Springer Series in Optical Sciences 213

Arthur McGurn

Nanophotonics

Springer

Springer Series in Optical Sciences

Volume 213

Springer Series in Optical Sciences

The Springer Series in Optical Sciences, under the leadership of Editor-in-Chief William T. Rhodes, Georgia Institute of Technology, USA, provides an expanding selection of research monographs in all major areas of optics: lasers and quantum optics, ultrafast phenomena, optical spectroscopy techniques, optoelectronics, quantum information, information optics, applied laser technology, industrial applications, and other topics of contemporary interest.

With this broad coverage of topics, the series is of use to all research scientists and engineers who need up-to-date reference books.

The editors encourage prospective authors to correspond with them in advance of submitting a manuscript. Submission of manuscripts should be made to the Editor-in-Chief or one of the Editors. See also www.springer.com/series/624

More information about this series at http://www.springer.com/series/624

Arthur McGurn

Nanophotonics

Springer

Arthur McGurn
Department of Physics
Western Michigan University
Kalamazoo, MI
USA

ISSN 0342-4111 ISSN 1556-1534 (electronic)
Springer Series in Optical Sciences
ISBN 978-3-319-77071-0 ISBN 978-3-319-77072-7 (eBook)
https://doi.org/10.1007/978-3-319-77072-7

Library of Congress Control Number: 2018933512

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to my wife, Maria.

Preface

This book is meant as an introduction to nanophotonics for students. It covers a number of topics that are important to the subject and which supply a basis for continuing on to a more advanced undertaking in the field. References to the recent literature and reviews of that literature are provided to direct the student on to a more advanced treatment.

The focus in the presentations is on analytical treatments. These often provide insight into the principles operating in a phenomenon. Such insights are not as readily available in computer simulation methods. However, work in nanophotonics is often simulation oriented so that the commonly used methods of computer simulation in electrodynamics are explained.

A chapter on mathematical preliminaries discusses common methods used in the study of composite and periodic media, providing an introduction to material science techniques. Simulation methods are also addressed, undertaking a basic consideration of the standard simulation techniques.

A focus is on presenting an introduction to photonic crystals, plasmonics, and metamaterials as foundations of many nanophotonic studies. Discussions of optical waveguides, circuitry, and impurity systems are presented for these types of media. In addition, topics of negative refraction, perfect lenses, and the propagation of radiation in metamaterials are introduced. Enhanced transmission is discussed along with a variety of other device-related properties in these types of engineered media.

Forces in nanosystems are discussed, involving a variety of magnetic, electric, and electrodynamic effects. These have various technological applications in biology, nanoscience systems, and nanomechanical devices. Included in these discussions are diverse nanoparticle mechanical interactions, the optical tweezer, Penning and Paul ion traps, and the Casimir force between surfaces.

A brief review of the properties of lasers is given. The approach is based on the ideas of nonlinearity and the similarity of the laser transition to that of a second-order phase transition. A focus is on discussions of the vertical column laser and on spasers. Unlike lasers, the spaser involves the generation of coherent surface plasmon polaritons in nanooptical systems.

Basic principles of near-field microscopy are presented. This technique introduces the possibility of sub-wavelength resolution of optical images. As such it provides for a significant advance in optical imaging technology.

A final topic is a discussion of the Einstein–Podolsky–Rosen paradox, the Bell's inequality, and an introduction to quantum computing. These have been of great current interest and provide a potential for the application of many of the ideas in nanophotonics.

In the history of the development of nanophotonics, a variety of different systems of units have been used to present new results. In the presentation given here, the development of the various results has been made in the original system of units in which they were formulated.

The author would like to thank the Department of Physics at the University of California, Riverside, for extending the use of the University Library. I also thank Ms. Robin De Haan in our Physics Department at Western Michigan University for help with Word. I thank Western Michigan University for providing the opportunity to write this book.

Rancho Mirage, California Arthur McGurn

Contents

Chapter 1
Introduction

In the last thirty years many new ideas focused on the nanosciences have been introduced into the field of optics [1–23]. These involve efforts in the development of: optical materials, technologies for the manipulation of light, techniques for the manipulation of atoms and nano- systems by means of light, applications of light in imagining and focusing with subwavelength resolutions, technologies of computer design and computation, new types of lasers and laser technologies, and realizations of nano-optical device designs. This is a partial list of the developing ideas that have rapidly advanced from year to year. In addition, each year new areas of the nanoscience applications of light are advanced and added to the list.

This book will focus on the theory of the operations of some of the above technological applications of light to the development of nanosciences. A point of particular interest will be on the basics of the earlier listed fields, giving a presentation of some of the theoretical ideas needed to understand the elementary functions proposed, applications of optical principles to the systems studied, and designs formulated in these fields.

The mathematical techniques that have been applied in nanophotonics will be introduced, developed, and illustrated with some applications to simple examples of the earlier listed applications of nanophotonics. In addition, some review of the experimental results for nanophotonics systems will be given. The book, however, is an introductory text to the field and is not meant to act as a comprehensive review of the fields presented here in outline. Rather efforts will be made to guide the student to the scientific literature in order for the student to begin on a fuller understand of the fields that are found of interest.

Some of the topics that are covered include: photonic crystals [1–6], metamaterials [7–10], plasmonics [1, 12], subwavelength focusing [7, 8], near-field scanning optical microscopy [17, 18], optical tweezers [19], some useful topics of quantum optics [20–23], trapped atoms [23], and ideas of quantum computing [21–23]. These fields have in common ideas for the manipulation of light on the nanoscale or the investigation of the interaction of nanoscale systems with light or the manipulation of nanoscale system though the application of light.

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_1

1.1 Mathematical Preliminaries and Examples of Specific Techniques

To begin with, some mathematical preliminaries will be developed which are most necessary for the study of nano-optical systems. An important initial division of the mathematical methods involves those techniques formulated to handle composite media and those formulated for the study of photonic crystal media. The first type of materials are disordered whereas the second type of materials involve ordered arrangement of media.

Mathematical techniques for the treatment of the dielectric response of general composites are treated [24–26]. Specific techniques are developed for the study of the refractive effects of composite systems on wavelengths of light which are large compared to the basic composite structure. The idea is to formulate the response of the composite system in an effective medium format.

In this approach the large scale properties of the composite are approximated and represented as properties of an homogeneous medium. The idea is then how best to choose the homogeneous medium with the effective composite response. As will be discussed later an effective medium treatment is most useful when the optical probe of the composite has a spatial variation which is large on the typical length scales defining the dielectric variation of the composite material.

The ideas used to treat composites are important in the study of many types of materials generated for technological applications and in particular can be of use in the study of metamaterials. Metamaterials, as discussed later, are media composed of artificially engineered nano-features [7–11]. They are designed to operate as homogeneous optical materials when interacting with the electromagnetic fields they are engineered to moderate. This is essentially the idea encountered in the refraction of light at visible wavelengths by glass. Glass is composed of a crystalline arrangement of atoms which on the atomic scale appears to be a discrete structure, however, on the scale of the wavelength of visible light the crystal is a homogeneous medium described by an index of refraction [10, 27, 28].

Techniques are also developed for the treatment of periodically varying dielectric systems. Specifically, these are systems designed to interact with light through their periodic properties. For this, the dielectric properties of such materials are chosen to be periodic over length scales of order of the wavelengths of the light that interacts with them.

Systems of this type are found in the study of photonic crystals which interact with light through the mechanism of diffraction. While metamaterials are designed to interact with light refractively, photonic crystals interact with light diffractively [1–5]. This difference in the interaction of these two classes of materials with light gives rise to great differences in the nanoscience applications of these two different types of media.

Important mathematical techniques in the study of new types of nanomaterials are computer simulation techniques [29–33]. In this regard, discussion of the basics of commonly used methods such as the finite difference time domain method, the method of moments, and finite element methods will be presented and discussed. These form the basis of many of the simulation studies that are published in the scientific literature and will be the source of some of the theoretical results on nanphotonics presented in the course of this book.

Other mathematical methods necessary for and focused on in the study of nonlinear optical systems will be presented [34, 35]. These include discussions of the mathematical properties and approaches necessary to discuss soliton modes and the generation of higher harmonics in nonlinear systems. In the treatment of solitons, some of the properties of bright, dark, and grey solitons in Kerr nonlinear media are studied. In addition, the propagation characteristics of these basic soliton modes will be treated in both metamaterial and photonic crystal models.

The origins, properties, and problems associated with the generation of second harmonics of radiation will also be discussed [36–40]. Refocusing on the materials side of these problems of nonlinear dynamics, the origins of the nonlinear polarization in systems exhibiting solitons and systems exhibiting second harmonic generation are explained [36–40]. Discussion are also given of the restrictions placed on the nonlinear polarization associated with these mechanisms and the symmetry properties of the crystal structure of the generating media as related to these restrictions [36–40].

The mathematics needed to understand the forces generated by electromagnetic fields on atoms and nanoparticles of matter are developed. These are important is some of the applications of optics to biology, problems of nano-engineering, and the confinement of trapped atoms. The last of these mentioned techniques of atomic manipulation is of potential application in the design of quantum computers. Some discussions of the elementary principles of quantum computing will be developed at the end of this work [41, 42].

1.2 Photonic Crystals

Photonic crystals are important to optical technology as they provide a means of diffractively molding the flow of light through space [1–5]. The photonic crystal is an optical system which has dielectric properties that vary periodically in space. Light with wavelengths of order of the spatial periodicity is diffracted by the photonic crystal, just as electrons moving in a semiconductor are diffracted by the periodic potential of the positively charged ions that binds the electrons and affects their motion in the semiconductor [43].

The effects of the periodic positive ion background on the electron motion in semiconductors is to alter the electron dispersion relation, opening a series of pass and stop bands in the semiconductor. The pass bands are regions of electron energy in which the electrons can propagate through the system, while the stop bands are regions of the electron energy in which the electrons cannot propagate through the system.

Similarly, the periodic dielectric constant of photonic crystals changes the dispersion relation of light in the photonic crystal, opening up a series of pass bands in which light propagates through the system and a series of stop bands in which light does not propagate through the photonic crystal. As in the case of the electronic semiconductors, the important effect here is the energy regions in which light does not propagate within the photonic crystal. These energy stop bands allow for molding the flow of light through space.

Photons with stop band energies cannot propagate into the bulk of a photonic crystal. For these photons the photonic crystals acts in a way similar to the action displayed by stones when they are placed in the bed of a stream. Stones in a stream can be used to rechannel the flow of water as it moves through the path of a stream. This follows as water cannot pass through the stones but must find a course around the stones. Similarly, light at stop band energies will not pass into the bulk of a photonic crystal so that the light is constrained to move only in the region outside of the photonic crystal. This constraint, arising from the electromagnetic band structure of photonic crystals, is nicely used in many device applications.

Photonic crystals can be designed to function as one-, two-, or three-dimensional devices. An example of a one-dimensional photonic crystal is a periodic layering of dielectric slabs [1–5]. Light incident perpendicular or nearly perpendicular to the interfaces of the slabs of the layered photonic crystal exhibits a band structure in the system. As a result only light at pass band energies is allowed to travel through the layering. Light with energies in a stop band is reflected from the layering. This effect is commonly used in laser mirrors and in dielectric coatings. In these designs an advantage of photonic crystals is that they can be made of layers of low conductivity dielectrics so that the system exhibits low energy losses. Such energy considerations are particularly important is the design of laser mirrors [2, 40].

Two-dimensional photonic crystals are formed of media with a spatially periodic dielectric variation in two-dimensions [2–5]. These types of photonic crystals have been used in various optical circuit applications [14, 44–53] and in the design of surface emitting lasers [53, 54]. A typical geometry of interest in these applications is a thin dielectric slab waveguide which has a periodic spatial dielectric pattern written into it. The periodicity pattern is chosen to be periodic in the large planar surfaces of the slab.

In the design of optical circuits the light acted upon by the photonic crystal is taken to move within the slab where it is confined to the slab by internal reflection at the slab surfaces and manipulated by its encounter with the patterning. In laser applications the slab photonic crystal acts as a Fabry-Perot resonator for a light source placed within a cavity contained within the photonic crystal slab. The periodic patterning of the slab is used to confine light that would otherwise propagate in the plane of the slab patterning, while the light is confined within the slab by internal reflection from the dielectric mismatch at the slab surfaces. These ideas can also be extended to three-dimensional systems.

In three-dimensional systems the band structure effects are important in suppressing the propagation of radiation in all spatial directions. Three-dimensional photonic crystals are important for additional applications to those discussed above for one- and two-dimensional systems. Some examples of these are in the enhancement or in the suppressing of thermal radiation emitted from matter and in various antenna design applications [52].

For all of the discussed photonic crystals technologies mentioned earlier, there are limitations placed on the functioning of photonic crystals in their proposed applications. Techniques for building nanostructures with photonic crystal patterning have improve rapidly through the years, but there is a continuing efforts to improve the quality of photonic crystals made in the laboratory. Due to the artificial nature of photonic crystals and the restrictions of current engineering practice, most of the photonic crystal applications have been applied to light between the microwave and the optical spectrum.

Some of the applications of photonic crystals are now qualitatively discussed. First some ideas of the Purcell effect from atomic physics are introduced, these are followed by discussions related to engineering and device applications of photonic crystals [48–56].

An early idea put forth was that photonic crystals could be used to suspend atoms in their excited states [2–5, 54–56]. This idea is based on the stop bands present in photonic crystals. If the energy of the photon emitted by an atom in the transition from its excited state to its ground state is in the stop band of a photonic crystal, the excited atom would not be able to emit the photon into the photonic crystal. Consequently, if it were in the bulk photonic crystal, it would be suspended in its excited state [56].

Under these condition, there would be no photon states available in the photonic crystal for the photon generated in the decay process to enter into and propagate away from the atom. This outcome is also readily seen from the Fermi Gold Rule transition rate equation in which the rate of atomic decay by photons from an excited state is proportional to the density of final photonic states available to accommodate the atomic transition.

Continuing this line of reasoning, remember that in semiconductor physics the electron density of states at the upper and lower edges of an electron stop band are enhanced [43]. Consequently, the electron density of states have maxima just outside of the stop band but near to both the upper and lower edges of the stop band.

In photonic crystals the same type of enhancement of density of states occurs near photonic stop bands in the density of states of the photonic crystal. As a consequence, atomic transitions with frequencies within the stop band are not allowed, but as the density of states is enhanced at the upper and lower edges of the photonic stop band the atomic transition rates can be increased for frequencies at these maxima [55, 56]. This follows from the proportionality of the transition rate to the density of photonic modes available for the decay photon to enter. Photonic crystals, consequently, offer mechanisms for both the suppression and for the enhancement of atomic excited state decay.

These ideas can be extended to the design of electromagnetic cavity resonators, waveguides, and circuits formed as networks of interconnecting waveguides [2–5]. A cavity resonator can be formed within the bulk of a photonic crystal by removing

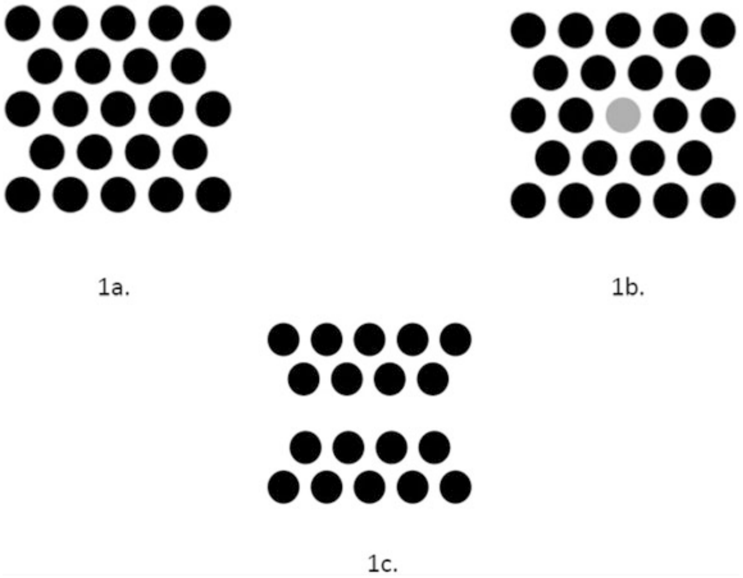


Fig. 1.1 Schematic drawings of a two-dimensional photonic crystal composed of parallel axis dielectric cylinders. The cylinder array is represented in the plane perpendicular to the cylinder axes for: **a** a triangle lattice patterning of photonic crystal, **b** a triangle lattice patterning with a resonant cavity, and **c** the triangle patterning surrounding a waveguide channel

a closed region of photonic crystal from the bulk of the photonic crystal. (See Fig. 1.1 for schematic drawings of the example of a two dimensional photonic crystal formed as a triangle lattice and the introduction into the photonic crystal of a resonator cavity.)

The resultant cavity will act as a cavity resonator for modes in the stop band of the bulk photonic crystal. For appropriately chosen materials the dissipative losses of such a cavity can be made to be much smaller than those found for cavities based on other technologies [2–5].

In a similar fashion waveguides are introduced into the bulk of a photonic crystal by surrounding a propagating channel by photonic crystal and sending electromagnetic waves at the frequencies of the stop bands of the photonic crystal to propagate down the channel. (See Fig. 1.1c for a schematic of a photonic crystal waveguide.) This is similar to the guiding applications of optical fibers. In fiber optics light is confined to the optical fiber through total internal reflection at the interface between the media of the optical fiber and the outside air. In photonic crystals the confining mechanism is the stop band effect of the photonic crystal surrounding the guiding channel. The photonic crystal confining mechanism is often more effective than the mechanism of total internal reflection found in fiber optics [2–5, 57]. For example, photonic crystals can exhibit low loses in waveguides with larger channel bends that are not possible in fiber optics technology.

Optical fibers often suffer high losses when they are bent through large angles and this is a limitation on their applications [57].

Schematics of a simple two-dimension photonic crystal formed as a periodic array of dielectric cylinders and a photonic crystal waveguide formed within it by removing a channel of dielectric cylinders are shown in Fig. 1.1. Light injected into the waveguide channel, propagating within the plane of the page at stop band energies, will propagate within the confining waveguide channel. In practical applications, two-dimensional photonic crystals are made in a slab geometry involving placing a periodic patterning parallel to the surfaces of the slab. This type of system has been used in various circuit applications and in the design of surface emitting lasers.

In circuit applications the slab is designed with various intersecting waveguide channels which allow the electromagnetic guided waves to pass through a variety of branchings. Light traveling this these circuits can be offloaded from the slab by means of optical multiplexers.

The slab geometry is not only effective in problems involving light traveling confined within the slab but has applications for light outside the slab. Light at normal incidence to slab photonic crystals [58, 59] can display enhanced transmission effects and filtering effects associated with the surface modes of the photonic crystal slab.

For two-dimensional photonic crystals the band structure is found to be dependent on the polarization of the modes propagating in the system [2, 60]. For photonic crystals patterned as an array of dielectric cylinders or as an array of cylindrical holes in an otherwise uniform dielectric medium, the modes are found to be polarized with electric fields polarized parallel to the cylinder axis or with their magnetic fields parallel to the cylinder axis. These two different polarization exhibit different dispersion properties in the system.

Consequently, the stop bands of the two polarization may not overlap one another. Early on the non-overlap was noted, and after some investigation it was found that in certain triangle lattice arrangements of dielectric cylinder or holes some of the stop bands can partially overlap. For these reasons, generally, triangular lattice have been used in the development of surface emitting lasers [44, 61].

A variety of three-dimensional structures have been investigated for photonic crystal applications. Again in three-dimensional systems the band structure can be polarization dependent and the modes may have similar polarization properties to those found in the harmonic modes of an atomic crystal. The diamond lattice of dielectric spheres was one of the earliest structures to exhibit complete photonic crystal stop bands in all directions of space [2–5]. Photonic band structures, however, have been computed for photonic crystals with a wide variety of three-dimension Bravais lattices. Some other three-dimensional arrangements have included three-dimensional layerings formed from nanoscopic strips or nanowires. In these types of photonic crystals the layering is built up as layers in which all of the strips of the layer are aligned in one direction. The strips of neighboring layers are arranged to be aligned in different directions. This is a type of system that can be built by various deposition processes which makes it of great experimental interest.

The nice property of three-dimensional photonic crystal is the ability to design systems with stop bands. This finds application in various antenna, sensor, and solar cell problems [63–65] and in controlling the thermal emission or the thermal signature of materials and shielded devices [63–65].

Photonic crystals have also been applied to the enhancement of some already existing technologies. An important example of this is in fiber optics technologies [57]. Fiber optical systems are often designed of a variety of materials so as to create a change in dielectric constant going from the center of the fiber to its interface with the air surrounding the fiber. This is done to improve the confining characteristic of the electromagnetic waves traveling along the fiber. One way to do this is to put a cladding or coating layer around the inner fiber forming the core of the optical fiber. In some recent applications the ideas of photonic crystals have been applied in the design of fiber cladding applied to optical fiber [66–68]. These are the so-called photonic crystal fibers [66–68].

In such systems a periodic photonic crystal pattern is introduced into the cladding of the optical fibers. The patterning is perpendicular to the axis of the fiber and is otherwise translational invariant along the axis of the fiber [66–68].

Enhancement of the fiber properties occur through two different mechanism in two different types of photonic crystal fibers. In a first type of photonic crystal fiber, the pass band-stop band properties of the photonic crystal patterning are directly applied. The frequencies of the guided modes are chosen to be within the stop bands of the cladding. This absence of transmission through the stop bands offers a more effective mechanism of confine light to the fiber than in the dielectric mismatch approach [66–68].

In a second type of fiber, the pass band properties of the photonic crystal fiber are not directly employed. Rather the photonic crystal patterning is used to manipulate the dielectric properties of the cladding materials, treating the photonic crystal cladding as a type of composite material that controls the system through its average dielectric properties [66–68].

Aside from their waveguide properties, photonic crystals fibers have found important applications in the design of sensors and fiber lasers. The laser design is based on doping the fiber so that the system can be pumped and operated as a laser [66–68].

Additional applications that have been suggested for photonic crystals are in the design of antennas [65]. Here they can be used to focus and improve the efficiency of antenna design. Other applications will be the focus of the Chapter on photonic crystals.

1.3 Metamaterials

Metamaterials are engineered materials that are designed to display, at certain frequencies of radiation, particular properties of permittivity and permeablitiy [7–11]. They are typically composite materials, formed by the inclusion of nano-features consisting of resonant structures, wires, etc., and are set to display the response of an homogeneous material at the wavelength at which they are designed to operate. This means that the engineered features forming the composite are generally small compared to the wavelength of the light with which the material interacts.

One of the original motivations for the study of metamaterials was in the design of materials that exhibit a negative refractive index. Naturally occurring substances are only found with positive indices of refraction, i.e., in Snell's law [69].

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{1.1a}$$

the indices n_1 and n_2 are only positive. (See Fig. 1.2 for a schematic of Snell's law for light incident in the first quadrant at a planar interface between two media with n_1 and n_2.)

This seems to be a fundamental limitation of nature, arising from the properties of the frequency dependent permittivity, $\varepsilon(\omega)$, and permeabilities, $\mu(\omega)$, in naturally occurring materials and their relation to the index of refraction [7–11, 14, 69]

$$n(\omega) = \sqrt{\varepsilon(\omega)\mu(\omega)}. \tag{1.1b}$$

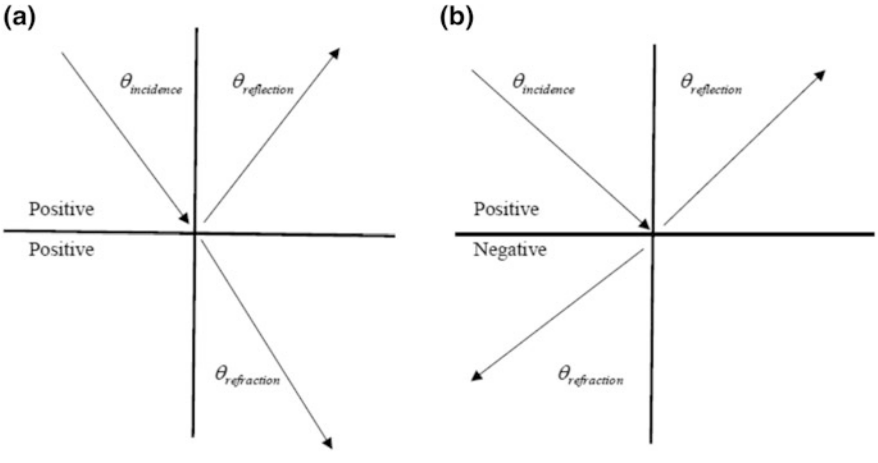


Fig. 1.2 Schematic showing the refraction of light at the planar interface between two different optical media where n_1 is the index of refraction of the upper medium and n_2 is the index of refraction of the lower medium. In **a** the refraction from the second to fourth quadrant is shown for light passing from a positive index medium to another positive index medium. In **b** the refraction from the second to third quadrant is shown for light passing from a positive index medium to a negative index medium. In the figures, $\theta_{incidence}$ is the angle of incidence, $\theta_{reflection}$ is the angle of reflection, and $\theta_{refraction}$ is the angle of refraction. In the figure the positive sense of $\theta_{incidence}$ is measure anti-clockwise from the vertical, the positive sense of $\theta_{reflection}$ is measure clockwise from the vertical, and the positive sense of $\theta_{refraction}$ is measure in the anti-clockwise sense from the vertical

While both the $\varepsilon(\omega)$ and $\mu(\omega)$ in naturally occurring materials can be positive or negative, materials have not been found to occur in which both $\varepsilon(\omega)$ and $\mu(\omega)$ are simultaneously negative at the same frequency.

From a study of the Maxwell equations and a subsequent derivation of the electromagnetic wave equations, it is found that in the case that both $\varepsilon(\omega)$ and $\mu(\omega)$ are negative the natural definition of the index of refraction occurring in Snell's law becomes [2–11, 14]

$$n(\omega) = -\sqrt{\varepsilon(\omega)\mu(\omega)}. \tag{1.1c}$$

This causes new physical effects to arise in the application of Snell's law which shall now be illustrated.

To illustrate the difference between refraction effects between positive and negative index media, consider the schematic illustrations in Fig. (1.2). In the figure n_1 is the index of refraction of the upper medium and n_2 is the index of refraction of the lower medium.

Figure (1.2a) qualitatively describes the refraction of light incident at a planar interface from the second quadrant of a positive index media, $n_1 > 0$, into the fourth quadrant of a second positive index media, $n_2 > 0$. Figure (1.2b), however, describes the refraction of light incident at a planar interface from the second quadrant of a positive index media, $n_1 > 0$, into the third quadrant of a second negative index media, $n_2 < 0$.

The difference between the refraction into the positive and negative index materials at the planar interface is the difference in the quadrant into which the refracted light enters. In the past, optics was limited in that light could only be refracted into paths similar to those shown in Fig. (1.2a). With the new metamaterials, however, light can now be refracted into paths similar to those shown in Fig. (1.2b). These features greatly expand the possibilities of optical design.

The earlier observations can be continued to the study of the planar interface between negative and positive indexed materials. In Fig. 1.3, if $n_1 < 0$ and $n_2 > 0$, (i.e., if the sign of the indices of refraction are reversed.) light incident on the interface in the second quadrant of n_1 will be refracted into the third quadrant n_2. This again is a new refractive response available in the study of optics [7–11, 14].

The expansion of the refractive properties of optical materials was then, in part, a great motivating factor in the search for designer materials that would exhibit negative index properties. An initial proposal for a metamaterial design was made by Pendry et al. [70–72] and requires the introduction of features of nano-technology.

An essential idea of the proposal is to generate a negative $\mu(\omega)$ by including resonant nano-features in the design of the metamaterials. The design of the nano-feature is made so that a negative $\mu(\omega)$ response is generated at frequencies for which $\mu(\omega)$ is only positive in naturally occurring materials. This extends the range of $\mu(\omega)$ and allows it to be adjusted. The nano-features used are known as split ring resonators and are basically designed to act as nano-circuits with the characteristics of inductor-resistor-capacitor, LRC, resonator circuits studied in

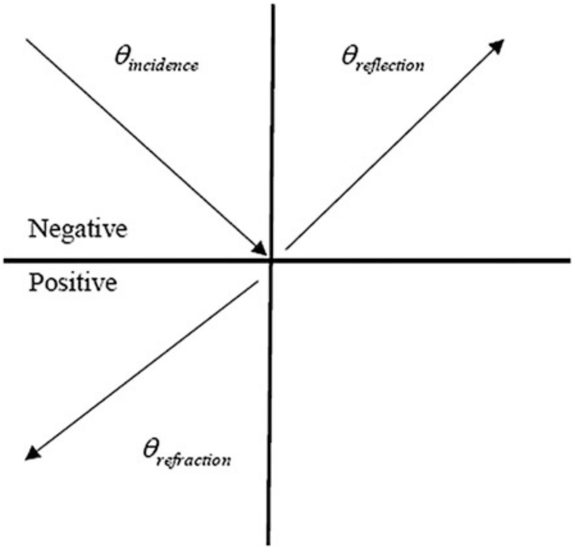


Fig. 1.3 Schematic showing the refraction of light at the planar interface between two different optical media where n_1 is the index of refraction of the upper medium and n_2 is the index of refraction of the lower medium. The figure shows the refraction from the second to third quadrant for light passing from a negative index medium to a positive index medium. In the figures, $\theta_{incidence}$ is the angle of incidence, $\theta_{reflection}$ is the angle of reflection, and $\theta_{refraction}$ is the angle of refraction. In the figure the positive sense of $\theta_{incidence}$ is measure anti-clockwise from the vertical, the positive sense of $\theta_{reflection}$ is measure clockwise from the vertical, and the positive sense of $\theta_{refraction}$ is measure in the anti-clockwise sense from the vertical

elementary physics courses. The resonators are tune so that $\mu(\omega) < 0$ at the same frequencies that the metamaterial has also been set to exhibit $\varepsilon(\omega) < 0$ [7–11, 14]. In nature such an arrangement has not been observed.

