$ indicates that this optical feedback corresponds to mirror reflectances given by $R_{m1}R_{m2} = e^{-4} = 0.02$. This is not very far from the thresholds required for cleaved cavity lasers where for the GaAs/air interfaces, $R_{m1}R_{m2} = 0.32^2 = 0.1$. Clearly, the required gain at threshold $\gamma_{\text{threshold}}$ can be diminished by increasing the coupling constant g, as this would correspond to using higher reflectivity mirrors. The coupling constant can only be increased by bringing the gratings closer to the active region of the laser. This can lead to significant losses, however, as positioning of the grating too close to the active region will diffract light into radiative ('leaky') guide modes. The optimization of such structures generally requires the use of numerical simulations.

13.B Strained quantum well lasers

The maximum gain $\gamma_{\max}$ for an amplifying medium employing quantum wells was given by (13.57) to be:

$$\gamma_{\max} = \alpha_{2D}(1 - e^{-n_s/n_c} - e^{-n_s/R_{cv}n_c}) \tag{13.B.1}$$

where α_{2D} is the cold cavity absorption for the medium (i.e. without carrier injection, $J = 0$) given by (13.46); n_s is the two-dimensional carrier density (in cm^{-2}) in the quantum well conduction and valence bands; n_c is the *critical two-dimensional density* in the conduction band (Eq. (13.54)); and R_{cv} is the ratio of the effective masses m_v/m_c, where m_v is the effective mass of states in the *first valence band* filled by current injection. This subband was justifiably assumed in Chapter 13 to be the heavy hole subband (with effective mass m_{hh}), as we saw in Chapter 8 that the lifting of degeneracy placed this subband above the others (i.e. above the light hole and spin orbit subbands). As is indicated in Fig. 13.23, the transparency density and, hence, the transparency current, both lessen as the ratio of the effective masses decreases. Figure 13.21 explains this effect: for equivalent carrier densities, the quasi-Fermi level penetrates more readily into the valence subband as its effective mass (and hence its associated density of states) becomes smaller. In other words, the smaller the effective mass in the valence subband, the more easy it is to satisfy the Bernard–Durrafourg condition.

We thus have a vested interest in working with effective hole masses that are as small as possible. Such flexibility is not possible, however, within the framework of a particular semiconductor system where, by construction, the heavy hole subband always dominates. Yablonovitch and Kane therefore proposed a method of inverting the heavy hole and light hole bands using mechanical strain to decrease the transparency current in semiconductor quantum well laser diodes. This can be achieved by incorporating *strained quantum well layers* into the device structure.

As an example, we consider an $In_{1-u}Ga_uAs/InP$ quantum well with a compositional fraction u that deviates from the lattice matched value for InP (i.e. $\delta a = a_0 - a(u) = 0$ for $u = 0.468$). During growth, the InP substrate forces the deposited InGaAs to assume the same positions along the growth surface as would normally be occupied by a continuation of the InP lattice. In the case of non-lattice matched layers, this accommodation is possible as long as the elastic strain energy remains less than the InGaAs/InP binding energy. The layer thickness for which both of these energies become equal to one another is referred to as the *critical thickness* l_c.

For layer thicknesses less than l_c, the mechanical strain resulting from the lattice mismatch is a tensor with elements given by:

$$\varepsilon = \varepsilon_{xx} = \varepsilon_{yy} = \frac{\delta a}{a} = \frac{a_0 - a(u)}{a_0}$$
$$\varepsilon_{zz} = -2\frac{\sigma}{1-\sigma}\varepsilon \tag{13.B.2}$$

where Ox and Oy are both directions in the plane of the quantum well and Oz lies

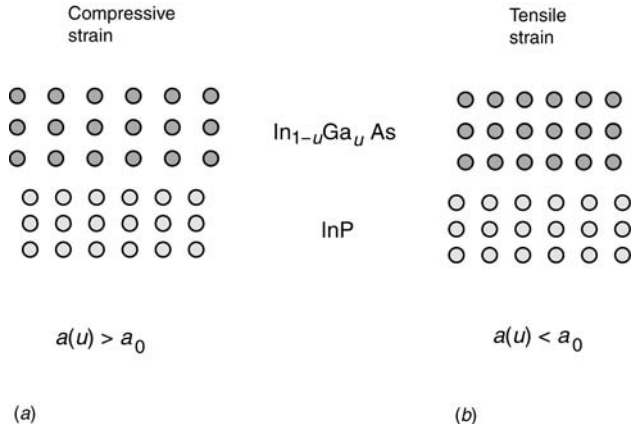


Fig. 13.B.1. (*a*) If the lattice constant of bulk $In_{1-u}Ga_uAs$ (dependent upon the particular value of u) is larger than that of InP, this material, when deposited epitaxially upon an InP substrate, will be compressively strained. In the inverse case involving $a(u) < a_0$, the resulting InGaAs layer would be tensely strained. These conclusions hold only as long as the layer thickness for a particular composition of $In_{1-u}Ga_uAs$ remains inferior to the critical thickness l_c.

along the growth direction. The Poisson modulus, σ (equal to $\sim 1/3$ for most III–V semiconductors), figures in (13.B.2) and describes the elasticity of the medium. When $a(u) > a_0$, the well is *compressively* strained; and when $a(u) < a_0$, the well is *tensely* strained (see Fig. 13.B.1).

Mechanical strain in InGaAs forces the constituent atoms to occupy positions that are not accessible in the unstrained structure. As a result, the overlap integrals between the atomic orbitals change and the band structure of the strained InGaAs differs from that of the unstrained material.

Without strain, the forbidden gap in $In_{1-u}Ga_uAs$ at 300 K is given by:

$$E_g(u) = 0.324 + 0.7u + 0.4u^2\,(\text{eV}) \tag{13.B.3}$$

The quadratic correction, which adds to the linear interpolation between the InAs and GaAs bandgaps, is referred to as the *bowing parameter*. In the presence of strain, perturbation calculations show that the bottom of the conduction band shifts by δE_c:

$$\delta E_c(u) = e_c(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) = 2e_c\frac{\delta a(u)}{a_0}\left(1 - \frac{\sigma}{1-\sigma}\right) \tag{13.B.4}$$

where e_c is the *deformation potential* for the conduction band and happens to be negative ($e_c = -5\,\text{eV}$) for InGaAs. Similarly, the top of the valence band shifts in energy, but by differing amounts for the heavy hole and light hole valence bands:

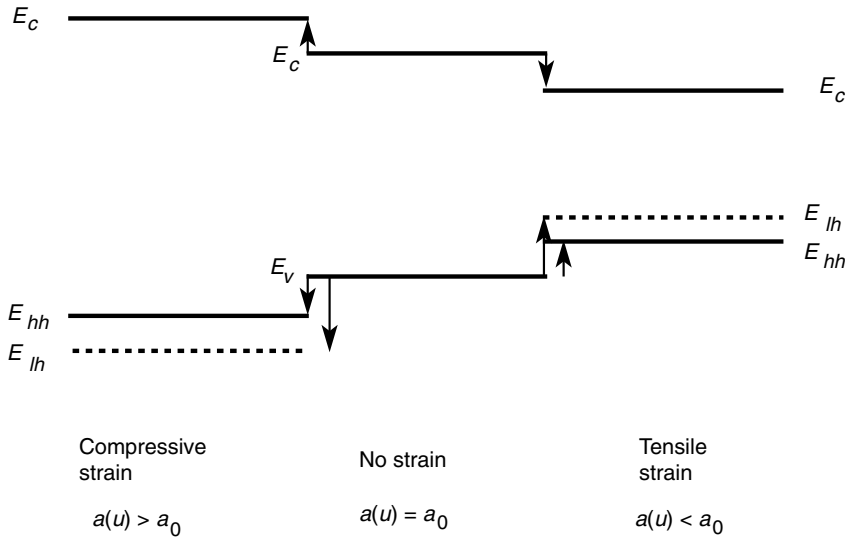


Fig. 13.B.2. In the $In_{1-u}Ga_uAs/InP$ system, depending on the sign and the magnitude of the strain, the light hole band may sit below, at, or above the position of the heavy hole band.

$$\delta E_{hh}(u) = -P(u) - Q(u)$$
$$\delta E_{lh}(u) = -P(u) + Q(u) \tag{13.B.5a}$$

with

$$P(u) = -e_v(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) = -2e_v\frac{\delta a(u)}{a_0}\left(1 - \frac{\sigma}{1-\sigma}\right)$$
$$Q(u) = -\frac{e_s}{2}(\varepsilon_{xx} + \varepsilon_{yy} - 2\varepsilon_{zz}) = -e_s\frac{\delta a(u)}{a_0}\left(1 + 2\frac{\sigma}{1-\sigma}\right) \tag{13.B.5b}$$

where e_v and e_s are the valence band deformation and shear deformation potentials, respectively, for InGaAs (with $e_v = 1.2\,\text{eV}$ and $e_s = -1.8\,\text{eV}$). Expression (13.B.5a) with the physical constants given above shows that under tensile strain, the heavy hole and light hole bands will be inverted relative to each other within the band structure. Figure 13.B.2 summarizes the different possibilities.

Thus, one of the main effects of strain is to lift the heavy hole/light hole degeneracy at $\mathbf{k} = 0$ for bulk InGaAs. The situation is a little more complex in quantum wells, however, as the lifting of the degeneracy due to the quantum confinement effect is larger for light holes than for heavy holes. Inversion of heavy hole and light hole subbands is a product of a subtle equilibrium between these two degeneracy lifting mechanisms. Figure 13.B.3 shows the subband structure obtained for an InGaAs/InP quantum well as a function of strain.

In fact, the final effect on the transparency current falls short of being spectacu-

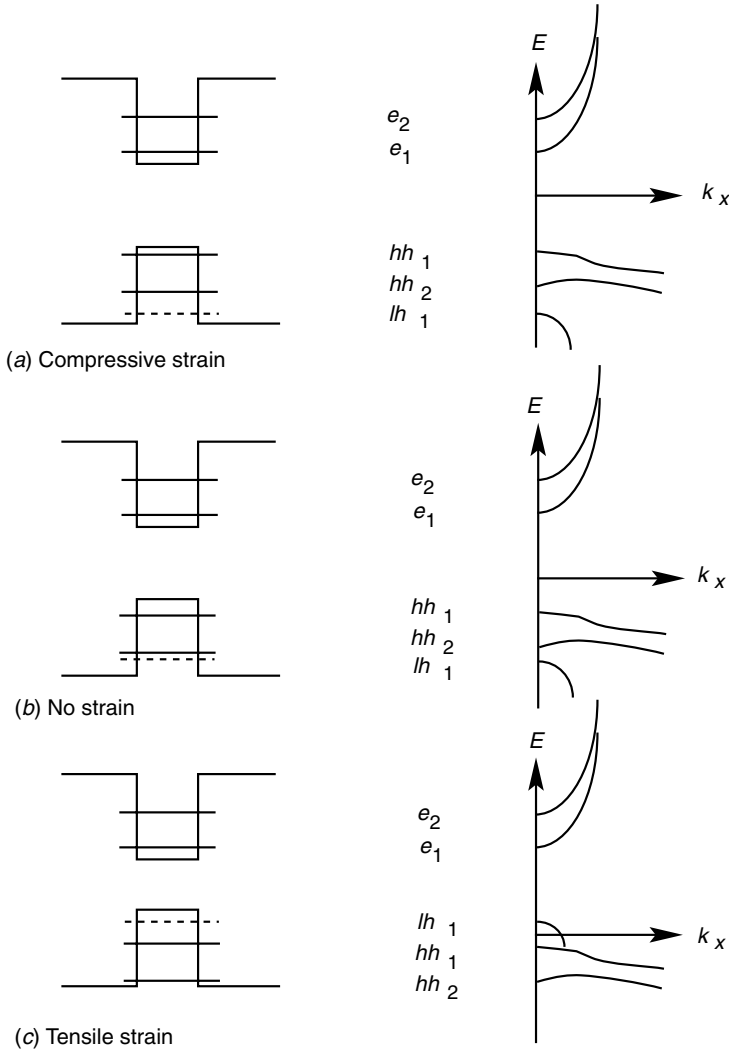


Fig. 13.B.3. As the strain in the quantum changes from compressive to tensile, the light hole subband gradually moves from sitting below hh_2 to above hh_1.

lar (a few tens of a per cent) but can be important nonetheless for long wavelength devices (e.g. for telecommunications applications at 1.55 μm).

Example

(Consult the associated table on p. xvii.)

We consider a 110 Å thick $In_{0.47}Ga_{0.53}As/InP$ quantum well. As $a(0.53) < a(0.47) = a_0$, the quantum well is tensely strained. In fact, by linear interpolation:

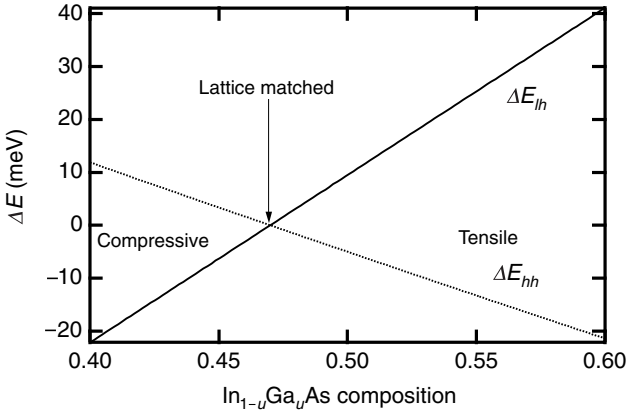


Fig. 13.B.4. Variation in the positions of the heavy hole and light hole bands in the $In_{1-u}Ga_uAs/InP$ material system.

$a_{InAs} = 6.0584\,\text{Å}$; $a_{GaAs} = 5.6533\,\text{Å}$
$a(0.53) = 0.53 \times 5.6533\,\text{Å} + 0.47 \times 6.0584\,\text{Å} = 5.8437\,\text{Å}$
as $a_{InP} = 5.8688\,\text{Å}$, $\delta a/a_0 = [a_0 - a(u)]/a_0 = 4.277 \times 10^{-3}$

The effective masses can also be obtained by linear interpolation between InAs and GaAs:

$m_c = 0.53 \times 0.067 + 0.47 \times 0.023 = 0.0463$
$m_{hh} = 0.53 \times 0.5 + 0.47 \times 0.4 = 0.453$
$m_{lh} = 0.53 \times 0.087 + 0.47 \times 0.026 = 0.0583$

which leads to $R_{c,hh} = m_{hh}/m_c = 9.78$ and $R_{c,lh} = m_{lh}/m_c = 1.259$.

The bandgap for unstrained bulk InGaAs is given by (13.B.3) to be $E_g = 0.807\,\text{eV}$. The mechanical physical parameters for this system are:

$\sigma = 0.33$, $e_c = -6.2\,\text{eV}$, $e_v = 1.1\,\text{eV}$, $e_s = -1.7\,\text{eV}$

The conduction band therefore shifts by $2 \times (-6.2) \times 4.277 \times 10^{-3}\,\text{eV} = -0.053\,\text{eV}$. The coefficients $P(0.53)$ and $Q(0.53)$ are, respectively, $-2 \times 1.1\,\text{eV} \times 4.277 \times 10^{-3}\,\text{eV} = -9.4\,\text{meV}$ and $2 \times 1.7\,\text{eV} \times 4.277 \times 10^{-3}\,\text{eV} = 14.5\,\text{meV}$. The displacements in the heavy hole and light hole valence subbands due to the strain are given by (13.B.4), i.e. $\delta E_{hh} = -\text{P} - \text{Q} = -5.1\,\text{meV}$ and $\delta E_{lh} = -\text{P} + \text{Q} = +23.9\,\text{meV}$ (see Fig. 13.B.4). The light hole subband lh_1 therefore resides above the heavy hole subband hh_1. The effective bandgap is then given by the energy separation $e_1 - lh_1 = 0.800\,\text{eV}$, allowing laser emission at $\lambda = 1.55\,\mu\text{m}$.

A reduction in the transparency carrier density can be obtained by comparing the solutions to the implicit Eq. (13.B.1) for $R_{c,hh}$ and $R_{c,lh}$, or $(n/n_c)_{c,hh} = 1.79$ and $(n/n_c)_{c,lh} = 0.78$. This results in a lowering of the transparency density by a factor of 2.3.

For emission at 1.55 μm, the unstrained quantum well should be of the order of 60 Å in thickness. The transparency current gain between the strained and unstrained structures is then only $60/110 \times 2.3$ or 1.3 and is in agreement with experiment.

FURTHER READING

S. L. Chuang, *Physics of Optoelectronic Devices*, Wiley Interscience, New York (1995).
E. Yablonovitch and E. O. Kane, *J. Lightwave Technol* **LT-4**, 504 (1986).

13.C Vertical cavity surface emitting lasers (VCSELs)

Edge emitting semiconductor lasers present a certain number of inconveniences. To begin with, they tend to be multimodal (principally due to spatial hole burning – see Fig. 4.11). This behaviour generates noise, most notably as a result of mode hopping due to small thermal fluctuations. Additionally, the laser emission is divergent making efficient coupling to optical fibres, for example, a significant technological challenge (see Fig. 13.31). Some of these inconveniences can be side-stepped by using *vertical cavity surface emitting lasers* (VCSELs – pronounced 'vixels').

The underlying concept is quite simple. A laser cavity is fashioned along the epitaxial growth axis allowing controlled deposition of extremely thin layers (with thicknesses less than a micrometre) and perfectly parallel cavity mirrors (precise to within a monolayer). The mirrors are Bragg reflectors formed by epitaxially depositing alternating semiconductor layers of appropriate thickness and composition. Metallic mirrors are not employed for several reasons: sufficiently high quality epitaxial deposition of metals onto III–V semiconductors is not possible, metals absorb infrared light and would adversely affect the threshold currents in such devices. Figure 13.C.1 shows a typical VCSEL consisting of: a Fabry–Pérot cavity (see Complement 9.D) defined by a bottom Bragg mirror with N_1 *quarter-wave* GaAs/AlAs bilayers doped *n* type and possessing a very high reflectance (>99.9%), a half-wave cavity $\lambda/2$, and an upper Bragg output mirror with N_2 quarter-wave GaAs/AlAs bilayers doped *p* type and possessing a high reflectance ($\approx$99%).

13.C.1 Conditions for achieving threshold in a VCSEL

We can easily estimate the performance characteristics of VCSELs starting with a few simple considerations. A VCSEL can be approximated as a Fabry–Pérot cavity of thickness $L = \lambda/2n_{sc}$ sandwiched between two metallic mirrors. The

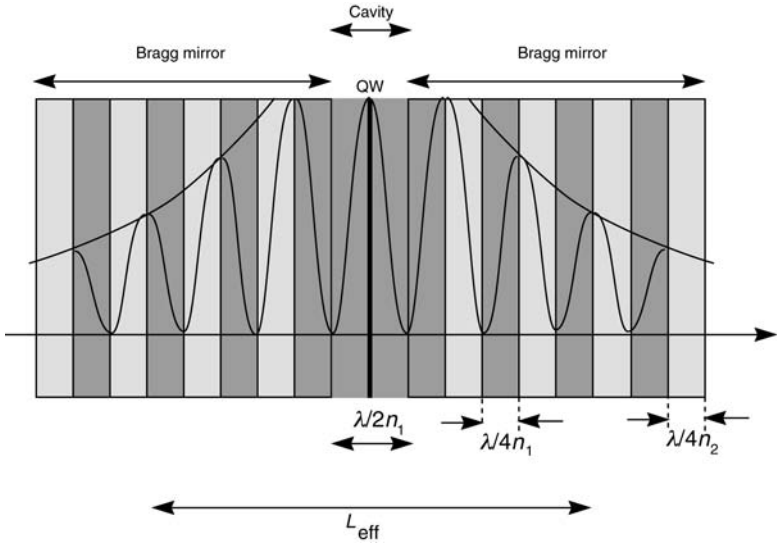


Fig. 13.C.1. A VCSEL consisting of quantum wells grown in a Fabry–Pérot cavity sandwiched between two Bragg mirrors. The figure shows the distribution of the electromagnetic field within the cavity.

cavity mode spacing is in this case $\Delta v = c'/2L = c'/\lambda$ ($c' = c/n_{sc}$), i.e. $\Delta hv = 1.3\,\text{eV}$ for $L = 0.16\,\mu\text{m}$ assuming typical values of $n_{sc} = 3$ and $v = 1\,\mu\text{m}$. As the mode spacing is much larger than the gain bandwidth, the laser operates in a *single longitudinal mode* (see Fig. 13.C.2).

The threshold for laser oscillation is then given by:

$$\Gamma\gamma_{\text{threshold}} = \alpha_p + \frac{1}{2L_{\text{eff}}}\ln\frac{1}{R_{m1}R_{m2}} \approx \alpha_p + \frac{1}{2L_{\text{eff}}}(T_{m1} + T_{m2}) \tag{13.C.1}$$

where R_{m1}, T_{m1}, R_{m2}, and T_{m2} are the reflectances and transmittances for the upper and lower mirrors, respectively; α_p is the parasitic absorption; and Γ is the overlap coefficient between the electromagnetic wave and the quantum well. L_{eff} is the effective cavity length: in the case of a VCSEL, this includes the distance over which the wave penetrates into the Bragg mirror or, alternatively, the distance over which the photons enter into the mirrors through tunnelling. While the calculation of L_{eff} is not simple, we will see that the determination of its value is not primordial in being able to estimate the performance characteristics of VCSELs.

The principal aspects of VCSELs are resumed in (13.C.1): while the overlap coefficient Γ is fairly weak, this can be compensated for by using mirrors with low transmittances. We are in a position to estimate each of these parameters. The transmittances T_m for Bragg mirrors were calculated in Complement 9.D (see (9.D.39)):

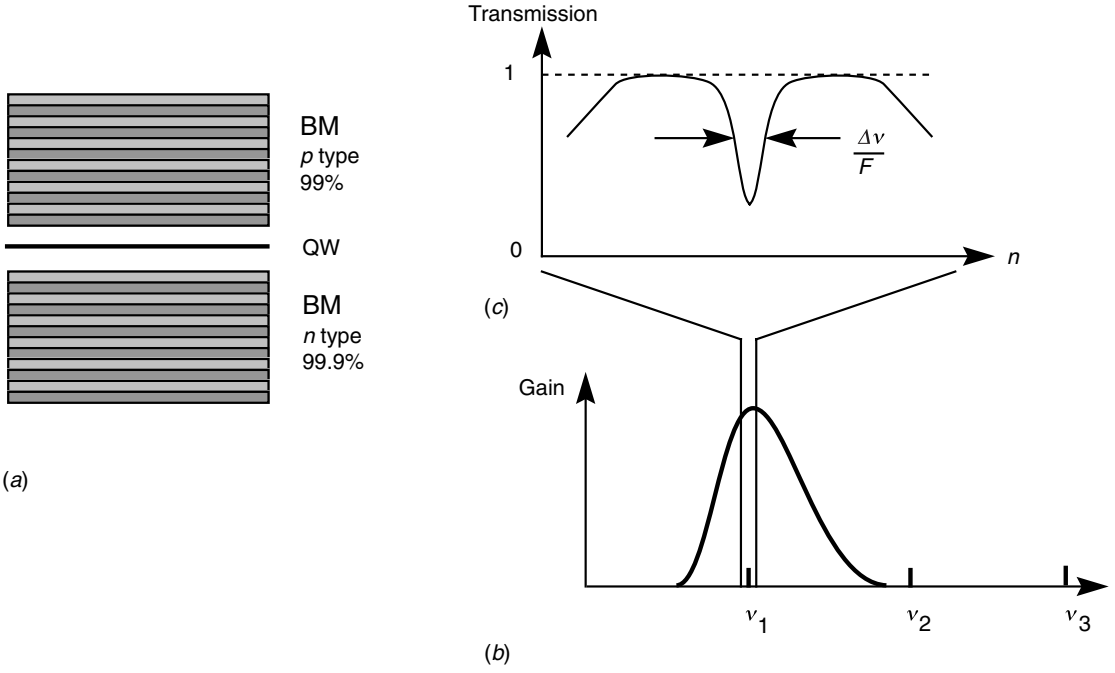


Fig. 13.C.2. In a VCSEL, the mode spacing is greater than the width of the gain spectrum. A VCSEL will therefore lase over a single longitudinal mode.

$$T_m = 1 - \left[\frac{1 - \left(\frac{n_2}{n_1}\right)^{2N}}{1 + \left(\frac{n_2}{n_1}\right)^{2N}} \right]^2 \tag{13.C.2}$$

where n_1 and n_2 are the optical indices of the Bragg mirrors' constituent layers. The overlap coefficient Γ is given by the integral (9.27):

$$\Gamma = \frac{\int_{-d/2}^{d/2} |E(z)|^2 \mathrm{d}z}{\int_{-\infty}^{+\infty} |E(z)|^2 \mathrm{d}z} \tag{13.C.3}$$

where d is the thickness of the active region. In order to estimate this integral, we have made use of the approximation to (9.D.26) that consists of taking the field $E(z)$ as a stationary wave:

$$E(z) = E_0 \cos\left(\frac{\pi z}{L_{\text{eff}}}\right), \qquad -L_{\text{eff}}/2 < z < L_{\text{eff}}/2 \tag{13.C.4}$$

and $E(z) = 0$ otherwise. This last approximation neglects the electromagnetic field beyond the tunnelling distance of the photon in the Bragg mirrors. Integral (13.C.3) can then be easily calculated:

$$\Gamma = \frac{dE_0^2}{\int_{-\infty}^{+\infty} E_0^2 \cos^2(\pi z/L_{\text{eff}}) \mathrm{d}z} \tag{13.C.5}$$

and leads to:

$$\Gamma = \eta \frac{d}{L_{\text{eff}}} \tag{13.C.6}$$

where $\eta = 2$ for approximation (13.C.4) but in general resides between 1 and 2. Integral (13.C.3) shows the importance of the location of the amplifier relative to the antinode of the electromagnetic mode. If the quantum well was situated at a node, rather than at an antinode, the effective gain $\Gamma\gamma$ would be virtually null. This result can also be obtained using (4.C.12) (derived during the study of electrodynamic laser equations in Complement 4.C) and by taking (13.C.4) for $E(z)$ as the only allowed mode in the cavity. Substituting (13.C.5) into (13.C.1) and neglecting parasitic losses, the required gain at threshold is given by:

$$T_{m1} + T_{m2} = 2\eta d\gamma_{\text{threshold}} \tag{13.C.7}$$

Threshold condition for a VCSEL

We are now ready to determine the typical conditions for VCSEL operation. We will focus on the GaAs/AlGaAs system as it is by far the most commonly used material system for implementing these devices. The active medium of our VCSEL will be assumed to be a 100 Å ($d = 10^{-6}$ cm) thick quantum well. We want the threshold current to be 1.2 mA for a $30 \times 30\,\mu\text{m}^2$ device (i.e. a threshold current density $J_{\text{threshold}} = 136\,\text{A cm}^{-2}$). Assuming a lifetime t_{tot} of 1.6 ns, this results in a two-dimensional carrier density of $n_s = J_{\text{threshold}} t_{\text{tot}}/q$, or $1.36 \times 10^{12}\,\text{cm}^{-2}$ (see (13.3)). The threshold gain is then given by (13.57) to be:

$$d\gamma_{\text{threshold}} = A_{2D}(1 - \mathrm{e}^{-n_s/n_c} - \mathrm{e}^{-n_s R_{cv}/n_c}) \tag{13.C.8}$$

where the coefficients $\alpha_{2D} d = A_{2D} = 0.55\%$, $n_c = 7.25 \times 10^{11}\,\text{cm}^{-2}$, and $R_{cv} = 6.8$ is as obtained in the example in Section 13.6.1. This gives $d\gamma_{\text{threshold}} = 4.9 \times 10^{-4}\alpha_{2D}$, i.e. $\gamma_{\text{threshold}} = 490\,\text{cm}^{-1}$. Equation (13.C.7) then yields a sum for $T_{m1} + T_{m2}$ of 9.8×10^{-4}. Taking $T_{m1} = T_{m2}$ and using (13.C.2), we see that we can select a stack comprising 30 GaAs/AlAs bilayers for the bottom mirror and 30 bilayers for the upper mirror (further assuming $\lambda \approx 1\,\mu\text{m}$, $n_1 = 3.4$, and $n_2 = 2.96$).

Fig. 13.C.3. Variation of a GaAs/AlAs VCSEL threshold current as a function of the number of Bragg mirror pairs ($S = 30 \times 30\,\mu m^2$).

Figure 13.C.3 shows the calculated decrease of threshold current as a function of the number of Bragg mirror pairs.

13.C.2 VCSEL performance

In addition to single mode operation, VCSELs possess other advantages, such as low threshold currents and reduced output beam divergence. The divergence of the output beam can be easily calculated using (4.E.5) for Gaussian beams:

$$\theta = \frac{\lambda}{\pi W_0} \tag{13.C.9}$$

where W_0 is the width of the output beam. For a VCSEL with a diameter of 10 μm (or $W_0 = 5\,\mu m$) operating at a wavelength of 1 μm, we obtain a divergence of $1/5\pi$ radians, or 4° in all directions.

The low threshold current results primarily from the small volume of material which needs to be inverted. In the case of a ridge waveguide laser without antireflection coatings on the facets, a 200 μm long cavity is generally required for sufficient gain. Assuming a quantum well thickness of 100 Å, and a device width of 3 μm, this corresponds to an active volume of 6 μm^3. This is to be compared with an active volume of $100\,Å \times (10\,\mu m)^2$ or 1 μm^3 in the case of a VCSEL. The threshold currents can be further reduced by oxidizing the device side walls to increase the lateral electrical and optical confinement. VCSELs with 3 μm diameter cavities fabricated in this manner have demonstrated threshold currents of the

order of a few μA (still corresponding, however, to threshold current densities of the order of 100 A cm^{-2}!).

FURTHER READING

Special issue on semiconductor lasers, *IEEE J. Quantum Electron.* June (1991).

13.D Thermal aspects of laser diodes and high power devices

One of the particularities of laser diodes in comparison with gas or ion lasers in solid matrices (such as Nd: YAG) is that the active material in which the carriers recombine (radiatively or otherwise) occupies a minute fraction of the total structure. A laser diode with a quantum efficiency $\eta = 1 - \xi < 1$, therefore dissipates a portion ξP of the total power as heat in a small volume of material. This can lead to significant heating of the material, which can perturb the operational characteristics of the device and even result in device failure (chemical degradation, melting, etc.). We will now sketch the theoretical framework that will allow us to examine this heating and its effects on the device characteristics of laser diodes. For simplicity, we will limit our treatment to a one-dimensional model.

Our starting point will be the *Newton–Fick law*, which states that the heat flux Φ (W) across a boundary in a material possessing an inhomogeneous temperature distribution (see Fig. 13.D.1) is proportional to the temperature gradient at that location, i.e:

$$\Phi(x) = -\kappa \nabla T(x) \tag{13.D.1}$$

The Newton–Fick law

where κ is the *thermal conductivity* in W cm^{-1} K^{-1}. This heat flux leads to a variation in the *internal energy* E_{int} (J cm^{-3}) of each volume element, and can be expressed in terms of energy conservation as:

$$\frac{\partial}{\partial t} E_{\text{int}}(x,t) + \nabla \Phi(x,t) = 0 \tag{13.D.2}$$

From thermodynamics, this internal energy may be written:

$$E_{\text{int}} = \rho c_p T \tag{13.D.3}$$

where ρ is the volumetric density (in g cm^{-3}) and c_p is the *specific heat* of the material (J K^{-1} g^{-1}). If the material possesses internal heat sources distributed according to $S(x,t)$ (W cm^{-3}), the ensemble of Eqs. (13.D.1)–(13.D.3) leads to Fourier's *heat equation*:

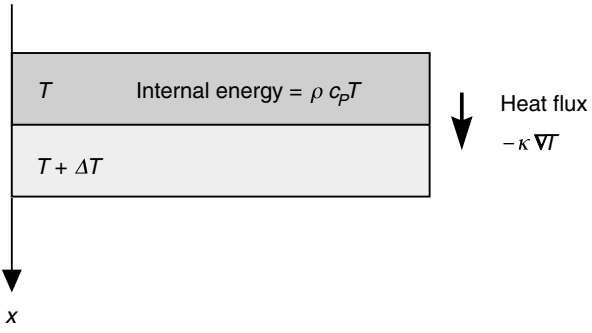


Fig. 13.D.1. Illustration of Newton–Fick law.

$$\frac{\partial T}{\partial t} - D\frac{\partial^2 T}{\partial x^2} = \frac{S(x,t)}{\rho c_p} \tag{13.D.4}$$

The Fourier heat equation

where D is the *thermal diffusion coefficient* ($\mathrm{cm^2\,s^{-1}}$) given by:

$$D = \frac{\kappa}{\rho c_p} \tag{13.D.5}$$

Thermal diffusion coefficient

We will now seek solutions of the Fourier equation for two illustrative examples, which will enable us to understand the thermal characteristics of laser diodes better.