A difficulty with this approach can be seen in the requirement of a resonator circuit in the design of the metamaterial. This comes from the split ring resonator nano-features. In this regard, it is know from the Kramers-Kronig relations [69] that a resonance in the response of an electromagnetic system will exhibit an energy loss, and the loss will be greatest near resonance. This, in itself, of course, affects the propagation characteristics of the metamaterial, causing an energy decay as the waves propagate through the electromagnetic medium. Energy loss then presents a problem as it is near the resonances of the split ring resonators that the optimal effect for the generation of $\mu(\omega) < 0$ are found.

Another difficulty is that the resonance producing the $\mu(\omega) < 0$ is generated only over a narrow band of frequencies so that materials based on this mechanism tend to have a small frequency band of negative refractive index. The problems of fabrication of these materials have also restricted most applications to the microwave and terahertz regions [7, 12, 14]. These problems of loss, limited frequency bands, and fabrication are the focus of many current research efforts.

Returning to a consideration of the Maxwell equations in the case that both $\varepsilon(\omega)$ and $\mu(\omega)$ are negative, other important qualitative differences between the optics of positive and negative index materials are found. These follow from a direct application of $\varepsilon(\omega) < 0$ and $\mu(\omega) < 0$ within the Maxwell equations.

In the electromagnetic solutions of the Maxwell equation for positive index media, the Poynting vector, $\vec{S} = \frac{1}{\mu}\vec{E} \times \vec{B}$, of plane wave solutions is parallel to the wave vector, $\vec{k}$, of the radiation. In negative index media, however, the Poynting vector, $\vec{S} = \frac{1}{\mu}\vec{E} \times \vec{B}$, of plane wave solutions is anti-parallel the wave vector, $\vec{k}$, of the radiation [10].

For the plane wave solutions, the differences in the relationship between the wave vector and Poynting vector in the positive and negative index materials account for the unusual refractive properties observed in Fig. 1.2b. Because of the translational symmetry of the interface between the two materials in Fig. 1.2b, the wave vectors of the positive and negative index plane wave solutions must be conserved across the planar interface between the two media. The energy flow in the positive index medium flows in the direction of its wave vector so that for energy to travel towards the interface the wave vector of the plane wave must point towards the interface. In the negative index material, however, to have an energy flow away from the interface the wave vector of the plane wave solution must be opposite the direction of the energy flow. The wave vector must point towards the interface. For the wave vector in the negative index medium to point towards the interface and have a component of wave vector at the interface equal to that of the incident wave in the positive index medium the refracted wave must be in the third quadrant. Because of the translational symmetry of the interface, the reflected wave in the positive index medium, following the usual argument, is such that the angle of incidence is equal to the angle of reflection.

Similar arguments can be made for the refraction at a planar interface of a wave incident from a media with a negative index of refraction into a media with a positive index of refraction. (See Fig. 1.3 for a schematic for this case.) The energy flows in Fig. 1.3 follow from this.

An interesting consequence follows from the refractive properties illustrated in Fig. 1.2b for light going from a positive index medium into a negative index medium and in Fig. 1.3 for light going from a negative index medium into a positive index medium. From these properties a focusing lens can be designed which is a slab of negative index materials surrounded by a universe of positive index medium. Figure 1.4 illustrates the function of such a lens. The rays from an object in the positive index medium to the left of the slab lens pass through the lens and are focused to an image in the positive medium to the right of the lens. This is a very unusual property as a slab of positive index material does not form a focusing lens [7–11, 14].

In the optics of positive index media, the function of lenses to focus light is based not just on the refractive properties of the materials forming the lens but also on the curved surfaces of the lens. In positive index media the curved surfaces of the lens are an essential part of its operation, and lenses without curvature will not

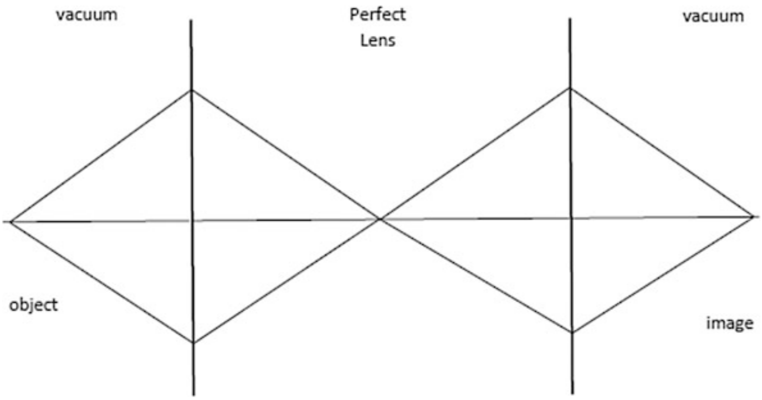


Fig. 1.4 The focusing properties of a perfect lens. The perfect lens is a slab of negative refractive index material with parallel planar surfaces. Three rays pass through the lens from the object to be reassembled at the image. The details of the properties and workings of the ideal lens will be discussed in Chap. 4

focus light [69]. In lens based on positive index media, the curvature of the surfaces are used to make up for the fact that the positive index medium refracts light only from the second to the fourth quadrant. Only materials which refract light from the second to the third quadrant are able to form a focus utilizing flat surfaces alone. Such materials bend light through a greater change in path, accommodating the addition path changes that are provided by the curvature of the surfaces in positive index lens.

A problem that arises in the design of positive index lens arises from the need to employ curved surfaces in their design. The curved surfaces of the lens must intersect, forming a lens with a finite lens aperture. The finite aperture of the lens introduces a fundamental limitation on its ability to resolve images at its focus. The light emanating from an object is composed of a complete spectrum of Fourier components of the space-time components of light carrying information away from the object. Due to the finite size of the lens aperture only a restricted set of Fourier components are allow to pass through the lens. This is a basis of the Rayleigh criterion for the resolution of lenses with apertures and would not be restriction in the case of slabs which have infinite apertures [69].

This, however, in not the only property of slab lens in the new optics based on negative index of refraction materials. The slab lens of negative index of refraction can also project evanescent wave between the object on one side of the slab and the image formed on the opposite side of the lens. Consequently, all of the optical components emitted by the object can be potentially assembled in the focused image. The image would be a perfect replica of the object. Because of this property, the slab of negative index medium is often referred to as a perfect lens or a superlense [71].

Negative index materials have also been applied in so-called optical cloaking devices [70]. These are based on the ability afforded by the new metamaterials of designing media in which the dielectric constant of the material can assume values ranging over the entire set of real numbers. A ray of light traveling in such a system can be bent through any angle within it forward quadrants as it passes through the medium. In principle it is then possible to create a metamaterial with a spatially varying dielectric constant, exhibiting any positive or negative index of refraction value at any point in space. This is done by spreading different types of nano-features throughout the volume of the material, choosing the feature appropriate for the desired index of refraction.

The idea of the cloaking device is to create a metamaterial with a spatially varying index of refraction designed to gradually refract rays of light around an object and send them out from metamaterial along paths the light would have taken if the cloak and object were not present. In Fig. 1.5 a schematic of such a device is shown, for a two-dimensional system. An object is placed within a cylinder of metamaterial in the form of a hollow cylinder surrounding the object to be hidden. The metamaterial is assumed to be designed with a position dependent index of refraction engineered for the function now proposed.

In the figure, light is incident on the system from the left as a series of parallel rays. As the light encounters the cylinder of metamaterial it is successively refracted around the hidden object and passes out of the metamaterial along the parallel paths that it would have taken in the absence of the metamaterial and the object it cloaks [70].

Experiments and computer simulation studies on such devices were originally made by Pendry et al. [70]. The systems designed involved the distribution of split ring resonators through a cylinder of materials, varying the resonator configuration along the circuit of the cylinder. This provided the spatially varying index of refraction of the cylinder. The limitations on the functioning of the system designed in this way are, consequently, due to the resonant nature of the split ring resonator

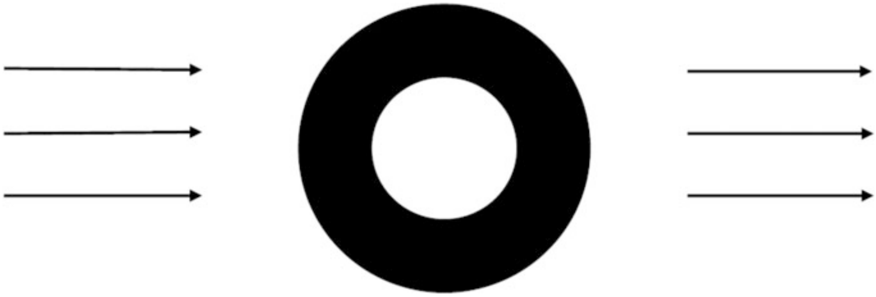


Fig. 1.5 Schematic of the application of a metamaterial in the form of a hollow cylinder to hide an object placed within the cylinder. Parallel rays of light incident from the left are sent through the system, exiting as parallel rays on the right of the system. The object within the cylinder is not evident in the light passing to the right of the cylinder

function. These include the restriction of the effect to a narrow band of frequencies associated with the resonators and the problems of loss again associated with the resonator operation [7–11, 70].

With the freedom to engineer materials with arbitrarily spatial varying dielectric properties many types of system applications have been suggested. Some of these include the design of metamaterials to exhibit properties found in systems of general relativity [73–75], these are based on spatial coordinate transforms [74] that can be related to spatially dependent permittivities and permeablities.

Other topics in metamaterials are the so-called hyperbolic materials [76] which are introduced as a means to achieve negative refractive properties. These materials exhibit unusual properties due to their optical dispersion relations, e.g., strong enhancement of spontaneous emission, negative index of refraction, and enhanced superlensing effects. Some of these properties have found applications in the design of metamaterial surfaces formulated for the modulation of light incident on them. The ideas of metamaterials have also been extended to the design of acoustic materials used to manipulate the properties of phonon systems.

1.4 Plasmonics

Plasmonics technologies are based on the excitation and use of surface electromagnetic waves known as surface plasmnon-polaritons to perform device functions [82–89]. Surface plasmons-polaritions occur at the interface of certain types of materials and exhibit a dispersive propagation along the interface. The nature of the dispersion depends on the dielectric properties of the media forming the interface and the nature of the interface geometry over which the plasmons travel.

Surface plasmon-polaritons are a type of electromagnetic plane wave that is bound to and travels along the interface with electromagnetic fields that have concentrated field intensities at the interface. Away from the interface the fields decrease to zero at infinite separation from the interface [82]. Surface plasmon-polariton modes exist on a wide range of planar and curved surface geometries and are responsible for a range of important optical phenomena, some of which are of technological importance. Applications based on surface plasmon excitations include: circuit applications, waveguides, sensors, enhanced field transmission, and other particular device based designs [82–89].

Commonly studied examples are the plasmon-polaritons at the interface of a metal and dielectric, but these excitations may also be found on the interface between two different dielectrics or on metal and dielectric slabs and thin films [82]. In the absence of dielectric losses from the media forming the interface, surface plasmon-polaritions exist indefinitely on ideal planar interfaces. With the introduction of surface corrugation or roughness, however, the plasmon-polaritions tend to scatter into electromagnetic modes that propagate away from the interface and into the bulk.

Surface plasmon-polaritions are closely related to the study of the refraction of light at an interface between two media. They first appear in the elementary treatments of the properties of light at the planar interface between media. Here they show up as a set of optical solutions which are distinctly different from those involving reflected and refracted waves of light incident on the interface [82].

In this regard, it is a common text book study to treat the electrodynamics of the refraction of light at a planar interface between two different media [69]. (See the schematic in Fig. 1.6a.) The Maxwell equation solution of the problem considers a plane wave incident from one media onto the interface. Upon encountering the interface, part of the incident wave is reflected back into the media through which the incident wave traveled and part of the incident wave is transmitted into the media on the other side of the interface as a refracted wave traveling away from the interface. These are, respectively, the reflected and refracted waves and represent one important class of electrodynamic solutions at the interface.

However, depending on the dielectric properties of the two media a second class of important solutions often exist on the interface. (See the schematic in Fig. 1.6b.) From a similar treatment to that in the treatment of refraction, one can often obtain the surface plasmon-polariton solutions at the planar interface between the two media. These propagate parallel to the interface between the two media and have fields which decrease in intensity with the separation from the interface. They are no more difficult to study than the solutions for refraction at an interface but are usually omitted in standard treatments of the electrodynamics of the planar interface.

The solutions of the surface plasmon-polartion modes at the interface involve treating the electromagnetic boundary conditions for waves traveling parallel to the interface in the two media [82]. In addition, to complete the solution an additional boundary condition is required that the fields in the media decay to zero at infinite separation from the interface. From these arise the surface plasmon-polariton modes

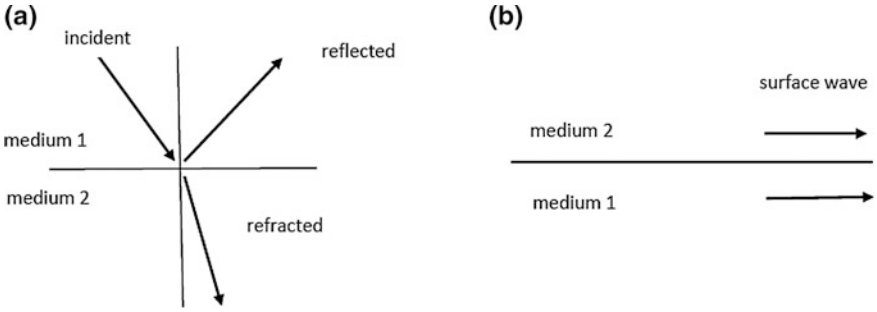


Fig. 1.6 Schematics for: **a** the refraction of light at a planar interface between two media and **b** the surface plasmon-polariton modes traveling at a planar interface. The surface plasmon-polaritions propagate parallel to the interface between the two media and have fields which decrease in intensity with the separation from the interface

which along with the refractive solutions are required to understand all of the electrodynamics of the media and the interface.

A particular example of the importance of surface plasmon-polaritons is in the study of light scattering from rough surfaces supporting surface plasmon-polaritons [83, 84]. For such systems, it is shown that the diffuse scattering from the rough surface involves the coupling of bulk electromagnetic waves into and out of the surface plasmon-polariton modes. This coupling is provided by the surface roughness. As an example, it can be shown that the Anderson localization of surface electromagnetic waves at the interface is responsible for certain important backscattering enhancement from the interface.

In a similar way, the excitation of surface plasmon-polaritons is often important is surface enhanced Raman spectroscopy [85]. In this phenomenon, the intense fields of surface plasmon-polaritons excited on an interface can increase the spectroscopic signals detected from molecules on the surface. An incident wave on the surface couples to and excites surface plasmon-polaritons which are then used to create spectroscopic transitions in molecules bound to the interface.

Related phenomena that utilized the enhanced fields of surface electromagnetic waves to enhance physical effects are enhanced transmission phenomena of screens and systems developed for near field microscopy [86, 87]. Thin films with a periodic patterning of subwavelength holes can exhibit an enhanced optical transmission. The enhancement is due to surface plasmon polaritons which travel through the holes to given them an enhancement over the transmission observed in the absence of surface electromagnetic waves.

Near field microscopes also utilize surface plasmon-polaritons [17, 18, 88]. In this case the surface electromagnetic waves increase the resolution of the near field microscope significantly over that of far field systems. This is due to the increased information carried in surfaces waves and the conversion of surface wave information to bulk electromagnetic waves by probe scattering. As a result, the system is found to offer a significant subwavelength resolution. This resolution increase represents a fundamental increase over that of far field microscopic techniques.

A final important example to nano-circuit technology is found in the formulation of laser-like systems involving surface-plasmon polaritons. These are the basis of the design of spasers [89]. Where the laser operates on the simulated emission of photons to produce an amplified coherent beam of photons, the spaser does this for the creation of a coherent beam of plasmon-polaritons. These provide for an easier and more natural coupling of light into plasmonic circuits and nano-devices.

1.5 Nonlinear Properties of Nano-optical Systems

The ideas of metamaterials, photonic crystals, plasmonic surfaces, and other nano-systems can be extended to include designs utilizing materials of nonlinear optics [34–40, 45–50, 81]. Nonlinear optical materials are of interest as the properties that they display depend on the intensity of the electromagnetic fields interacting with them. This shows up in a change of the index of refraction as the intensity of the electric fields applied to the nonlinear material are increased. It can also show up within some materials in the generation of higher harmonics of an initially applied harmonic electromagnetic mode [40, 81].

Nonlinear effects ultimately arise from the field dependence in electromagnetic materials of the electric polarization on the fields applied to them. In linear materials the electric polarization has a simple relationship to the applied electric field, i.e., it is found to be directly proportional to the applied electric fields. For non-ferroelectric materials the linear polarization is the first term in a Taylor expansion of the electric polarization in the applied field, and it is generally the dominant term of the expansion. The remaining terms of the expansion of the polarization, containing higher powers of the field, are responsible for the effects studied in nonlinear optics. These include the Kerr effect, and the generation and mixing of higher harmonics of radiation within the material [34–40].

In a crystalline material, the polarization and the applied field are defined relative to a crystal lattice containing chemical features which are repeated throughout the lattice [34–40]. Both the lattice and the chemical features possess certain spatial symmetries which are important in determining the properties of the polarization and its response to an applied field. In particular, the symmetry properties of the material are essential in determining the nonlinear effects displayed by the material. As a result, it is found that the tensors describing how the polarization vector is related to the various components and products of components of the applied electric field are accordingly symmetry restricted.

An important example of symmetry restrictions involves second harmonic generation [40]. Crystalline symmetry considerations are particularly important in the study of the generation of second harmonics of radiation, accounting for the absence of the phenomena in many materials. In particular, as a technologically important example, symmetry considerations are responsible for the absence of second harmonic generation in the bulk of metals. Metals generally have symmetry groups that are inconsistent with second harmonic generation. It is only at the surface of a metal were the surface breaks the symmetry of the bulk metal that second harmonics can be generated in metallic systems [40].

A particularly interesting class of materials are those exhibiting Kerr nonlinearity [34–40]. In these types of systems the dielectric constant of the nonlinear media depends on the intensity of the electric field applied to the material. This forms the basis of optoelectronic interest in Kerr media where it offers considerations of possible use in the design of optical switches, transistor, and diodes. Here the intensity of the applied electric field can cause a change in the dielectric response of a nonlinear material that is functioning as part of an optical device. The change in the dielectric response is used to modulate the output of the device, leading to a type of switching functionality. In addition, multiple beam of light can be caused to interact with each other in a Kerr material. This allows for the beams of light to modulate one another's transmission or reflection from the material.

The Kerr nonlinearity also leads to a number of interesting nonlinear optical excitations to exist in some nonlinear systems [34–40]. In systems formed of linear media the excitations are the linear optical modes. These linear modes usually arise as solutions of eigenvalue problems generated from the Maxwell equations. From these the general solutions of the system are, consequently, written as a linear combination of linear eigenmodes. This idea of a general solution expressed in linear combinations of modal solutions is no longer the case in nonlinear systems.

In the limit of small nonlinearity there are solutions which look like renormalized versions of the linear modes of the linear limit of the system, but because of the nonlinearity in the system, these solutions do not form linear combinations which are also solutions of the nonlinear system [34]. In addition, there exist completely new types of excitations in nonlinear systems that do not have counterparts in linear systems [35]. The nonlinear system is much more difficult to approach theoretical than the linear system, and often there are no standardized methods for finding and generating solutions for the excitations in nonlinear systems.

New, technologically important, types of excitations are found in nonlinear media solutions which are seen to vanish from the system in its linear limit. These are unlike the modes of the linear media systems as they have unique propagation properties unlike modes of the linear media. The new class are soliton modes and multiple solitons modes. These have been objects of study in many types of nonlinear systems, and they have a long history in many branches of physics, engineering, and applied mathematics. The soliton modes that will be treated in this book and that are of primary interest in nonlinear optics are bright, dark, and grey solitons [35–40].

Solitons arise in nonlinear media because the dielectric constant of the media depends on the intensity of the fields applied to it. In systems formed of linear media it is usually possible to introduce localized dielectric impurities to the system in such a way that the impurity media can support electromagnetic modes that are bound to the impurity media. These are localized modes of the linear system which are bound to the region in the vicinity of the impurity. In nonlinear optical media, the intensity of an electromagnetic solution changes the dielectric response of the media that binds the electromagnetic solution to the system. If the electromagnetic solution is of the form of a localized pulse, it is possible for the field intensity of the pulse to generate a dielectric response of the nonlinear medium so as to support the pulse as a solution of the system. This is the origin of pulses known as bright solitons. It represents a pulse of electromagnetic energy that is propagated through the nonlinear media.

In addition, it is sometimes found that an intensity dip in the electromagnetic fields can, through intensity modification of the dielectric of the nonlinear medium, in the same way be supported by the nonlinearity of the system [35]. This type of solution is known as a dark soliton. It represents a decrease in the electromagnetic field and energy which is propagated through the system. A solution generated as a partial dip in field intensity can also exist and propagate through the nonlinear media. This known as a grey soliton. It is similar to the dark soliton but never has an extinction of the field energy density in space. The bright, dark and grey solitons are the single soliton solutions encountered in the study of nonlinear optics.

Solutions representing combinations of the various solitons and extended nonlinear wave forms may also be generated and studied in nonlinear media [35]. These are multiple soliton and nonlinear wave interactions. It is important to note that while the components of these multiple soliton modes have similar appearances to the single solitons and/or the multiple mode solutions, they are not linear combinations of the single soliton and nonlinear wave solutions. The multiple soliton solutions are of interest in the study of the scattering of the various components of the system from one another as the system evolves in time. In this way, all of the solitons and multiple solitons generated in Kerr nonlinear media arise though the mechanism of self-consistent interaction of the solution with the nonlinear medium which in turn supports the solution. In this book only nonlinear wave and single solitons will be treated.

The solutions of systems formed of nonlinear optical media allow for important new physics and engineering properties to be displayed for device applications. The possibility of excitations to interact with themselves and with other excitations of the nonlinear media provides the basis for the design and functioning of optical transistor, diodes, and switches [36–40, 45–50]. In addition, nonlinearity expands the excitations possible in the system adding new features to the response of the system.

The transmission and reflection properties of devices formed of nonlinear media are found to exhibit anomalies associated with the various excitations that can be found within the nonlinear media. This may be an important element in the determination of the functioning characteristics of devices employing these types of materials. In addition, bright soliton modes find interest in system designs as the field intensity peaks that they display can be used as a means of generating enhance electromagnetic fields in nonlinear materials. Such enhanced fields may have important applications in the generation of second harmonics of radiation and in the design of optical switches.

In fiber optics systems bright solitons may also provide an effective means of transmission of information as they offer energy efficiency and stability [90]. This should also extend to the current study of these solutions in applications of photonic crystals and metamaterial devices. These type of applications are currently developing as areas of interest in telecommunications and offer great promise.

Both photonic crystal and metamaterial technologies have also been applied to the enhancement of second harmonic generation [81]. This is another field of technological potential in the applications of the ideas of nonlinear optics. Second harmonic generation in these materials is now discussed.

A second class of materials that are important in nano-science applications are those used in the generation of second harmonics of radiation [36–38]. Second harmonics generation results when an harmonic of light is introduced into a nonlinear medium which has the appropriate crystal symmetry for second harmonic generation. Once introduced into the medium, the light interacts with the nonlinearity of the medium so as to create a new component of light in the system at twice the frequency of the introduced harmonic. The generation of second harmonics has applications in lasers technologies, microscopy of biological systems, and in spectroscopies of molecules and surfaces. In these applications the composition of some of the laser components or of the biological samples or molecules studied have second harmonic generating components associated with them.

The generation of second harmonics has many technological applications, but there are also a number of problems associated with the technology employed for second harmonic generation. As shall be shown later many of these difficulties are naturally solved using photonic crystal and metamaterial technologies.

The problems of creating high intensity fundamental harmonics to enhance the terms of the frequency doubled time components of the nonlinear polarization has already been mentioned. Along with this problem is a material science consideration of finding materials that, during their use in the generation process, will handle intense fundamentals applied to them without having a breakdown of the material. Solutions to these problems are topics of much research. The focus is on endeavoring through material science to find new materials with more efficient conversion properties and to understand what features of materials will assist in the formulation of more efficient conversions. Even if these problems are overcome, however, it does not necessarily result in an efficient generation process.

An additional important consideration during the generation of second harmonics is the problem of phase matching. As the fundamental propagates through a uniform medium of nonlinear material, it can continuously generate second harmonic radiation all along its trajectory. The waves generated at different points along the trajectory will have different phases. This means that waves from particular regions of the trajectory will add constructively with waves generated along other parts of the trajectory. However, waves generated in particular regions along the trajectory will also encounter waves from other parts along the trajectory that will add destructively with them. When considered as a whole, the destructive and constructive inference effects result in second harmonic generation which is of low efficiency in uniform homogeneous systems [36–38, 40].

1.6 Forces

A topic of interest in nanophotonics involves the application of force to nano-particles and atoms [41, 91, 92]. This can provide a means of manipulating them or trapping them is space. Both applied electric and magnetic fields can exert forces on nanoparticles [41, 91]. These may be static fields with a spatial gradient or time dependent fields. The interaction of electromagnetic waves with individual particles, in itself, provides a means of spatially manipulating the particles [92].

For static fields, the energy of interaction with particles is described by the field energy density so that a particle experiences an electromagnetic force as it moves in space to lower its energy [41]. This also applies to the interaction with electromagnetic waves [92]. In this case, it may be seen from the conservation of momentum and the change in the radiation pressure of an electromagnetic wave as it scatters with a nano-particle.

These are very basic electrodynamic features of the interaction of particles and fields, but they have interesting applications for investigations in biology and other nano-particle systems. Here they have been used to direct the motions of bacterium and in combination with chemical reactions to create the motion of mano-projectiles [41]. This is a form of nano-machine [41].

At the atomic level, an important problem is the trapping in space of individual ions [91]. This requires the application of time dependent electromagnetic fields as it is a theorem of electrostatics that a static trapping potential cannot be formulated in three-dimensional space. The time-dependent fields are adjusted so that, on time average, they provide a trapping mechanism in three-dimensions. This is done in one of two basic trapping schemes which form the basis of the Penning and Paul traps.

An important application of trapped atoms, aside from studies in quantum optics, is in the proposed designs for quantum computers [91]. Here there are, however, many other types of quantum systems that have been of interest in quantum computing.

A final important topic in nanoscience forces is the Casimir force [93]. This is a force arising between surfaces due to the vacuum fluctuations in quantum electrodynamics. The force arises from the change in zero-point energy with changes in the surface configuration and the associated field boundary conditions. It is a short-ranged interaction which can be significant on the scale of nanoscience systems.

A similar type of force is observed in the attractive force between two closely parked boats arising from the wave motion on the water between the two vessels. The force again arises from the changing energy of the waves as the boundary conditions change with the vessel separation. In this analogy, it should not be forgotten that the Casimir force is only present in quantum systems whereas the boat analogy is solely a classical mechanics based phenomenon.

1.7 Near-Field Microscopy

Another important technology based on nanoscience phenomena is near-field microscopy [94–98]. This is the key to subwavelength imaging which allows for the study of the features of systems which cannot be imaged in far-field microscopy. It is a recent development bases on a proposed system suggested by Synge in 1928 [94, 95]. As it required a close, measured, approach to the surface being studied, it required many years of development for its implementation [96–98].

Due to a fundamental limitation from the principles of microscopy, the far-field microscope has a limited resolution that it can achieve. This is a limitation fixed by the wavelength of the radiation and arises from the nature of the dispersion relation of light and the far-field nature of the collection process.

An important aspect of the formulation of the near-field system is that the near-field microscope measures the components of the evanescent waves from the object being imaged. The measurement of these components is required to accurately represent the image generated within the microscopy process. The evanescent components decay exponentially with separation from the object so that they are absent in the collected fields of far-field devices.

While far field microscopes involve the use of lenses in order to collect and process the collected radiation into an image, this is not the case with the near-field microscope. An important component of a near-field microscope is a probe which transforms the evanescent field waves of the object into propagating waves carrying the subwavelength features of the information gathered from the object.

The development, understanding, and interpretation of the images obtained by a near-field microscopy scheme is a recent, ongoing, project [94–98]. To this end many different probe and collection arrangement have been employed.