Emissive surface in an infinite medium: transient response

This example describes the transient response of a laser diode in the absence of a heat sink. We define an effective surface heat source, which includes all the dissipative mechanisms occurring in the structure (ohmic contacts, free carrier absorption, non-radiative recombination, . . .) as a result of a current pulse:

$$S(x,t) = \xi P \delta(x) H_\tau(t) \tag{13.D.6}$$

where $\delta(x)$ is the Dirac delta function ($\mathrm{cm^{-1}}$) and $H_\tau(t)$ is the boxcar function, which equals 1 between 0 and τ and zero otherwise (dimensionless). For all time t, the variation in the slope of $T(x)$ anywhere across the emissive surface can be obtained by integrating (13.D.4) between $-\varepsilon$ and $+\varepsilon$ and by recalling that T must be continuous at $x = 0$:

$$\int_{-\varepsilon}^{+\varepsilon} \frac{\partial T}{\partial t}\mathrm{d}x - D\int_{-\varepsilon}^{+\varepsilon} \frac{\partial^2 T}{\partial x^2}\mathrm{d}x = \frac{\xi P}{\rho c_p} H_\tau(t) \int_{-\varepsilon}^{+\varepsilon} \delta(x)\mathrm{d}x \tag{13.D.7}$$

or

$$2D\frac{\partial T}{\partial x}\bigg|_{x=0} = -\frac{\xi P}{\rho c_p} H_\tau(t) \tag{13.D.8}$$

where we have assumed the medium to be symmetric about $x = 0$. This last expression implies that the heat flux ξP is shared equally by both sides of the junction as a heat flux $D\partial T/\partial x$.

In order to solve (13.D.4), we take the temporal Laplace transform:

$$D\frac{\partial^2}{\partial x^2}\tilde{T}(x,s) - s\tilde{T}(x,s) = -\frac{\xi P}{\rho c_p}\delta(x)\left[\frac{1-\mathrm{e}^{-s\tau}}{s}\right] \tag{13.D.9}$$

where the term in square brackets is the Laplace transform of the boxcar function. The homogeneous solution to the left-hand side of the equation is of the form (for $x > 0$):

$$\tilde{T}(x,s) = A(s)\mathrm{e}^{-\sqrt{s/D}x} \tag{13.D.10}$$

The function $A(s)$ can then be obtained from condition (13.D.8), such that:

$$A(s) = \frac{1}{2}\frac{\xi P}{\rho c_p\sqrt{D}}\frac{1-\mathrm{e}^{-s\tau}}{s^{3/2}} \tag{13.D.11}$$

This allows us to write the temporal Laplace transform for the temperature distribution as:

$$\tilde{T}(s) = \frac{1}{2}\frac{\xi P}{\rho c_p\sqrt{D}}\,\mathrm{e}^{-\sqrt{s/D}x}\,\frac{1-\mathrm{e}^{-s\tau}}{s^{3/2}} \tag{13.D.12}$$

The inverse Laplace transform of this last expression exists but is fairly complicated (as we will see later on). We will interest ourselves for the time being with heating of the junction, i.e. with the situation at $x = 0$. The inverse Laplace transform of (13.D.12) for $x = 0$ is trivial and is given by:

$$T(t) = \frac{\xi P}{\rho c_p\sqrt{\pi D}}\left(\sqrt{t} - \sqrt{t-\tau}\right) \tag{13.D.13}$$

The maximum temperature $T_{\max}$ therefore occurs at the end of the current pulse (at $t = \tau$), which is not surprising:

$$T_{\max} = \frac{\xi P\sqrt{\tau}}{\rho c_p\sqrt{\pi D}} = \frac{\xi E_P}{\sqrt{\pi}\rho c_p L_D} \tag{13.D.14}$$

In writing (13.D.14) we introduced the energy of the light pulse $E_P = P\tau$ and a *thermal diffusion length* $L_D = (D\tau)^{1/2}$. Written in these terms, (13.D.14) is easy to understand: the heating that takes place results from the dissipation of a quantity

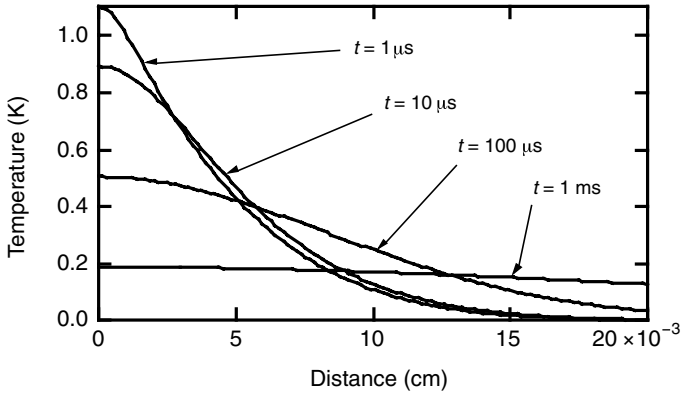


Fig. 13.D.2. Evolution of the temperature increase in a GaAs laser diode as a function of distance from the junction for various times after a 200 W cm^{-2} thermal pulse of 100 μs duration.

of heat ξE_P into a layer of material with thickness $L_D\sqrt{\pi}$. This equation also shows that in this configuration, the diode heats continuously as a function of the pulse duration τ and leads inevitably to device destruction if the pulse is too long. It is therefore necessary to find a means of dissipating heat from the junction region. Figure 13.D.2 shows the complete solution for the thermal distribution as a function of time as given by the inverse Laplace transform of (13.D.12), i.e:

$$T(t) = \frac{\xi P}{2\rho c_p \sqrt{D}}[f(t) - f(t-\tau)]$$
$$f(t) = 2\sqrt{\frac{t}{\pi}}\exp\left(-\frac{x^2}{4Dt}\right) - \frac{x}{\sqrt{D}}\operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \qquad (13.D.15)$$

where erfc is the complementary error function 1-erf, for the conditions given in the following example.

Example

The relevant physical constants for GaAs from Table 13.D.1 are:

$\rho = 5.32\,\mathrm{g\,cm^{-3}}$
$c_p = 0.35\,\mathrm{J\,g^{-1}\,K^{-1}}$
$\kappa = 0.46\,\mathrm{W\,cm^{-1}\,K^{-1}}$
$D = 0.25\,\mathrm{cm^2\,s^{-1}}$

For a diode dissipating an amount of heat ξP (W cm^{-2}), with a pulse length of τ (μs), the temperature increase is given by:

$$T_{\max} = 6 \times 10^{-4}\xi P(\tau(\mu s))^{1/2}$$

Table 13.D.1. *Thermal constants for GaAs and InP*: ρ *is the mass density*, c_p *the specific heat*, κ *the thermal conductivity, and* D *the thermal diffusion coefficient*

Material	$\rho\,(\mathrm{g\,cm^{-3}})$	$c_p\,(\mathrm{J\,g^{-1}\,K^{-1}})$	$\kappa\,(\mathrm{W\,cm^{-1}\,K^{-1}})$	$D\,(\mathrm{cm^{-2}\,s^{-1}})$
GaAs	5.32	0.35	0.46	0.25
InP	4.787	0.34	0.67	0.41

A 100 W laser bar has a width of 1 cm and a length of 500 μm. Assuming a heating coefficient ξ of 10% implies a heat load of $\xi P = 2 \times 10^2\,\mathrm{W\,cm^{-2}}$. A 1 ms pulse width leads to a temperature increase of 3.8 K, while after 1 s the junction temperature has risen to 120 K, leading to irreversible damage of the laser diode.

Thermal dissipation on both sides of the junction under continuous operation

In order to prevent heat build up, thermal dissipaters must be placed on either side of the junction, at for instance $x = +L$ and $x = -L$. The heat equation, (13.D.4), which describes the thermal response, remains the same except that different boundary conditions are used ($T(L,t) = T(-L,t) = 0$ for all time t). Expecting such a system to possess a stationary solution, the stationary Fourier equation for T may be written:

$$-\frac{\partial^2 T_\infty}{\partial x^2} = \frac{\xi P}{\kappa}\delta(x) \tag{13.D.16}$$

Thanks to the first integral in (13.D.8), this last equation can be integrated without difficulty, yielding $T(x) = (\xi P/2\kappa)(L - x)$ for $x > 0$. The steady state temperature drop between the source and the thermal dissipater is therefore linear. The increase in the junction temperature relative to the heat sink is therefore:

$$\Delta T_{\max} = \frac{\xi P L}{2\kappa} \tag{13.D.17}$$

which may be interpreted as the capacity of the heat dissipater to extract a power of $\xi P/2$ over a distance L given a thermal conductivity κ for the material. It is then clearly advantageous to be able to minimize the distance between the heat dissipater and the junction.

In general, (13.D.17) may be written as:

$$\Delta T_{\max} = R_{\mathrm{th}}\xi P \tag{13.D.18}$$

where R_{th} is the *thermal impedance* of the system ($\mathrm{cm^2\,K^{-1}\,W^{-1}}$).

For the case of a heat sink situated a distance L from the junction (see Fig. 13.D.3), the thermal resistance is given by (13.D.18) to be:

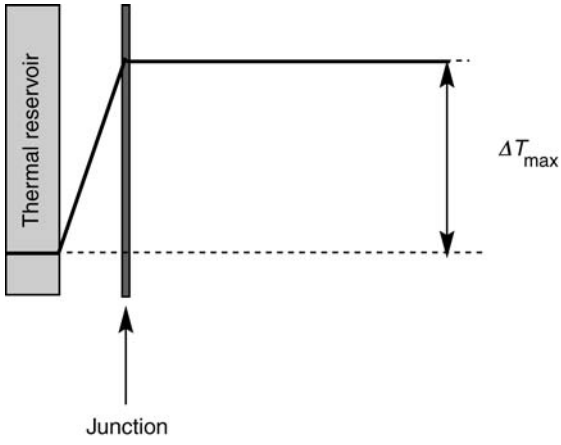


Fig. 13.D.3. The presence of a constant temperature heat sink located a distance L from the junction imposes a steady state temperature on the junction.

$$R_{th} = \frac{L}{\kappa} \tag{13.D.19}$$

Thermal impedance for a heat sink situated a distance L from a junction ($cm^2\,K^{-1}\,W^{-1}$)

Example

Continuing from the previous example: we calculate the stationary temperature increase for the same junction, this time with the presence of a heat dissipater located 5 μm from the junction. We calculate a thermal resistance of $5 \times 10^{-4}\,cm/0.46\,W\,cm^{-1}\,K^{-1}$ or $1.1 \times 10^{-3}\,cm^2\,K^{-1}\,W^{-1}$. The associated temperature increase, ΔT_{max} equals $2 \times 10^2\,W\,cm^{-2} \times 1.1 \times 10^{-3}\,cm^2\,K^{-1}\,W^{-1}$ or 0.22 K. Comparing this quantity to the previous value of 3.8 K, gives some indication of the advantage gained by using heat dissipaters.

Aside from the danger of melting the laser diode, the consequences from allowing the junction temperature to increase unchecked are numerous. First, given thermal expansion of the semiconductors, the bandgaps decrease with increasing temperature and the emission wavelength of the device redshifts. For a typical GaAs/AlGaAs laser diode, the emission wavelength will redshift by approximately $3.2\,Å\,K^{-1}$. This may have a dramatic influence in applications in which the laser diodes are used to pump a laser material; the redshift will lead to a decrease of the overlap between the laser diode wavelength and the ion optical cross-section (see Complement 4.E).

Another effect associated with increasing temperatures, is a rise in the required overall threshold current, $I_{threshold}$. The physical mechanisms responsible for this are innumerable (Auger effect, thermal activation of non-radiative recombination mechanisms, . . .). The dependence of $I_{threshold}$ on the device temperature can be

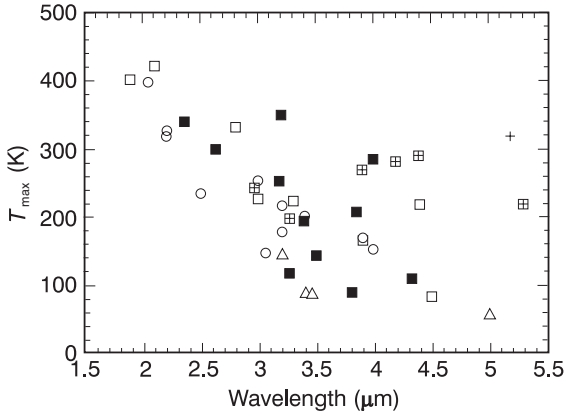


Fig. 13.D.4. Maximum operating temperatures for semiconductor lasers emitting between 1.5 and 5.5 μm: (○) type I, III–V Sb heterostructures, (□) type I, III–V Sb multiple quantum well structures, (■) type II and III, III–V Sb heterostructures, (⊞) lead salts, (△) HgCdTe heterostructures, (+) quantum cascade lasers. (A. Joullié, private communication.)

modelled according to the phenomenological relationship:

$$I_{\text{threshold}} = I_0 e^{T/T_0} \tag{13.D.20}$$

where the phenomenological parameter T_0 is referred to as the *characteristic temperature* of the laser. As a rule of thumb, the larger the bandgap of the semiconductor and the deeper the wells, the higher the corresponding characteristic temperature T_0 and the more stable the device will be in regards to temperature. Table 13.D.2 shows typical T_0s obtained for a variety of material systems. Laser diodes fabricated from GaAs/AlGaAs heterostructures possess characteristic temperatures of ~100 K, whereas mid-infrared lasers fashioned from InGaSb/InAs have T_0 values of ~30 K indicating that they only operate at low temperatures. Figure 13.D.4 shows the maximum operational temperatures for laser diodes with emission wavelengths situated between 1.5 and 5.5 μm.

Table 13.D.2. *Characteristic temperatures for laser diodes implemented in various material systems*

Heterostructure	Wavelength (μm)	T_0 (K)
GaN/InGaN	0.40	180
GaAs/AlGaAs	0.85	150
InGaAs/GaAs	0.98	90
InGaAs/AlGaAs/InP	1.55	60
InGaSb/InAs/GaSb	3–4	35

13.E Spontaneous emission in semiconductor lasers

In all preceding derivations, we have neglected the influence of spontaneous emission on the behaviour of semiconductor lasers. As semiconductors provide ample gain, significant amplification of spontaneous emission occurs (referred to as *ASE* or *amplified spontaneous emission*). It turns out that ASE plays a fundamental role in determining the elementary device characteristics of laser diodes. We devote the present section to examining its influence on laser diode operation.

The system of equations (13.69) describes the coupling between the carrier density n and the photon density s in the cavity. As we indicated in Complement 4.A, to take spontaneous emission into account, we need only add a term (independent of the number of photons s) for the additional photon generation R_{spon}. Equation (13.69) then becomes:

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= \frac{J}{qd} - \gamma_{\text{tot}} n - c' g s (n - n_{\text{tr}}) \\ \frac{\mathrm{d}s}{\mathrm{d}t} &= c' g s (n - n_{\text{tr}}) \Gamma - \gamma_c s + R_{\text{spon}} \end{aligned} \tag{13.E.1}$$

The radiative emission rate was calculated in Chapter 7 and is given by Eq. (7.61) $R_{\text{spon}} = \gamma_{\text{spon}} n$. In this formula, γ_{spon} is the spontaneous emission rate given by $\gamma_{\text{spon}} = Bp$, where $p\,(=n$ if we hold to our assumption that all radiative recombination occurs in the undoped region) is the density of holes and B is the bimolecular recombination coefficient. We must nonetheless take into account the fact that all the photons emitted in the active region are not released into guided modes (see Fig. 13.E.1). Only a fraction β of all spontaneously emitted photons (referred to as the *spontaneous emission factor*) will be coupled into the waveguide and amplified. The calculation for obtaining this β coefficient is rather involved and not particularly reliable. It is therefore customary to treat it as an adjustable parameter (i.e. determined from curve fitting) with a value lying between 10^{-4} and 10^{-5} for ridge waveguide lasers.

At stationary state, (13.E.1) provides the electron and photon densities:

$$\begin{aligned} s &= \frac{R_{\text{spon}}}{\gamma_c - \Gamma c' g (n - n_{\text{tr}})} = \beta \frac{\gamma_{\text{spon}}}{\gamma_c - \Gamma c' g (n - n_{\text{tr}})} n \\ \frac{J}{qd} &= \gamma_{\text{tot}} n + c' g (n - n_{\text{tr}}) s \end{aligned} \tag{13.E.2}$$

The first of these two equations is particularly useful and expresses the fact that the laser oscillations build up from spontaneous emission once the gain begins to

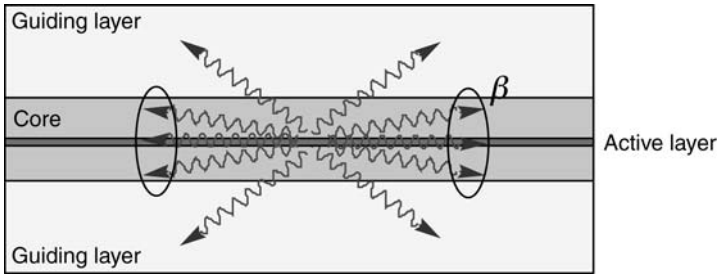


Fig. 13.E.1. Of all spontaneously emitted photons in the structure, only a fraction β are released into the guided mode and subject to amplification.

exceed the cavity losses. The threshold density $n_{\text{threshold}}$ is then given by:

$$n_{\text{threshold}} = n_{\text{tr}} + \frac{\gamma_c}{\Gamma c' g} \tag{13.E.3}$$

The second equation in (13.E.2) can be expressed more concisely by introducing the dimensionless variables:

$$x = \frac{n}{n_{\text{tr}}};\ y = \frac{J}{J_{\text{tr}}} \text{ and } \alpha = \frac{n_{\text{threshold}}}{n_{\text{tr}}} = 1 + \frac{\gamma_c}{\Gamma c' g n_{\text{tr}}} \tag{13.E.4}$$

where $J_{\text{tr}} = qd\gamma_{\text{tot}} n_{\text{tr}}$ is the transparency current density. We thus obtain an implicit relationship between the carrier density and the pump current, written in this more compact notation as:

$$y = x + \frac{\beta}{\Gamma}\eta \frac{x(x-1)}{\alpha - x} \tag{13.E.5}$$

where η is the quantum efficiency given by $\eta = \gamma_{\text{spon}}/\gamma_{\text{tot}} = \tau_{\text{tot}}/\tau_{\text{spon}}$. To find the dependence $s(J)$ of the photon density on the pump current, we need only solve this implicit equation and substitute its solution into the first equation in (13.E.2), yielding:

$$s = n_{\text{tr}}\beta(\alpha - 1)\frac{\gamma_{\text{spon}}}{\gamma_c}\frac{x}{\alpha - x} \tag{13.E.6}$$

Equations (13.E.5) and (13.E.6) are the two parametric equations in x, which will allow us to calculate the characteristics $s(J)$. If $\beta = 0$, we recover the laser diode response described in the bulk of Chapter 13, where the carrier density n increased in proportion to the electrical current $J = q\gamma_{\text{tot}} dn$ below threshold and clamped abruptly to $n = n_{\text{tr}}$ above threshold. If $\beta \neq 0$, the behaviour is more gradual as depicted in Fig. 13.E.2.

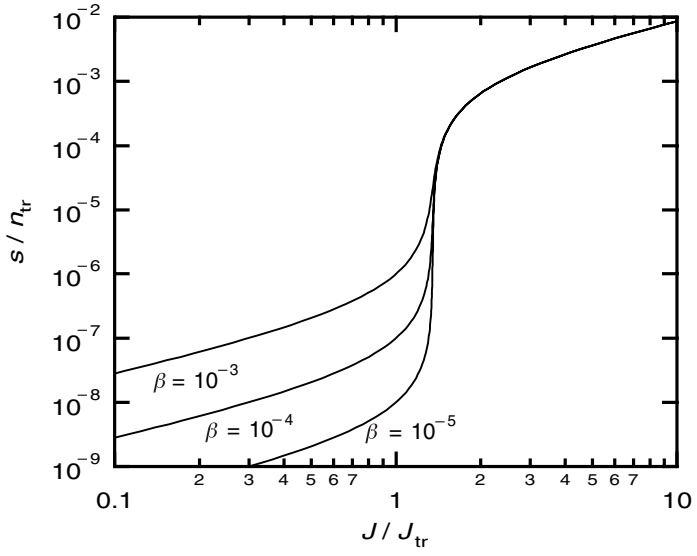


Fig. 13.E.2. Evolution of the photon density s (normalized by the transparency threshold carrier density n_{tr}) as a function of pump current density (normalized by the transparency current density J_{tr}) for different values of the spontaneous emission factor β. Other implicit physical parameters are as listed in the following example.

Example

The curves in Fig. 13.E.2 were calculated assuming a quantum efficiency η of 0.5, a cavity lifetime τ_c of 2 ps, a spontaneous lifetime τ_{spon} of 2 ns, a confinement factor Γ of 0.5, a transparency density n_{tr} of $1.25 \times 10^{18}\,\text{cm}^{-3}$, and a dynamic gain g of $2.5 \times 10^{-16}\,\text{cm}^2$ or $\alpha = 1.352$. The MATHEMATICA program given below can be used to obtain the laser characteristics near threshold.

```
beta=.;gamma=0.5;tauc=1/(2. 10^-12);taus=1/(2. 10^-9);
eta=0.5;g=2.5 10^-16; ntr=1.25 10^18;nop=3.3;
c'=3. 10^10/nop;
alpha=1+ tauc/(gamma*c'*g*ntr);
R=taus/tauc;
betap=(alpha-1)*beta*R;
eq1= yp == alpha*zp/(zp+betap)+beta/betap*zp*
(alpha*zp/(zp+betap)-1);
sol=Last[Solve[eq1, zp]];
t=zp/.sol;
Needs[''Graphics'Graphics'']
beta=10^1-4;
plot1=LogPlot[t,{yp,.01,2}]
beta=10^-3;
plot2=LogPlot[t,{yp,.01,2}]
beta=10^-5;
```

```
plot3=LogPlot[t,{yp,.01,2}]
Show[plot1,plot2,plot3]
```

Figure 13.E.2 can be interpreted as follows: below threshold, photons are spontaneously released into the guided mode of the cavity (in this regime the laser diode behaves as an LED). Above threshold, the photon density grows spectacularly (as previously described in Complement 4.A). Furthermore, in Fig. 13.E.2 the threshold effect becomes increasingly subdued as β tends towards 1.

This picture has been further refined by Yamamoto and colleagues. This author pointed out that the dynamic gain g is itself given by $g = \beta S/\tau_{spon}$, where S is the surface area of the sample. Introducing this expression into (13.E.5) and (13.E.6), he was able to show that, as β tends towards 1, the laser threshold tends towards 0. From this observation evolved the concept of the *zero threshold laser* in which all spontaneously emitted photons contribute to laser emission in the amplified mode.

It is interesting to use the present model to describe the multimodal behaviour of semiconductor lasers near threshold. To do so, we will assume a certain number of modes m are simultaneously available in the cavity. These modes possess a frequency spacing $\Delta\nu_{cav}$ of $c'/2L$. Each mode m is then described by an equation of type (13.E.1), with the additional possibility that the parameters g and β vary according to the mode index:

$$\begin{aligned}\frac{dn}{dt} &= \frac{J}{qd} - \gamma_{tot} n - \sum_m c' g_m s_m (n - n_{tr}) \\ \frac{ds_m}{dt} &= c' g_m s_m (n - n_{tr})\Gamma - \gamma_c s_m + \beta_m \gamma_{spon} n\end{aligned} \tag{13.E.7}$$

Laser equations for homogeneous gain

Note that in these equations all the modes share the same inversion population n; which, by definition, means that the laser medium is assumed *homogeneous*. We assume a Lorentzian gain curve, see Fig. 13.E.3:

$$g = \frac{g_0}{1 + [(\nu - \nu_0)/\Delta\nu]^2} \approx g_0 \left[1 - \left(\frac{\nu - \nu_0}{\Delta\nu}\right)^2\right] \tag{13.E.8}$$

Clearly, as seen in Chapters 7 and 13, semiconductor gain media intrinsically possess more complex gain curves. This formula therefore only offers an approximation to the exact situation (it compares, however, quite favourably with experiment). Writing the total number of amplified modes as $2M + 1$ (with $M = \Delta\nu/\Delta\nu_{cav}$), i.e. those for which $g > 0$, Eq. (13.E.8) gives the amplification factors for each of the modes m (where m ranges from $-M$ to $+M$):

$$g_m = g_0 \left[1 - \left(\frac{m}{M}\right)^2\right] \tag{13.E.9}$$

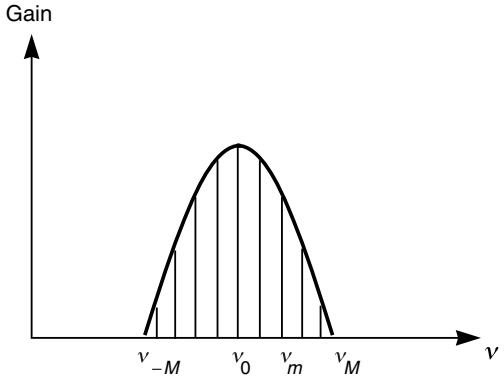


Fig. 13.E.3. Parabolic envelope approximating the gain curve for a semiconductor laser. The number of amplified modes is $2M + 1$.

The system of equations in (13.E.7) can be solved in exactly the same manner as before, leading to an implicit equation for the normalized current density $y = J/J_{\mathrm{tr}}$ as a function of normalized carrier density $x = n/n_{\mathrm{tr}}$:

$$y = x + \beta\frac{\eta}{\Gamma}x(x-1)\sum_{m=-M}^{m=M}\frac{1-(m/M)^2}{(\alpha-1)-(x-1)[1-(m/M)^2]} \tag{13.E.10a}$$

Allowing the number of modes $2M + 1$ to tend towards infinity, this last equation can be rewritten as:

$$y = x + \beta\frac{\eta}{\Gamma}x\int_{-1}^{+1}\frac{1-u^2}{[(\alpha-1/x-1)-1]+u^2}\,\mathrm{d}u \tag{13.E.10b}$$

which can be integrated exactly to give:

$$y = x + 2\beta\frac{\eta}{\Gamma}x\left[\frac{1}{\varepsilon}(1+\varepsilon^2)\arctan\left(\frac{1}{\varepsilon}\right)-1\right] \tag{13.E.11a}$$

with ε being defined as:

$$\varepsilon^2 = \frac{\alpha - x}{x-1} \tag{13.E.11b}$$

These equations are to be interpreted in terms of the absolute values of complex numbers if $\varepsilon^2 < 0$. Similarly, the photon density s_m for the mth mode is obtained according to the stationary value of the second equation in (13.E.7), i.e:

$$s_m = s_0\frac{1}{1+\dfrac{x-1}{\alpha-x}\left(\dfrac{m}{M}\right)^2} \tag{13.E.12}$$

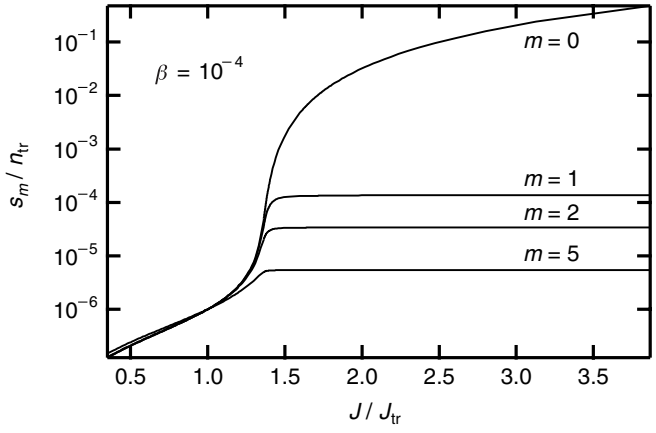


Fig. 13.E.4. Evolution of photon density s_m for the mth mode (normalized by the transparency threshold carrier density n_{tr}) as a function of pump current density J (normalized by the transparency current density J_{tr}) for a spontaneous emission factor β of 10^{-4}. Other relevant physical parameters are given in the following example.

where s_0 is the photon density at the peak of the gain curve given by (13.E.6) or:

$$s_0 = \beta \frac{\gamma_{spon}}{\gamma_c}(\alpha - 1)n_{tr}\frac{x}{\alpha - x} \tag{13.E.13}$$

Here, to simplify matters, we have assumed that the spontaneous emission factor β is mode independent.

Several conclusions can be drawn from this approach. To begin, although we have assumed that this system is homogeneously broadened, we note (see Fig. 13.E.4) that several laser modes can coexist in the cavity. Emission in these modes is seeded by spontaneous emission (as described by β, note that if $\beta = 0, s_m = 0$). Furthermore, as the pump current becomes large ($y \to \infty$), the normalized carrier density x tends towards α and becomes clamped. Whereas the photon density s_0 in the central mode tends towards infinity, the photon density in the mth mode saturates according to:

$$s_m^{\infty} = \beta \frac{\gamma_{spon}}{\gamma_c}(\alpha - 1)n_{tr}\frac{\alpha}{\alpha - 1}\left(\frac{M}{m}\right)^2 \tag{13.E.14}$$

The relative intensity in the lateral modes therefore falls to naught in comparison with the central mode intensity. This behaviour is reproduced experimentally and is presented in Fig. 13.E.5.

Example

As an extension to the previous example, the MATHEMATICA program listed below can be used to calculate the response of a multimode laser as a function of pump current near threshold.

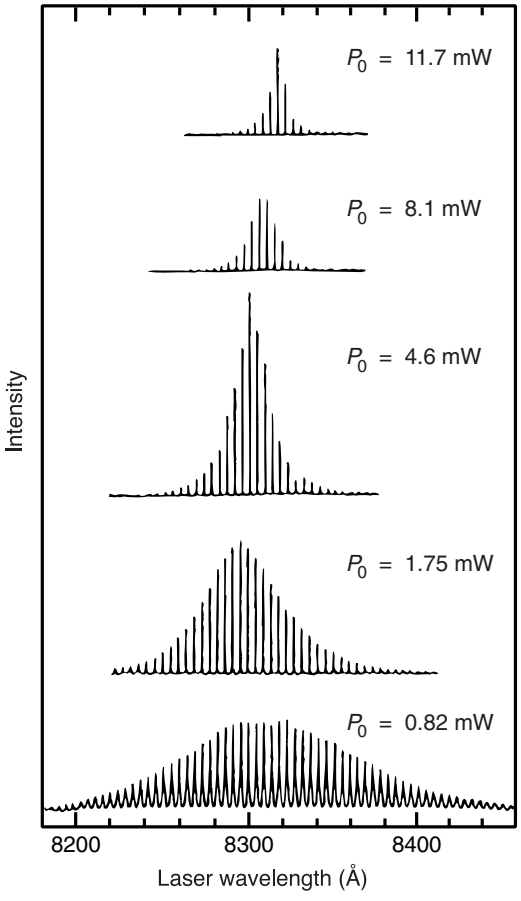


Fig. 13.E.5. Longitudinal mode spectrum for an InGaAs/InP heterostructure laser for various pump currents. (From W. Streifer, D. R. Scifres, and R. D. Burnham, *Appl. Phys. Lett.* **40**, 305 (1982).)

```
M=10;
beta=10^-3;
x=alpha-u;
z=betap*(alpha/u-1);
eps=Sqrt[u/(alpha-1-u)];
y=Abs[x+2*beta*x*((1+eps^2)/eps*ArcTan[1/eps]-1)];
m=1;
z1=z*1/(1+((x-1)/(alpha-x)*(m/M)^2));
tabley1=Table[{y,z1},{u,10^-6,10^-4,10^-6}];
tabley2=Table[{y,z1},{u,10^-4,10^-2,10^-4}];
tabley3=Table[{y,z1},{u,10^-2,1.,10^-2}];
tablep=Join[tableyp1,tableyp2,tableyp3];
p2=LogListPlot[tablepy]
```

FURTHER READING

G. P. Agrawal and N. K. Dutta, *Long Wavelength Semiconductor Lasers*, Van Nostrand Reinhold, New York (1986).

G. Björk and Y. Yamamoto, *IEEE J. Quantum Electron.* **QE 27**, 2386 (1991).

13.F Gain saturation and the *K* factor

We saw in Chapter 4 (Eq. (4.10)) that the gain in a gas or ion laser is saturable, i.e. that the gain as a function of photon density s in such a cavity may be written $\gamma(s) = \gamma_0/(1 + s/s_{\text{sat}})$, where γ_0 is the cold cavity gain (cm^{-1}). This saturation serves to stabilize the laser gain to its steady state value (see Fig. 4.6). By setting the losses at threshold $\alpha_c (= 1/c'\tau_c)$ equal to the above equation for $\gamma(s)$, the photon density as a function of pump current can be derived. This saturation results from competition between the pump rate R and the two available mechanisms for depopulating the excited level: stimulated emission and carrier recombination (radiative or non-radiative). The same effect occurs in semiconductor lasers where these two last mechanisms involve interband transitions (see Fig. 13.F.1). We easily find the photon saturation density to be $s_{s,\text{inter}} = 1/c'gt_{t,\text{inter}}$, where c' is the speed of light in the material, g is the differential gain, and $t_{t,\text{inter}} = t_{\text{tot}}$ is the interband recombination time (see Eq. (13.69)). Comparing this expression with (4.9), we see the similarity between the roles played by the differential gain g in semiconductor lasers and the optical cross-section σ_{op} in atomic lasers.