1.8 Quantum Computers

The development of new means of computation has been an important aspect of nanoscience. Much of this work has been directed at applying new ideas of optics to develop conventional computers based on the ideas formulated in classical studies of computation. Some of the work on computers, however, has been focused on new ideas of quantum computation [99–104]. These involve the development of the unique properties of quantum mechanical systems in the design of massively parallel computational arrangements. For now, most of this work is theory, but there is a continuing process focused on the experimental implementation of these ideas.

The essential property in the development of quantum computers is that of the ideas of superposition of quantum states and entanglement [99–101]. These are essentially quantum mechanical properties which arise as the quantum mechanical formulation is based on studying the dynamics of probability amplitudes rather than probability distributions. Entanglement properties are at the basis of the Einstein-Podolsky-Rosen paradox, and it was shown by Bell to be responsible for the difference in the idea of probability in quantum and classical mechanics. Ideas of superposition of states and entanglement are fundamental in developing the parallel algorithms which would function on quantum computers.

In a classical computer, the operations are based on switches which can register as on or off. By arranging a sequence of a number of such switchings, information can be stored and processed. In these processes, however, a switch can only be on or off. This is not the case in quantum computing.

In a quantum computer, the switch can be represented by an atom so that in the ground state the atomic switch is off and in the excited state the atomic switch is on [102–104]. In the quantum mechanics of an atom, however, the atom can be in a superposition of the ground and excited states. This and the entangled state involved in the probability amplitudes of states of multiple atomic systems is a new feature of quantum systems. It does not occur in classical mechanics, and these new features are found to allow for novel types of parallel computation.

In a quantum computer, a superposition of states or entangled state of quantum switches can be developed as an initial state of data input. The mixed input should consist of equality weight states of all possible input data to the computer. The input state is then acted upon by an algorithm to generate an output of superposed or entangled states composed from states each consisting of an original input state associated with its output from the computation.

The output contains the results from the calculation made on each of the inputs making up the input superposed state. It is a superposed state composed of both input and their associated output states with each such composite state in a mixture of all input-output composite states. The focus of the program for this processing is to design an algorithm which develops an output superposed state which predominantly contains the input-answer data which is of interest to the programmer. Consequently, by redoing the computation a few times it should be possible to determine the correct answer to the problem being solved.

This type of process is shown theoretically to be effective, for example, in the factorization of large integers. Using conventional methods of classical computation this factorization problem quickly becomes intractable. It is found in a number of computation processes that the parallel nature of the quantum computation represents an increase in efficiency over the traditional processes of classical computation.

1.9 The Focus of the Book

The outline of the topics given above will form the focus of the topics presented in the following chapters. The object is to present the basic ideas involved in the phenomena which form the study of nanophotonics. Rather than aiming at a comprehensive presentation or a formal review, the materials are meant as an initial help to the student to progress in the student's interests in the field of nanophotonics.

Reviews have been cited in each of the areas addressed. These should facilitate the initial steps in each of the fields discussed.

References

1. P.N. Prasad, *Nanophotonics* (Wiley-Interscience, Wiley, Inc., Hoboken, New Jersey, 2004)
2. J.D. Joannopoulos, P.R. Vilenueve, S. Fan, *Photonic Crystals* (Princeton University Press, Princeton, 1995)
3. K. Sakoda, *Optical Properties of Photonic Crystals* (Springer, Berlin, 2001)
4. P.N. Favennec, *Photonic Crystals: Towards Nanoscale Photonic Devices* (Springer, Berlin, 2005)
5. A.R. McGurn, in *Survey of Semiconductor Physics,* ed. by W. Boer (Wiley, New York, 2002) Chapter 13

6. R. Waser (ed.), *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices*, 2nd edn. (Wiley-VCH, Weinheim, 2005)
7. W. Cai, V. Shalaev, *Optical Metamaterials: Fundamental and Applications* (Springer, New York, 2010)
8. N. Engheta, R.W. Ziolkowski (eds.), *Metamaterials: Physics and Engineering Explorations* (IEEE Press, Wiley-Interscience, Wiley, Canada, 2006)
9. S.A. Ramakrishna, Physics of negative index materials. Rep. Prog. Phys. **68**, 449 (2005)
10. V.G. Veselago, The electrodynamics of substances with simultaneously negative values of ε and μ. Sov. Phys. Usp. **10**, 509 (1968)
11. P. Markos, C.M. Soukoulis, *Wave Propagation* (Princeton University Press, Princeton, 2008)
12. Y.-S. Lee, *Principles of Terahertz Science and Technology* (Springer, Berlin, 2009)
13. W.J. Smith, *Modern Optical Engineering*, 4th edn. (McGraw-Hill Education, New York, 2007)
14. A.A. McGurn, *Nonlinear Optics of Photonic Crystals and Meta-Materials (IOP Concise Physics)* (Morgan and Claypool Publishers, San Rafeal, 2015)
15. R. De La Rue, J.-M. Lourtioz, S. Yu, *Compact Semiconductor Lasers* (Wiley, London, 2014)
16. H. Benistry, *Confined Photon Systems: Fundamentals and Applications* (Springer, Berlin Heidelberg, 2007)
17. M.A. Paesler, P.J. Moyer, *Near-Field Optics* (Wiley-Interscience Publication, Wiley Inc., London, 1996)
18. M. Ohtsu, K. Kobayashi, *Optical Near Fields* (Springer, Berlin, 2004)
19. D.L. Andrews, *Structured Light and Its Applications* (Academic Press, London, 2008)
20. A. Griggin, T. Nikuni, E. Zaremba, *Bose-Condnesated Gases at Finite Temperatures* (Cambridge University Press, Cambridge, 2009)
21. M. Nakahara, O. Tetsuo, *Quantum Computing: From Linear Algebra to Physical Realizations* (CRC Press, Hoboken, 2008)
22. S. Barnett, *Quantum Information* (Oxford University Press, Oxford, 2009)
23. C.C. Gerry, P. Knight, *Introductory Quantum Optics* (Cambridge University Press, Cambridge, 2005)
24. R.J. Elliott, J.A. Krumhansl, P.L. Leath, The theory and properties of randomly disordered crystals and related physical systems. Rev. Mod. Phys. **46**, 465 (1974)
25. D.J. Bergman, D. Stroud, Physical properties of macroscopically inhomogeneous media, in *Solid State Physics*, vol. 46, ed. by H. Ehrenreich, D. Turnbull (Academic Press, Boston, 1992), pp. 146–269
26. D.J. Bergman, The dielectric properties of composite materials-a problem in classical physics. Phys. Rep. **43**, 378 (1978)
27. V.M. Agranovich, YuN Gartstein, Spatial dispersion and negative refraction of light. Phys. Usp. **49**, 1029 (2006)
28. V.M. Agranovich, Hybrid organic-inorganic nanostructures and light-matter interaction. in *Problems of Condensed Matter Physics,* ed. by A.L. Ivanov, S.G. Tikhodeev (Clarendon Press, Oxford, 2006), Chapter 2
29. A. Taflove, *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House, Boston, 1995)
30. W.C. Gibson, *The Method of Moments in Electromagnetics* (Chapman and Hall/CRC, Boca Raton, 2008)
31. H.M. El Misilmani, K.Y. Kabalan, M.Y. Abou-Shahine, M. Al-Husseini, A method of moment approach in solving boundary value problems. J. Electromagn. Anal. Appl **7**, 61 (2015)
32. S. Humphries, *Finite-Element Methods for Electromagnetics* (Field Precission LLC, Albuquerque, 2010, CRC Press, 1997)
33. X.-Q. Sheng, W. Song, *Essentials of Computational Electromagnetics* (IEEE Wiley, Singapore, 2012)

34. A.H. Nayfeh, *Introduction to Perturbation Technique* (Wiley, New York, 1981)
35. T. Dauxois, M. Peyrard, *Physics of Solitons* (Cambridge University Press, Cambridge, 2006)
36. R.W. Boyd, *Nonlinear Optics*, 2nd edn. (Academic, Amsterdam, 2003)
37. D.L. Mills, *Nonlinear Optics* (Springer, Berlin, 1998)
38. P.P. Banerjee, *Nonlinear Optics* (Dekker, New York, 2004)
39. Y.S. Kivshar, G.P. Agrawal, *Optical Solitons* (Academic, Amsterdam, 2003)
40. A. Yariv, *Introduction to Optical Electronics*, (Holt, Rinehart, and Winston, Inc., New York, 1971) and A.R. McGurn, T.A. Leskova, V.M. Agranovich, Weak localization effects in the generation of second harmonics of light at a randomly rough vacuum—metal grating. Phys. Rev. **B44**, 11441 (1991)
41. R.S.M. Rikken, R.J.M. Nolte, J.C. Maan, J.C.M. van Hest, D.A. Wilson, P.C.M. Christianen, Manipulation of micro- and nanostructure motion with magnetic fields. Soft Matter **10**, 1295–1308 (2014)
42. R.F. Ismagilov, A. Schwartz, N. Bowden, G.M. Whitesides, Autonomous movement and self-assembly. Angew. Chem. Int. Ed. **41**, 652 (2002)
43. M.P. Marder, *Condensed Matter Physics*, 2nd edn. (Wiley, Hoboken, 2010)
44. H. Benistry, V. Berger, J.-M. Gerard, D. Maystre, A. Tchelnokov, *Photonic Crystals: Towards Nanoscale Photonic Devices*, 2nd edn. (Springer, Berlin, 2008)
45. M. Scalora, J.P. Dowling, C.M. Bowden, M.J. Bloemer, Optical limiting and switching of ultrashort pulsesin nonlinear photonic band gap materials. Phys. Rev. Lett. **73**, 1368 (1994)
46. M. Scalora, J.P. Dowling, C.M. Bowden, M.J. Bloemer, The photonic band edge optical diode. J. Appl. Phys. **76**, 2023 (1994)
47. R.E. Slusher, B.J. Eggleton, *Nonlinear Photonic Crystals* (Springer, Berlin, 2013)
48. M.F. Yanik, S. Fan, M. Soljacic, J.D. Joannopoulos, All-optical transistor action with bistable switching in a photonic crystal cross-waveguide geometry. Opt. Lett. **28**, 2506 (2003)
49. M. Soljacac, S.G. Johnson, S. Fan, M. Ibanescu, E. Ippen, J.D. Joannopoulos, Photonic-crystal slow-light enhancement of nonlinear phase sensitivity. J. Opt. Soc. Am. B **19**, 2052 (2002)
50. M. Soljacic, M. Ibanescu, S.G. Johnson, Y. Fink, J.D. Joannopoulos, Optimal bistable switching in nonlinear photonic crystals. Phys. Rev. E **66**, 055601(R) (2002)
51. V.G. Arkhipkin, S.A. Myslivet, All Optical transistor Using photonic-crystal cavity with an active Raman gain medium. Phys. Rev. A **88**, 033847 (2013)
52. Q. Gong, X. Hu, *Photonic Crystals: Principles and Applications* (CRC Press, Hoboken, 2013)
53. T. Kamiya, B. Monemar, H. Yenghaus, Y. Yamamoto (eds.), *Vertical-Cavity Surface-Emitting Laser Devices* (Springer, Berlin, 2003)
54. E.M. Purcell, Spontaneous emission possibilities at radio-frequencies. Phys. Rev. **69**, 681 (1946)
55. M. Boroditsky, R. Vrijen, T.F. Krausss, R. Coccioli, R. Bhat, E. Yablonovitch, Spontaneous emission extraction and purcell enhancement from thin-film 2-D Photonic crystals. J. Lightwave Technol. **17**, 2096 (1999)
56. S. Noda, M. Fujita, T. Asano, Spontaneous-emission control by photonic crystals and nanocavities. Nat. Photon. **1**, 449 (2007)
57. G.P. Agrawal, *Fiber-Optics Communication Systems: Edition 4* (Wiley, New York, 2012)
58. W. Suh, S. Shanhui, All-pass transmission or flattop reflection filters using a single photonic crystal slab. Appl. Phys. Lett. **84**, 4905 (2004)
59. E. Moreno, F.J. Garcia-Vidal, L. Martin-Moreno, Enhanced transmission and beaming of light via photonic crystal surface modes. Phys. Rev. B **69**, 121402(R) (2004)
60. A.A. Maradudin, A.R. McGurn, The photonic band structure of a truncated, two-dimensional, periodic medium. J. Opt. Soc. Am. B **10**, 307 (1993)
61. A.R. McGurn, Impurity mode techniques applied to the study of light sources. J. Phys. D App. Phys. **38**, 2338 (2005)

62. A. Andreone, A. Cusano, A. Cutolo, V. Galdi (eds.), *Selective Topics in Photonic Crystals and Metamaterials* (World Scientific, Singapore, 2011)
63. C. Luo, A. Narayanaswamy, G. Chen, J.D. Joannopoulos, Thermal radiation form photonic crystals: a direct calculation. Phys. Rev. Lett. **93**, 213905 (2004). I.Z. Ye, J.-M. Park, K. Constant, T.-G. Kim, K.-M. Ho, Photonic crystal: energy-related applications, J. Photon. Energy **2**(1), 021012 (2012). M. Florescu, H. Lee, I. Puscasu, M. Pralle, L. Florescu, D.Z. Ting, D.J. Dowling, Improving solar cell efficiency using photonic band-gap materials. Solar Energy Mater. Solar Cells **91**, 1599 (2007)
64. B.J. O'Regan, Y. Wang, T.F. Krauss, Silicon photonic crystals thermal emitter at near-infrared wavelengths. Sci. Rep. **5**, 13415 (2015). D.L.C. Shen, M. Soljacic, J.D. Joannopoulos, Thermal emission and design in 2D-periodic metallic-photonic crystal slabs. Optics Express **14**, 8785 (2006)
65. E.R. Brown, C.D. Parker, E. Yablonovitch, Radiation properties of a planar antenna on a photonic-crystal substrate. J. Opt. Soc. Am. B **10**, 404 (1993)
66. A.M.R. Pinto, M. Lopez-Amo, Photonic crystal fibers for sensing applications. J. Sens. **2012**, Article ID 598178, 21 (2012). https://doi.org/10.1155/2012/598178
67. F. Poli, A. Cucinotta, S. Selleri, *Photonic Crystal Fibers: Properties and Applications* (Springer, Berlin, 2007)
68. J.C. Knight, Photonic crystal fibers and fiber lasers. J. Opt. Soc. Am. **B24**, 1661 (2007)
69. J.D. Jackson, *Classical Electrodynamics*, 3rd edn. (Wiley, New York, 1999)
70. J.B. Pendry, D. Schurig, D.R. Smith, Controlling electromagnetic fields. Science **312**, 1780 (2006)
71. J.B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966 (2000)
72. J.B. Pendry, A. Holden, W. Stewart, I. Youngs, Extremely low frequency plasmons in metallic mesostructures. Phys. Rev. Lett. **76**, 4773 (1996)
73. U. Leonhardt, T.G. Philibin, General relativity in electrical engineering. New J. Phys. **8**, 247 (2006)
74. U. Leonhardt, T.G. Philibin, Transformation optics and the geometry of light. Prog. Optics **52**, 69 (2009)
75. J. Van Bladel, *Relativity and Engineering* (Springer, Belin, 1984)
76. A. Poddubny, I. Lorsh, P. Belov, Y. Kivshar, Hyperbolic metamaterials. Nat. Photon. **7**, 948 (2013)
77. L.-P. Peng, C.-C. Hsu, Y.C. Shih, Second-harmonic green generation from two-dimensional nonlinear photonic crystal with orthorhombic lattice structure. Appl. Phys. Lett. **83**, 3447 (2003)
78. Y.R. Shen, Surface properties probed by second-harmonic and sum-frequency generation. Nature **337**, 519 (1989)
79. J.L. Lee, M. Tymchenko, C. Argyropoulos, P.-Y. Chen, F. Lu, F. Demmerie, G. Boehm, M.-C. Amann, A. Alu, M.A. Belkin, Giant nonlinear response from plasmonic metasurfaces coupled to intersubband transitions. Nature **511**, 65 (2014)
80. J. Matroell, R. Corbalan, Enhancement of second harmonic generation in a periodic structure with a defect. Opt. Commun. **108**, 319 (1994)
81. M. Scalora, M.J. Bloemer, A.S. Manka, J.P. Dowling, C.M. Bowden, R. Viswanathan, J.W. Haus, Pulsed second-harmonic generation in nonlinear, one-dimensional, periodic structures. Phys. Rev. A **56**, 3166 (1997)
82. R.F. Wallis, G.I. Stegeman (eds.), *Electromagnetic Surface Excitations* (Springer, Berlin, 1986)
83. A.R. McGurn, Enhanced retroreflectance effects in the reflection of light from randomly rough surfaces. Surf. Sci. Rep. **10**, 357 (1990)
84. A.R. McGurn, A.A. Maradudin, An analogue of enhanced backscattering in the transmission of light through a thin film with a randomly rough surface. Optics Commun. **72**, 279 (1989)
85. M. Baibarac, M. Cochet, M. Lapkowski, L. Mihut, S. Lefrant, I. Baltog, SERS spectra of plyaniline thin film s deposited on rough Ag, Au, and Cu Polymer film thickness and roughness parameter dependence of SERS spectra. Synth. Mater. **96**, 63–70 (1998)

86. F.I. Baida, M. Boutria, R. Oussaid, R. Van Labeke, Enhanced transmission metamaterials as anisotropic plates. Phys. Rev. B **84**, 035107 (2011)
87. J. Wang, W. Zhou, E.-P. Li, Enhanceing the light transmission of plasmonic metamaterials through polygonal aperture arrays. Opt. Express **17**, 20349–20354 (2009)
88. A. Dereux, C. Girard, J.-C. Weeber, Theoretical principles of near-field optical microscopies and spectroscopies. J. Chem. Phys. **112**, 7775–7789 (2000)
89. M.I. Stockman, Spasers explained. Nat. Photonics **2**, 327–329 (2008)
90. G. Agrawal, *Nonlinear Fiber Optics*, 5th edn. (Academic Press, Oxford, 2013)
91. R.E. March, An introduction to quadrupole ion trap mass spectrometry. J. Mass Spectro. **32**, 351–369 (1997)
92. M.S. Rocha, Optical tweezers for undergraduates: theoretical analysis and experiments. Am. J. Phys. **77**, 704–712 (2000)
93. W.M.R. Simpson, *Surprises in Theoretical Casimir Physics: Quantum Forces in Inhomogeneous Media* (Springer, Heidelburg, 2015)
94. E.H. Synge, Suggested method for extending microscopic resolution into the ultra-microscopic region. Phil. Mag. **6**, 356–362 (1928)
95. E.H. Synge, An application of piezo-electricity to microscopy. Phil. Mag. **13**, 297–300 (1932)
96. M.A. Paesler, P.J. Meyer, *Near-Field Optics: Theory, Instrumentation, and Applications* (Wiley, Inc., London, 1996)
97. D. Courjon, *Near-Field Microscopy and Near-Field Optics* (Imperial College Press, London, 2003)
98. E. Betzig, J.K. Trautman, Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. Science **257**, 189–195 (1992)
99. L. Maccone, A simple proof of Bell's inequality. Am. J. Phys. **81**, 854–859 (2013)
100. R.B. Griffiths, EPR, Bell, and quantum locality. Am. J. Phys. **79**, 954–967 (2016)
101. G. Blaylock, The EPR paradox, Bell's inequality, and the question of locality. Am. J. Phys. **78**, 111–122 (2010)
102. C.P. William, *Explorations in Quantum Computing* (Springer, New York, 2011)
103. G.P. Berman, G.D. Doolen, R. Mainieri, V.I. Tsifrinovich, *Introduction to Quantum Computers* (World Scientific, Singapore, 1998)
104. A. Steane, Quantum computing. Rep. Prog. Phys. **61**, 117–173 (1998)

Chapter 2
Mathematical Preliminaries

In this chapter, the mathematics needed to understand the basic properties of nanophotonic systems is reviewed. These are basic techniques which have been developed for general applications in condensed matter physics and in studies of electrical engineering problems. They are important in nanophotonics as many of the systems of nanophotonics are similar to systems studied in the general physics of material science and in engineering applications. The primary difference being the length scales relevant to the definition of the problems being posed.

First the mathematics of photonic crystals and metamaterials is treated. This involves the study of periodic structures and the treatment of the response of composite media that appear homogeneous on the scales of interest for their applications. On atomic scales, these problems have been studied in solid state physics and material science, and a number of techniques have been developed in their treatment. It should not be surprising that ideas formulated at the atomic scale can be generalized to systems engineered at other scales.

This is followed by an introduction to some of the basic points of the finite difference time domain method for numerically integrating the Maxwell equations along with the related approaches of the method moments and the finite element method. The method of moments and the finite element methods are effective for the numerical study of electromagnetic frequency modes in complex systems, while the finite difference time domain method is applied to obtain any of the various types of time dependent solutions. These three numerical approaches are commonly used methods, applied to the solution of electrodynamic problems in complicated materials.

The first treatment is of the general properties of composite media and photonic crystals [1–11]. As shall be seen later, metamaterials may be viewed as a type of composite material [1–3]. The approach to composite media which is of interest for metamaterials is that which determines the properties of composites in an effective medium approach [1–3]. This approach determines the response of the medium to fields which change slowly in space compared to the spatial variations of the

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_2

composite medium. The response of the composite material is then viewed as that of a homogeneous medium.

Photonic crystals, on the other hand, are important for their periodically spatially varying properties [4–9]. These allow for a variety of properties which are used to confine and direct the flow of electromagnetic energy through space. The approach to the study of photonic crystals is to treat them as a medium with periodically varying dielectric properties [4–11]. For these systems the interest is in the response to electromagnetic fields with spatial variations that are of the same order as those of the periodicity of the dielectric properties. In this regard, the theory of photonic crystals is closely related to that of the theory of electrons in metals and semi-conductors [4].

2.1 Dielectric Properties of Composites and Photonic Crystals

In their basic forms photonic crystals and metamaterials are based on periodic arrays of dielectric structures [1–11]. Photonic crystals interact with waves with wavelength of order of the basic periodicity of the dielectric properties while metamaterials interact with waves having wavelength larger than the order of the basic periodicity of the features which are periodically arrayed. For these materials it is necessary to review some of the basic mathematical properties of periodic media and some of the responses of composite materials [1–9].

2.1.1 General Theory for Composites

Some general considerations can be given regarding the nature of the effective dielectric constant for composite media [1–3]. The focus in the following will be on composites formed of two different media, though the discussions are subject to a continuation to include composites of an arbitrary number of different dielectric media. Only a system of two different dielectrics with permittivities ε_1 and ε_2 is considered in detail.

A. **Effective Media Approaches**

Effective media approaches are popular in handling composites in which the granules forming the medium are much smaller that the wavelengths of the electromagnetic modes interacting with the system. They are based on treating the granular medium response to an interaction with electrodynamic modes as that of a uniform averaged effective medium.

The simplest type of approximation of this type is the virtual crystal approximation. In the virtual crystal approximation a composite of grains with dielectric constants $\{\varepsilon_i\}$ and volume fractions $\{p_i\}$ is replaced by an effective medium with an effective dielectric constant [1]

$$\varepsilon_{eff} = \sum_i p_i \varepsilon_i. \tag{2.1}$$

The response of the systems is that of a system with the volume averaged dielectric constant of the composite granules. As shown later this can be a very poor approximation under certain circumstances.

A better approximation is the Maxwell Garnett or Effective Media Approximation (EMA) [1, 2]. In this approach, the granules of a three dimensional system with dielectric constants $\{\varepsilon_i\}$ and volume fractions $\{p_i\}$ are approximated by spheres. To develop the theory, a sphere of dielectric constants ε_i is set within an infinite otherwise homogeneous effective medium of effective dielectric constant ε_{eff}. A uniform electric field, $\vec{E}_{app} = E_0 \hat{k}$, is then applied to the effective medium-single sphere system, and the field generated within the sphere is determined as the response of the granule to the average system.

From classical electrodynamics the field inside the sphere is uniform and given by [1, 2]

$$\vec{E}_{in}(i) = \frac{3\varepsilon_{eff}}{2\varepsilon_{eff} + \varepsilon_i} \vec{E}_{app}. \tag{2.2}$$

It then follows from this that

$$\vec{E}_{in}(i) = \left[1 + \frac{\varepsilon_i - \varepsilon_{eff}}{3\varepsilon_{eff}}\right]^{-1} \vec{E}_{app} \tag{2.3}$$

which is rewritten as

$$\vec{E}_{in}(i) = [1 - \Gamma \delta\varepsilon_i]^{-1} \vec{E}_{app} \tag{2.4}$$

where $\Gamma = -\frac{1}{3\varepsilon_{eff}}$ and $\delta\varepsilon_i = \varepsilon_i - \varepsilon_{eff}$. This approximates the field in the ε_i granules as they interacting with the effective medium average of the granular system. An important point to note is that the radii of the sphere do not enter into the relationships between the fields in (2.3) and (2.4) and are, consequently, irrelevant to the effective medium discussions presented here. This is a nice feature of the sphere approximation [1, 2].

The effective medium is then determined by computing the average of the electric fields, $\bar{E}_{in}(i)$, and the average of the displacement vectors, $\vec{D}_{in}(i)$ over the ε_i of the composites system [1, 2]. The averages are computed using $\{p_i\}$ as weight functions, taking advantage of the fact that both the electric and displacement vector fields within the spheres are constant. The effective medium dielectric constant is chosen so that the average electric field is related to the average displacement vector through ε_{eff} in the standard way.

Within the sphere the average of the electric fields and of the displacement vectors are, respectively,

$$\left\langle \vec{E}_{in} \right\rangle = \sum_i p_i \vec{E}_{in}(i), \tag{2.5}$$

$$\left\langle \vec{D}_{in} \right\rangle = \sum_i p_i \varepsilon_i \vec{E}_{in}(i) \tag{2.6}$$

so that from $\left\langle \vec{D}_{in} \right\rangle = \varepsilon_{eff} \left\langle \vec{E}_{in} \right\rangle$ or

$$\sum_i p_i \varepsilon_i \vec{E}_{in}(i) = \varepsilon_{eff} \sum_i p_i \vec{E}_{in}(i). \tag{2.7}$$

Substituting (2.3) and (2.4) in (2.7) then gives [1, 2]

$$\sum_i p_i \frac{\varepsilon_i - \varepsilon_{eff}}{1 - \Gamma \delta \varepsilon_i} = 0, \tag{2.8}$$

which is then solved for ε_{eff} describing the response of the effective medium.

B. Analytic Forms Describing the Response of a Composite Medium

In the following, some general theoretical expressions describing the effective response of a two component dielectric composite are developed [1–3]. These are obtained based on the theory of the electrostatics of composite materials considered in the absence approximations. They offer a less restrictive approach to composites than those of the virtual crystal and Maxwell Garnett theories.

The theory developed has many useful applications in electrostatics. It is also important in studying the quasi-static limit of electrodynamics, in which the electromagnetic wavelength is larger than the typical length scales of the composite granules [1–3].

Consider a composite medium consisting of granules of dielectric constants ε_1 and ε_2 that is the dielectric between the plates of an infinite parallel plate capacitor [2, 3]. The capacitor plates are located at $z = 0$ and $z = L$, and a potential is applied across the capacitor so that $\phi(z = 0) = 0$ and $\phi(z = L) = V_0$ where $\phi(\vec{r})$ is the potential between the plates (see the schematic in Fig. 2.1).

In the static and quasi-static limits the potential in the capacitor is a solution of

$$\nabla \cdot (\varepsilon \nabla \phi) = 0, \tag{2.9}$$

subject to the earlier given capacitor boundary conditions. The composite dielectric in (2.9) is given by

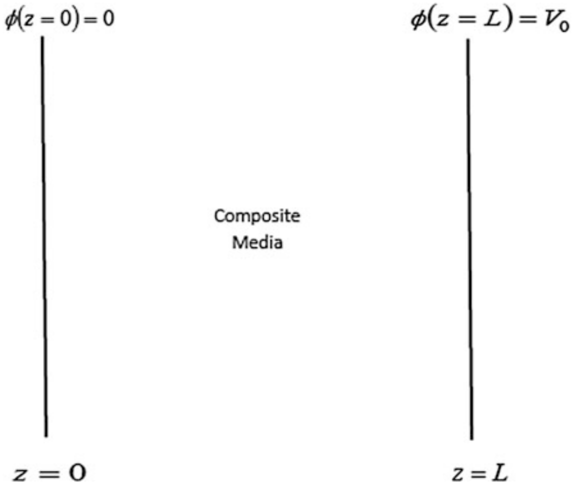


Fig. 2.1 Schematic of the capacitor geometry. The composite material is contained between the capacitor plates which are located at $z = 0$ and $z = L$. A potential is applied across the capacitor so that $\phi(z = 0) = 0$ and $\phi(z = L) = V_0$ where $\phi(\vec{r})$ is the potential between the plates

$$\varepsilon(\vec{r}) = \varepsilon_1\theta(\vec{r}) + \varepsilon_2[1 - \theta(\vec{r})] = \varepsilon_2 + (\varepsilon_1 - \varepsilon_2)\theta(\vec{r}) \tag{2.10}$$

where

$$\begin{aligned} \theta(\vec{r}) &= 1 \quad \text{in regions of } \varepsilon_1 \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{2.11}$$

The object in the following discussions is to develop the response of the composite system to the applied potential in terms of an effective permittivity ε_{eff} and effective electric field E_{eff} [2, 3].