This *interband saturation* is only a saturation in terms of power in optical amplifiers (as in MOPA amplifiers, for example). In laser diodes, however, the output power in this approach remains linear as a function of the pump current. Taking typical values of $c' = 10^{10}\,\text{cm}\,\text{s}^{-1}$, $g = 10^{-15}\,\text{cm}^2$, and $t_{\text{inter}} = 1\,\text{ns}$, we obtain an interband saturation density of the order of $10^{14}\,\text{cm}^{-3}$. There is another saturation mechanism which results from *intraband relaxation* (see Chapter 6 and Fig. 13.F.1). The characteristic intraband relaxation times are of the order of a picosecond ($t_{\text{intra}} = 1\,\text{ps}$). Using the expression $s_{s,\text{intra}} = 1/c'gt_{t,\text{intra}}$, this leads to intraband saturation densities of the order of $s_{s,\text{intra}} = 10^{17}\,\text{cm}^{-3}$. This last type of saturation can be considered as *instantaneous* in dynamic equations for semiconductor lasers and is given by diverse expressions such as:

$$\gamma(n,s) = \frac{\gamma(n)}{1 + s/s_{s,\text{intra}}} = \frac{\gamma(n)}{1 + \varepsilon s} \tag{13.F.1a}$$

or

$$\gamma(n,s) = \gamma(n)(1 - \varepsilon s) \tag{13.F.1b}$$

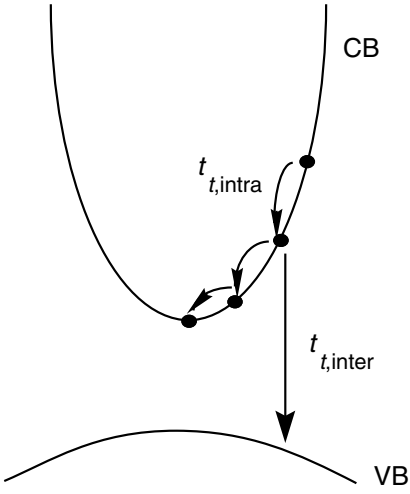


Fig. 13.F.1. Interband and intraband relaxation mechanisms giving rise to the relaxation times $t_{t,\text{inter}}$ and $t_{t,\text{intra}}$: CB, conductance band; VB, valence band.

where ε is the *gain suppression coefficient* and is of the order of a few 10^{-17} cm^3. The photon density is therefore always small in comparison to $1/\varepsilon$. We will therefore use the second formula (13.F.1b) as it leads to simpler results.

In order to include this novel effect in describing the function of a laser, we return to (13.69) and include in it an expression for saturated gain (13.F.1b), i.e:

$$\begin{aligned}\frac{\mathrm{d}n}{\mathrm{d}t} &= \frac{J}{qd} - \gamma_{\text{tot}} n - G(n,s)s \\ \frac{\mathrm{d}s}{\mathrm{d}t} &= [\Gamma G(n,s) - \gamma_c]s\end{aligned} \tag{13.F.2}$$

where $G(n,s) = c'\gamma(n,s)$ is an emission rate (s^{-1}), whereas $\gamma_c = 1/\tau_c$ and $\gamma_{\text{tot}} = 1/t_{\text{tot}}$ are the cavity loss and non-radiative recombination rates, respectively. To illustrate the influence of the saturation term ε on the stationary behaviour of the laser, we choose a model for the gain in the laser medium such as $\gamma(n) = g(n - n_{\text{tr}})$, where we recall that n_{tr} is the transparency density.

At stationary state, the number of photons s_0 is given by the non-trivial solution to the system of equations in (13.F.2) obtained by writing $\mathrm{d}n/\mathrm{d}t = \mathrm{d}s/\mathrm{d}t = 0$ for $s = s_0$ and $n = n_0$, i.e:

$$\begin{aligned}\frac{J_0}{qd} &= \frac{n_0}{t_{\text{tot}}} + c'g(n_0 - n_{\text{tr}})(1 - \varepsilon s_0)s_0 \\ \Gamma c'g(n_0 - n_{\text{tr}})(1 - \varepsilon s_0) &= \frac{1}{\tau_c}\end{aligned} \tag{13.F.3}$$

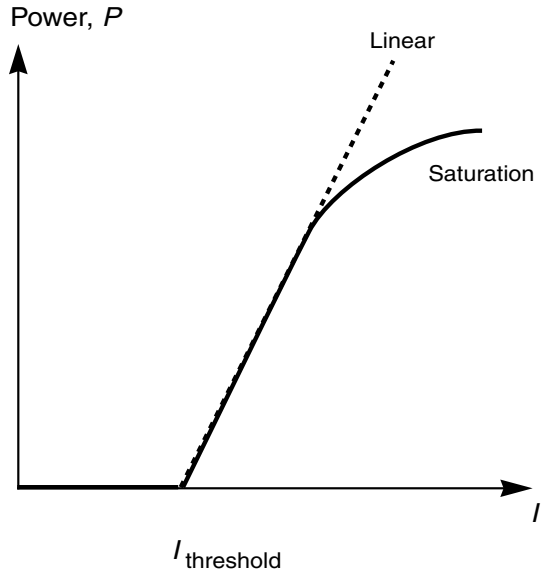


Fig. 13.F.2. Characteristic output power versus injection current plot for a laser diode driven beyond its linear regime.

Eliminating n_0 from both equations, we obtain the second degree equation:

$$s_0 + \frac{\varepsilon}{t_{\text{tot}} c' g} \frac{s_0}{1 - \varepsilon s_0} = \frac{\Gamma \tau_c}{qd} (J - J_{\text{threshold}}) \tag{13.F.4}$$

where we have used for $J_{\text{threshold}}$:

$$J_{\text{threshold}} = \frac{qd}{t_{\text{tot}}} n_{\text{threshold}} = \frac{qd}{t_{\text{tot}}} \left(n_{\text{tr}} + \frac{1}{\Gamma c' g \tau_c} \right) \tag{13.F.5}$$

When ε is small, we return to the situation described in (13.72), i.e. a linear dependence of s on the current density J, once J exceeds $J_{\text{threshold}}$. When the current J becomes large, so that εs_0 is no longer negligible in comparison to 1, the photon density s_0 becomes sub-linear, as required by the (easily obtained) solution to the second degree equation (13.F.4). Figure 13.F.2 presents the resulting dependence of output power on pump current.

The dynamic behaviour of the laser can be obtained by writing $n = n_0 + \delta n$, $J = J_0 + \delta J$, and $s = s_0 + \delta s$, where the variations in the parameters a (either n, J, or s) are harmonic functions, i.e. $\sim \delta a e^{i\omega t}$. The recombination term $G(n, s) = c' g_0 (n - n_{\text{tr}})(1 - \varepsilon s)$ can be expanded as:

$$G = G_0 + G_n \delta n + G_s \delta s \tag{13.F.6a}$$

where the expansion terms are given by:

$$G_n = c'g_0;\ G_s = -\varepsilon c'\gamma_0 \tag{13.F.6b}$$

We therefore expand (13.F.2) by only keeping the first-order terms and by applying the equilibrium conditions (13.F.3), which can be written here as $\Gamma G_0 = \gamma_c$ and $J_0/qd = \gamma_{\text{tot}} n_0 + G_0 s_0$, giving:

$$\begin{aligned} \frac{\delta J}{qd} &= \delta n(\mathrm{i}\omega + \Gamma_n) + \delta s(G_s s_0 + G_0) \\ \delta n \Gamma G_n s_0 &= \delta s(\mathrm{i}\omega - \Gamma G_s s_0) = \delta s(\mathrm{i}\omega + \Gamma_s) \end{aligned} \tag{13.F.7}$$

where we have introduced the damping terms:

$$\Gamma_n = \gamma_{\text{tot}} + c'g_0 s_0;\ \Gamma_s = \varepsilon \Gamma c'\gamma_0 s_0 \tag{13.F.8}$$

Γ_n is the damping already obtained in Section 13.7 (Eq. (13.77b)) and describes the losses in the system (i.e. photon leakage from the cavity and non-radiative carrier recombination). Γ_s is a new term and is due to the saturation. The system of equations in (13.F.7) can be solved simply, as in (13.75), to obtain the temporal response for a laser diode:

$$\frac{\delta s}{\delta J} = \frac{\Gamma G_n s_0/qd}{(\mathrm{i}\omega + \Gamma_n)(\mathrm{i}\omega + \Gamma_s) + \Gamma G_n G_0 s_0} \tag{13.F.9}$$

where we have neglected $G_s s_0$ in comparison with G_0. Being interested only in the response of the amplitude of the laser diode, (13.F.9) can be put into the form:

$$\left|\frac{\delta s}{\delta J}\right| = \left(\frac{\Gamma G_n s_0}{qd}\right)\left|\frac{1}{(\Gamma_n\Gamma_s + \Gamma G_n G_0 s_0) - \omega^2 + \mathrm{i}\omega(\Gamma_n + \Gamma_s)}\right| \tag{13.F.10}$$

and may be approximated as:

$$\left|\frac{\delta s}{\delta J}\right|^2 = \left(\frac{\Gamma\tau_c}{qd}\right)^2 \frac{\omega_R^4}{(\omega_R^2 - \omega^2) + (2\omega\Gamma_R)^2} \tag{13.F.11}$$

In this last expression, we have introduced the following notation:

- the (angular) relaxation frequency ω_R given by:

$$\omega_R \approx (\Gamma G_n G_0 s_0)^{1/2} = \sqrt{\frac{c'g_0 s_0}{\tau_c}} \tag{13.F.12}$$

- Γ_R, or the *damping coefficient* for the laser diode:

$$\Gamma_R = \frac{1}{2}(\Gamma_n + \Gamma_s) \tag{13.F.13}$$

which we express as a function of the resonance frequency f_R as:

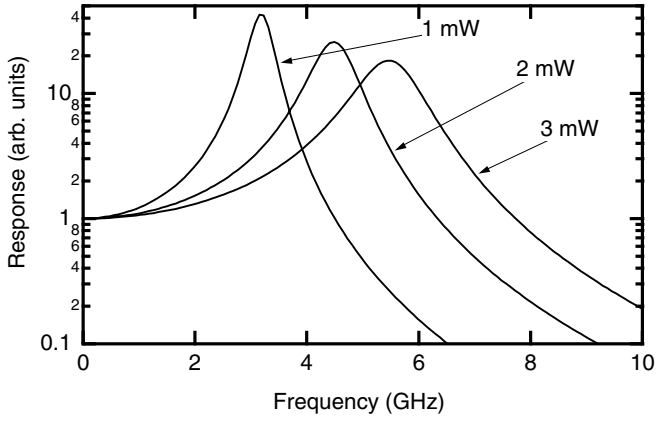


Fig. 13.F.3. Calculated frequency response for a laser diode with parameters given in Table 13.F.1. As a result of saturation, the maximum emitted power by the laser diode decreases and the resonance relaxation curves begin to broaden.

$$\Gamma_R = \frac{1}{2}\left(Kf_R^2 + \frac{1}{t_{\text{tot}}}\right) \tag{13.F.14}$$

by means of the *K factor*, which is a time (generally expressed in ns) given by:

$$K = 4\pi^2\left(\tau_c + \frac{\varepsilon}{c'g_0}\right) \tag{13.F.15}$$

Expressions (13.F.11)–(13.F.15) show that the laser diode resonance relaxation curves broaden in proportion to Γ_R, which in turn increases more rapidly with the internal photon density s_0 (i.e. with the internal power) than is predicted by the simple theory of Section 13.7 (see Fig. 13.F.3). The experimental plot of 'resonance width Γ_R as a function of resonance frequency f_R' allows one to determine the carrier lifetime t_{tot} and the gain suppression coefficient ε. The experimental data, shown in Fig. 13.F.4, clearly shows that the K factor is dominated by the gain saturation effect. Assuming a typical value for ε of $5 \times 10^{-17}\,\text{cm}^3$, $\varepsilon/c'g_0$ equals $5 \times 10^{-17}\,\text{cm}^3/(3 \times 10^{-16}\,\text{cm}^2 \times 9 \times 10^9\,\text{cm s}^{-1})$ or 18 ps, which clearly dominates over the cavity lifetime τ_c (about 1 to 2 ps) in the expression for the K factor (13.F.15).

Table 13.F.1. *Parameter values used for the calculations presented in Fig. 13.F.3*

Output power	p_{mW}		
Wavelength	$h\nu$	0.8 eV	$\lambda = 1.55\,\mu\mathrm{m}$
Cavity length	L	250 μm	
Cavity width	Z	2 μm	
Mode Width	d_{mode}	0.2 μm	
Total Well Width	d	0.05 μm	
Mode Volume	V_{mode}	$1.2 \times 10^{-10}\,\mathrm{cm}^3$	$V_{\mathrm{mode}} = LZd_{\mathrm{mode}}$
Active Volume	V_{active}	$2.5 \times 10^{-11}\,\mathrm{cm}^3$	$V_{\mathrm{active}} = LZd$
Overlap Factor	Γ	0.1	
Optical Index	n_{sc}	3.4	$c' = c/n_{\mathrm{sc}}$
Mirror Reflectance	R_m	0.3	$R_m = (n_{\mathrm{sc}} - 1)^2/(n_{\mathrm{sc}} + 1)^2$
Spontaneous emission factor	β	5×10^{-5}	
Mirror absorption coefficient	α_m	$48\,\mathrm{cm}^{-1}$	$\alpha_m = -\ln(R_m)/L$
Parasitic absorption coefficient	α_P	$20\,\mathrm{cm}^{-1}$	
Photon			
Density (cm^{-3})	s_0	$3 \times 10^{14}\,p_{\mathrm{mW}}$	$p_{\mathrm{mW}} = s_0 V_{\mathrm{mode}} \times h\nu\alpha_m c'/2$
Number	S_0	$3.6 \times 10^{4}\,p_{\mathrm{mW}}$	$S_0 = s_0 \times V_{\mathrm{mode}}$
Lifetime	τ_c	1.6 ps	$1/\tau_c = (\alpha_m + \alpha_P)c/n_{\mathrm{sc}}$
Differential gain	g_0	$2.5 \times 10^{-16}\,\mathrm{cm}^2$	
Transparency density	n_{tr}	$1 \times 10^{18}\,\mathrm{cm}^{-3}$	
Threshold density	n_s	$3.7 \times 10^{18}\,\mathrm{cm}^{-3}$	$\Gamma g_0(n_s - n_{\mathrm{tr}}) = (\alpha_m + \alpha_P)$
Number of carriers at threshold	N_s	9.2×10^{7}	$N_s = n_s V_{\mathrm{active}}$
Stimulated emission coefficient	B	$10^{-10}\,\mathrm{cm}^3\,\mathrm{s}^{-1}$	
Non-radiative lifetime	t_{tot}	2 ns	
Radiative lifetime	τ_{rad}	2.7 ns	$\tau_{\mathrm{rad}}^{-1} = Bn_s$
Spontaneous emission rate	r_{spon}	$1.7 \times 10^{12}\,\mathrm{s}^{-1}$	$r_{\mathrm{spon}} = \beta V_{\mathrm{active}}/\tau_{\mathrm{rad}}$ $= r_{\mathrm{spon}} V_{\mathrm{active}}$
Saturation coefficient	ε	$10^{-17}\,\mathrm{cm}^{-3}$	
Damping coefficient	Γ_n	$(5 + 6.6\,p_{\mathrm{mW}})10^8\,\mathrm{s}^{-1}$	$\Gamma_n = 1/t_{\mathrm{tot}} + c' g_0 s_0$
	Γ_s	$1.9 \times 10^{9}\,p_{\mathrm{mW}}\,\mathrm{s}^{-1}$	$\Gamma_s = \varepsilon s_0/\tau_c$
Relaxation			
Frequency	f_R	$3.2 \times 10^{9}\,(p_{\mathrm{mW}})^{1/2}\,\mathrm{s}^{-1}$	$\omega_R = 2\pi f_R = (c' g_0 s_0/\tau_c)^{1/2}$
Damping	Γ_R	$1.3 \times 10^{9}\,p_{\mathrm{mW}}\,\mathrm{s}^{-1}$	$\Gamma_R = (\Gamma_n + \Gamma_s)/2$