The effective dielectric constant and electric field are defined such that the energies of the composite and the effective medium systems are the same. This requires [2, 3]

$$\varepsilon_{eff} = \frac{1}{V}\int_V dV\varepsilon(\vec{r})\frac{E^2(\vec{r})}{E_{eff}^2} \tag{2.12}$$

where V is the volume between the capacitor plates. Specifically, in obtaining (2.12) the energies of the composite system capacitor and the capacitor in which the composite medium is replaced by an effective medium are equated. From $\vec{E} = -\nabla\phi$ and (2.12) it then follows that

$$\varepsilon_{eff} = -\frac{1}{VE_{eff}^2}\int_V dV\varepsilon\vec{E}\cdot\nabla\phi. \tag{2.13}$$

which introduces a connection to the potential between the plates.

The form in (2.13) is important in obtaining an expression for ε_{eff} in terms of a surface integral over the capacitor plates. This is done by applying the vector identity $\nabla \cdot (\varphi \vec{a}) = \vec{a} \cdot \nabla \varphi + \varphi \nabla \cdot \vec{a}$ and Gauss's law to (2.13) so that

$$\varepsilon_{eff} = -\frac{1}{VE_{eff}^2} \int_V dV \left[\nabla \cdot \left(\phi \varepsilon \vec{E} \right) - \phi \nabla \cdot \left(\varepsilon \vec{E} \right) \right] = -\frac{1}{VE_{eff}^2} \int_V dV \nabla \cdot \left(\phi \varepsilon \vec{E} \right). \quad (2.14)$$

An application of the Divergence theorem to the far right form in (2.14) then gives ε_{eff} as a surface integral over the capacitor plates. In applying the Divergence theorem only the surfaces corresponding to the two capacitor plates contribute to the surface integral. The other surfaces are infinitely smaller than the surfaces of the two plates. In addition, since the potential is only non-zero at $z = L$. Equation (2.13) then becomes an integral only over the right capacitor plate [2, 3]

$$\varepsilon_{eff} = \frac{1}{AE_{eff}^2} \int_{RP} dA \varepsilon \vec{E}_{eff} \cdot \vec{E}. \quad (2.15)$$

In (2.15), $\vec{E}_{eff} = -\frac{V_0}{L} \hat{k}$ where, for an effective field in the positive z-direction, $V_0 < 0$ is the potential on the right capacitor plate. A third and final important relationship for the effective medium is then given by [2, 3]

$$\varepsilon_{eff} = \frac{1}{V} \int_V dV \varepsilon(\vec{r}) \frac{\vec{E}_{eff} \cdot \vec{E}(\vec{r})}{E_{eff}^2}. \quad (2.16)$$

Equation (2.16) can be shown to be equivalent to (2.15) and consequently to (2.12). This is done using the relationships $\phi_{eff} = -E_{eff} z$ and $\vec{E}_{eff} = -\nabla \phi_{eff}$ between the effective permittivity and effective potential of the effective media capacitor to rewrite (2.16)

$$\varepsilon_{eff} = -\frac{1}{VE_{eff}^2} \int_V dV \varepsilon(\vec{r}) \vec{E}(\vec{r}) \cdot \nabla \phi_{eff}. \quad (2.17)$$

Equations (2.13) and (2.17) are similar in form and applying the same steps in going from (2.13) to (2.15) to (2.17) reduces (2.17) to (2.15). Key to this result in that the effective permittivity in (2.17) only involves a surface integral over the right hand plate of the capacitor. Consequently, (2.16) is equivalent to (2.12) and (2.15). These three relationships are of value in the formulation of a general treatment of the electromagnetism of composite materials.

Combining (2.9) and (2.10) the potential between the capacitor plates is a solution of [2, 3]

$$\nabla^2\phi = \left(1 - \frac{\varepsilon_1}{\varepsilon_2}\right)\nabla \cdot [\theta(\vec{r})\nabla\phi]. \tag{2.18}$$

This Poisson equation for ϕ is useful as it allows the boundary value problem for the capacitor containing a composite medium to be reformulated into an easier to handle, equivalent, integral equation.

The conversion is accomplished in the standard way, using the Green function $G(\vec{r}, \vec{r}')$ solution of

$$\nabla^2 G(\vec{r}, \vec{r}') = -\delta^d(\vec{r} - \vec{r}'). \tag{2.19}$$

Here d is the dimension of the space in which the composite medium is defined, and the Green function is subject to $G = 0$ boundary conditions on the closed surface enclosing the infinite parallel plate capacitor.

Before making the reformulation it is important to note that the solution to (2.18) satisfying the capacitor boundary conditions can be written in the form [2, 3]

$$\phi = \phi_h + \phi_{inh}. \tag{2.20}$$

In (2.20) ϕ_h is a solution of $\nabla^2\phi_h = 0$ with boundary conditions $\phi_h(z = 0) = 0$, $\phi_h(z = L) = V_0$, and ϕ_{inh} is a solution of

$$\nabla^2\phi_{inh} = \left(1 - \frac{\varepsilon_1}{\varepsilon_2}\right)\nabla \cdot [\theta(\vec{r})\nabla\phi], \tag{2.21}$$

with boundary conditions $\phi_{inh}(z = 0) = \phi_{inh}(z = L) = 0$. The combined solutions in (2.20) satisfy both (2.18) and the capacitor boundary conditions.

Using the Green function for (2.19) and the considerations of (2.20), (2.18) with its boundary conditions are expressed by the more tractable integral equation [2, 3]

$$\phi(\vec{r}) = -E_{eff}z - \left[1 - \frac{\varepsilon_1}{\varepsilon_2}\right]\int\limits_V dV' G(\vec{r}, \vec{r}')\nabla' \cdot [\theta(\vec{r}')\nabla'\phi(\vec{r}')]. \tag{2.22}$$

Substituting (2.22) into the left side of (2.18) and using (2.19) yields the form on the right side of (2.18). This shows that (2.22) represents an equivalent integral equation reformulation of the differential equation boundary value problem.

Equation (2.22) is an inhomogeneous integral equation which is formally treated using Hilbert- Schmidt theory to give a general idea of the physics of the composite material. The formal solution of (2.22) in the Hilbert-Schmidt is now considered.

To begin, the identity $\nabla \cdot (\varphi\vec{a}) = \vec{a} \cdot \nabla\varphi + \varphi\nabla \cdot \vec{a}$ is applied to the integral in (2.22) so that [2, 3]

$$\begin{aligned}&\int_V dV' G(\vec{r},\vec{r}')\nabla'\cdot[\theta(\vec{r}')\nabla'\phi(\vec{r}')]\\&\quad=\int_V dV'\{\nabla'\cdot[\theta(\vec{r}')G(\vec{r},\vec{r}')\nabla'\phi(\vec{r}')]-\theta(\vec{r}')\nabla'\phi(\vec{r}')\nabla'G(\vec{r},\vec{r}')\}.\end{aligned}\tag{2.23}$$

Using the Divergence theorem, the first integral on the right in (2.23) is converted to a surface integral, but $G = 0$ on the surface enclosing the volume of the capacitor so that the first integral is zero. This leaves

$$\int_V dV' G(\vec{r},\vec{r}')\nabla'\cdot[\theta(\vec{r}')\nabla'\phi(\vec{r}')] = -\int_V dV'\theta(\vec{r}')\nabla'G(\vec{r},\vec{r}')\cdot\nabla'\phi(\vec{r}').\tag{2.24}$$

Defining the integral operator [2, 3]

$$\Gamma\phi = \int_V dV'\theta(\vec{r}')\nabla'G(\vec{r},\vec{r}')\cdot\nabla'\phi(\vec{r}'),\tag{2.25}$$

Equation (2.24) and (2.25) applied to (2.22) result in the operator form

$$\begin{aligned}\phi &= -E_{eff}z + \left[1-\frac{\varepsilon_1}{\varepsilon_2}\right]\Gamma\phi\\&= \phi_h + \phi_{inh}\end{aligned}\tag{2.26}$$

to be solved for ϕ.

The operator equation in (2.26) is studied using the Hilbert-Schmidt theory. In this approach the solution to the inhomogeneous integral equation is expressed in terms of the eigenvalues and eigenvectors of the kernel of the integral equation. To apply the method in the following discussions, the eigenvalue problem for the kernel is first treated, followed by a discussion of the orthogonally properties of the eigenvectors, and finally the solution of (2.26) is developed as an expansion in the eigenfunctions.

Specifically, consider the eigenvalue problem [2, 3]

$$\Gamma\phi_n = s_n\phi_n\tag{2.27}$$

where the eigenfunctions $\{\phi_n\}$ are equal to zero on the surface of the capacitor. The problem in (2.27) is shown to give an orthonormal set, $\{\phi_n\}$.

To this end, consider the tautological form obtained for the real operator Γ from (2.27)

$$\int_V dV\theta\nabla\phi_m\cdot\nabla\left(\Gamma\phi_n^*\right)-\int_V dV\theta\nabla\phi_n^*\cdot\nabla(\Gamma\phi_m)=\left(s_n^*-s_m\right)\int_V dV\theta\nabla\phi_m\cdot\nabla\phi_n^*. \tag{2.28}$$

This is just the integral equation form of that used to show the orthogonality properties of differential equation eigenvalue problems. The left side of (2.28) will be shown to be equal to zero, and from this the orthonormality properties of the eigenfunctions will follow.

To see this, from (2.25) and (2.28) it follows that

$$\begin{aligned}\int_V dV\theta\nabla\phi_m\cdot\nabla\left(\Gamma\phi_n^*\right)&=\int_V dVdV'\theta(\vec{r})\nabla\phi_m\theta(\vec{r}')\cdot\nabla\nabla' G(\vec{r},\vec{r}')\cdot\nabla'\phi_n^*(\vec{r}')\\&=\int_V dVdV'\theta(\vec{r})\nabla\phi_n^*(\vec{r})\cdot\nabla\nabla' G(\bar{r},\vec{r}')\cdot\nabla'\phi_m(\vec{r}')\\&=\int_V dV\theta\nabla\phi_n^*\cdot\nabla(\Gamma\phi_m)\end{aligned} \tag{2.29}$$

so that the integral operators on the far right and far left of (2.29) are self-adjoint. Using (2.29) in (2.28) it follows that

$$\left(s_n^*-s_m\right)\int_V dV\theta\nabla\phi_n^*\cdot\nabla\phi_m=0. \tag{2.30}$$

so that $\{s_n\}$ are real and for $s_n\neq s_m$ the eigenfunctions ϕ_n and ϕ_m are orthogonal. For properly normalized functions the orthogonality condition is

$$\int_V dV\theta\nabla\phi_n^*\cdot\nabla\phi_m=\delta_{n,m}. \tag{2.31}$$

For a cases in which a complete set of functions $\{\phi_n\}$ exist, (2.31) can be used in (2.22) to express the kernel, the inhomogeneity, and the solution ϕ as an expansion in $\{\phi_n\}$. This allows for (2.22) to be formally solved in terms of the eigenstates of (2.27).

Specifically, for the complete set $\{\phi_n\}$

$$\phi=\sum_n a_n\phi_n, \tag{2.32}$$

where from (2.32) and (2.31) $a_n = \int_V dV \theta \nabla \phi_n^* \cdot \nabla \phi$ so that [2, 3]

$$\phi(\vec{r}) = \sum_n \left[\int_{V'} dV' \theta(\vec{r}') \nabla' \phi_n^*(\vec{r}') \cdot \nabla' \phi(\vec{r}') \right] \phi_n(\vec{r}). \tag{2.33}$$

A formal solution for ϕ of the operator equation in (2.26) begins by rewriting (2.26) in the form [2, 3]

$$\phi = -E_{eff} z + u\Gamma\phi \tag{2.34}$$

for $u \equiv 1 - \frac{\varepsilon_1}{\varepsilon_2}$. From this

$$[1 - u\Gamma]\phi = -E_{eff} z, \tag{2.35}$$

with a solution given by

$$\phi = -\left[1 + u\Gamma + (u\Gamma)^2 + \cdots + (u\Gamma)^n + \cdots . \right] E_{eff} z. \tag{2.36a}$$

(Note: one can add to (2.36a) a solution of the homogeneous equation $[1 - u\Gamma]\phi = 0$, but these are analytic and not important to the present discussions.)

If the infinite series in (2.36a) exists, it is seen by direct substitution to be a solution of (2.35). In problems with a complete set $\{\phi_n\}$, a are more tractable form of (2.36a) is obtained using the completeness properties of $\{\phi_n\}$ to write $z = \sum_n b_n \phi_n$ which, upon substitution into (2.36a) and applying (2.27), gives

$$\phi = -E_{eff} \sum_n \frac{1}{1 - us_n} b_n \phi_n. \tag{2.36b}$$

An expression for the effective dielectric constant in terms of the solutions of the eigenvalue problem is obtained from (2.36b). This is done by considering the expression for the effective dielectric constant in (2.16) rewritten in the form [2, 3]

$$\begin{aligned} \varepsilon_{eff} &= \frac{1}{VE_{eff}^2} \int_V dV \varepsilon \nabla \phi_{eff} \cdot \nabla \phi \\ &= -\frac{\varepsilon_2}{VE_{eff}} \int_V dV [1 - u\theta(\vec{r})] \nabla z \cdot \nabla \phi \\ &= \frac{\varepsilon_2}{VE_{eff}} \int_V dV [1 - u\theta(\vec{r})] E_z. \end{aligned} \tag{2.37}$$

Here (2.10) and $\phi_{eff} = -E_{eff} z$ have been used on the second equality on the right. It then follows from (2.37) that for $s \equiv \frac{1}{u}$

$$F(s) = 1 - \frac{\varepsilon_{eff}}{\varepsilon_2} = \frac{1}{E_{eff}}\frac{1}{sV}\int\limits_V dV\theta(\vec{r})E_z(\vec{r})$$
$$= -\frac{1}{E_{eff}}\frac{1}{sV}\int\limits_V dV\theta(\vec{r})\nabla z \cdot \nabla\phi, \tag{2.38}$$

gives a relationship between the solution for the field and the effective dielectric constant. From (2.36b) and (2.38) it follows that

$$F(s) = \sum_n \frac{F_n}{s - s_n}, \tag{2.39}$$

where $F(s) = \frac{1}{V}b_n^2$.

Equation (2.39) shows an interesting structure for the effective dielectric constant. The effective dielectric constant as a function of $s = \frac{\varepsilon_2}{\varepsilon_2 - \varepsilon_1}$ consists of a series of terms that exhibit singularities at the eigenvalues, $\{s_n\}$, of (2.27).

In general it is found, however, that for the eigenvalue problem in (2.27) to have solutions the condition $\frac{\varepsilon_1}{\varepsilon_2} < 0$ must hold. This range of $\frac{\varepsilon_1}{\varepsilon_2}$ is commonly associated with the quasi-static limit, discussed later, rather than electrostatic problems.

For cases in which $\frac{\varepsilon_1}{\varepsilon_2} > 0$, (2.36a) cannot be summed using eigen-solutions but a term by term treatment of the series should be handled. When they exist, the poles of $F(s)$ show the strong dependence of the effective medium dielectric constant on s, revealing that simple approximations such as the virtual crystal approximation may often be far off the mark in approximating the average response of the composite.

Later discussions of the simplified nature of the poles in (2.39) for various effective medium theories and simple composite geometries will be addressed [2, 3].

Quasi-Static Limit

The above theory has focused on the static system, but in some cases it can be extended to frequency dependent systems. This is done in the so-called quasi-static limit of electrodynamics. For an electric field between the capacitor plates of the form $\vec{E}(\vec{r}, t) = \vec{E}(\vec{r})e^{-i\omega t}$ the Maxwell equations become [2, 3]

$$\nabla \cdot \left(\varepsilon' \vec{E}\right) = 4\pi\rho \tag{2.40a}$$

$$\nabla \cdot \vec{B} = 0 \tag{2.40b}$$

$$\nabla \times \vec{E} = \frac{i\omega}{c}\vec{B} \tag{2.40c}$$

$$\nabla \times \vec{B} = \frac{4\pi}{c}\sigma'\vec{E} - \frac{i\omega}{c}\varepsilon'\vec{E}. \tag{2.40d}$$

In these equations σ' is the conductivity of the free charge which contributes to the DC conductivity and ε' is the dielectric constant of the bound charge which contributes to the DC dielectric constant of the system.

It is important to note that this separation of the charges into the σ' and ε' responses of the system is known to be to some extent arbitrary and can be remade in a number of different ways. As long as one develops a theory that is consistent in how the separation is made, the electrodynamics in a consistent approach are the same.

In addition to the Maxwell equations the system dynamics must satisfy the continuity equation

$$\nabla \cdot (\varepsilon' E') - i\omega\rho = 0. \tag{2.41}$$

This is a statement of the conservation of charge for the system. In the quasi-static limit of (2.40) and (2.41) the equations for the electric field are rewritten to look like those in electrostatics. As now shown, this involves a renormalization of the field equations in the long wavelength limit.

From (2.40a) and (2.41) it follows that [2, 3]

$$\nabla \cdot \left[\left(\varepsilon' + \frac{4\pi i\sigma'}{\omega}\right)\vec{E}\right] = 0, \tag{2.42}$$

and defining the permittivity $\varepsilon = \varepsilon' + \frac{4\pi i}{\omega}\sigma'$ and displacement vector $\vec{D} = \varepsilon\vec{E}$, (2.42) becomes

$$\nabla \cdot D = 0. \tag{2.43}$$

In the case that the wavelength of the electromagnetic wave generated in the capacitor medium is much greater than the linear dimensions of the grains forming the composite medium, $i\frac{\omega}{c}\vec{B} \approx 0$ so that (2.40c) becomes [2, 3]

$$\nabla \times \vec{E} = 0. \tag{2.44}$$

This last condition is also an important consideration in developing an effective medium description of the average response of the granular system.

Equations (2.43) and (2.44) which now describe the electric field in the system are the standard equations of electrostatics, but now the dielectric constant is complex. Once the electric field is obtained as a solution of (2.43) and (2.44), the magnetic field corresponding to these electric fields is obtained as a solution of (2.40b) and (2.40d). The resulting electric and magnetic fields give the complete field solutions in the quasi-static limit.

Applications

An interesting use of the effective media approaches come in understanding the properties of a material with various artistic applications. It is well known that when small gold particles are dissolved in glass, the dilute mixture develops a beautiful red color. The color arises from plasmon resonances of the gold particles and their interaction with light sent through the glass.

To understand this effect, consider the effective dielectric constant of a weakly dilute suspension of gold particles in a background medium of unit dielectric constant (approximating that of glasses). For the metal particles take the dielectric constant to be of the form [2, 3]

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2}, \tag{2.45}$$

where ω_p is the plasma frequency. In (2.45) the response of the conduction electrons of the metal to the frequency dependent electric fields are completely described by a dielectric constant so that the conductivity effects are included in (2.45).

An expression for the effective dielectric constant of the dilute mixture of gold particles in glass is obtained by applying (2.8) to the two media system. Specifically, consider

$$\varepsilon_1(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \tag{2.46}$$

with a volume fraction $p \ll 1$ and

$$\varepsilon_2(\omega) = 1 \tag{2.47}$$

with volume fraction $1 - p$.

From (2.8) with $i = 1, 2$ and $\Gamma = -\frac{1}{3\varepsilon_{eff}}$ (computed for spheres) it then follows [2, 3]

$$p\frac{\varepsilon_1 - \varepsilon_{eff}}{1 + \frac{1}{3\varepsilon_{eff}}\left(\varepsilon_1 - \varepsilon_{eff}\right)} + (1-p)\frac{\varepsilon_2 - \varepsilon_{eff}}{1 + \frac{1}{3\varepsilon_{eff}}\left(\varepsilon_2 - \varepsilon_{eff}\right)} = 0. \tag{2.48}$$

Notice that (2.48) does not depend on the radii of the sphere but only on their dielectric constants.

For $p \ll 1$ with $\varepsilon_{eff} = \varepsilon_2 + \delta\varepsilon_{eff}$ it follows from (2.48) that to first order in p and $\delta\varepsilon_{eff} = \Theta(p)$

$$\varepsilon_{eff} = \varepsilon_2 + 3p\varepsilon_2\frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1 + 2\varepsilon_2} \tag{2.49}$$

It is interesting to note that, as an illustration of the general theory in (2.39), with a little rewrite (2.49) can be put into the form generated in (2.39). Following some algebra it is found that [2, 3]

$$F(s) = 1 - \frac{\varepsilon_{eff}}{\varepsilon_2} = \frac{p}{s - \frac{1}{3}} \tag{2.50}$$

where $s = \frac{\varepsilon_2}{s_2 - s_1}$.

A further useful expression for the frequency dependence of (2.49) is obtained using (2.46) and (2.47) in (2.49). This gives [2, 3]

$$\varepsilon_{eff} = 1 - 3p\frac{\omega_p^2}{\omega^2}\frac{1}{3 - \frac{\omega_p^2}{\omega^2}}, \tag{2.51}$$

which now offers an explanation of the red glow of the diluted gold particulate in glass.

To understand the red glass consider the Maxwell field equations [2, 3]

$$\nabla \cdot \left(\varepsilon_{eff}\vec{E}\right) = 0. \tag{2.52a}$$

$$\nabla \cdot \vec{B} = 0 \tag{2.52b}$$

$$\nabla \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t} \tag{2.52c}$$

$$\nabla \times \vec{B} = \frac{\varepsilon_{eff}}{c}\frac{\partial \vec{E}}{\partial t} \tag{2.52d}$$

Applying the standard treatment from electrodynamics, it follows from (2.52c) and (2.52d) that

$$\nabla\left(\nabla \cdot \vec{E}\right) - \nabla^2\vec{E} = -\frac{\varepsilon_{eff}}{c^2}\frac{\partial^2\vec{E}}{\partial t^2} \tag{2.53}$$

is the wave equation in the effective medium.

By substitution of a plane wave form, it is found that a longitudinal wave solution of (2.52) and (2.53) occurs in the case that $\varepsilon_{eff} = 0$. The condition, then, for a longitudinal wave in the system is that [2, 3]

$$\varepsilon_{eff} = 1 - 3p\frac{\omega_p^2}{\omega^2}\frac{1}{3 - \frac{\omega_p^2}{\omega^2}} = 0, \tag{2.54}$$

Solving in the $p \ll 1$ limit gives

$$\omega = \frac{1}{\sqrt{3}}\left(1 + \frac{3}{2}p\right)\omega_p. \tag{2.55}$$

Light in the effective medium at this frequency will resonantly interact with the media so as to create the red glow in the dilute gold-glass material.

As an interesting point in regard to this result: The effective medium treatment of the interaction of light with the plasma resonances, resulting in (2.55), is similar to that of the Rayleigh scattering of light from molecules in the atmosphere. The sky blue coloration of the atmosphere and the interaction of light with the atmosphere is treated in an effective medium theory based on the molecular polarizability of atmospheric gases. In the result for the atmosphere, the blue coloration is from the ω^4 dependence of the molecular elastic scattering cross section. For the glass problem the scattering of light is from the plasmons of the gold particles.

Some other exact result of the effective medium treatment that are worth mentioning are applications of (2.39) to layered media between the plates of the capacitor. Consider a layered media of slabs formed from a first medium of dielectric constant ε_1 and volume fraction p_1 and a second medium of dielectric constant $\varepsilon_2\varepsilon_2$ and volume fraction $p_2 = 1 - p_1$.

In the case that the slab surfaces are parallel to the planes of the capacitor, it is shown that [2, 3]

$$\frac{1}{\varepsilon_{eff}} = \frac{p_1}{\varepsilon_1} + \frac{p_2}{\varepsilon_2}. \tag{2.56}$$

From (2.39) and (2.56) it follows that

$$F(s) = 1 - \frac{\varepsilon_{eff}}{\varepsilon_2} = \frac{p_1}{s - p_2} \tag{2.57}$$

where $s = \frac{\varepsilon_2}{\varepsilon_2 - \varepsilon_1}$.

In a second case, if the slab surfaces are perpendicular to the planes of the capacitor, it is shown that [2, 3]

$$\varepsilon_{eff} = p_1\varepsilon_1 + p_2\varepsilon_2 \tag{2.58}$$

where again $p_2 = 1 - p_1$. From (2.39) and (2.58) it follows that [2, 3]

$$F(s) = 1 - \frac{\varepsilon_{eff}}{\varepsilon_2} = \frac{p_1}{s} \tag{2.59}$$

where $s = \frac{\varepsilon_2}{\varepsilon_2 - \varepsilon_1}$. These are simple cases which illustrate the general form for the theory of composites given in (2.39).

2.2 General Theory for Periodic Media

The theory of periodic media begins with the complication introduced by the periodic variation in space of the media properties and the constraints relating to these properties put on spatial functions of the medium [4–11]. To describe the periodicity it is natural to introduce a lattice with the periodicity of the medium and to define functions describing the properties of the media relative to this lattice. This is done in three dimensions by choosing a set of three smallest linearly independent vectors of the lattice $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ which translate the lattice into itself as well as the periodic medium into itself (see the schematic drawing in Fig. 2.2).

From these vectors any lattice translation taking the lattice into itself is of the form [4–11]

$$\vec{T}_i = n_{i,1}\vec{a}_1 + n_{i,2}\vec{a}_2 + n_{i,3}\vec{a}_3 \tag{2.60}$$

for integers, $(n_{i,1}, n_{i,2}, n_{i,3})$, and the set of lattice translation vectors $\{\vec{T}_i\}$ describe the translational symmetry group of the lattice. The chosen vectors $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ are known as basis vectors, and in general the position vector of an arbitrary point in space is written in terms of them as

$$\vec{r} = x_1\vec{a}_1 + x_2\vec{a}_2 + x_3\vec{a}_3 \tag{2.61}$$

for (x_1, x_2, x_3) real.

In particular, for a function $f(\vec{r})$ with the periodicity of the medium it follows that

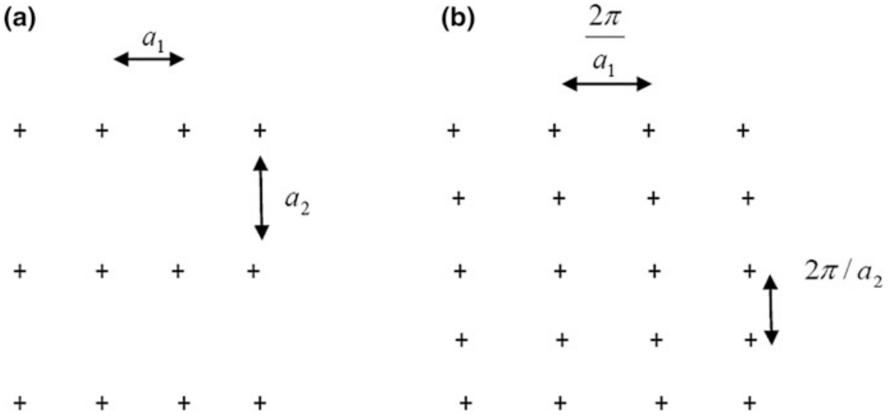


Fig. 2.2 Schematic of: **a** the *x*-*y* plane of a cubic lattice indicating a set of smallest translation vectors $\vec{a}_1$, $\vec{a}_2$, and **b** the corresponding k-space lattice showing the *x*-*y* plane with smallest k-space translation vectors $\vec{b}_1 = \frac{2\pi}{a_1}\hat{\imath}$, $\vec{b}_2 = \frac{2\pi}{a_2}\hat{\jmath}$

$$f\left(\vec{r}+\vec{T}_i\right)=f(\vec{r}). \tag{2.62}$$

The restriction in (2.62) applies to all spatially dependent physical properties of the medium and places a set of constraints on the form of the Fourier series of $f(\vec{r})$. Specifically, considering the standard form of Fourier series

$$f(\vec{r})=\sum_i f\left(\vec{k}_i\right)\mathrm{e}^{i\vec{k}_i\cdot\vec{r}}, \tag{2.63}$$

the translational symmetry requirement in (2.63) restricts $\{\vec{k}_i\}$ to satisfy

$$\vec{k}_i\cdot\vec{T}_l=2\pi n \tag{2.64}$$

for some integer, n. In the absence of (2.64) the function given by the Fourier series is no longer periodic in the lattice.

For a three dimensional system (2.64) has solutions [4–11]

$$\vec{k}_i=m_{i,1}\vec{b}_1+m_{i,2}\vec{b}_2+m_{i,3}\vec{b}_3 \tag{2.65}$$

where

$$\vec{b}_1=2\pi\frac{\vec{a}_2\times\vec{a}_3}{\vec{a}_1\cdot\vec{a}_2\times\vec{a}_3} \tag{2.66a}$$

$$\vec{b}_2=2\pi\frac{\vec{a}_3\times\vec{a}_1}{\vec{a}_1\cdot\vec{a}_2\times\vec{a}_3} \tag{2.66b}$$

$$\vec{b}_3=2\pi\frac{\vec{a}_1\times\vec{a}_2}{\vec{a}_1\cdot\vec{a}_2\times\vec{a}_3} \tag{2.66c}$$

for integers $\left(m_{i,1},m_{i,2},m_{i,3}\right)$ with $n=m_{i,1}n_{l,1}+m_{i,2}n_{l,2}+m_{i,3}n_{l,3}$ an integer (see Fig. 2.2 for a schematic of these vectors for the cubic lattice).