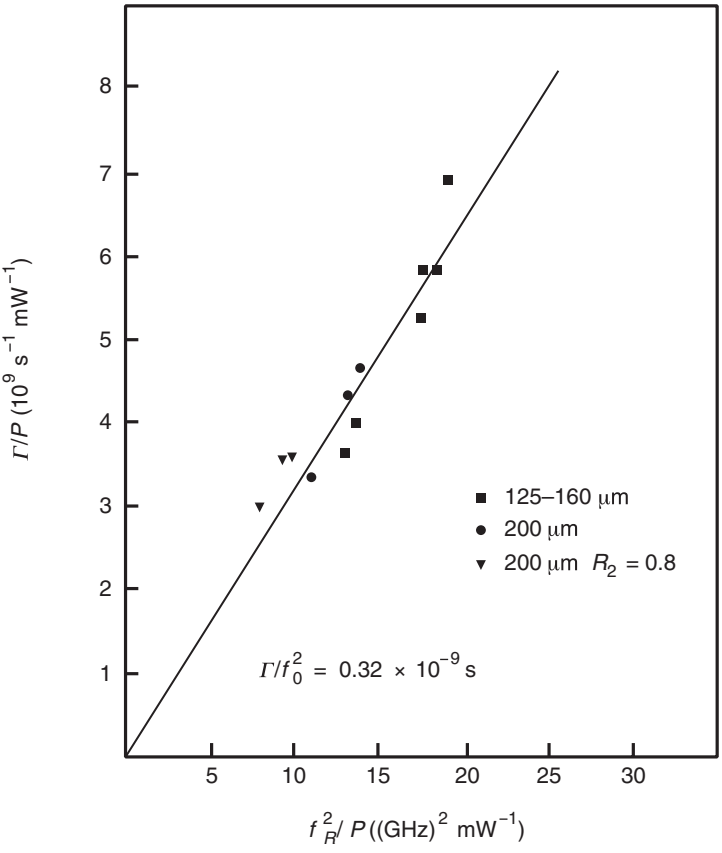


Fig. 13.F.4. Dependence of resonance width Γ_R of the frequency response for a laser diode as a function of the square of the resonance frequency f_R. The slope of this graph yields the K factor from which the gain suppression coefficient ε may be derived, while the y-intercept allows determination of the carrier lifetime. (From R. Olshansky, P. Hill, V. Lanzirera, and W. Powasinik, *IEEE J. Quantum Electron.* **QE–23**, 1410 (© 1987 IEEE).

13.G Laser diode noise and linewidth

We saw in Complement 4.D how spontaneous emission, being the product of a quantum transition and hence intrinsically unpredictable, can be represented within the dynamic laser equations by means of *Langevin forces*. These forces are random temporal processes (or *stochastic processes*) which possess an autocorrelation function amplitude obtained by reconciling (see 4.D.12b) the dynamic equations in (4.D.3) with the rate equations in (4.A.13). In this latter case, spontaneous emission is taken into account by replacing the emission term $c'\sigma_{op}sn$ by $c'\sigma_{op}(s+1)n$. Two decoupled differential equations are thus obtained for the random intensity I and phase ϕ variables (see (4.D.15)). The same formalism can be

applied to laser diodes. The calculated spectral widths, however, are much narrower than those observed experimentally. It was C. Henry who was able to explain this phenomenon successfully by noting that a semiconductor laser diode is a *mismatched resonator* by virtue of the fact that the random intensity and phase variables *are correlated through coupled variations in the gain and optical index*. In other words, fluctuations in the optical index of the semiconductor also cause a fluctuation in the resonance wavelength of the cavity leading to a broadening of laser linewidth. This effect is unique to semiconductor materials. We continue here with a presentation of C. Henry's model.

Anticipating the importance of relative permittivity on propagation of the electromagnetic field, we can no longer content ourselves with studying the electromagnetic fluctuations at a single point (at $x = 0$, for instance) as was the case in (4.D.3). We begin, therefore, with a description of the electromagnetic field in terms of a *random temporal envelope* (see (4.D.14)):

$$E(z,t) = E(t)\mathrm{e}^{\mathrm{i}(kz-\omega t)} \tag{13.G.1}$$

We seek the Langevin equation, for a laser medium with relative permittivity $\varepsilon_{\mathrm{sc}}$, which this temporal envelope must satisfy. We write for the wave equation:

$$\frac{\partial^2}{\partial z^2}E(z,t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\varepsilon_{\mathrm{sc}}E(z,t) \tag{13.G.2}$$

Within the temporal envelope function approximation, $\partial^2 E/\partial t^2$ can be neglected, and (13.G.1) and (13.G.2) lead to:

$$2\mathrm{i}\omega\frac{\varepsilon_{\mathrm{sc}}}{c^2}\frac{\partial}{\partial t}E(t) = -\left(\frac{\omega^2}{c^2}\varepsilon_{\mathrm{sc}} - k^2\right)E(t) \tag{13.G.3}$$

where the relative permittivity $\varepsilon_{\mathrm{sc}}$ is by definition:

$$\varepsilon_{\mathrm{sc}} = (n_{\mathrm{sc}} + \mathrm{i}\, n_{\mathrm{Im}})^2 \tag{13.G.4}$$

and n_{sc} and n_{Im} refer to the real and imaginary optical indices, respectively. The real index n_{sc} is related to the propagation factor k through:

$$k = \frac{\omega}{c}n_{\mathrm{sc}} = \frac{\omega}{c'} \tag{13.G.5}$$

The imaginary portion of the index n_{Im} describes the decay of the wave amplitude in the semiconductor medium. The absorption coefficient given by (3.36) can then be written:

$$\alpha = 2\frac{\omega}{c}n_{\mathrm{Im}} \tag{13.G.6}$$

(We recall in this case that the amplitude will vary as $\mathrm{e}^{-\alpha/2z}$ and the power as $\mathrm{e}^{-\alpha z}$.)

In the laser medium, the effective absorption coefficient α in fact results from competition between parasitic absorption α_P and mirror losses α_m ($\alpha_m = -1/2L\ln(R_1R_2)$ and $\alpha_c = \alpha_P + \alpha_m$) on the one hand and the gain γ on the other so that:

$$\frac{\omega}{c} n_{\text{Im}} = \frac{1}{2}(\alpha_c - \gamma) \tag{13.G.7}$$

At laser threshold, the gain matches the losses so that $n_{\text{Im}} = 0$ and the optical index becomes purely real. In spite of this, fluctuations in the number of carriers Δn lead to fluctuations in the permittivity ε_s through:

$$\begin{aligned} \varepsilon_{\text{sc}} &= (n_{\text{sc}} + \Delta n_{\text{R}} + \text{i}\Delta n_{\text{Im}})^2 \\ &\approx n_{\text{sc}}^2 + 2\text{i}n_{\text{sc}}\Delta n_{\text{Im}}(1 - \text{i}\beta_e) \end{aligned} \tag{13.G.8}$$

where β_e is the *linewidth broadening factor* defined as:

$$\beta_e = \frac{\Delta n_{\text{R}}}{\Delta n_{\text{Im}}} \tag{13.G.9}$$

Linewidth broadening factor

Substituting the expression for ε_{sc} (13.G.8) into the evolution equation, (13.G.3), employing (13.G.5), and keeping only the first-order terms, we find that the envelope function $E(t)$ is a solution of the differential equation:

$$\frac{\partial}{\partial t} E(t) = -\omega \frac{\Delta n_{\text{Im}}}{n_{\text{sc}}}(1 - \text{i}\beta_e)E(t) \tag{13.G.10}$$

which can also be written as (see (13.G.7)):

$$\frac{\partial}{\partial t} E(t) + \frac{1}{2} c'(\alpha_c - \gamma)(1 - \text{i}\beta_e)E(t) = 0 \tag{13.G.11}$$

The reader will easily recognize this equation as being identical to (4.D.3) (where the effect of fluctuations in the permittivity ε_{sc} were ignored) with the exception of an additional multiplication factor $(1 - \text{i}\beta_e)$ which takes the fluctuations in ε_{sc} into account.

This coefficient β_e is null for atomic transitions, as in this case; laser oscillation occurs at the peak of the Lorentzian gain curve, where Δn_{R} is zero (see Fig. 13.G.1). The Kramers–Kronig relation stipulates that the real optical index n_{R} is proportional to the derivative of the absorption coefficient, which is itself tied to n_{Im}. On the other hand, in condensed media, the gain curve is no longer Lorentzian, and even at the peak of the gain curve, Δn_{R} is non-zero (see Fig. 13.G.1).

In step with the formalism developed in Complement 4.D, we introduce the Langevin force $F(t)$, which *drives* the electric field as:

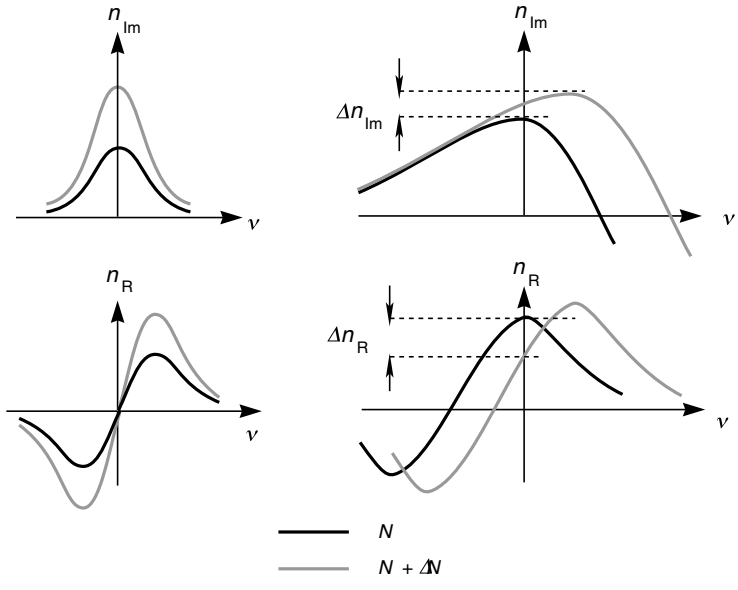


Fig. 13.G.1. (*a*) In atomic systems, the variation in optical index as a function of carrier density $\Delta n_R/\Delta N$ is null given the symmetry of the Lorentzian gain curve. (*b*) In a semiconductor medium, the gain curve is not symmetric and $\Delta n_R/\Delta N$ is non-zero leading to a non-zero linewidth broadening factor $\Delta n_R/\Delta n_{Im}$.

$$\frac{\partial}{\partial t}E(t) + \frac{1}{2}c'(\alpha_c - \gamma)(1 - \mathrm{i}\beta_e)E(t) = F(t) \tag{13.G.12}$$

Introducing as in Complement 4.D the intensity I and the phase ϕ, $E(t) = I^{1/2}\mathrm{e}^{\mathrm{i}\phi}$, this differential equation can be separated into:

$$\frac{\mathrm{d}}{\mathrm{d}t}I + c'(\alpha_c - \gamma)I = 2\sqrt{I}\mathrm{Re}[F(t)\mathrm{e}^{-\mathrm{i}\phi}] \tag{13.G.13a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi - \frac{1}{2}c'(\alpha_c - \gamma)\beta_e = \frac{1}{\mathrm{i}\sqrt{I}}\mathrm{Im}[F(t)\mathrm{e}^{-\mathrm{i}\phi}] \tag{13.G.13b}$$

We now add to these two equations, the variation in the carrier number n, and return to the notation introduced in Complement 13.F (see (13.F.2)). We thus obtain the *Langevin equations* for a semiconductor laser medium:

$$\frac{\mathrm{d}}{\mathrm{d}t}s - (\Gamma G - \gamma_c)s + R_{\mathrm{spon}} = F_s(t) \tag{13.G.14a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi - \frac{1}{2}(\Gamma g - \gamma_c)\beta_e = F_\phi(t) \tag{13.G.14b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}n + \gamma_{\mathrm{tot}}n + Gs = \frac{1}{qd}\frac{\mathrm{d}}{\mathrm{d}t}J + F_n(t) \tag{13.G.14c}$$

Langevin equations for a semiconductor laser

Alongside (4.D.15), we have added to the evolution equation for the photon number s (13.G.14a) the deterministic spontaneous recombination term R_{spon}, in such a manner as to make the average of the random term $\langle F_s(t)\rangle$ zero. This term R_{spon} (in $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$) is the radiative recombination rate obtained in Section 7.5 and was already introduced in Complement 13.E: $R_{\mathrm{spon}} = \beta Bnp = \beta n/\tau_{\mathrm{rad}}$, where B is the radiative recombination coefficient, τ_{rad} is the radiative lifetime (see Eq. (7.61)), and β is the spontaneous emission coefficient (see Complement 13.E).

We still need to determine the autocorrelation functions for the Langevin functions $F_s(t)$, $F_n(t)$, and $F_\phi(t)$. The procedure for the calculation is the same as that developed in Complement 4.D. As the calculation is fairly involved, we will limit ourselves here to a summary of the results:

$$\langle F_s(t)F_s(t')\rangle = 2D_{ss}\delta(t-t');\ D_{ss} = R_{\mathrm{spon}}s \tag{13.G.15a}$$

$$\langle F_n(t)F_n(t')\rangle = 2D_{nn}\delta(t-t');\ D_{nn} = R_{\mathrm{spon}}s + \gamma_{\mathrm{tot}}n \tag{13.G.15b}$$

$$\langle F_\phi(t)F_\phi(t')\rangle = 2D_{\phi\phi}\delta(t-t');\ D_{\phi\phi} = \frac{R_{\mathrm{spon}}}{4s} \tag{13.G.15c}$$

where we have neglected the cross terms in D_{sn}. A fairly complete description of this last calculation can be found in C. H. Henry (1982).

13.G.1 Linewidth broadening

Integration of equation (13.G.14b) yields:

$$\phi(t) - \phi(t_0) = \frac{1}{2}\frac{R_{\mathrm{spon}}}{s_0}\beta_e(t-t_0) + \int_{t_0}^{t} F_\phi(u)\mathrm{d}u \tag{13.G.16}$$

where we have made use of the fact that $\Gamma G - \gamma_c = R_{\mathrm{spon}}/s_0$ at stationary state (Eq. (13.G.14a)). Carrying out the same calculation as in Complement 4.D, we obtain a laser emission linewidth $\Delta\nu$ of:

$$\Delta\nu = \frac{h\nu}{2\pi\tau_c^2 P_{\mathrm{out}}}(1+\beta_e^2) \tag{13.G.17}$$

Emission linewidth for a semiconductor laser

where P_{out} is the output power of the laser. We now understand the justification for referring to be as a 'linewidth broadening factor'. With a typical value of 5 for β_e at room temperature, the linewidth is increased by a factor of 26 relative to the

Schawlow–Townes linewidth obtained in (4.D.31).

Example
We consider a semiconductor laser cavity ($h\nu = 1\,\text{eV}$) with a length of 300 μm, a parasitic loss coefficient α_P of $10\,\text{cm}^{-1}$, and two mirrors with reflectances of 0.3. The mirror loss α_m is then $1/(3 \times 10^{-2}\,\text{cm}) \times \ln 0.3$ or $40\,\text{cm}^{-1}$. If the semiconductor possesses an optical index of 3.3, the photon lifetime $\tau_c = 1/c'\alpha_c$ is then $3.3/(3 \times 10^{10}\,\text{cm s}^{-1} \times 50\,\text{cm}^{-1})$ or 2.2 ps. Assuming an output power of 1 mW and a value for β_e of 5, the laser diode will exhibit an emission linewidth of:

$$\Delta\nu = \frac{1.6 \times 10^{-19}\,\text{J}}{2\pi(2.2 \times 10^{-12}\,\text{s})^2\ 10^{-3}\,\text{W}}(1 + 5^2) = 140\,\text{MHz}$$

13.G.2 Relative intensity noise (RIN) and optical link budget

In order to determine the noise amplitude in laser diodes resulting from Langevin force fluctuations, we expand (13.G.14a) and (13.G.14c) in the same fashion as in Complement 13.F (see (13.F.7)). As we are not interested in fluctuations due to current, we can set $\delta J = 0$ in these equations leaving:

$$\begin{aligned}&(\mathrm{i}\omega + \Gamma_n)dn + G_0\delta s = \tilde{F}_n\\ &-\Gamma G_n s_0 \delta n + (\mathrm{i}\omega + \Gamma_s)\delta s = \tilde{F}_s\end{aligned} \tag{13.G.18}$$

where $\tilde{F}_n$ and $\tilde{F}_s$ are the Fourier transforms of the Langevin forces. We recall, however, that these Fourier transforms do not possess a mathematical existence *per se*, but are rather an intermediary step in the calculation as exposed in Complement 4.D. Only the quantities $\langle F_s(t)^2\rangle = \langle \tilde{F}_s\tilde{F}_s{}^*\rangle\Delta f = 2D_{ss}\Delta f$ and $\langle F_n(t)^2\rangle = \langle \tilde{F}_n\tilde{F}_n{}^*\rangle\Delta f = 2D_{nn}\Delta f$ are in fact defined and represent the noise power contained in a bandwidth Δf.

The system of equations in (13.G.18) can be solved immediately to find:

$$\tilde{\delta s} = \frac{\Gamma G_n s_0 \tilde{F}_n + (\mathrm{i}\omega + \Gamma_n)\tilde{F}_s}{(\omega_R^2 - \omega^2) + 2\mathrm{i}\omega\Gamma_R} \tag{13.G.19}$$

The variance $\langle \delta s\delta s^*\rangle$ of δs can be obtained by writing:

$$\langle \delta s\delta s^*\rangle = \frac{(\Gamma G_n s_0)^2\langle \tilde{F}_n\tilde{F}_n^*\rangle + (\omega^2 + \Gamma_n^2)\langle \tilde{F}_s\tilde{F}_s^*\rangle}{(\omega_R^2 - \omega^2)^2 + (2\omega\Gamma_R)^2} \tag{13.G.20}$$

where the relaxation frequency ω_R and the laser damping coefficient Γ_R are as introduced in (13.F.12) and (13.F.13), respectively.

We define the ratio of the variance $\langle \delta s^2\rangle$ with s_0^2 as the *relative intensity noise* (RIN). The RIN relates to (13.G.20) according to:

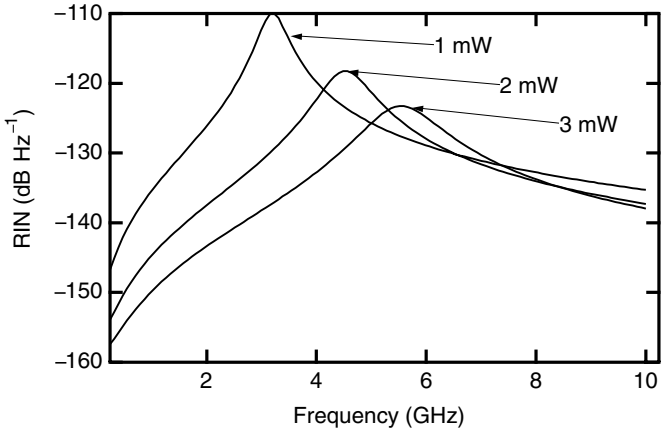


Fig. 13.G.2. Relative intensity noise (RIN) for a typical 1.55 μm laser diode using the parameters listed in Table 13.F.1.

$$\text{RIN} = \frac{\langle \delta s^2 \rangle}{s_0^2 \Delta f} = \frac{\langle \delta s \delta s^* \rangle}{s_0^2} \tag{13.G.21}$$

Given the variances in (13.G.15), the RIN may be written as:

$$\text{RIN} = 2\frac{R_{\text{spon}}}{s_0} \frac{(\omega^2 + \Gamma_n^2) + (\Gamma c' g_0 s_0)^2 (1 + \gamma_{\text{tot}} n_s / R_{\text{spon}} s_0)}{(\omega_R^2 - \omega^2)^2 + (2\omega\Gamma_R)^2} \tag{13.G.22a}$$

Relative intensity noise or RIN for a laser diode

Dimensional analysis of this ratio shows that the RIN is expressed in units of Hz^{-1} (and in accordance with definition (13.G.21)). As this number is generally weak, it is customary to express it in terms of dB Hz^{-1}. The relative intensity fluctuations $\langle \delta s^2 \rangle / s_0^2$ over a bandwidth Δf can then be obtained by the formula:

$$\frac{\langle \delta s^2 \rangle}{s_0^2} = 10^{\text{RINdB}/10} \Delta f = \text{RIN}\, \Delta f$$

Therefore, a RINdB of $-160\,\text{dB Hz}^{-1}$ corresponds to a RIN of $10^{-16}\,\text{Hz}^{-1}$. Figure 13.G.2 shows the result of a RINdB calculation using Eq. (13.G.22). The parameters used for this calculation are those presented in Table 13.F.1 for a typical laser diode used in telecommunications applications. Several points are worth noting from these results:

- the noise is at a maximum at resonance, which is also the general result from electrical circuit theory;
- the noise decreases as the injection current and laser power increase;
- the noise levels drop rapidly at reduced modulation frequencies.

An interesting limiting condition for RIN occurs at very low modulation frequencies ($\omega = 0$) and large photon densities s_0 ($c' g_0 s_0 \gg \gamma_{\text{tot}}$ and R_{sp} on $s_0 \gg n_s \gamma_{\text{tot}}$):

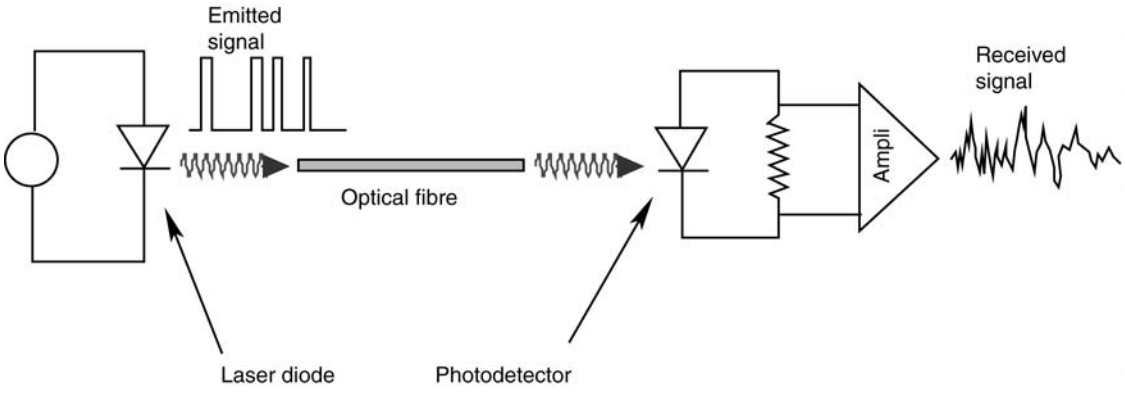


Fig. 13.G.3. Optical budget considerations in optical telecommunications.

$$\mathrm{RIN} \approx 2\frac{R_{\mathrm{spon}}\tau_c^2}{s_0} \tag{13.G.22b}$$

where we recall that τ_c is the photon lifetime in the cavity. For typical laser diode parameters given in Table 13.F.1, we find a RIN, assuming a power of 1 mW, of $2 \times 1.7 \times 10^{12}\,\mathrm{s}^{-1} \times (1.6 \times 10^{-12}\,\mathrm{s})^2/3.6 \times 10^4 = 2.4 \times 10^{-16}\,\mathrm{Hz}^{-1}$ or -156 dB.

The RIN parameter is an essential component for calculating the *optical budget* in telecommunications applications. The current $i(t)$ generated in a photodetector by a laser output signal $P(t)$ is given by (11.18) or (11.41) to be:

$$i(t) = \eta q\frac{P(t)}{h\nu} \tag{13.G.23}$$

where η is the total efficiency, given the link and detector losses (see Fig. 13.G.3). The fluctuations in $P(t)$, will also cause the total diode current to fluctuate. Given the proportionality relation (13.G.23), the relative variance $\langle \Delta i(t)^2 \rangle / i_0^2$ in the current is identical to the relative variance in the signal $\langle \Delta P(t)^2 \rangle / P_0^2$ contributed by the laser diode. As we have shown on several occasions (see Complements 4.D and 11.A), variance of the temporal fluctuations and the frequency spectrum of the fluctuations $S_P(f)$ (i.e. the average noise power in the frequency bandwidth Δf) are related according to:

$$\langle \Delta P(t)^2 \rangle = \int_{-\infty}^{+\infty} \Delta P(t)^2 \mathrm{d}t = \int_{f}^{f+\Delta f} S_P(f)\mathrm{d}f = S_P(f)\Delta f \tag{13.G.24}$$

This allows one to relate the variance of the photodetector noise to the laser diode RIN:

$$\mathrm{RIN} = \frac{\langle \Delta i(t)^2 \rangle}{i_0^2 \Delta f} = \frac{S_i(f)}{i_0^2} = \frac{S_P(f)}{P_0^2} = \frac{S_s(f)}{s_0^2} \tag{13.G.25}$$

In an optical budget calculation, (13.G.25) therefore allows one to calculate the photodetector reception noise due to the laser diode. The laser RIN must then be compared with the noise contributed by the detector itself (by employing the concept of detectivity as explained in Complement 11.B), and by the optical fibre (e.g. polarization noise, amplification noise, . . .).

Example
We consider a laser diode with the device parameters listed in Table 13.F.1. Assuming an operational power P_0 of 3 mW and a modulation frequency of 1 GHz ($\Delta f = 10^9$ Hz), the laser diode will have a RIN of -150 dB (see Fig. 13.G.2). If the global link efficiency $\eta = 0.1$, the photocurrent i_0 would be:

$$i_0 = \eta \frac{P_0}{(h\nu/q)} = 0.1 \frac{3 \times 10^{-3}\,\mathrm{W}}{0.8\,\mathrm{eV}} = 0.37\,\mathrm{mA}$$

The relative current fluctuation in the photodetector is then:

$$\sqrt{\langle \Delta i(t)^2 \rangle} = i_0 (10^{-15} \times 10^9)^{1/2}$$

implying a variance of $10^{-3} \times i_0$ or 3.7×10^{-7} A.

FURTHER READING

G. P. Agrawal and N. K. Dutta, *Long Wavelength Semiconductor Lasers*, Van Nostrand Reinhold, New York (1986).
C. H. Henry, *IEEE J. Quantum Electron.* **QE18**, 259 (1982).
C. H. Henry, *IEEE J. Quantum Electron.* **QE19**, 1391 (1983).

13.H Unipolar quantum cascade lasers

One characteristic of semiconductor laser diodes is that two carrier types (electrons and holes) come into play in a forward biased diode. This makes traditional laser diodes *bipolar devices*. Another class of lasers exists whose operation is founded uniquely upon electrons – these are *unipolar quantum cascade lasers*. In this case, the idea is to make use of intersubband transitions in quantum wells (see Section 11.6). Just as intersubband absorption can be used in quantum well detectors, intersubband gain is employed in this type of laser. The principal concepts which enable the operation of this device are illustrated in Fig. 13.H.1:

- A quantum cascade laser is a four-level laser. Electrons coming from level $|4\rangle$

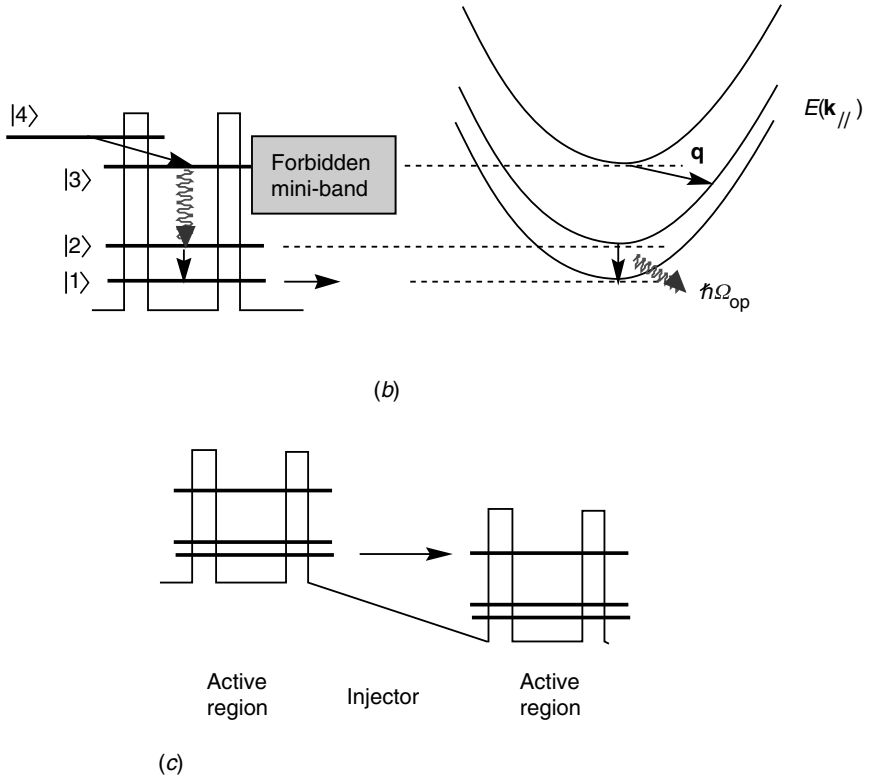


Fig. 13.H.1. (*a*) Schematic representation of intersubband transitions in a quantum well based quantum cascade laser. (*b*) Condition on momentum transfer in the $E(\mathbf{k})$ diagram. (*c*) Bias conditions allowing resonant tunnelling between successive quantum wells.

(actually a level $|1\rangle$ belonging to an adjacent quantum well), tunnel across the barrier material into level $|3\rangle$ of the quantum well. Due to the existence of a forbidden energy *mini-band* (resulting from the periodic succession of potential barriers provided by the wells, which act as an 'electron Bragg grating'), the electrons cannot easily tunnel into continuum states and can only recombine in the quantum wells (Fig. 13.H.1*a*).