In the case of a two dimensional lattice system with basis vectors $\{\vec{a}_1,\vec{a}_2\}$, (2.63) and (2.64) give solutions [4–11]

$$\vec{k}_i=m_{i,1}\vec{b}_1+m_{i,2}\vec{b}_2 \tag{2.67}$$

$$\vec{b}_1=2\pi\frac{\vec{a}_2\times\hat{n}_\perp}{\vec{a}_1\cdot\vec{a}_2\times\hat{n}_\perp} \tag{2.68a}$$

$$\vec{b}_2=2\pi\frac{\hat{n}_\perp\times\vec{a}_1}{\vec{a}_1\cdot\vec{a}_2\times\hat{n}_\perp} \tag{2.68b}$$

for $n = m_{i,1}n_{l,1} + m_{i,2}n_{l,2}$ an integer and where the unit vector $\hat{n}_{\perp}$ is perpendicular to both $\vec{b}_1$ and $\vec{b}_2$. With the representations of the $\{\vec{k}_i\}$ in (2.65) and (2.66) or (2.67) and (2.68), (2.63) yields a three or two-dimensional periodic function which has the periodicity of the, respective, media.

In the case of the electrodynamics of periodic media the field equations are given by [4–11]

$$\nabla\left(\nabla \cdot \vec{E}\right) - \nabla^2 \vec{E} = -\frac{\varepsilon(\vec{r})}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} \tag{2.69a}$$

and

$$\nabla \times \left(\frac{1}{\varepsilon(\vec{r})}\nabla \times \vec{B}\right) = -\frac{1}{c^2}\frac{\partial^2 \vec{B}}{\partial t^2}. \tag{2.69b}$$

Equation (2.69a) is interesting as the differential form $\nabla(\nabla\cdot) - \nabla^2$ exhibits complete translational symmetry in space so that the restriction of the operators to the translational symmetry of the periodic medium enters through the dielectric constant, $\varepsilon(\vec{r})$. Similarly, (2.69b) for the magnetic induction is periodic in the medium. The magnetic induction can also be obtained from the electric fields by an application of Faraday’s law so that its symmetry properties are intimately related to those of the electric fields.

For electromagnetic fields of the form $\vec{E} = \vec{E}_a \mathrm{e}^{-i\omega t}$, $\vec{B} = \vec{B}_a \mathrm{e}^{-i\omega t}$ (2.69) become

$$\left\{\nabla(\nabla\cdot) - \nabla^2 - \frac{\varepsilon(\vec{r})\omega^2}{c^2}\right\}\vec{E} = 0 \tag{2.70a}$$

and

$$\left\{\nabla \times \left(\frac{1}{\varepsilon(\vec{r})}\nabla\times\right) - \frac{\omega^2}{c^2}\right\}\vec{B} = 0 \tag{2.70b}$$

and the operators in the $\{\}$ are seen to exhibit the translational symmetry of the periodic lattice and medium. It is important to note, however, that even though the operators in (2.70) are invariant under the translational symmetry, this does not mean that the solutions $\vec{E}_a$, $\vec{B}_a$ are necessarily invariant under the translational symmetry.

Due to the translational invariance of the operator in the $\{\}$, it follows that the translated fields $\vec{E}_a\left(\vec{r} + \vec{T}_l\right)$, $\vec{B}_a\left(\vec{r} + \vec{T}_l\right)$ for any given translation vector $\vec{T}_l$ must also be solutions of the operators in (2.70). These solutions generally are not the same as the original untranslated solutions $\vec{E}_a(\vec{r})$, $\vec{B}_a(\vec{r})$, but they are new solutions with the same frequency ω. From the conditions of translational symmetry, along with the boundary conditions for (2.70), an important statement about the general form of $\vec{E}_a(\vec{r})$, $\vec{B}_a(\vec{r})$ can be inferred. These conditions are now discussed.

To define the $\vec{E}_a$, $\vec{B}_a$ solutions of (2.70) in infinite space, periodic boundary conditions over a parallelepiped of edge $L \to \infty$ are applied. This is done to obtain a set of solutions of the operators in (2.70) which are useful in discussing transport properties of electromagnetic waves within the periodic dielectric system. If all of space is partitioned into parallelepipeds of edge $L \to \infty$ the periodic boundary conditions require [4–11]

$$\vec{E}_a\left(\vec{r}+\vec{T}_L\right) = \vec{E}_a(\vec{r}), \tag{2.71a}$$

$$\vec{B}_a\left(\vec{r}+\vec{T}_L\right) = \vec{B}_a(\vec{r}) \tag{2.71b}$$

for $\vec{T}_L$ a translation taking one of the covering parallelepipeds of space into any of the others in the spatial partition.

A general form for solutions satisfying these conditions is given by

$$\vec{E}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{U}_{\vec{k},n,a}(\vec{r}), \tag{2.72a}$$

$$\vec{B}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{V}_{\vec{k},n,a}(\vec{r}) \tag{2.72b}$$

where $\vec{U}_{\vec{k},n,a}(\vec{r})$ and $\vec{V}_{\vec{k},n,a}(\vec{r})$ are periodic functions of the lattice, and in which the $\{\vec{k}\}$ satisfy

$$\vec{k}\cdot\vec{T}_L = 2\pi n, \tag{2.72c}$$

for some integer, n, where $\vec{T}_L$ is a translation vector of the parallelepiped spatial partition. It should be noted that the periodic functions $\vec{U}_{\vec{k},n,a}(\vec{r})$ and $\vec{V}_{\vec{k},n,a}(\vec{r})$ may be different for different $\vec{k}$, and the subscript n is a band index as there may be multiple solutions for a given $\vec{k}$ corresponding to different eigenvalue solutions, ω^2.

A simple example of solutions for periodic boundary conditions is the case in which $\varepsilon(\vec{r}) = \varepsilon_0$ is a constant independent of position in space, i.e., the case of complete translational symmetry. Equations (2.70) and (2.71) become

$$\left\{\nabla(\nabla\cdot) - \nabla^2 - \frac{\varepsilon_0\omega^2}{c^2}\right\}\vec{E} = 0 \tag{2.73a}$$

and

$$\left\{\nabla(\nabla\cdot) - \nabla^2 - \frac{\varepsilon_0\omega^2}{c^2}\right\}\vec{B} = 0. \tag{2.73b}$$

where now the operators have complete translational symmetry. Under these symmetry conditions the general from of (2.72) for the solutions becomes

$$\vec{E}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{E}_0, \tag{2.74a}$$

$$\vec{B}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{B}_0 \tag{2.74b}$$

where $\vec{U}_{\vec{k},n,a}(\vec{r}) = \vec{E}_0$ and $\vec{V}_{\vec{k},n,a}(\vec{r}) = \vec{B}_0$ are constant due the invariance of the operators to translations of any length scale. Consequently, translating (2.74) through $\vec{T}$ gives

$$\vec{E}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{T}}\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{E}_0, \tag{2.75a}$$

$$\vec{B}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{T}}\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{B}_0. \tag{2.75b}$$

Upon substituting either (2.74) or (2.75) into (2.73) gives

$$\left\{-\vec{k}\left(\vec{k}\cdot\right) + k^2 - \frac{\varepsilon_0\omega^2}{c^2}\right\}\vec{E}_a = 0 \tag{2.76a}$$

and

$$\left\{-\vec{k}\left(\vec{k}\cdot\right) + k^2 - \frac{\varepsilon_0\omega^2}{c^2}\right\}\vec{B}_a = 0. \tag{2.76b}$$

These equations generate the dispersion relations of the electromagnetic waves in the uniform medium. Note that the first terms on the left side of the equation, from the divergence Maxwell equations, are required to be zero in a uniform homogeneous medium.

This example gives an illustration of the general forms expected in the theory due to translational symmetry. It will now be shown that spatially periodic systems also exhibit periodicity properties of their dispersion relations in $\vec{k}$-space.

As a consequence of the translational symmetry of the lattice and the periodic boundary conditions the electromagnetic fields are of the general forms in (2.72). From (2.72) it is now shown that the dispersion relations $\omega(\vec{k})$ has periodicity properties in $\vec{k}$-space. Specifically, consider substituting (2.72) into (2.70). This gives

$$\begin{aligned}&\left\{\nabla(\nabla\cdot)-\nabla^2\right\}\vec{U}_{\vec{k},n,a}\\&\quad+i\left\{\vec{k}(\nabla\cdot)+\nabla\left(\vec{k}\cdot\right)-2\vec{k}\cdot\nabla\right\}\vec{U}_{\vec{k},n,a}\\&\quad+\left\{-\vec{k}\left(\vec{k}\cdot\right)+k^2\right\}\vec{U}_{\vec{k},n,a}-\frac{\varepsilon(\vec{r})\omega^2\left(\vec{k}\right)}{c^2}\vec{U}_{\vec{k},n,a}=0\end{aligned}\tag{2.77a}$$

where the forms for the magnetic induction can be obtained from Faraday's Law

$$\vec{B}=\frac{-ic}{\omega}\nabla\times\vec{E}.\tag{2.77b}$$

The solution of these equations yields the periodic envelops $\vec{U}_{\vec{k},n,a}(\vec{r})$ and $\vec{V}_{\vec{k},n,a}(\vec{r})$ of the forms in (2.72) for a specified $\vec{k}$.

Now consider obtaining solutions for the case [4–11]

$$\vec{E}_a(\vec{r})=\mathrm{e}^{i\left(\vec{k}+\vec{k}_i\right)\cdot\vec{r}}\vec{U}_{\vec{k}+\vec{k}_i,n,a}(\vec{r}),\tag{2.78a}$$

$$\vec{B}_a(\vec{r})=\mathrm{e}^{ik\left(\vec{k}+\vec{k}_i\right)\cdot\vec{r}}\vec{V}_{\vec{k}+k_i,n,a}(\vec{r})\tag{2.78b}$$

where $\vec{k}_i=m_{i,1}\vec{b}_1+m_{i,2}\vec{b}_2+m_{i,3}\vec{b}_3$ are defined in (2.65) and (2.66) for a three-dimensional system or for the case of a two-dimensional system are specialized in (2.67) and (2.68). If (2.72) is rewritten as

$$\vec{E}_a(\vec{r})=\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{U}_{\vec{k}+\vec{k}_i,n,a}(\vec{r})\mathrm{e}^{i\vec{k}_i\cdot\vec{r}}=\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{U}'_{\vec{k},n,a}(\vec{r}),\tag{2.79a}$$

$$\vec{B}_a(\vec{r})=\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{V}_{\vec{k}+\vec{k}_i,n,a}(\vec{r})\mathrm{e}^{i\vec{k}_i\cdot\vec{r}}=\mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{V}'_{\vec{k},n,a}(\vec{r}),\tag{2.79b}$$

it is seen that $\vec{U}'_{\vec{k},n,a}(\vec{r})$ and $\vec{V}'_{\vec{k},n,a}(\vec{r})$ are again periodic functions of the lattice and as such must also satisfy (2.77). It then follows that the solution set of eigenvalues $\left\{\omega^2(\vec{k})\right\}$ and $\left\{\omega^2(\vec{k}+\vec{k}_i)\right\}$ are the same. The eigenvalues are periodic in $\vec{k}$ -space with the periodicity of the lattice defined by the $\vec{k}$ -space translation vectors $\{\vec{k}_i\}$.

As an example of the translational symmetry of the dispersion problem, consider the one-dimensional problem of a wave moving in a direction perpendicular to the layers of a one-dimensional periodic dielectric medium. For a wave traveling along the z-direction, the form of the wave equation for the electric field polarized perpendicular to the direction of propagation is given by [4–9, 11]

$$\frac{\partial^2 \vec{E}}{\partial z^2} - \varepsilon(z)\frac{1}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} = 0. \tag{2.80}$$

Here $\varepsilon(z)$ is the position dependent dielectric constant of the periodic layered medium.

Substituting

$$\vec{E}(z,t) = \vec{E}_a(z)\mathrm{e}^{-i\omega t} = \mathrm{e}^{ikz}\vec{U}_{k,n,a}(z)\mathrm{e}^{-i\omega t} \tag{2.81}$$

yields

$$\left\{-k^2 + 2ik\frac{d}{dz} + \frac{d^2}{dz^2} + \varepsilon(z)\frac{\omega^2}{c^2}\right\}\vec{U}_{k,n,a}(z) = 0. \tag{2.82}$$

For the periodic permittivity let a be the smallest distance such that

$$\varepsilon(z+a) = \varepsilon(z) \tag{2.83}$$

then from (2.63) the Fourier series of the $\varepsilon(z)$ is

$$\varepsilon(z) = \sum_m \varepsilon_m \mathrm{e}^{ik_m z} \tag{2.84}$$

where $k_m = \frac{2\pi}{a}m$ for integers, m. Likewise, $\vec{U}_{k,n,a}(\vec{r})$ is periodic in z with the Fourier series

$$\vec{U}_{k,n,a}(z) = \sum_m \vec{U}(k,n,a)_m \mathrm{e}^{ik_m z} \tag{2.85}$$

Substituting (2.84) and (2.85) in (2.82) reduces the eigenvalue problem to an algebraic form

$$(k+k_m)^2\vec{U}(k,n,a)_m - \frac{\omega^2}{c^2}\sum_p \varepsilon_p \vec{U}(k,n,a)_{p-m} = 0. \tag{2.86}$$

Applying the same considerations to [4–11]

$$\vec{E}(z,t) = \mathrm{e}^{i(k+k_i)z}\vec{U}_{k+k_i,n,a}(z)\mathrm{e}^{-i\omega t} = \mathrm{e}^{ikz}\vec{U}_{k+k_i,n,a}(z)\mathrm{e}^{i(k_i z-\omega t)} \tag{2.87}$$

where

$$\vec{U}_{k+k_i,n,a}(z)\mathrm{e}^{ik_i z} = \sum_m \vec{U}'(k+k_i,n,a)_m \mathrm{e}^{ik_m z} \tag{2.88}$$

reduces to the eigenvalue problem

$$(k+k_m)^2\vec{U}'(k+k_i,n,a)_m - \frac{\omega^2}{c^2}\sum_p \varepsilon_p \vec{U}'(k+k_i,n,a)_{p-m} = 0. \tag{2.89}$$

This is algebraically the same as that in (2.86) so the two problems have the same solution sets.

A simple example of (2.86) is the case of a constant dielectric $\varepsilon(z) = \varepsilon$. For this dielectric, it follows that $\varepsilon_p = \varepsilon\delta_{p,0}$ so that

$$(k+k_m)^2\vec{U}(k,n,a)_m - \frac{\omega^2}{c^2}\varepsilon_0\vec{U}(k,n,a)_m = 0 \tag{2.90}$$

where $\vec{U}$ is an even function of the subscript m. This has the standard solution of a plane wave solution in a uniform homogeneous medium.

To summarize: the spatial periodicity of the dielectric function gives rise to frequency modes of the form

$$\vec{E}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{U}_{\vec{k},n,a}(\vec{r}), \tag{2.91a}$$

$$\vec{B}_a(\vec{r}) = \mathrm{e}^{i\vec{k}\cdot\vec{r}}\vec{V}_{\vec{k},n,a}(\vec{r}) \tag{2.91b}$$

where $\vec{U}_{\vec{k},n,a}(\vec{r})$ and $\vec{V}_{\vec{k},n,a}(\vec{r})$ are periodic functions with the same periodicity as the dielectric. Due to the translational symmetry of the operator eigenvalue problem, the eigenvalue solution sets are invariant under a translation in $\vec{k}$-space by $\vec{k}_i = m_{i,1}\vec{b}_1 + m_{i,2}\vec{b}_2 + m_{i,3}\vec{b}_3$. Consequently, $\left\{\omega^2(\vec{k})\right\}$ and $\left\{\omega^2\left(\vec{k}+\vec{k}_i\right)\right\}$ are the same.

Example of a One-Dimensional Photonic Crystal

A simple example of the properties of photonic crystals is given by the treatment of a periodically layered medium. This problem is solved exactly for all of the earlier derived formal properties arising in periodic systems. As will be discussed later, it has many practical applications in designs of laser mirrors, optical coatings, optical transistors, and other types of devices.

Consider a periodic system of slabs of thickness d with the interfaces between the slabs taken to be perpendicular to the z-axis. The layering consists of slabs of dielectric constants ε and vacuum alternating in their layering along the z-axis (see Fig. 2.3 for a schematic drawing of the layered system).

Specifically, the periodic dielectric, $\varepsilon(z)$, of the layering is defined by [10, 11]

$$\varepsilon(z) = \varepsilon \quad \text{for} \quad 2nd \le z \le (2n+1)d \tag{2.92a}$$

$$\varepsilon(z) = 1 \quad \text{for} \quad (2n+1)d \le z \le (2n+2)d \tag{2.92b}$$

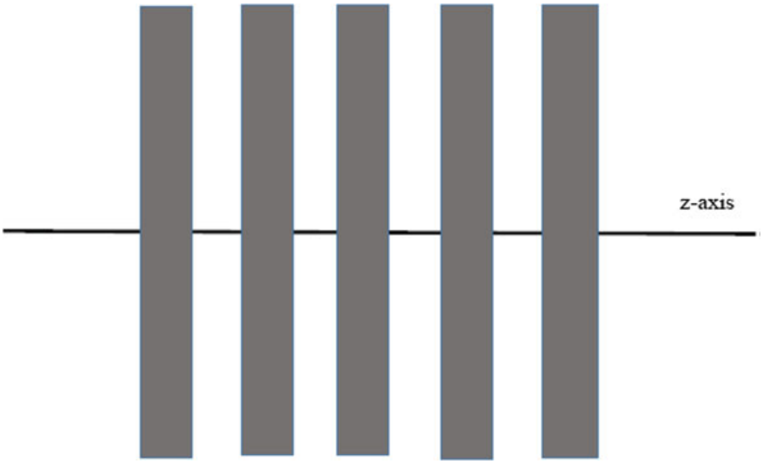


Fig. 2.3 Schematic drawing of the one-dimensional layered media of slabs of thickness d periodically layered with vacuum and dielectric of dielectric constant n. The horizontal line is the z-axis and the slabs are infinite in the $x-y$ plane [10]

where n is an integer. As a simple illustration of the periodic properties of the medium, the system is studied for light propagating along z-axis, obtaining the band structure, the form of the wave functions, and the periodicity properties mentioned in the earlier discussions.

To begin, it is helpful to consider a single slab of dielectric constant ε surrounded by vacuum and to determine the solutions of the waves propagating along the z-axis. Again, the interfaces of the slab are perpendicular to the z-axis. For generality, the surfaces of the slab are taken at $z = z_0$ and $z = z_1$ where $z_0 < z_1$.

With this geometry, the electric field in the vacuum satisfies [10, 11]

$$\left[\frac{d^2}{dz^2} + \frac{\omega^2}{c^2}\right]E = 0 \tag{2.93a}$$

and within the dielectric satisfies

$$\left[\frac{d^2}{dz^2} + \varepsilon\frac{\omega^2}{c^2}\right]E = 0. \tag{2.93b}$$

From (2.93) the frequency dependent solutions of the fields outside and within the slab have the following forms: To the left of the slab the fields are [10]

$$E_L(z,t) = \left[A\mathrm{e}^{ik_0 z} + B\mathrm{e}^{-ik_0 z}\right]\mathrm{e}^{-i\omega t}, \tag{2.94a}$$

within the slab the fields are

$$E_S(z,t) = \left[C\mathrm{e}^{ikz} + D\mathrm{e}^{-ikz}\right]\mathrm{e}^{-i\omega t}, \tag{2.94b}$$

and to the right of the slabs the fields are

$$E_R(z,t) = \left[E\mathrm{e}^{ik_0z} + F\mathrm{e}^{-ik_0z}\right]\mathrm{e}^{-i\omega t}. \tag{2.94c}$$

In (2.94)

$$k_0 = \frac{\omega}{c}, \tag{2.95a}$$

and

$$k = \sqrt{\varepsilon}\frac{\omega}{c}. \tag{2.95b}$$

The boundary conditions at the interfaces between the vacuum and slab are the continuity of the electric fields,

$$E_{vac} = E_{slab}, \tag{2.96a}$$

and of the field derivatives,

$$\frac{\partial E_{vac}}{\partial z} = \frac{\partial E_{slab}}{\partial z}. \tag{2.96b}$$

Applying the boundary conditions in (2.96) at the right interface of the slab gives the matrix equation

$$\begin{vmatrix} \mathrm{e}^{ikz_1} & \mathrm{e}^{-ikz_1} \\ \mathrm{e}^{ikz_1} & -\mathrm{e}^{-ikz_1} \end{vmatrix} \begin{vmatrix} C \\ D \end{vmatrix} = \begin{vmatrix} \mathrm{e}^{ik_0z_1} & \mathrm{e}^{-ik_0z_1} \\ \frac{k_0}{k}\mathrm{e}^{ik_0z_1} & -\frac{k_0}{k}\mathrm{e}^{-ik_0z_1} \end{vmatrix} \begin{vmatrix} E \\ F \end{vmatrix}, \tag{2.97a}$$

and at the left interface of the slab gives the matrix equation

$$\begin{vmatrix} \mathrm{e}^{ik_0z_0} & \mathrm{e}^{-ik_0z_0} \\ \mathrm{e}^{ik_0z_0} & -\mathrm{e}^{-ik_0z_0} \end{vmatrix} \begin{vmatrix} A \\ B \end{vmatrix} = \begin{vmatrix} \mathrm{e}^{ikz_0} & \mathrm{e}^{-ikz_0} \\ \frac{k}{k_0}\mathrm{e}^{ikz_0} & -\frac{k}{k_0}\mathrm{e}^{-ikz_0} \end{vmatrix} \begin{vmatrix} C \\ D \end{vmatrix} \tag{2.97b}$$

After some algebra, it follows from these two matrix equations that the field coefficients in the vacuum at the left and right of the slab are related by [10]

$$\begin{vmatrix} A\mathrm{e}^{ik_0z_0} \\ B\mathrm{e}^{-ik_0z_0} \end{vmatrix} = \begin{vmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{vmatrix} \begin{vmatrix} E\mathrm{e}^{ik_0z_1} \\ F\mathrm{e}^{-ik_0z_1} \end{vmatrix}. \tag{2.98}$$

Here

$$M_{11}(z_0 - z_1) = \cos[k(z_0 - z_1)] + \frac{i}{2}\left(\frac{k_0}{k} + \frac{k}{k_0}\right)\sin[k(z_0 - z_1)] \tag{2.99a}$$

$$M_{12}(z_0 - z_1) = \frac{i}{2}\left(\frac{k}{k_0} - \frac{k_0}{k}\right)\sin[k(z_0 - z_1)] \tag{2.99b}$$

$$M_{21}(z_0 - z_1) = -\frac{i}{2}\left(\frac{k}{k_0} - \frac{k_0}{k}\right)\sin[k(z_0 - z_1)] \tag{2.99c}$$

and

$$M_{22}(z_0 - z_1) = \cos[k(z_0 - z_1)] - \frac{i}{2}\left(\frac{k_0}{k} + \frac{k}{k_0}\right)\sin[k(z_0 - z_1)] \tag{2.99d}$$

From (2.99) it is seen that any two of the field coefficient A, B, E, F determine the other two. These relations can now be applied to obtain the solution of the periodic layering of the photonic crystal [10].

Equations (2.98) and (2.99) can be applied to the periodic layering in (2.92), treating the fields outputted by one slab of the layering as the fields inputted at its nearest neighbor slabs. In this way, the fields across a repeat segment of the periodic layering are used to generate the solutions of the entire periodic system. The repeat units of the periodic layering in (2.92) are defined such that the nth vacuum-dielectric repeat unit is taken as the slabs between the $z_1 = (2n + 1)d$ interface and the $z_1' = z_1 - 2d = (2n - 1)d$ interface. It is a repeat unit containing two slabs, a vacuum slab on the left of the unit and a dielectric slab on the right of the unit. This forms the basic vacuum-dielectric unit from which the entire periodic photonic crystal layering is generated.

Applying (2.98) between the two points z_1 and z_1', and introducing some notational changes meant to facilitate the study of the periodic system, gives the relation [10]

$$\begin{vmatrix} A_{n,0}\mathrm{e}^{ik_0(2n-1)d} \\ B_{n,0}\mathrm{e}^{-ik_0(2n-1)d} \end{vmatrix} = \begin{vmatrix} \mathrm{e}^{-ik_0d} & 0 \\ 0 & \mathrm{e}^{ik_0d} \end{vmatrix} \begin{vmatrix} M_{11}(-d) & M_{12}(-d) \\ M_{21}(-d) & M_{22}(-d) \end{vmatrix} \begin{vmatrix} A_{n+1,0}\mathrm{e}^{ik_0(2n+1)d} \\ B_{n+1,0}\mathrm{e}^{-ik_0(2n+1)d} \end{vmatrix}, \tag{2.100}$$

Here the fields in the left vacuum slab of the nth dielectric-vacuum repeat unit have been, from (2.94a), written in the form [10]

$$E_{L,n}(z,t) = \left[A_{n,0}\mathrm{e}^{ik_0z} + B_{n,0}\mathrm{e}^{-ik_0z}\right]\mathrm{e}^{-i\omega t}, \tag{2.101a}$$

where the subscripted n indicates that the fields and coefficients are those of the vacuum slab in the nth repeat unit. The field in the vacuum slab to the right of the dielectric slab in the nth vacuum-dielectric repeat unit, $E_{R,n}(z,t)$, is just the field in the right vacuum slab of the $(n+1)$th vacuum-dielectric repeat unit, i.e., $E_{L,n+1}(z,t)$. Consequently, in our new notation, from (2.94) it follows that

$$E_{R,n}(z,t) = E_{L,n+1}(z,t) = \left[A_{n+1,0}\mathrm{e}^{ik_0 z} + B_{n+1,0}\mathrm{e}^{-ik_0 z}\right]\mathrm{e}^{-i\omega t}, \tag{2.101b}$$

where the $A_{n+1,0}$, and $B_{n+1,0}$ coefficient notation used here has been applied in (2.101a). The new notation developed here is found to aid in treating the periodic nature of the layering.

In terms of (2.101a) the right and left propagating fields at $z = (2n-1)d$ are, respectively,

$$A_{n,0}\mathrm{e}^{ik_0(2n-1)d} = A_n\mathrm{e}^{-ik_0 d} \tag{2.102a}$$

$$B_{n,0}\mathrm{e}^{-ik_0(2n-1)d} = B_n\mathrm{e}^{ik_0 d} \tag{2.102b}$$

and from (2.101b) the right and left propagating fields at $z = (2n+1)d$ are, respectively,

$$A_{n+1,0}\mathrm{e}^{ik_0(2n+1)d} = A_{n+1}\mathrm{e}^{-ik_0 d} \tag{2.103a}$$

$$A_{n+1,0}\mathrm{e}^{-ik_0(2n+1)d} = B_{n+1}\mathrm{e}^{ik_0 d}. \tag{2.103b}$$

In (2.102) and (2.103) the notation $A_n = A_{n,0}\mathrm{e}^{i2nk_0 d}$, $B_n = B_{n,0}\mathrm{e}^{-i2nk_0 d}$, $A_{n+1} = A_{n+1,0}\mathrm{e}^{i2(n+1)k_0 d}$, and $B_{n+1} = B_{n+1,0}\mathrm{e}^{-i2(n+1)k_0 d}$ is introduced.

Applying this notation and using (2.100) it follows that

$$\begin{vmatrix} A_n \\ B_n \end{vmatrix} = \begin{vmatrix} \mathrm{e}^{-ik_0 d}M_{11}(-d) & \mathrm{e}^{ik_0 d}M_{12}(-d) \\ \mathrm{e}^{-ik_0 d}M_{21}(-d) & \mathrm{e}^{ik_0 d}M_{22}(-d) \end{vmatrix} \begin{vmatrix} A_{n+1} \\ B_{n+1} \end{vmatrix} \tag{2.104}$$

The resulting (2.104) relates the right and left propagating electric field components at the right and left interfaces of the nth dielectric-vacuum repeat unit. In terms of the newly defined coefficients A_n, B_n, A_{n+1}, and B_{n+1} the fields in the vacuum adjacent to the left of the nth dielectric slab are [10]

$$E_{L,n}(z,t) = \left[A_n\mathrm{e}^{ik_0[z-2nd]} + B_n\mathrm{e}^{-ik_0[z-2nd]}\right]\mathrm{e}^{-i\omega t}, \tag{2.105a}$$

and the fields in the vacuum adjacent to the right of the nth dielectric slab are

$$E_{R,n}(z,t) = E_{L,n+1}(z,t) = \left[A_{n+1}\mathrm{e}^{ik_0[z-2(n+1)d]} + B_{n+1}\mathrm{e}^{-ik_0[z-2(n+1)]d}\right]\mathrm{e}^{-i\omega t}. \tag{2.105b}$$

From the relation between the field coefficients on the right and left of the repeat unit of the periodic layering a transfer matrix can be defined. This is a matrix that can be used by its successive applications to generate the entire set of coefficients $\{A_n\}$ $\{B_n\}$ along the periodic layering. It is seen from (2.104) that the transfer matrix is

$$\overset{\leftrightarrow}{T} = \begin{vmatrix} \mathrm{e}^{-ik_0 d} M_{11}(-d) & \mathrm{e}^{ik_0 d} M_{12}(-d) \\ \mathrm{e}^{-ik_0 d} M_{21}(-d) & \mathrm{e}^{ik_0 d} M_{22}(-d) \end{vmatrix} = \begin{vmatrix} \alpha & \beta \\ \beta^* & \alpha^* \end{vmatrix}, \tag{2.106a}$$

where

$$\alpha = \mathrm{e}^{-ik_0 d}\left[\cos kd - \frac{i}{2}\left(\frac{k_0}{k} + \frac{k}{k_0}\right)\sin kd\right] \tag{2.106b}$$

and

$$\beta = -\mathrm{e}^{ik_0 d}\,\frac{i}{2}\left(\frac{k}{k_0} - \frac{k_0}{k}\right)\sin kd. \tag{2.106c}$$

Applying the transfer matrix, for example, the field between the 0th and nth repeat units are given by [10]

$$\begin{vmatrix} A_0 \\ B_0 \end{vmatrix} = \overset{\leftarrow}{T}^{n} \begin{vmatrix} A_n \\ B_n \end{vmatrix}. \tag{2.107}$$

Consequently, all of the $\{A_n\}$, $\{B_n\}$ coefficients of the infinite system are generated through the repeated application of (2.107).

From (2.107) it is seen that the matrix $\overset{\leftrightarrow}{T}$ must generate sets of coefficients which are bounded over the extent of the layering. If this is not the case, the wave function solutions will blow up at some point along the periodic layering. Solutions which approach infinite limits along the layering exhibit unacceptable physical properties as they do not represent either propagating or bound states within the layering. The requirement set upon $\overset{\leftrightarrow}{T}$ for it to generate bound solutions are now addressed. In the course of this discussion the dispersion relation of the propagating modes in the layering is obtained.

Since $\overset{\leftrightarrow}{T}$ is diagonalizable the general properties of the $\{A_n\}$, $\{B_n\}$ solutions, generated through the application of (2.107), can be deduced in terms of the eigenvalue and eigenvectors of $\overset{\leftrightarrow}{T}$. The eigenvalue problem of $\overset{\leftrightarrow}{T}$ and how its solution affects the physics of the periodic layering is now addressed. From (2.106) the eigenvalue problem involving $\overset{\leftrightarrow}{T}$ has the form [10]

$$\begin{vmatrix} \alpha - \lambda & \beta \\ \beta^* & \alpha^* - \lambda \end{vmatrix} \begin{vmatrix} A \\ B \end{vmatrix} = 0 \tag{2.108}$$

This is solved applying standard methods of linear algebra to find that the eigenvalues and eigenvector of (2.108). The eigenvalues are shown to be given by

$$\lambda_\pm = \frac{R \pm \sqrt{R^2 - 4}}{2}, \tag{2.109a}$$

where

$$R = 2\cos k_0 d\, \cos kd - \left(\frac{k_0}{k} + \frac{k}{k_0}\right) \sin k_0 d\, \sin kd. \tag{2.109b}$$

The eigenvectors associated with these $\lambda_\pm$ are from (2.108) of the form

$$A_\pm = \frac{-\beta}{\alpha - \lambda_\pm} B_\pm. \tag{2.109c}$$

Here $B_\pm$ can be chosen arbitrarily or to meet some form of normalization conditions, and the coefficients (A_+, B_+) (corresponding to λ_+) and (A_-, B_-) (corresponding to λ_-) are a complete set in the space of possible (A_n, B_n) field coefficients. In addition, for the following discussions, it should be noted from (2.106) and (2.109) that $A_\pm$ and $B_\pm$ only depend on the variables k_0, k, and d so that they are independent of position along the photonic crystal layering.

Since the eigenvectors of (2.108) are a complete set, the behavior of the transfer matrix relation in (2.106) is set by the properties of the eigenvalues of $\overleftrightarrow{T}$. In the coefficient generating (2.106), if $(A_n, B_n) = (A_+, B_+)$ it is found that the transfer matrix is of the form

$$\overleftrightarrow{T} = \begin{vmatrix} \lambda_+ & 0 \\ 0 & \lambda_+ \end{vmatrix} \tag{2.110a}$$

Similarly taking $(A_n, B_n) = (A_-, B_-)$ in (2.107) implies that

$$\overleftrightarrow{T} = \begin{vmatrix} \lambda_- & 0 \\ 0 & \lambda_- \end{vmatrix} \tag{2.110b}$$

Consequently, the properties of the coefficients in the layered system of slabs are determined by the integer powers of λ_+ and λ_-. The eigenvectors corresponding to these eigenvalues, in fact, are the modes of the layered system, corresponding to modal solutions of distinct frequency and wave vector. As noted later, the dispersion relations of the modes are obtained by studying the eigenvalues in (2.108). It is necessary to turn to the discussion of eigenvalue properties to see how these affect the behavior of the system, determine the dispersion relation of the modes, and how the eigenvalues are related to the modes of the periodic layering.

For the generation of a nonzero and bounded sets of $\{A_n\}$, $\{B_n\}$ coefficients along the layering, a set of highly restrictive conditions are required on λ_+ and λ_-. For bounded solutions of (2.107), the $\lambda_\pm$ need to be of the form

$$\lambda_\pm = \mathrm{e}^{\pm i2Kd}. \tag{2.111}$$

for real K. The form of λ_+ and λ_- in (2.109) means that the coefficients and their integral powers will maintain themselves at unit modulus all along the layering. This keeps the coefficient sets $\{A_n\}$, $\{B_n\}$ to bound solutions along the layering. The consequences of this are seen from (2.107).

In addition, the trace of $\overleftrightarrow{T}$ is the sum of its eigenvalues so that from (2.109) it is found that K must satisfy [10]

$$2\cos 2Kd = \alpha + \alpha^* = 2\cos k_0 d\, \cos kd - \left(\frac{k_0}{k} + \frac{k}{k_0}\right) \sin k_0 d\, \sin kd. \tag{2.112}$$

This equation is seen to relate K to the general physical parameters of the system.

To examine (2.112) for the general layering, it is found from (2.95) that $k = \sqrt{\varepsilon} k_0$ so that (2.112) relates either k_0 or k to the phase argument K in an expression that also depends on d and ε. In later discussions it is shown that, for fixed d and ε, there are values of k_0 or k for which K has no real solutions.

In these cases the system does not have propagating solutions. This causes the solution to exhibit an energy band structure similar to that observed in the dispersion relation of electrons in semiconductors. Consequently, propagating solutions of light only exist for real solutions of K. This has a great consequence for the dispersion of light in the layered medium. Before some numerical results are presented, demonstrating the band structure of the system, a discussion of the properties of the wave functions is now given.

To complete the solution for the modes and the modal dispersion relation of the infinite photonic crystal it is necessary to take into account the boundary conditions on the solutions of the differential equations in (2.93) at the ends of the infinite photonic crystal. The boundary conditions that are usually applied are periodic boundary conditions between the ends at the left and right $z \rightarrow \pm\infty$ infinite edges of the layering. These boundary conditions give traveling waves type of solutions that are found to be most effective in treating the transport properties of the system. The boundary conditions also place restrictions on the form of K entering into the dispersion relation in (2.112).

If there are $N \rightarrow \infty$ repeat units in the photonic crystal layering, the length of the photonic crystal layering is $L = 2Nd$ where $2d$ is the length of a repeat unit of the periodic layering. For N layerings in the photonic crystal, periodic boundary conditions on the system can be viewed as taking the $N + 1$ layer of the photonic crystal to be equivalent to the first layering of the photonic crystal. Considering the form of the fields in (2.101), this means that

$$E_{L,1}(z=d,t)=E_{L,N+1}(z=(2N+1)d,t) \tag{2.113a}$$

so that from (2.105a) and the definitions of A_n and B_n in (2.102) and (2.103) it follows that

$$A_1\mathrm{e}^{-ik_0d}+B_1\mathrm{e}^{ik_0d}=A_{N+1}\mathrm{e}^{-ik_0d}+B_{N+1}\mathrm{e}^{ik_0d}. \tag{2.113b}$$

Examining this relationship for each of the two complete set of eigenvector states, i.e., for each of the choices

$$(A_n,B_n)=(A_\pm,B_\pm), \tag{2.113c}$$

it follows from (2.107), (2.110), and (2.113b) that

$$\lambda_\pm^N=1 \tag{2.113d}$$

is required. From (2.113d) K must be of the form

$$K=\frac{2\pi n}{2Nd}, \tag{2.114}$$

where n is an integer. This condition fixes the allowed values of K and consequently of the k_0 and k consistent with the periodic boundary conditions.

Next, consider the wave functions corresponding to the dispersion relations. The vacuum fields in the nth repeat unit can now be written in terms of the eigenvector solutions of (2.108). This shall be done for the λ_+ solutions of (2.108). For the eigenvalue $\lambda_+=\mathrm{e}^{i2Kd}$ the eigenvectors can be denoted $(A_+(K),B_+(K))$, and from (2.105a)

$$\begin{aligned}E_{L,n}(z,t)&=\left[A_n\mathrm{e}^{ik_0[z-2nd]}+B_n\mathrm{e}^{-ik_0[z-2nd]}\right]\mathrm{e}^{-i\omega t}\\&=\mathrm{e}^{i2nKd}\left[A_+(K)\mathrm{e}^{ik_0[z-2nd]}+B_+(K)\mathrm{e}^{-ik_0[z-2nd]}\right]\mathrm{e}^{-i\omega t}\end{aligned} \tag{2.115}$$

represents the field within the region $(2n-1)d\le z\le(2nd)$. Equation (2.115) can be rewritten as

$$E_{L,n}(z,t)=\mathrm{e}^{iKz}U_{K,n}(z)\mathrm{e}^{-i\omega t} \tag{2.116a}$$

where

$$U_{K,n}(z) = \mathrm{e}^{-ikK[z-2nd]}\left\{A_+(K)\mathrm{e}^{ik_0[z-2nd]} + B_+(K)\mathrm{e}^{-k_0[z-2nd]}\right\} \tag{2.116b}$$

defined over the region $(2n-1)d \leq z \leq (2nd)$. The general expression for the fields within all of the vacuum slabs in then given by

$$E_{vac}(z,t) = \mathrm{e}^{iKz}U_K(z)\mathrm{e}^{-i\omega t} \tag{2.117a}$$

for

$$U_K(z) = \mathrm{e}^{-ikK[z-2nd]}\left\{A_+(K)\mathrm{e}^{ik_0[z-2nd]} + B_+(K)\mathrm{e}^{-k_0[z-2nd]}\right\} \tag{2.117b}$$

where n is the smallest integer for which $\frac{z}{2d} \leq n$.

The wave function in (2.117) is of the form given in (2.72) needed to satisfy the symmetry restrictions on the system. Similar calculations can be done for the λ_-, $(A_-(K), B_-(K))$ eigenvalue-eigenvector results. The same can be done for the fields within the dielectric slabs for a complete solution of the fields throughout the entire layering.

Once the model solutions of the layered systems are obtained. Expansions of the general non-modal solutions for the electromagnetic fields can be written in terms of them. This is done using the orthogonality properties of the eigensolutions given by

$$\int\limits_{-\infty}^{\infty} dz\varepsilon(z)E_K^*(z)E_{K'}(z) = 2\pi\delta(K-K') \tag{2.118}$$

where the eigenmodes are of the form discussed earlier

$$E_K(z,t) = E_K(z)\mathrm{e}^{-i\omega t} = \mathrm{e}^{iKz}U_{K,layers}(z)\mathrm{e}^{-i\omega t}. \tag{2.119}$$

Here $U_{K,layers}(z)$ is defined over the entire z-axis as the periodic function part of the K modal solution of the one-dimensional photonic crystal.

To illustrate some of the properties of the one-dimensional photonic crystals, the dispersion relation of a layering is now discussed, with a presentation of numerical data evaluated for a specific set of parameters. The results are found to exhibit a series of pass and stop bands and to display periodicity of the dispersion relation in wave vector space. In the previous section, these properties were of a type noted to be common to all periodic systems.

The dispersion relation of the system defined in (2.92) is obtained from (2.109) and (2.112) as

$$\cos 2Kd = \cos\frac{\omega}{c}d\,\cos\sqrt{\varepsilon}\frac{\omega}{c}d - \frac{1}{2}\left(\frac{1}{\sqrt{\varepsilon}} + \sqrt{\varepsilon}\right)\sin\frac{\omega}{c}d\,\sin\sqrt{\varepsilon}\frac{\omega}{c}d, \qquad (2.120)$$

where (2.95) relating k_0 and k to $\frac{\omega}{c}$ have been applied. Equation (2.120) is a transcendental equation with solutions yielding $\frac{\omega}{c}$ as a function of K for the modal solutions of the layered media. It can be solved numerically for specific examples.

In Fig. 2.4a, a plot from (2.120) of the dispersion relation is generated for a systems with dielectric slabs of refractive index $n = \sqrt{\varepsilon} = 10$. The plot shows $\frac{\omega}{c}d$ determined as a function of Kd and is observed to consist of a series of stop and pass bands. Only modes at pass band frequencies are allowed in the system, and modes at stop band frequencies can be shown to decay exponentially in the system. Consequently, solutions at stop band frequencies are not stable solutions. Notice that in Fig. 2.4a only the plot of positive Kd are shown as (2.120) is invariant under the change of sign $Kd \rightarrow -Kd$.

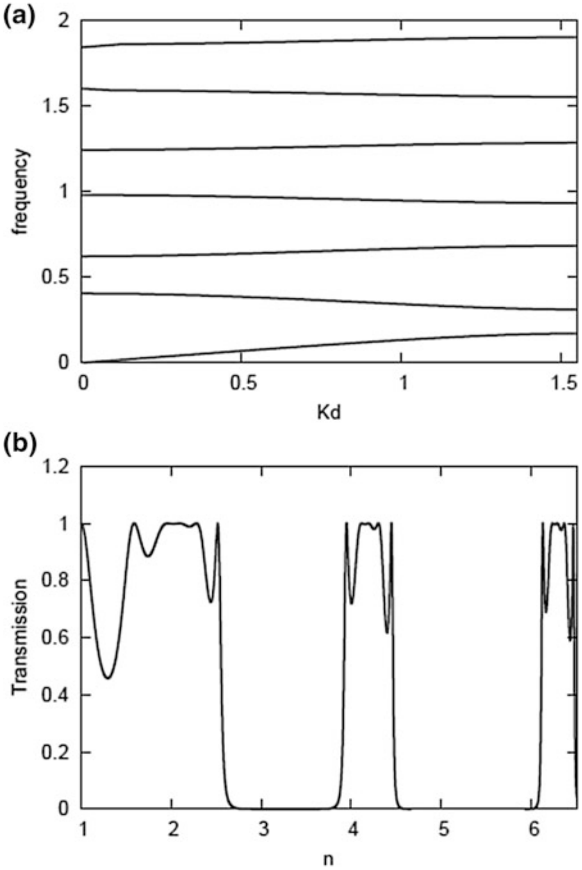


Fig. 2.4 Plots of the properties of the one-dimensional photonic crystal showing: **a** the dispersion relation, $\frac{\omega}{c}d$ as a function of Kd from (2.120) for dielectric slabs of refractive index $n = 10$, and **b** the transmission coefficient versus slab index of refraction for a wave with $k_0 d = 1.5$ incident on a layering of five dielectric-vacuum slabs. The plots are present for dielectric and vacuum slabs each of width d [10]

In accordance with the discussions in the earlier section on the properties of periodic systems, the dispersion is periodic in Kd with a periodicity of $\Delta Kd = \pi$. The periodicity extends over the entire range of $-\infty < Kd < \infty$ and shall be seen as a limit on the general properties of the solutions of the system.

From the discussions in (2.115) through (2.119), the form of the fields in (2.119) are found to also obey the symmetry

$$E_K(z,t) = E_{K+n\Delta K}(z,t) \tag{2.121}$$

for n an integer. Consequently, both the dispersion relation and the wave functions of the modes exhibit the same periodicity in Kd. The solutions found in a length ΔK of the K axis represent a set of unique solution of the modal problem of the layer system. Solutions outside this interval are replicas of those within the length ΔK.

Coatings on Interfaces and Mirrors

An interesting variation of the one-dimensional photonic crystal problem which is relevant to the study of surface coatings is a treatment of the transmission properties of a barrier composed as a finite layering of photonic crystal [10]. It is surprising that a finite layering of even a small number of dielectric slabs exhibits many of the properties of the infinite layered system. In the following such a comparison of finite and infinite systems is discussed.

Consider the system defined in (2.92), but now restrict to a finite number of layers. For a system of five layers [10]

$$\varepsilon(z) = \varepsilon \quad \text{for } 2nd \le z \le (2n+1)d \tag{2.122a}$$

$$\varepsilon(z) = 1 \quad \text{for } (2n-1)d \le z \le 2nd \tag{2.122b}$$

for $n = 0, 1, 2, 3, 4$ and

$$\varepsilon(z) = 1 \quad \text{for } z < -d \text{ an } 9d < z. \tag{2.122c}$$

A schematic of this system can be represented by the five slabs shown in Fig. 2.3.

The fields within the vacuum layers defined in (2.122) are of the form given in (2.105a) and their coefficient are again related by (2.104). Outside the layering the fields for transmission boundary conditions are given by

$$E_{inc,refl}(z,t) = \left[U\mathrm{e}^{ik_0z} + V\mathrm{e}^{-ik_0z}\right]\mathrm{e}^{-i\omega t} \tag{2.123a}$$

for $z < -d$ and

$$E_{trans}(z,t) = T\mathrm{e}^{ik_0z}\mathrm{e}^{-i\omega t} \tag{2.123b}$$

for $9d < z$. Equations (2.123a) and (2.123b), respectively, represent the incident and reflected waves on the left of the barrier and the transmitted wave to the right of the barrier.

The solutions of (2.122) and (2.123) are matched up using the boundary conditions in (2.96). This follows from the theory presented in (2.97). Using these matrix relations in combination with the relations in (2.104) the coefficient U, V are expressed in terms of T.

Once the amplitude of the incident, reflected, and transmitted waves are found the energy flow in the system can be studied. The Poynting vectors of the incident, reflected, and transmitted waves are, respectively,

$$\vec{S}_U = \frac{c}{8\pi}\vec{E}_U \times \vec{H}_U^* \tag{2.124a}$$

$$\vec{S}_V = \frac{c}{8\pi}\vec{E}_V \times \vec{H}_V^* \tag{2.124b}$$

$$\vec{S}_T = \frac{c}{8\pi}\vec{E}_T \times \vec{H}_T^* \tag{2.124c}$$

where $\vec{E}_U(z,t) = \vec{U}\mathrm{e}^{ik_0 z}\mathrm{e}^{-i\omega t}$, $\vec{E}_V(z,t) = \vec{V}\mathrm{e}^{-ik_0 z}\mathrm{e}^{-i\omega t}$, $\vec{E}_T(z,t) = \vec{T}\mathrm{e}^{ik_0 z}\mathrm{e}^{-i\omega t}$ have been rewritten to make their vector nature manifest and $\vec{H}_U(z,t)$, $\vec{H}_V(z,t)$, $\vec{H}_T(z,t)$ are the associated magnetic fields. In terms of these Poynting vectors the refection and transmission coefficients of the incident wave are

$$R_{refl} = \frac{|\vec{S}_V|}{|\vec{S}_U|} \tag{2.125a}$$

$$T_{Trans} = \frac{|\vec{S}_T|}{|\vec{S}_U|} \tag{2.125b}$$

respectively.

For normal incidence the reflection and transmission coefficients are independent of the polarization of the electromagnetic waves. Once the polarization of the electric field is chosen, however, the polarization is of magnetic field in the electromagnetic wave is determined relative to that of the electric field. This determines the flow of energy in the system.

Numerical results for the transmission of the system of five dielectric layers plotted as a function of the dielectric constant are presented in Fig. 2.4b. The plot is made for a wave with $k_0 d = 1.5$ incident on the array of dielectric-vacuum slabs each of which is of width d. Even for this small number of layers the regions of high and low transmission in the plot compare well with the results for the stop and pass bands in the dispersion relation of the infinite layered system [10].

As the dielectric constant of the dielectric slabs in the array is varied the transmission is seen to go through a series of regions of near zero transmission. These regions of near zero transmission for incident waves with $k_0 d = 1.5$ correspond fairly well with the stop bands of the $k_0 d = 1.5$ waves in the associated infinite array of dielectric-vacuum slabs at that particular slab dielectric constant [10].

For example, in the infinite system the stop bands for the $k_0 d = 1.5$ system are obtained (2.120) to be located in the regions $1.054 < n < 1.234$, $2.611 < n < 3.876$, and $4.497 < n < 6.081$. The lowest stop band shows the poorest correlation, with a rather incomplete transmission minimum in the region $1.054 < n < 1.234$. In this case the dielectric slabs in the region have dielectric constants which contrast poorly with the vacuum layers. The higher bands, however, offer greater contrasts between the dielectric slabs and the vacuum and the correlation is very good.

In general, the array of five slabs already displays many of the filtering properties of the infinite system, refusing to transmit or propagate modes through the systems in the stop bands. Likewise, the pass band modes of the associated infinite array are generally allowed to pass through the finite array with high transmission coefficients.

The types of one-dimensional photonic crystal layerings treated above find many applications in confining or filtering radiation. Examples are fiber Bragg gratings and distributed Bragg reflectors. These types of systems are used as filters in fiber optics having applications in telecommunications, optical sensors, and in the design of lasers [10].

Another type of application of periodic layerings is in the design of coatings for mirrors [10]. Here the coatings are applied to modulate the reflection of light or the fields excited near the surface of mirror. The last application usually involves mirrors composed with surface features that break the translational symmetry of the otherwise planar surface of the mirror. This allows for the excitation of surface plasmon-polaritons at the mirror surface.

As an example of such an application, consider a finite layering of dielectric-vacuum slabs on the surface of a mirror formed as a planar surface of perfect conductor. The fields in the layering are computed. The focus is on determining the field properties within the layering.

Consider the system defined in (2.92) restricted to a finite number of layers. For a system of five layers on a perfect conducting mirror, the layering will be described by [10]

$$\varepsilon(z) = \varepsilon \quad \text{for } 2nd \le z \le (2n+1)d \tag{2.126a}$$

$$\varepsilon(z) = 1 \quad \text{for } (2n-1)d \le z \le 2nd \tag{2.126b}$$

for $n = 0, 1, 2, 3$, with

$$\varepsilon(z) = 1 \quad \text{for } 7d \le z \le 8d \tag{2.126c}$$

$$\varepsilon(z) = \varepsilon \quad \text{for } 8d \le z \le \frac{17}{2}d \tag{2.126d}$$

and

$$\varepsilon(z) = 1 \quad \text{for } z < -d \, \text{an} \, \frac{17}{2} d < z. \tag{2.126e}$$

The surface of the perfect conducting mirror is located on the right side of the layering at $z = \frac{17}{2} d$. Here the width of the slab adjacent to the mirror has been taken to have a width $\frac{d}{2}$ so that the resulting layering and its image within the mirror maintain the periodic layering of the earlier discussed barrier. This assures that the coating on the perfect conducting surfaces retains the pass and stop band structure of the barrier and its related infinite array (see Fig. 2.5 for a schematic figure which gives an example of the system).

The fields within the vacuum layers defined in (2.126) are again of the form given in (2.105a) and their coefficient are related by (2.104). Outside the layering the fields for incident and reflected waves from the mirror are given by

$$E_{inc,refl}(z,t) = \left[U e^{ik_0 z} + V e^{-ik_0 z}\right] e^{-i\omega t} \tag{2.127a}$$

for $z < -d$ and at the perfection conducting mirror

$$E_{mir}\left(z = \frac{17}{2} d, t\right) = 0 \tag{2.127b}$$

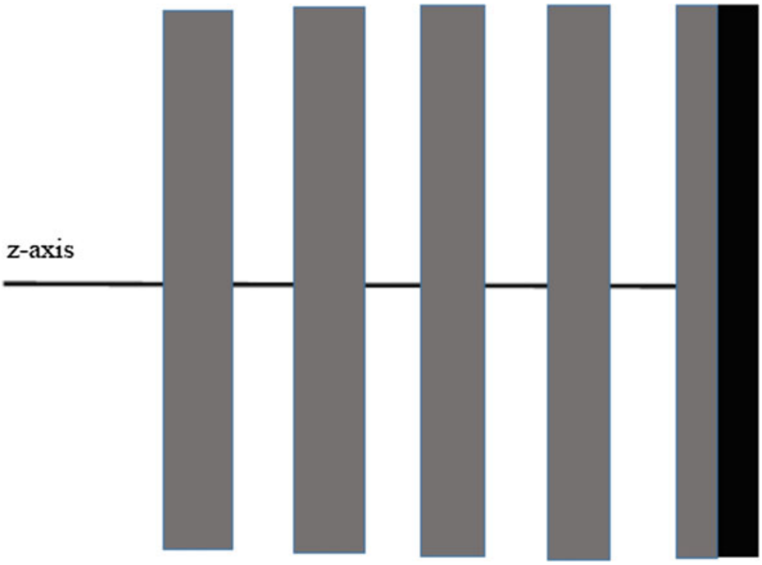


Fig. 2.5 Schematic figure of a layered coating on a perfect reflecting mirror. The mirror is represented in black on the far right side of the figure

This is essentially the statement that the transmission amplitude at the right edge of the layering is zero while leaving the derivative of the field at the perfect conducting surface unspecified.

Equations (2.127a) represents the incident and reflected waves on the left of the barrier, and the reflection amplitude for the mirror can be obtained by applying (2.125a). In the absence of dielectric losses, the refection coefficient is one. The incident and reflected wave amplitude in (2.127a), however, experience a phase shift from the layered media and this phase shift shows up in the distribution of the electromagnetic fields within the coating.

The properties of the coating on the perfect conductor mirror are illustrated with a numerical example based on the same parameters as those used in generating the results in Fig. 2.4b. To illustrate the behavior of the fields within the coating, the field amplitudes are determined at the left hand vacuum-dielectric interface of each slab.

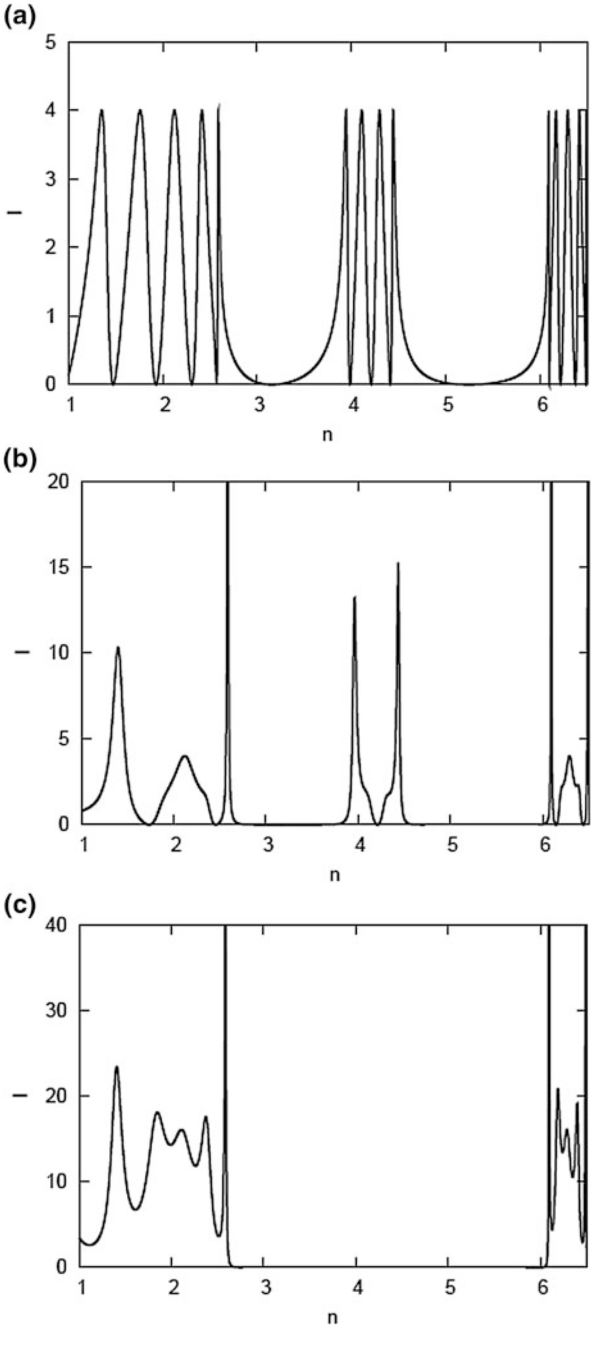
Numerical results are presented in Fig. 2.6 for the amplitude of the waves at some of the interfaces of the array as a function of the dielectric constant. The plot shows result for the field intensity at the left surfaces of three different dielectric slabs of the array.

The plots are made for a wave with $k_0 d = 1.5$ incident on the array of dielectric-vacuum slabs described by (2.126). As the dielectric constant of the dielectric slabs in the array is varied the amplitudes of the waves are seen to go through a series of regions of large and small amplitudes corresponding with the stop (low, near zero, transmissions) and pass (high, significantly greater than zero transmissions) band regions of the system in Fig. 2.4b.