- In order to ensure that the lifetime τ_{12} for transition $|2\rangle \rightarrow |1\rangle$ is significantly less than the lifetime τ_{32} for $|3\rangle \rightarrow |2\rangle$ (i.e. a requirement for population inversion – see Chapter 4), the energy separation E_{12} is made resonant with the optical phonon energy in the host semiconductor ($E_{12} = \hbar\omega_{op} = 34$ meV for GaAs, see Complement 6.B) which leads to $\tau_{12} \approx 0.4$ ps (Fig. 13.H.1*b*).
- The transition $|3\rangle \rightarrow |2\rangle$ must involve significant momentum transfer $\mathbf{q}$ in terms of optical phonon emission so that the lifetime τ_{32} becomes as long as possible (see Complement 6.B and Eq. (6.B.39) where the transition rate was shown to vary as $1/q^2$) (Fig. 13.H.1*b*).
- The barrier material which separates the quantum wells must be of the

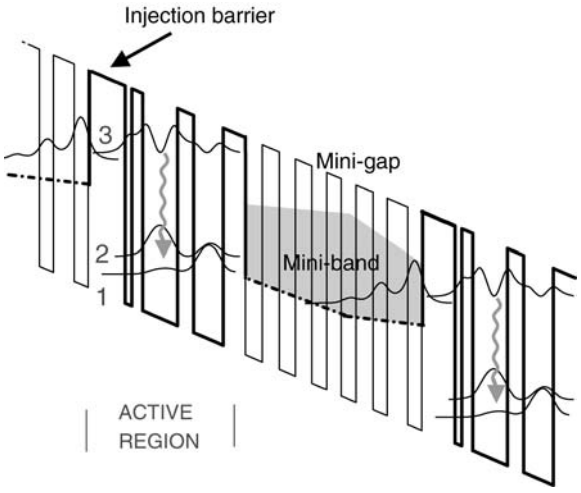


Fig. 13.H.2. Band structure for a quantum cascade laser. (C. Sirtori@THALES.)

appropriate composition and thickness so that under the influence of an *electric field*, level $|1\rangle$ in one well becomes resonant with level $|3\rangle$ in an adjacent 'downstream' well (Fig. 13.H.1*c*).

When all these conditions are satisfied, we obtain the complex structure shown in Fig. 13.H.2. The device structure is engineered so that the matrix elements z_{32} are as large as possible. The potential structure obtained under the influence of an electric field appears as a staircase across which electrons will cascade (hence the name for this type of laser). At each 'step' down the structure, an electron will release a photon. If optical feedback is made available, laser oscillation over these transitions becomes possible. We will calculate the operational characteristics for a quantum cascade laser with the following device parameters:

- Emission wavelength $\lambda = 9.3\,\mu\text{m}$ implying a radiative lifetime τ_R of 60 ns (see Eqs. (3.69) and (3.71));
- lifetime in level $|2\rangle$ $\tau_2 = 0.23\,\text{ps}$ (including the combined effects of optical phonons and tunnelling);
- lifetime in level $|3\rangle$ $\tau_3 = 1.5\,\text{ps}$ resulting in a quantum electroluminescence efficiency $\eta = \tau_3/\tau_R$ of 2.5×10^{-5} (see Eq. (13.13));
- 1 mm cavity length;
- effective optical index in the guide $n_{sc} = 3.26$ resulting in a reflectance R_m of 0.28 and a mirror loss coefficient α_m of $1/L\ln R = 18\,\text{cm}^{-1}$;
- a parasitic absorption coefficient α_p of $45\,\text{cm}^{-1}$ due to free carrier absorption, which is extremely efficient at these long wavelengths (see Complement 7.C and Eq. (7.C.12), where α_p is proportional to λ^2);
- the confinement factor for N quantum wells Γ_N is given by $N\Gamma_1$, where Γ_1 is the

single well confinement factor ($\Gamma_1 = d/d_{\text{mode}}$, where d is the quantum well thickness and d_{mode} is the mode width) so that $\Gamma_N = 0.33$ for 25 wells.

The condition for laser oscillation is then given by setting the gain and the losses at threshold equal to each other (13.26):

$$\Gamma\gamma_{\text{threshold}} = \alpha_m + \alpha_p \tag{13.H.1}$$

resulting in a gain for the quantum wells of 190 cm^{-1}. The gain γ (cm^{-1}) due to the interband transitions $|3\rangle \rightarrow |2\rangle$ is given by either (8.87b) or (11.57) to be:

$$\gamma \approx \frac{(qz_{23})^2}{\hbar}\frac{Z_0}{n_{\text{sc}}}\frac{h\nu}{\hbar\Gamma_{32}}\frac{\Delta n_{32}}{d} \tag{13.H.2}$$

where d is the well thickness (10 nm), Z_0 is the vacuum impedance (377 Ω), $\hbar\Gamma_{32}$ is the broadening of the transition (typically 10 meV), $h\nu$ is the resonant photon energy (130 meV), and z_{23} is the dipolar matrix element (of the order of 2 nm). This results in a threshold density of $2 \times 10^{10}\,\text{cm}^{-2}$ – weak in comparison with that obtained for interband lasers ($>10^{12}\,\text{cm}^{-2}$). This difference results from the larger oscillator strengths associated with intersubband transitions. The inversion density $\Delta n_{32} = n_3 - n_2$ is related to the pump current through (13.3), i.e:

$$\frac{J}{q} \approx \frac{\Delta n_{32}}{\tau_3} \tag{13.H.3}$$

We therefore obtain a threshold current of 5 kA cm^{-2}. This extremely large threshold current results from the short lifetime in the excited state. In spite of these elevated threshold currents, quantum cascade lasers possess a number of advantages. The three dominant ones are:

- These lasers can be designed to operate at any fixed wavelength between 4 and 13 μm without changing material systems. This can be achieved through *bandgap engineering* as described in Chapter 8.
- The output power emitted by an N well quantum cascade laser is extremely high (several hundred milliwatts) as a single electron can emit a total of N photons (one per well). The output power can be written (from (13.34)) as:

$$P_{\text{out}} = \eta_{\text{ext}} N \frac{h\nu}{q}(I - I_{\text{threshold}}) \tag{13.H.4}$$

  Clearly, as the applied voltage is $N \times h\nu$, the wallplug efficiency $P_{\text{out}}/P_{\text{el}}$ remains unchanged (see (13.37b)).
- We saw in Fig. 13.D.4 that the maximum operating temperatures were determined by Auger recombination. The Auger mechanism, however, is virtually non-existent in intersubband transitions as the subbands are parallel (see Complement 6.D and Fig. 6.D.1). As a result, the characteristic temperatures T_0 for

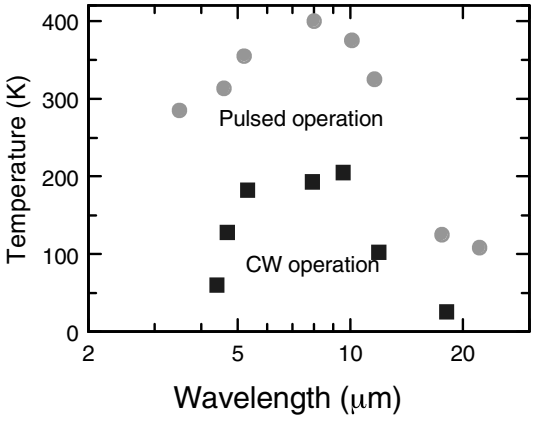


Fig. 13.H.3. Maximum operational temperatures for quantum cascade lasers at various wavelengths. (C. Sirtori@THALES.)

intersubband lasers are quite high. For example, for a laser that emits at 9.3 μm, $T_0 = 140$ K. Quantum cascade lasers can therefore be operated in pulsed mode all the way up to room temperature even at such long emission wavelengths (see Fig. 13.H.3).

FURTHER READING

J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, *Science* **264**, 553 (1994).

C. Sirtori, J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, *IEEE J. Quantum Electron.* **QE-33**, 89 (1997).

13.I Mode competition: cross gain modulators

We saw in Complement 11.E that near threshold, a semiconductor laser can oscillate in multimode fashion even though the gain medium is homogeneous. Driven sufficiently far from threshold so that spontaneous emission becomes negligible in comparison with stimulated emission, the laser will eventually recover its single mode behaviour. We will now study this phenomenon in greater detail. We will show that this effect can be understood in terms of cross-saturation, and see how this effect is harnessed in an optoelectronic device – the *cross gain modulator*.

We start with (13.E.7), which describes the dynamic equations for a laser cavity in a multimode regime. To simplify the calculations, we will assume that the laser is

operated far above transparency so that $n \gg n_{tr}$ and R_{spon} is negligible, and that only two modes are present. Equation (13.E.7) can then be written:

$$\begin{aligned} \frac{dn}{dt} &= \frac{J}{qd} - \frac{n}{t_{tot}} - c'g_1 n s_1 - c'g_2 n s_2 \\ \frac{ds_i}{dt} &= \Gamma c' g_i n s_i - \frac{s_i}{\tau_c},\ i = 1, 2 \end{aligned} \tag{13.I.1}$$

We recall that n and s are the electron–hole pair and photon densities, respectively, J is the pump current density; d is the thickness of the active region; t_{tot} is the total recombination time; τ_c is the photon lifetime in the cavity; g_i the dynamic gains; c' the speed of light in the medium; and Γ is the confinement factor. At stationary state, (13.I.1) gives the electron–hole pair density:

$$n = \frac{n_0}{1 + c'g_1 t_{tot} s_1 + c'g_2 t_{tot} s_2} \tag{13.I.2}$$

where $n_0 = Jt_{tot}/qd$ is the cold-cavity density of electron–hole pairs and leads to a medium gain G_i for the ith mode of:

$$G_i = \frac{G_{i0}}{1 + \varepsilon_1 s_1 + \varepsilon_2 s_2} \tag{13.I.3}$$

where $G_{i0} = \Gamma c' g_i n_0$ is the cold-cavity gain and $\varepsilon_i = c' g_i t_{tot}$. As in Complement 11.F, and following common practice, the ε terms designate the inverse photon saturation densities s_{sat} for which the magnitude of the gain decreases by a factor of 2 (see Chapter 4). In this last equation, (13.I.3), we see that the gain at one wavelength can be saturated by a pump beam at a different wavelength. This cross-gain saturation effect is exploited in *cross gain modulators*. We will now explain the operation of this device.

This device is able to transfer a modulated signal with carrier wavelength λ_2 into a modulated signal with carrier wavelength λ_1. To accomplish this, a semiconductor (MOPA type) optical amplifier is pumped by a continuous beam at λ_2 with an intensity which approaches the saturation density (see Fig. 13.I.1). The additional signal at λ_1 saturates the medium (see (13.I.3)) and the wave at λ_2 experiences a decrease in gain. The output power at λ_2 therefore decreases when the input power at λ_1 increases. Thus, a signal initially carried at λ_1 becomes 'negatively' mapped onto a light beam at λ_2 (see Fig. 13.I.1*b*).

We will now give a brief description of the dynamic behaviour of the laser due to competition between modes 1 and 2. To do so, we will assume that the laser gain G is in a stationary regime and that the optical modes are not. This comes down to having $t_{tot} < \tau_c$, which holds for atomic transition lasers but not for semiconductor lasers. While it is not absolutely necessary to make this approximation, not doing

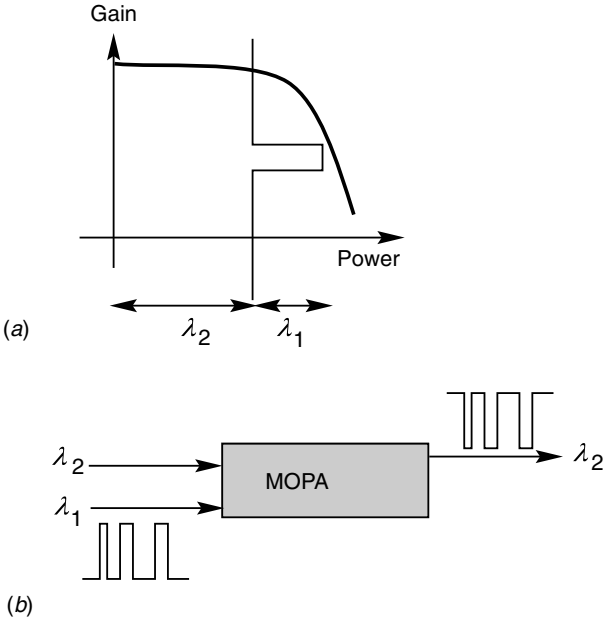


Fig. 13.I.1. Operation of a cross gain modulator.

so would bury the key physical aspects of the problem we wish to expose.

Substituting the expression for the saturated gain (13.I.3) into (13.I.1) leads to:

$$\dot{s}_i = \frac{G_{i0}}{1 + \varepsilon_1 s_1 + \varepsilon_2 s_2} s_i - \frac{s_i}{\tau_c},\ i = 1, 2 \tag{13.I.4}$$

We assume that the medium is only slightly saturated (meaning $\varepsilon_i s_i \ll 1$) allowing a limited expansion of (13.I.4) as:

$$\begin{aligned} \dot{s}_1 &= \alpha_1 s_1 - \kappa_1 s_1^2 - \theta_{12} s_1 s_2 \\ \dot{s}_2 &= \alpha_2 s_2 - \theta_{21} s_1 s_2 - \kappa_2 s_2^2 \end{aligned} \tag{13.I.5}$$

In this last expression, we have introduced the following notation:

$$\alpha_i = G_{i0} - \frac{1}{\tau_c},\ \kappa_i = G_i \varepsilon_i,\ \theta_{ij} = G_i \varepsilon_j \tag{13.I.6a}$$

which are, respectively, the global gain (resulting from competition between the cavity gain and losses), the *autosaturation*, and *cross-saturation* terms. Within the context of approximation (13.I.5), we see that there are two stationary solutions in terms of the photon density to the two equations:

$$\begin{aligned} (\alpha_1 - \kappa_1 s_1 - \theta_{12} s_2) s_1 &= 0 \\ (\alpha_2 - \theta_{21} s_1 - \kappa_2 s_2) s_2 &= 0 \end{aligned} \tag{13.I.6b}$$

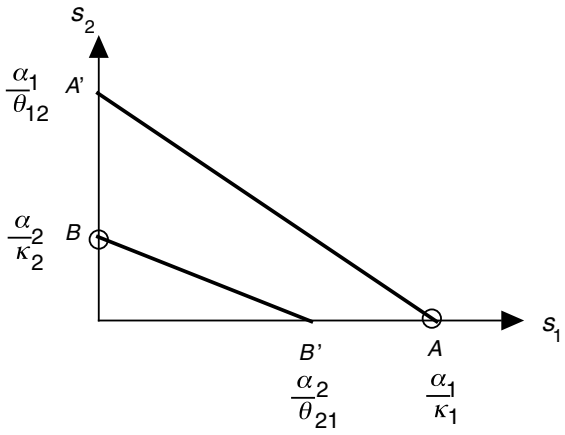


Fig. 13.I.2. Of the two stationary solutions to the mode competition equation, (13.I.5), indicated by circles, only the mode with the greater gain (i.e. that furthest from the origin) is stable.

It is straightforward to show that the lines $\alpha_i = \kappa_i s_i + \theta_{ij} s_j$ do not intersect. The solutions thus lie at the intersections of the lines defined by conditions (13.I.6b) and the co-ordinate axes, and are represented in Fig. 13.I.2: point A ($s_{20} = 0$ and $s_{10} = \alpha_1/\kappa_1$) and point B ($s_{10} = 0$ and $s_{20} = \alpha_2/\kappa_2$). We will see that only one of these two points constitutes a stable solution to (13.I.5).

Near point A, the photon density can be expanded in the form of $s_2 = \delta s_2$ and $s_1 = s_{10} + \delta s_1$. The differential Eqs. (13.I.5) are then linearized and can be written in matrix form as:

$$\begin{bmatrix} \delta\dot{s}_1 \\ \\ \delta\dot{s}_2 \end{bmatrix} = \begin{bmatrix} -\alpha_1 & -\theta_{12}\dfrac{\alpha_1}{\kappa_1} \\ \\ 0 & \alpha_2 - \theta_{21}\dfrac{\alpha_1}{\kappa_1} \end{bmatrix} \begin{bmatrix} \delta s_1 \\ \\ \delta s_2 \end{bmatrix} \tag{13.I.7}$$

We see that the solution is stable only if:

$$\alpha_2 < \theta_{21}\frac{\alpha_1}{\kappa_1} \tag{13.I.8}$$

which, given definitions (13.I.6), corresponds to:

$$G_{10} > G_{20} \tag{13.I.9}$$

In other words, although both wavelengths experience positive gain in the cavity, at stationary state the laser will preferentially oscillate in the mode with the highest gain (hence the term *mode competition*). Using the same formalism, we could have shown that an inhomogeneously broadened medium will exhibit stationary multi-modal operation.

FURTHER READING

G. Grynberg, A. Aspect, and C. Fabre, *Introduction aux lasers et à l'optique quantique*, Ellipses, Paris (1997).

A. E. Siegman, *Lasers*, University Science Book, Mill Valley, CA (1986).

I. Valiente, J. C. Simon, and M. LeLigne, *Electron. Lett*. **29**, 502 (1993).

Index

Numbers in italics indicate *tables* or *figures*

Optoelectronics

Emmanuel Rosencher and
Borge Vinter