For layerings made from dielectric slabs with index of refraction n within pass bands of Fig. 2.4b, the fields within the layered coating are found to be of the same order of magnitude as the incident and reflected fields in the vacuum to the left of the coating. This is due to the coating allowing the fields incident from the outside to pass to the perfect conducting surface, be reflected by the perfect conducting mirror, and pass back through the coating to return to the vacuum to the left of the coating.

For layerings made from dielectric slabs with index of refraction n within stop bands of Fig. 2.4b, the fields within the layered coating are found to be much less than the order of magnitude of the incident and reflected fields in the vacuum to the left of the coating. This is due to the coating not allowing the fields incident upon it from the outside to pass to the perfect conducting surface. Instead, the fields of the incident waves decay as it enters the coating, while being reflected back into the vacuum to the left of the coating.

Fig. 2.6 Plots of the field intensity at the interfaces of a five dielectric layer coating on a mirror as a function of n. Results are shown at the left surfaces of: **a** the left most dielectric slab, **b** the middle dielectric slab, and **c** the right most dielectric slab



2.3 Finite Difference Time Domain Simulations, Method of Moments, and Finite Element Simulation

In this section an outline will be given of some of the important points of the techniques of the finite difference time domain method [11–15], the method of moments [15–17] and the finite element method [18]. These are numerical methods which are commonly used to solve problems involving the propagation of electromagnetic waves through interactive media. The finite difference time domain method focuses on the motion of the fields through space and time while the method of moments and the finite element methods are focused on the frequency modes of the system. These three simulation techniques will be successively outlined in the following.

2.3.1 Computer Simulation Methods

Computer simulation methods are approaches that are commonly employed to obtain the solutions of problems in electrodynamics and have, in particular, formed a standard basis for the study of many of the diverse systems encountered in the fields of nanophotonics. All of the simulation methods are based on discretizing the four Maxwell equations in space, time, or frequency to form sets of algebraic difference equations. The resulting algebraic equations are then treated by computer to generate what is often an essentially exact solutions to the electrodynamics problems being considered.

As shall be seen the computer treatment, though a simplification over the difficulties involved in obtaining an exact solution to the set of differential equations, introduces new sets of obstacles which must be overcome in order to obtain an accurate representation of solutions. These include difficulties associated with the finite memory available to the computer, the finite computational time available, and the speed at which the computer operates. A balance must be struck in computer methods between the accuracy of the generated solution and the efficient management of the computer resources.

For the study of nanophotonics the most commonly employed computer methods are Finite Difference Time Domain methods [11–15], the Method of Moments [15–17], and Finite Element methods [15, 18]. These three different methodologies of computer simulation will be the focus of the following discussions. In the development of the simulation techniques, some of the basic advantages and disadvantages in their application to different types of problems in electrodynamics will be noted.

The basics of each of the three techniques will be discussed in the context of the study of the propagation and scattering of electromagnetic waves. In these treatments, systems will be considered in the absence of net electric and magnetic charges for problems involving regions containing dielectric and magnetic materials. This requires the introduction into the simulation program of the spatial organization of the various dielectric and magnetic materials and the definition of the regions in which the incident radiation is scattered and in which it leaves the scattering media.

Once the geometry and dielectric and magnetic nature of the scattering problem are defined, the next important aspect of the problem is the development of the Maxwell equations representing the scattering system. This includes a determination of the appropriate boundary conditions at the interfaces between the different regions composing the scattering volume acted upon by the computer simulation.

Following these considerations, the specified differential equations must be effectively discretized in order to reduce the problem to a manageable algebraic form. In handling the discretization a number of important considerations are, then, needed for writing an effective computer algorithm to efficiently process a set of inputted data so as to output a solution.

To begin these programing considerations, an important point to note is that the form of the Maxwell equations used in computer simulation studies is often modified from the standard set of Maxwell equations encountered in classical electrodynamics. The reason for this modification is to aid in the correct treatment of scattering boundary conditions at large separations from the scattering structures being studied. The boundary conditions in these regions, for example, must account for the need to solve a scattering problem in infinite space as it is approximated by a finite space represented within a computer memory.

In particular, the standard form of Maxwell's equations relates the electrical charges and currents and the electrical polarization and the magnetizations to the four fields of electrodynamics. The generalization of these used in computer simulation studies is often made by introducing a set of fictitious magnetic charges and magnetic currents into the standard form of the Maxwell equations. As shall be seen later, the fictitious magnetic charges and currents are useful in devising scattering boundary conditions.

Consequently, for many of the following discussions, it is convenient to take Maxwell's equations in the form [11, 14]:

$$\frac{\partial \vec{B}}{\partial t} = -\nabla \times \vec{E} - \vec{J}_m, \tag{2.128a}$$

$$\frac{\partial \vec{D}}{\partial t} = \nabla \times \vec{H} - \vec{J}_e, \tag{2.128b}$$

$$\nabla \cdot \vec{D} = 0, \tag{2.128c}$$

$$\nabla \cdot \vec{B} = 0, \tag{2.128d}$$

with constituent relations

$$\vec{B} = \mu\vec{H}, \tag{2.129a}$$

$$\vec{D} = \varepsilon\vec{E}, \tag{2.129b}$$

$$\vec{J}_m = \sigma^m\vec{H}, \tag{2.129c}$$

$$\vec{J}_e = \sigma^e\vec{E}. \tag{2.129d}$$

As per the earlier remarks, these equations include the possibility of both a fictitious magnetic current, $\vec{J}_m$, as well as a real electric current, $\vec{J}_e$. In addition, for a linear medium the electric and magnetic currents are both related to the magnetic and electric fields through the magnetic conductivity, σ^m, and the electrical conductivity, σ^e. Written in this form, the above equations will be seen later to lead to successful treatment of computer simulation studies of scattering from structures within linear electrical and magnetic media.

The formulation in (2.128) and (2.129) involving both electric and magnetic currents is useful in simulations meant to study the generation and propagation of scattered waves by linear scattering media. In the later discussions of boundary conditions it will be found that by effectively arranging the electrical and magnetic conductivities of an outer boundary layer of the finite scattering region, a perfectly absorbing region which does not reflect radiation incident upon it can be arranged. This makes the finite scattering region look infinite in extent. In this way, the modified form of Maxwell equations used in computer studies are made to allow for the approximation of the scattering in an infinite region of space by that in a finite sub-region of space. Away from the perfectly absorbing boundary layer and within the finite scattering region of the simulation, the magnetic charge and current are zero and the simulation reverts to the standard forms of the Maxwell equations.

The arrangement of the absorbing region at the outer boundary layer of the simulation is known as Perfectly Match Layer (PML) type of Absorbing Boundary Conditions (ABC) [11, 14]. It is useful at the outer edges of the necessarily bounded spatial region of a computer simulation, keeping radiation that arrives at the outer spatial edges of a simulation from being reflected back into the interior of the simulation. Such reflected components would give rise to spurious results in the scattering generated by the simulation. This would limit the effectiveness of the computer results as an approximation of the scattering in an infinite system.

Finite Difference Time Domain Method in a Two-Dimensional Medium

To develop the finite difference time domain method the basic ideas involved in its formulation as a technique of computer simulation are illustrated by some simple applications. Particularly useful illustrations of the method are applications to the study of the electrodynamics of the scattering of an incident wave in a two-dimensional electromagnetic medium. These provide important examples which can be easily extended to the study of higher and lower dimensional systems and to more general electrodynamics problems than those focused on simple scattering.

In this regard, two-dimensional problems are complicated enough that they illustrate many of the basic difficulties to be overcome in developing finite difference time domain programs. However, they are still easier to express mathematically than is the theory for a fully three-dimensional system. This is not an essential difficulty as many problems in the electrodynamics of nanophotonic materials are essentially two-dimensional in nature and benefit to a considerable extent from two-dimensional studies.

In the case of a two-dimensional system, the dielectric and magnetic properties are functions of the coordinates defined in a plane but do not vary in the direction perpendicular to that plane. For the following mathematical discussions, the dielectric and magnetic properties of the scattering media will be taken to vary in the *x*-*y* plane, but, otherwise, they will not depend on the coordinates along the z-axis. In addition, the electromagnetic waves propagating in the system will only travel in the *x*-*y* plane. These conditions then specify the standard formulation for studying electrodynamics in general two-dimensional systems.

For these type of problems, the equations in (2.128) and (2.129) reduce to the set of coupled partial differential equations [11, 14] given by

$$\frac{\partial H_x}{\partial t} = -\frac{1}{\mu}\left(\frac{\partial E_z}{\partial y} + \sigma^m H_x\right), \tag{2.130a}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_z}{\partial x} - \sigma^m H_y\right), \tag{2.130b}$$

$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} - \sigma^m H_z\right), \tag{2.130c}$$

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial y} - \sigma^e E_x\right), \tag{2.130d}$$

$$\frac{\partial E_y}{\partial t} = -\frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial x} + \sigma^e E_y\right), \tag{2.130e}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - \sigma^e E_z\right). \tag{2.130f}$$

In turn, due to the two-dimensional nature of the problem, these sets of equations can be decoupled into systems of equations describing two different types of modes, existing independent of one another in the electromagnetic system.

In this separation of modes, one set of modes are the TM modes. TM modes have their magnetic field polarized perpendicular to the z-axis. Another set of modes, which are separate from the TM modes, are the TE modes. These have their electric field polarized perpendicular to the z-axis.

The decoupled set of equations describing the TM modes are given by

$$\frac{\partial H_x}{\partial t} = -\frac{1}{\mu}\left(\frac{\partial E_z}{\partial y} + \sigma^m H_x\right), \tag{2.131a}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_z}{\partial x} - \sigma^m H_y\right), \tag{2.131b}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - \sigma^e E_z\right), \tag{2.131c}$$

and their solutions provide half of the modes needed to characterize the general behavior of the electrodynamic solutions of the system. From the same decoupling, the remaining equations describing the TE modes are given by

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial y} - \sigma^e E_x\right), \tag{2.132a}$$

$$\frac{\partial E_y}{\partial t} = -\frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial x} + \sigma^e E_y\right), \tag{2.132b}$$

$$\frac{\partial H_z}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} - \sigma^m H_z\right), \tag{2.132c}$$

with modal solutions representing the remaining half of the modes needed to characterize a general electrodynamic solution of the system.

For a general study of the two-dimensional problem both the TM and TE modal equations need to be discretized into the form of algebraic equations which are then separately studied by simulation methods. Here the purpose of the discussions is on obtaining a basic understanding of the ideas of the finite difference time domain techniques. Consequently, in the following a focus will be on how the difference equations for the TM modes are obtained and on a discussion of their solutions. This will serve as a means to develop in the reader a general understanding of the method of finite difference time domain techniques. Subsequently, the TE mode results are obtained in a similar manner to those of the TM, and it is left as an exercise for the reader to work them out.

The set of difference equations for the evolution in space and time of the TM modes are generated from the differential form of the Maxwell equations. This is accomplished by discretizing the space and time derivatives and other general functional forms encountered in the Maxwell differential equations on a space-time lattice composed of a fine mesh of isolated points [11, 14]. The continuum variables of the problem are then replaced by a discrete mesh which must be fine enough in space and time so that the algebraic solutions accurately represent the electrodynamic properties of the system treated. If the mesh is not fine enough the solutions of the simulation will not effectively approximate the behavior of the system.

To begin the discretization process, first focus on the transformation of the differential equation in (2.131c) into algebraic difference equations. To discretize the equation the discrete space-time lattice over which the electric field is defined can be taken to be of the form $(i\Delta x, j\Delta y, n\Delta t)$ for i, j, n integers.

To simplify the notation, in the following these space-time coordinates will be abbreviated by just listing the integers of the lattice sites, i.e., (i, j, n). Consequently, with these conventions the continuum electric field $E_z(x, y, t)$ in (2.131) is transformed into the electric field defined on the discrete lattice and is given as

$$E^n_{z\,i,j} = E_z(i\Delta x, j\Delta y, n\Delta t) \tag{2.133}$$

for the set of discrete space-time coordinates denoted by (i, j, n).

Writing the definition of the derivative in the format of the trapezoidal rule, it follows that the continuum time derivative in (2.131c), in discretized form, becomes

$$\left.\frac{\partial E_z}{\partial t}\right|_{x=i\Delta x, y=j\Delta y, t=\left(n+\frac{1}{2}\right)\Delta t} \approx \frac{E^{n+1}_{z\,i,j} - E^n_{z\,i,j}}{\Delta t}. \tag{2.134a}$$

It should be noted here that, due to the nature of the discretization of the electric field along the time axis and the form of the trapezoidal rule, it is necessarily found that the time derivative obtained is at $t = \left(n + \frac{1}{2}\right)\Delta t$, i.e., at the $n + \frac{1}{2}$ coordinate on the time lattice.

Considering (2.131c) it is found that a consequence of this is that, while the planes of constant time for the discretized $\vec{E}$ field are at $t = n\Delta t$, the planes of constant time for discretized $\vec{H}$ are at $t = \left(n + \frac{1}{2}\right)\Delta t$. As a result, the electric and magnetic fields in the space-time lattice are not defined at the same time points of the lattice. Rather they are seen to alternate in their time updates as the simulation proceeds forward in time. The electric fields at integer time lattice sites are used to find the magnetic fields at half-integer time lattice sites which in turn are used to determine the electric fields at integer time lattice sites.

The need to define integer and half-integer lattice sites on the time lattice in the discretization of (2.131c) and (2.134a) arose because of the nature of the trapezoidal rule. Similar considerations must also be extended in the treatment of the space derivatives in (2.131c). In particular, the electric and magnetic fields must be defined, respectively, on integer and half-integer space lattice points to obtain a successful set of difference equations. The discretization of the fields on the space lattice will now be discussed.

Next consider the space derivatives in (2.131c) for the case in which the electric fields are defined on the integer space lattice coordinates. Defining the electric field in this manner again is found to set the nature of the representation of the magnetic fields on the space lattice, requiring the magnetic field to be defined on half-integer space lattice sites. In this way, adopting the notation of half-integer space lattice points, it follows that

$$\begin{aligned}\left.\frac{\partial H_y}{\partial x}\right|_{x=i\Delta x, y=j\Delta y, t=\left(n+\frac{1}{2}\right)\Delta t} &\approx \frac{H_y\left(\left(i+\frac{1}{2}\right)\Delta x, j\Delta y, \left(n+\frac{1}{2}\right)\Delta t\right) - H_y\left(\left(i-\frac{1}{2}\right)\Delta x, j\Delta y, \left(n+\frac{1}{2}\right)\Delta t\right)}{\Delta x} \\ &= \frac{H_{y_{i+\frac{1}{2},j}}^{n+\frac{1}{2}} - H_{y_{i-\frac{1}{2},j}}^{n+\frac{1}{2}}}{\Delta x}\end{aligned} \tag{2.134b}$$

and similarly

$$\left.\frac{\partial H_x}{\partial y}\right|_{x=i\Delta x, y=j\Delta y, t=\left(n+\frac{1}{2}\right)\Delta t} \approx \frac{H_{x_{i,j+\frac{1}{2}}}^{n+\frac{1}{2}} - H_{x_{i,j-\frac{1}{2}}}^{n+\frac{1}{2}}}{\Delta y}. \tag{2.134c}$$

The forms in (2.134b) and (2.134c) are the derivatives defined at $(x, y) = (i\Delta x, j\Delta y)$. They occur on the right of (2.131c) and are seen to be written in terms of the magnetic fields at the sites $(x, y) = \left(\left(i+\frac{1}{2}\right)\Delta x, j\Delta y\right)$ and $(x, y) = \left(i\Delta x, \left(j+\frac{1}{2}\right)\Delta y\right)$.

Consequently, it is follows from (2.131c) that under these consideration it is most natural to discretize the H_x and H_y as

$$H_{x_{i,j+\frac{1}{2}}}^{n+\frac{1}{2}} = H_x\left(i\Delta x, \left(j+\frac{1}{2}\right)\Delta y, \left(n+\frac{1}{2}\right)\Delta t\right) \tag{2.135a}$$

and

$$H_{y_{i+\frac{1}{2},j}}^{n+\frac{1}{2}} = H_x\left(\left(i+\frac{1}{2}\right)\Delta x, j\Delta y, \left(n+\frac{1}{2}\right)\Delta t\right). \tag{2.135b}$$

In this discretization the magnetic fields on space-time lattices are shifted by half-integers relative to the discretization of the electric fields which are defined to be on the lattice at coordinates that are integer triplets.

The remaining spatial derivatives in (2.131) are handle in a similar manner. Applying the same discretization convention to these equations, the time derivatives of the magnetic fields in (2.131) are represented by

$$\left.\frac{\partial H_x}{\partial t}\right|_{x=i\Delta x, y=\left(j+\frac{1}{2}\right)\Delta y, t=n\Delta t} \approx \frac{H_{x_{i,j+\frac{1}{2}}}^{n+\frac{1}{2}} - H_{x_{i,j+\frac{1}{2}}}^{n-\frac{1}{2}}}{\Delta t}, \tag{2.136a}$$

$$\left.\frac{\partial H_y}{\partial t}\right|_{x=\left(i+\frac{1}{2}\right)\Delta x, y=j\Delta y, t=n\Delta t} \approx \frac{H_{y_{i+\frac{1}{2},j}}^{n+\frac{1}{2}} - H_{y_{i+\frac{1}{2},j}}^{n-\frac{1}{2}}}{\Delta t}. \tag{2.136b}$$

As a further note in the discretization process, it is found that in order to assure the consistency of the various discretization of the fields, the last term on the right in (2.131c) must take the form

$$\frac{E^{n+1}_{z\,i,j}+E^{n}_{z\,i,j}}{2}. \tag{2.137a}$$

In doing this it is assumed that there is an absence of free charges in the system and that the electric field is continuous.

Under similar considerations to those for the introduction of the term in (2.137a) into (2.131), it is found that in the last term on the right of (2.131a) the magnetic field can be discretized into the replacement form

$$\frac{H^{n+\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}+H^{n-\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}}{2}, \tag{2.137b}$$

and in the last term on the right of (2.131b) the magnetic field there can be discretized into the replacement form

$$\frac{H^{n+\frac{1}{2}}_{y_{i+\frac{1}{2},j}}+H^{n-\frac{1}{2}}_{y_{i+\frac{1}{2},j}}}{2}. \tag{2.137c}$$

Added to the earlier considerations in (2.133) through (2.136) for the discretization of the various terms in (2.131), (2.133) through (2.137) form the complete the set of relationship needed for a discretization procedure of the total set of TM equations in (2.131). The application of these relationships is found to provide a discretization of the entire set of differential equations in a consistent and satisfactory manner.

The difference equations for the advancement in time of the magnetic and electric fields of the TM modes on the space-time lattice are obtained by substituting (2.133) through (2.137) into (2.131) and applying a little algebra. In this way the magnetic fields are developed in time by the equations [11, 14]

$$\begin{aligned} H^{n+\frac{1}{2}}_{x_{i,j+\frac{1}{2}}} &= \frac{1}{1+\frac{\Delta t}{2}\frac{\sigma^{m}_{i,j+\frac{1}{2}}}{\mu_{i,j+\frac{1}{2}}}}\left[\left(1-\frac{\Delta t}{2}\frac{\sigma^{m}_{i,j+\frac{1}{2}}}{\mu_{i,j+\frac{1}{2}}}\right)H^{n-\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}+\frac{\Delta t}{\Delta y}\frac{1}{\mu_{i,j+\frac{1}{2}}}\left(E^{n}_{z\,i,j}-E^{n}_{z\,i,j+1}\right)\right] \\ &= A_0(i,j)H^{n-\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}+B_0(i,j)\left(E^{n}_{z\,i,j}-E^{n}_{z\,i,j+1}\right), \end{aligned} \tag{2.138a}$$

and

$$\begin{aligned} H^{n+\frac{1}{2}}_{y_{i+\frac{1}{2},j}} &= \frac{1}{1+\frac{\Delta t}{2}\frac{\sigma^m_{i+\frac{1}{2},j}}{\mu_{i+\frac{1}{2},j}}}\left[\left(1-\frac{\Delta t}{2}\frac{\sigma^m_{i+\frac{1}{2},j}}{\mu_{i+\frac{1}{2},j}}\right)H^{n-\frac{1}{2}}_{y_{i+\frac{1}{2},j}}+\frac{\Delta t}{\Delta x}\frac{1}{\mu_{i+\frac{1}{2},j}}\left(E^n_{z\,i+1,j}-E^n_{z\,i,j}\right)\right] \\ &= A_1(i,j)H^{n-\frac{1}{2}}_{y_{i+\frac{1}{2},j}}+B_1(i,j)\left(E^n_{z\,i+1,j}-E^n_{z\,i,j}\right), \end{aligned} \tag{2.138b}$$

which provide for the advancement of the magnetic fields to time $t = \left(n+\frac{1}{2}\right)\Delta t$ in terms of the magnetic and electric fields in the two earlier time steps, $t = \left(n-\frac{1}{2}\right)\Delta t$ and $t = n\Delta t$, respectively.

Similarly the electric fields are developed in time by the equations [11, 14]

$$\begin{aligned} E^{n+1}_{zi,j} &= \frac{1}{1+\frac{\Delta t}{2}\frac{\sigma^e_{i,j}}{\varepsilon_{i,j}}}\left[\left(1-\frac{\Delta t}{2}\frac{\sigma^e_{i,j}}{\varepsilon_{i,j}}\right)E^n_{zi,j}+\frac{\Delta t}{\Delta x}\frac{1}{\varepsilon_{i,j}}\left(H^{n+\frac{1}{2}}_{y_{i+\frac{1}{2},j}}-H^{n+\frac{1}{2}}_{y_{i+\frac{1}{2},j}}\right)\right. \\ &\qquad \left.+\frac{\Delta t}{\Delta y}\frac{1}{\varepsilon_{i,j}}\left(H^{n+\frac{1}{2}}_{x_{i,j-\frac{1}{2}}}-H^{n+\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}\right)\right] \\ &= A_2(i,j)E^n_{zi,j}+B_2(i,j)\left(H^{n+\frac{1}{2}}_{y_{i+\frac{1}{2},j}}-H^{n+\frac{1}{2}}_{y_{i-\frac{1}{2},j}}\right) \\ &\quad +C_2(i,j)\left(H^{n+\frac{1}{2}}_{x_{i,j-\frac{1}{2}}}-H^{n+\frac{1}{2}}_{x_{i,j+\frac{1}{2}}}\right) \end{aligned} \tag{2.138c}$$

which provide for the advancement of the electric fields to time $t = (n+1)\Delta t$ in terms of the magnetic and electric fields in the two earlier time steps, $t = \left(n+\frac{1}{2}\right)\Delta t$ and $t = n\Delta t$, respectively.

The equations in (2.138) are sequential applied in order to successively advance the electric and magnetic fields in time. Applying (2.138a) and (2.138b) advances the magnetic fields by one time step on the lattice so that (2.138c) can then be applied to advance the electric fields by one time step on the lattice. Following this the magnetic fields are again advanced a time step by (2.138a) and (2.138b), etc. The cycle is repeated over and over again to develop the entire fields in space and time.

In the formulation of the computer algorithm for the simulation, an important point to note about the coefficients $A_l(i,j)$ and $B_l(i,j)$ for $l = 0, 1, 2$ and $C_2(i,j)$ in (2.138) is that they do not depend on the update time. Consequently, they only need to be calculated once in the simulation program and may be stored for further applications throughout the time stepping process. In addition, the fields determined at any given time step are calculated from the fields of the two previous time steps, and, as a result of this, at any time during the simulation process the fields needed to be retained in storage are also limited. This facilitates the implementation of the finite difference time domain method in terms of the requirements on the computer memory needed for a computation.

It should also be noted in computing the field coefficients for (2.138) that care must be exercised in choosing the values of Δt, Δx, and Δy under which the simulation operates. For an accurate time integration, the time step must be much less than the period of the maximum frequency mode to be simulated. This, however, must be balanced with the finite time available to execute the simulation. enough to accurately represent the spatial variations of the modes being modeled by the simulation. In this regard, the finite time available to execute the simulation is again a factor.

A final important consideration in the design of a simulation program is the boundary conditions. This is a very important consideration as the boundary conditions select out solutions with the correct physics of the system being studied. The boundary conditions can also be used to effectively model an approximation of the physics of an infinite system with solutions obtained for a finite system. This facilitates the application of the restricted resources available on a computer.

There are two general considerations in setting the simulation boundary conditions [11, 14]. Initially it is necessary to specify the nature of the source of electromagnetic radiation entering or generated within the spatial region of the simulation. In the later discussions, for an illustration, the initial fields will be taken to be in the form of an incident plane wave. This is a common consideration for the study of scattering problems.

The next point to deal with in formulating the boundary conditions involves the finite spatial extent of the simulation region. Computer resources are limited to treat scattering within a finite region of space, but most scattering problems of interest occur within an infinite space. For the scattering from a localized target in infinite space the scattered waves propagate off to infinity. In the scattering from a localized target contained within a finite region of space, eventually the scattered radiation will reach the outer boundary of the finite spatial region of the simulation where it will be reflected back towards the target. For the finite spatial region of the simulation to represent a good approximation of the scattering of the localized system in infinite space, the radiation that reaches the outer boundary of the simulation must be absorbed rather than reflected back to the target. Some type of effective absorbing boundary conditions at the outer edges of the finite simulation region are required or the simulation will contain many boundary reflections as well as the target scattering it is meant to simulate.

For the simulation of a two-dimensional system the finite spatial region of the computer simulation can be thought of as a rectangular region with the scattering target located near the center of the rectangle (a schematic of the set up in the x-y plane is given in Fig. 2.7). In order to develop an approximation of the scattering in infinite space, along the outer edges of the rectangular region of the simulation is a thin strip that frames the region all along its outer perimeter. The frame can be used to apply both the boundary conditions of the incident plane wave fields and the absorbing boundary conditions used to remove the scattered fields that reach the framed region at the outer edges of the simulation rectangle. The incident waves are introduced in the region of the frame and move towards the center of the simulation region where they scatter from the target. In addition, the absorption feature in the frame region keeps the scattered fields from reflecting back into the target region of the simulation.

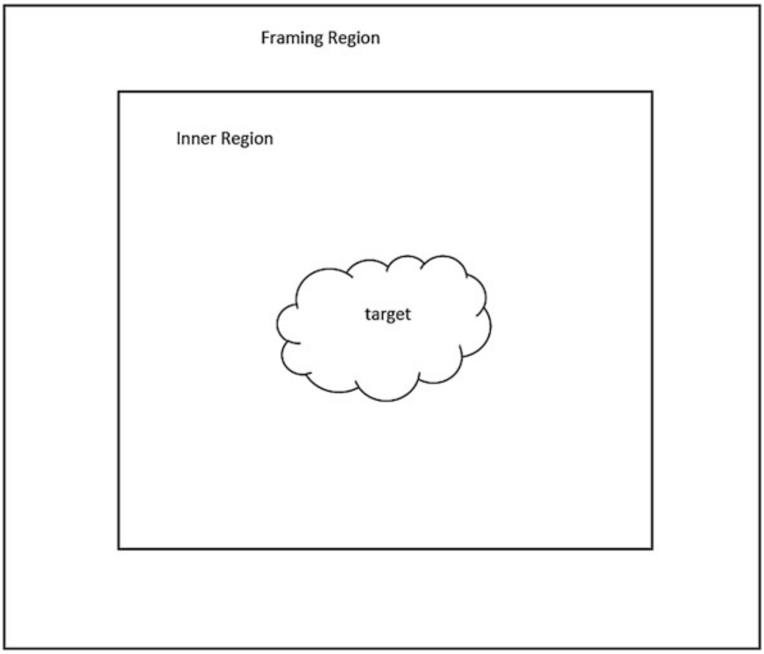


Fig. 2.7 Simulation inner region and framing regions and the outermost boundary of the simulation. The incident plane wave is introduced on the edge between the inner region and the framing region. The framing region contains only a scattered wave which must not be allowed to reflect back into the center of the simulation from the outer edge of the framing region (i.e., the outer most rectangle). Within the inner region the electric field is composed of the incident and scattered fields and the target

First consider how to simulate a plane wave incident on the inner rectangular region that is surrounded by the frame. This is the region containing the target media. The incident plane wave is introduced into the inner rectangle at the interface between the inner rectangle and the region of the frame. This is done by an application of appropriate boundary conditions at this interface. In the inner rectangular region containing the target, the total fields are composed of the incident wave and the scattered wave. In this inner region the total fields evolve in time by (2.138).

Within the surrounding framing region, which is outside the inner region containing the target, only the scattered wave exists. The reason for this is because the incident wave traveling to the target is introduced at the boundary between the framing region and the inner region containing the target. In the framing region, however, the scattered fields are allowed to enter, and in this region they again evolve in time by (2.138).

Boundary conditions are required at the interface between the framing region and the inner rectangle region containing the scattering media in order to match the conditions on the electromagnetic fields just described. To match the two different types of solutions in the framing region and in the inner rectangle it is necessary to make it appear that the incident plane waves in the inner region arise at the interface between the framing region and the inner region containing the target.

As an illustration of how the incident wave boundary conditions are formulated, consider the treatment at the lower edge interface between the frame and inner rectangle in Fig. 2.7. The lower edge is parallel to and below the x-axis in the figure. Its location can be identified by giving its y-coordinate which is of the form $y = j_l \Delta y$ for the integer coordinate j_l.

As discussed earlier, the fields on and above the edge are total fields while those below the edge are only scattered fields. Applying (2.138c) on the edge, taking to account the form of the electric fields in the regions separated by the edge, gives the time evolution equation at the edge [11, 14]

$$\begin{aligned} E^{T^{n+1}}_{z\,i,j_l} &= A_2(i,j_l)E^{T^n}_{z\,i,j_l} + B_2(i,j_l)\left(H^{T^{n+\frac{1}{2}}}_{y_{i+\frac{1}{2},j_l}} - H^{T^{n+\frac{1}{2}}}_{y_{i-\frac{1}{2},j_l}}\right) + C_2(i,j_l)\left(H^{S^{n+\frac{1}{2}}}_{x_{i,j_l-\frac{1}{2}}} - H^{T^{n+\frac{1}{2}}}_{x_{i,j_l+\frac{1}{2}}}\right) \\ &\quad + C_2(i,j_l)H^{I^{n+\frac{1}{2}}}_{x_{i,j_l-\frac{1}{2}}}. \end{aligned} \tag{2.139a}$$

Here superscripts have been introduced on the fields to indicate the total, scattered, and incident fields.

In (2.139) the incident fields have been separated out so that they can be fed into (2.139a) as inputs into the boundary conditions. In addition, (2.139a) contains the scattered and total fields which are generated from the simulation upon introduction of the incident fields. In this form, (2.139a) represents a boundary condition which introduces the incident wave into the system and propagates it into the inner region. At the same time the boundary conditions only allow scattered waves to propagate into the framing region. On the two sides of the lower edge the scattered and total fields in the equation are portrayed as evolving in time from the incident wave.

Additional equations for the magnetic fields at the lower edge of the inner rectangle are also required to input the incident plane wave. For these fields, at the same lower edge of the rectangle, (2.138a) becomes

$$\begin{aligned} H^{S^{n+\frac{1}{2}}}_{x_{i,j_l-\frac{1}{2}}} &= A_0(i,j_l-1)H^{S^{n-\frac{1}{2}}}_{x_{i,j_l-\frac{1}{2}}} + B_0(i,j_l-1)\left(E^{S^n}_{z\,i,j_l-1} - E^{T^n}_{z\,i,j_l}\right) \\ &\quad + B_0(i,j_l-1)E^{I^n}_{z\,i,j_l}. \end{aligned} \tag{2.139b}$$

As with (2.139a) the incident fields are again introduced into the simulation by the last term on the right hand side of the equation. The remaining (2.138b) does not changed its form at the lower edge, however, as it only couples terms along the x-axis, i.e.,

$$H_{y_{i+\frac{1}{2},j_l}}^{T^{n+\frac{1}{2}}} = A_1(i,j_l)H_{y_{i+\frac{1}{2},j_l}}^{T^{n-\frac{1}{2}}} + B_1(i,j_l)\left(E_{z\,i.+1,j_l}^{T^n} - E_{z\,i,j_l}^{T^n}\right). \tag{2.139c}$$

The scattered fields below and the total fields at and above the lower edge are coupled to one another and related to the incident wave source terms located on the edge by (2.139a) and (2.139b). Similar considerations to those in (2.139) are needed along the other three sides of the rectangular inner region.

In addition to the considerations of the boundary conditions on the four edges, some special considerations at the vertices of the inner rectangular region are also required. These all involve basically the same reasoning as applied in (2.139) and will not be treated here. Further details of the considerations of these addition boundary condition and the equations arising from them can be found in [11, 14].

The next set of boundary conditions that need to be addressed are the absorption boundary conditions for the scattered waves. These boundary conditions keep the scattered waves generated at the target from being reflected back at the target upon their encounter with the outer edges of the simulation region. Two basic approached have been developed to handle the removal of the scattered waves at the outer boundary of the simulation.

The first type of approach is based on placing a dissipative medium at the outer edges of the simulation in the framing region. In formulating the dissipative medium it is useful to include both electrical and magnetic dissipation by introducing a medium with an electrical conductivity, σ^e, and a magnetic conductivity, σ^m. This is where the form of the equations involving magnetic charges and magnetic currents enters into consideration. The other dielectric and magnetic properties of the dissipative medium match those of the medium containing the scattering target. Consequently, for the boundary conditions developed along these line, the theory in known as the perfect matching layer (PML) approach.

In the following, as an illustration of the technique, the development of the theory will be made for the TM system of equations in (2.131). The object of the theory is to generate an approximation of a non-reflective medium. To do this, in particular, start by finding the conditions required on σ^e and σ^m so that a reflected wave is not generated by TM waves at normal incidence to a planar interface composed of the dissipative medium.

Remember that only the conductivities of the dissipative medium differ from the medium with which it interfaces. Under these condition on the conductivities, the resulting dissipative medium is taken as the PML for the problem being studied. The conditions will now be worked out for a general planar interface between the two media.

Consider the planar interface to be the y-z plane with the nonconductive medium in the region $x < 0$ and the conductive medium in the region $x \geq 0$. A TM wave solution of (2.131) propagating along the x-axis in the nonconductive medium is then given by

$$E_z^{nc} = E_{z0}^{nc} \mathrm{e}^{i(kx-\omega t)}, \tag{2.140a}$$

$$H_y^{nc} = H_{y0}^{nc} \mathrm{e}^{i(kx-\omega t)}. \tag{2.140b}$$

where

$$\frac{\omega}{k} = \frac{1}{\sqrt{\varepsilon\mu}}, \tag{2.141a}$$

and

$$H_{y0}^{nc} = -\frac{\varepsilon}{\sqrt{\varepsilon\mu}} E_{z0}^{nc}, \tag{2.141b}$$

for the dispersion relation and the relationship between the electric and magnetic field amplitudes, respectively.

Similarly, a TM wave propagating along the x-axis in the conductive medium is given by

$$E_z^{c} = E_{z0}^{c} \mathrm{e}^{i(kx-\omega t)}, \tag{2.142a}$$

and

$$H_y^{c} = H_{y0}^{c} \mathrm{e}^{i(kx-\omega t)}. \tag{2.142b}$$

Substituting (2.142) into (2.131) for the conductive medium then yields

$$k^2 = \varepsilon\mu\left(\omega + i\frac{\sigma^e}{\varepsilon}\right)\left(\omega + i\frac{\sigma^m}{\mu}\right), \tag{2.143a}$$

$$H_{y0}^{c} = -\varepsilon \frac{\omega + i\frac{\sigma^e}{\varepsilon}}{k} E_{z0}^{c}, \tag{2.143b}$$

for the dispersion relation and relation between the electric and magnetic field amplitudes, respectively.

Under the condition that

$$\frac{\sigma^e}{\varepsilon} = \frac{\sigma^m}{\mu}, \tag{2.144}$$

it is found that the solutions at the $x = 0$ interface between the nonconductive media in (2.140) and (2.141) and the conductive media in (2.142) and (2.143) reduce to one another. In particular, under these condition both $E_{z0}^{nc} = E_{z0}^{c}$ and $H_{y0}^{nc} = H_{y0}^{c}$. The conductive and nonconductive media, consequently, are perfectly matched at their interface so that at normal incidence reflected waves are not radiated from the interface.

From (2.143a), however, the wave vector of the wave in the conductive medium is seen to be complex. As a result of this the wave is found to decay in the absorbing medium during it propagation along the x-axis. Consequently, it is effectively removed from the system.

While the above calculation only treats normal incidence and the conditions derived from it are only valid at normal incidence, in simulation work it is taken as an approximation for the absorption condition at general incident angles. For discussions regarding the accuracy of this method and generalizations and perturbations on it that can be made to improve its accuracy, the reader is referred to the literature [14, 15].

Another way of handling the problem of the reflection of the scattered waves from the outer boundary of the simulation region is to surround the simulation region with a layer of medium that only allows waves to propagate away from the inner scattering region. The layer is then a mathematically constructed type of one way medium for wave propagation, with the direction of one way propagation being away from the simulation region containing the target.

The one way boundary layer is chosen to have dielectric and magnetic properties matching those of the medium in which the scattered wave propagates towards the simulation boundary. A consequence of this choice is that no wave is reflected from the interface of the one way medium and the medium in the simulation region containing the target media. Notice that the one way medium is a directional medium and otherwise differs from the PML as it does not employ a dissipative electric and magnetic conductivity.

To develop the mathematics for the operation of the layer of one way medium, the TM system in (2.131) will be discussed. (A similar development for the TE equations is left to the reader to work through.) The equations for the propagation of the TM waves in a general medium can be rewritten in the form of second order partial differential equations [11, 14] given by

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} - \varepsilon\mu\frac{\partial^2}{\partial t^2}\right)\left\{\begin{matrix} H_x \\ H_y \\ E_z \end{matrix}\right\} = 0. \tag{2.145}$$

This equation describes the propagation of the plane wave solutions in the inner scattering region which contains the target material. In particular, it describes the scattered waves as they propagate away from the target media and towards the outer boundary of the simulation region. The idea now is to construct a layer of one way medium at the outer perimeter of the simulation region which will only allow the waves to propagate away from the target media. At the same time the layer of one way medium must generate no reflected wave propagating back towards the target media in the scattering region of the simulation.

To see how this is accomplished we will look at the right hand the edge of the outer boundary of the simulation (the treatment of the top, bottom, and left hand edges of the simulation region can all be worked out by the reader in a similar fashion to that presented here for the right hand edge). The goal is to create an absorbing layer at the right hand boundary of the simulation with the same dielectric and magnetic properties as the medium of the simulation region which contains the target media. After scattering from the target the scattered waves now approach the simulation boundary and the absorbing media placed on it.

Considering the differential Helmholtz operator in (2.145), a formal solution for the operator which gives the x-component of the wave vector of the waves moving in the scattering region is

$$-i\frac{\partial}{\partial x} = \pm\left[\frac{\partial^2}{\partial y^2} - \varepsilon\mu\frac{\partial^2}{\partial t^2}\right]^{\frac{1}{2}}. \tag{2.146}$$

The x-component of the wave vector operator in (2.146) has two types of eigenvalue solutions corresponding to the choice of sign made on the right hand side of (2.146).

In particular, the positive sign gives the solution for a wave moving towards the right hand boundary of the simulation and the solution for the negative sign gives the solution for a wave moving away from the right hand boundary of the simulation. Combining the solutions for both signs reproduces the entire plane wave solution set of (2.145). The choice of sign in (2.146) is the basis for defining the one way medium needed for the design of the outer boundary layer of the scattering medium.

To describe a wave that can move only to the right in the proposed boundary layer at the outer right hand edge of the simulation, the correct choice of operators in (2.146) is given by

$$i\frac{\partial}{\partial x} = -\left[\frac{\partial^2}{\partial y^2} - \frac{1}{\varepsilon\mu}\frac{\partial^2}{\partial t^2}\right]^{\frac{1}{2}}. \tag{2.147}$$

This should be used to determine the solutions in the boundary layer rather than (2.145) or the plus sign version of (2.146). Representing the wave as a solution of this equation only allows for motion of the scattered waves out of the system. Notice that in this treatment the medium in (2.147) has the same dielectric and magnetic properties as the medium described by (2.145) and (2.146). Consequently, there is no reflection at the interface between the layer and the scattering medium.

Placing the boundary layer within the framing region, it is seen that only waves of positive x-component of wave vector reach the one way boundary layer. Waves of negative x-component cannot reach the one way boundary layer as only the scattered wave solutions are present in the framing region. In addition, waves traveling towards the target media cannot originate in the one way layer, which does not support them, or at the interface between the framing medium and the one way layer. These considerations were applied at the right hand side of the simulation, but they can easily be generalized to considerations of the remaining edges and the corners at the outer boundary of the simulation.

A difficulty with the implementation of the proposed method is how to deal with the square root in (2.147). In the development of a program which handles the nonlinearity of the square root in the wave vector operator, a number of numerical problems are encounter. In particular, directly treating the nonlinear form leads to program complexities and inefficiencies which greatly slow the simulation process and requires too many computer resources to accommodate. In this regard, it is in general found that to develop an efficient simulation it is helpful if the operator in (2.147) could be approximated by a linear operator.

To overcome these difficulties a number of approximation methods have been which replace the square root by a power law expansion. In the following a basic approach will be developed as an illustration.

Consider applying the one way boundary layer method to treat radiation at the frequency, ω. For this particular frequency (2.147) becomes [11, 14]

$$i\frac{\partial}{\partial x} = -\left[\varepsilon\mu\omega^2 + \frac{\partial^2}{\partial y^2}\right]^{\frac{1}{2}}. \tag{2.148}$$

Next consider the Taylor series expansion of the square root in (2.148), and retain the first two terms of the series to obtain

$$i\frac{\partial}{\partial x} = -\sqrt{\varepsilon\mu}\omega\left[1 + \frac{1}{2}\frac{1}{\varepsilon\mu\omega^2}\frac{\partial^2}{\partial y^2}\right]. \tag{2.149}$$

The resulting differential form is now a linear form in the first and second order partial differential operations.

Consequently, (2.149) can be rewritten into the form

$$-i\omega\frac{\partial}{\partial x} = \sqrt{\varepsilon\mu}\omega^2 + \frac{1}{2}\frac{1}{\sqrt{\varepsilon\mu}}\frac{\partial^2}{\partial y^2}. \tag{2.150}$$

For the constant frequency problem, the frequency in (2.150) can now be rewritten as a time derivative, and (2.150) becomes

$$\frac{\partial^2}{\partial x\partial t} = -\sqrt{\varepsilon\mu}\frac{\partial^2}{\partial t^2} + \frac{1}{2}\frac{1}{\sqrt{\varepsilon\mu}}\frac{\partial^2}{\partial y^2}. \tag{2.151}$$

The result in (2.151) can now by take as the operator of the one way wave propagation in the directional medium, approximating the wave equation in the directional medium by the form [11, 14]

$$\left(\frac{\partial^2}{\partial x \partial t} - \frac{1}{2}\frac{1}{\sqrt{\varepsilon\mu}}\frac{\partial^2}{\partial y^2} + \sqrt{\varepsilon\mu}\frac{\partial^2}{\partial t^2}\right)\begin{Bmatrix} H_x \\ H_y \\ E_z \end{Bmatrix} = 0. \tag{2.152}$$

The wave equation in (2.152) describes waves with positive x-components of the wave vectors. In addition it has the advantage in its implementation in a computer algorithm that it is a linear equation, ready for application in one way boundary layer of the finite difference time domain method.

For details of the accuracy and implementation of the method based on a directional layer, the reader is referred to the literature [11, 14, 15].

Method of Moments

The method of moments is used to treat the spatial dependence of the constant frequency modes of an electrodynamic system [11, 15–17]. Since the frequencies of the modes are assumed to be already known, the discretization of the continuum limit of the electrodynamic equations only involves a treatment of the spatial coordinates of the field equations. Nevertheless, the ideas and problems in the discretization process are somewhat similar to those involved in the finite difference time domain method. In particular, for the method of moments treatment the same considerations must be extended to the space variables of the system as in that approach.

A difference in the method of moments approach, however, is that the focus of the study is now on an integral equation formulation of the problem. In this formulation the relevant integral equations are obtained directly from the differential forms of the Maxwell equations. This is facilitated in the constant frequency study because the frequency and consequently the time dependence of the solutions are already known.

The method of moments allows for the determination of the excitations of the system by discretizing the integral equations into the form of matrix equations, and the solutions of the matrix equations are subsequently studied using the techniques of linear algebra. As a result, instead of a time integration of a set of difference equations the problem is replaced by a matrix inversion.

As an example of the method of moments consider a two-dimensional scattering problem in which TM waves propagating in free space scatter from perfect conductor targets [11, 16]. The waves are considered to travel in the x-y plane and are polarized with their electric fields along the z-axis. They are incident on perfect conductor structures which are translational invariant along the z-axis.

In the region outside of the perfect conductors, the fields are of frequency, ω, and are obtained as solutions of the set of Helmholtz equations of the forms

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \varepsilon_0\mu_0\omega^2\right)\begin{Bmatrix} H_x \\ H_y \\ E_z \end{Bmatrix} = 0. \tag{2.153}$$

Due to the structure of the scattering problem, the fields in (2.153) can be separated into incident and scattered field components. For example, the electric field is written in the form

$$E_z(x, y, \omega) = E_z^I(x, y, \omega) + E_z^S(x, y, \omega). \tag{2.154}$$

where the superscripts I and S, respectively, refer to the incident and scattered field contributions to the total field.

Considering the case of the electric field, the total electric field is zero insider the perfect conductors. Consequently, at the surface of the perfect conductors the derived surface condition

$$E_z^I(x, y, \omega)_{Surface} = -E_z^S(x, y, \omega)_{Surface} \tag{2.155}$$

is a restriction on the incident and scattered field components. This restriction is a very important condition on the fields at the surface of the conductors, and it is essential in the following development of the integral equation formulation for the method of moments.

The method of moments is developed from (2.153) through (2.155) by replacing the above posed scattering problem involving the perfect conducting system by an equivalent antenna radiation problem. In this replacement, the geometry of the perfect conductors is used to design antennas which radiate fields that are the same as those of the scattered waves in the original scattering problem. The idea is to determine a set of surface currents on the geometry of the scattering surfaces to replicate the fields of the scattered waves in the scattering problem. In this approach the antennas are no longer perfect conductors, but rather they are surfaces supporting surfaces currents.

Before considering how to obtain the necessary surface currents, consider for the moment that the surface currents are known and are of the form, $J_{Surface}(x, y, \omega)$. A discussion will first be given as to how the radiated fields from these current distributions are expressed in terms of the electrodynamics of the radiating system. This will be followed by a treatment of the determination of the surface currents appropriate for generating the scattered fields within the context of the equivalent antenna problem.

The waves radiated by $J_s(x, y)$ can be written in terms of the Green's function of the operator in (2.153). In particular, the conversion of the Helmholtz equations in (2.153) to treat the radiation from the current sources of the replacement antenna problem is made by introducing a source term on the right hand side of (2.153).

The appropriate Green's function for the standard solution of this antenna source problem is given as a solution of the Green's equation

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \varepsilon_0 \mu_0 \omega^2\right] G(x, y; x', y', \omega) = -\delta(x - x')\delta(y - y'). \tag{2.156}$$

Here the solution is made for radiation boundary conditions in infinite space yielding a Green's function which can be written in terms of Hankel functions.

Once the Green's function is determined, the formal solutions for the scattered wave is expressed in terms of the retarded Green's function and surface current as

$$E_z^S(x, y.\omega) = i\mu_0\omega \int\limits_{Surfaces} dx'dy'G(x, y; x', y', \omega)J_{Surface}(x', y', \omega). \tag{2.157}$$

This represents the scattered field in terms of the surface current which will be obtained in the later discussions. It is a valid solution both outside and at the surface of the antenna array.

If the fields in (2.157) are evaluated at the scattering surface, taking account of (2.155), the following useful relationship is obtained

$$\begin{aligned} E_z^S(x, y.\omega)_{Surface} &= -E_z^I(x, y, \omega)_{Surface} \\ &= i\mu_0\omega \int\limits_{Surfaces} dx'dy'G(x, y; x', y', \omega)J_{Surface}(x', y', \omega), \end{aligned} \tag{2.158}$$

Here the first equality in (2.158) is due to the zero of the total electric field at the perfect conducting surface and the fact that the total electric field is composed as a sum of the incident and scattered waves.

The incident wave is then often taken to be of the form of a plane wave so that from the second equation in (2.158)

$$E_z^I(x, y, \omega)_{Surface} = -i\mu_0\omega \int\limits_{Surfaces} dx'dy'G(x, y; x', y', \omega)J_{Surface}(x', y', \omega). \tag{2.159}$$

This equation relates the known incident field to the unknown surface current.

Equation (2.159) is a very important relationship for determining the surface current. The antenna current of the equivalent antenna problem is obtained from (2.159) by inverting the equation to express the surface current in terms of the known electric field of incident wave. Once the surface current is obtained as a solution of (2.159) it can then be used in (2.157) to determine the scattered wave everywhere in space.

The inversion of (2.159) to find the surface current can be handled numerically by converting it into a matrix equation which is then solved for the surface currents by methods of linear algebra. A common approach to the discretization of the integral equations is to cover the conducting surfaces with a finite array of points.

In this formulation, for such an array of points on the surface, each point of the array, $\vec{r}_i = x_i\hat{i} + y_i\hat{j}$, has a weight function, $f_i(x, y)$, associated with it. The weight function is defined to be one at the array point, $\vec{r}_i = x_i\hat{i} + y_i\hat{j}$, and to be non-zero over a small region of space which approaches the nearest neighbor points of $\vec{r}_i$ on the surface. Otherwise, outside the small region about $\vec{r}_i$ the function remains zero. Consequently, the weight functions is characterized in part by

$$f_i(\vec{r}_j) = \delta_{i,j}. \tag{2.160}$$

The object is to choose the covering array of points on the surface so that the surface current can be approximated in terms of these functions by the form

$$J_{Surface}(x, y, \omega) = \sum_i a_i f_i(\vec{r}). \tag{2.161}$$

This requires that the number of points, $\vec{r}_i = x_i \hat{i} + y_i \hat{j}$, contained in the array must be large so as to accurately represent the surface current. In addition, the weight functions, $f_i(x, y)$, must also be such as to adequately represent the surface currents.

Upon substituting the surface current form in (2.161) into (2.159) gives [11, 16]

$$E^I(x, y, \omega)_{Surface} = -i\mu_0\omega \int_{Surfaces} dx'dy' G(x, y; x', y', \omega) \sum_j a_j f_j(x', y') \tag{2.162}$$

as the form of the scattered wave in terms of surface current form. A matrix equation can be generated from (2.162) by multiplying both sides of (2.162) by $f_i(x, y)$ and integrating over x and y.

In this way, a matrix equation is generated having the form

$$b_i = \sum_j M_{i,j} a_j. \tag{2.163}$$

where

$$M_{i,j} = -i\mu_0\omega \int_{Surface} dxdydx'dy' f_i(x, y) G(x_i, y_i; x', y', \omega) f_j(x', y'). \tag{2.164}$$

and

$$b_i = \int_{Surface} dxdy f_i(x, y) E^I(x, y, \omega)_{Surface}. \tag{2.165}$$

The unknown a_j's, needed to represent the surface currents, are then obtained in terms of the known b_i's from a straightforward solution of the matrix equation in (2.163). In terms of the solutions for the a_j's, the surface currents are obtained from (2.161) and the associated scattered fields are determined everywhere from (2.157).

In this method the determination of the solutions for the scattered fields relies on the ability to invert the matrix equation in (2.163). Most problems of interest involve the treatment of systems represented by large matrices. This is a complication, but many simplification have been developed for the study of problems in which the matrices involved can be written in the form of a positive symmetric matrix.

In particular, for these cases solutions are greatly facilitated using the conjugate gradient method [11, 16]. It shall now be briefly shown how the problem in (2.163) can be rewritten in the context of a symmetric matrix. For the conjugate gradient approach to the solution of symmetric matrix systems, however, the reader is referred to the literature [11, 16].

Writing (2.163) in operator form, the two vectors a and b are related to one another through the matrix M by the matrix equation

$$b = Ma. \tag{2.166}$$

The resulting matrix equation in (2.166) can be converted to the form of a symmetric matrix problem by using the matrix transpose. Multiplying on the left of both sides of the equation by the transpose of the matrix M, (2.166) becomes

$$M^T b = M^T Ma \tag{2.167}$$

where $M^T M$ is a positive symmetric matrix.

The equation in (2.167) can then be rewritten as

$$c = Aa \tag{2.168}$$

where $A = M^T M$ is a known matrix and $c = M^T b$ is a known vector. The problem now reduces to the inversion of the symmetric matrix in (2.168) to find the vector a.

Once a solution for a is obtained the currents and scattered fields are generated by (2.161) and (2.157).

Finite Element Method

The last of the three major simulation approaches to be described is the finite element method [11, 18]. This is a very general methodology which usually is associated with problems that are primarily focused on the spatial variables of the electromagnetic systems being treated. These types of problems often include either time-independent problems or problems associated with the determination of the properties of frequency dependent solutions.

As an illustration of the basic ideas involved in the application of the method, consider the use of the finite element method for the determination of the modal solutions of the two-dimensional Helmholtz equation. In this treatment the differential form of the Helmholtz equation is replaced by a variational problem expressed in terms of a functional integral.

In the finite element method the solutions of the differential equations are shown to be given as the solutions which are extrema of the functional integral. This is an advantage as often the search for an extrema of the functional integral is easier to deal with numerically than the direct numerical simulation of the differential equation.

The basic form of the two-dimensional Helmholtz equation is represented by

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\phi+k^2\phi=0. \qquad (2.169)$$

Here, for simplicity, the case in which $\phi(x,y)$ is a scalar field is studied as a function of the propagation parameter k, and the problem is defined over a finite region of space with boundary conditions on the edges of the spatial region.

Problems represented by (2.169) are often encountered in the electrodynamics of dielectric or metallic waveguides and in the study of photonic crystals. In the case that $k=0$, the Helmholtz problem in (2.169), in addition, reduces to the study of the two-dimensional Laplace equation. Again Laplace equation problems are often encountered in electrostatics and in fluid mechanic systems undergoing potential flow.

The differential form in (2.169) can be obtained by finding the scalar field $\phi(x,y)$ which is an extrema of the functional [11, 18]

$$\Gamma(\phi)=\frac{1}{2}\iint dxdy\left[\left(\frac{\partial\phi}{\partial x}\right)^2+\left(\frac{\partial\phi}{\partial y}\right)^2-k^2\phi^2\right], \qquad (2.170)$$

In particular, the solution of $\phi(x,y)$ from (2.169) can be shown to be an extrema of the functional in (2.170), and, conversely, the $\phi(x,y)$ which is an extrema of (2.170) is the solution of (2.169). Here, of course, it is assumed that appropriate boundary conditions are applied in determining these solutions.

As noted earlier, the idea of the finite element method is to numerically find an extrema of the functional in (2.170) as an easier approach to the numerical study of the solutions of (2.169). This is accomplished, applying a discretization process to the functional integral, by converting (2.170) into an algebraic form for which the extrema are determined algebraically.

The discretization of (2.170) is done by breaking the two-dimensional x-y space up into little non-overlapping triangles which nonetheless cover the region over which the two-dimensional problem is defined. In this division of the two-dimensional area into a set of covering triangles, the vertices of each triangle are labeled by i = 1, 2, 3 and the triangles are in turn numbered t = 1, 2, …, N. Here N is the total number of triangles in the covering set, and the set of triangles is very large so that the area cover by any triangle of the covering set is very small.

In the small region covered by the tth triangle the scalar field can be approximated by the linear form

$$\phi_t(x,y)=a_t+b_tx+c_ty. \qquad (2.171)$$

The ability to represent the scalar field within the triangle by this linear form is, in fact, one of the criterion used in the choice of the triangle covering for the system.

Given the general form in (2.171) it is then necessary to determine the coefficients a_t, b_t, c_t as these are unknowns which are used to determine the unknown scalar fields. This is accomplished by evaluating (2.171) at the three vertices, (x_i, y_i), of the tth triangle. The evaluation at the three vertices generates a system of three linear equations for the coefficients a_t, b_t, c_t.

From these three equations, solutions are obtained in the forms

$$a_t = \sum_{i=1}^{3} a_{ti}\phi_t(x_i, y_i), \tag{2.172a}$$

$$b_t = \sum_{i=1}^{3} b_{ti}\phi_t(x_i, y_i), \tag{2.172b}$$

$$c_t = \sum_{i=1}^{3} c_{ti}\phi_t(x_i, y_i). \tag{2.172c}$$

The coefficients are seen to be written in terms of the potentials at the vertices of the tth triangle.

Substituting the solutions in (2.172) into (2.171) it is found that the scalar field at a general point within the triangle is given by [11, 18]

$$\phi_t(x, y) = \sum_{i=1}^{3} (a_{ti} + b_{ti}x + c_{ti}y)\phi_t(x_i, y_i). \tag{2.173}$$

This formal solution gives an approximation for the scalar function within the tth triangle in terms of the potentials at the corners of the triangle. These potentials are unknown and are now to be determined algebraically.

The forms for each of the triangles in the covering, given by the set of $\phi_t(x, y)$ in (2.173), can now be substituted in (2.170). Upon integrating (2.170) over x and y, the functional integral then reduces to an algebraic form involving the unknown $\phi_t(x_i, y_i)$. Following this the set of $\phi_t(x_i, y_i)$ can be used as variables in the process of determining the extrema of (2.170) for the resulting algebraic form and are subsequently determined from this process of extrema determination.

In this conversion of the functional integral into an algebraic form, the space integrals are easily done because they only involve polynomials of the x and y coordinates over the entire two-dimensional region of the solution space. Performing the space integrals then yields directly an algebraic form in the $\phi_t(x_i, y_i)$ which is then searched for extrema. The resulting set of $\phi_t(x_i, y_i)$'s obtained from this search can then be used in (2.173) for each triangle to generated the scalar function $\phi(x, y)$ throughout the entire solution space.

References

1. R.J. Elliott, J.A. Krumhansl, P.L. Leath, The theory and properties of randomly disordered crystals and related physical systems. Rev. Mod. Phys. **46**, 465 (1974)
2. D.J. Bergman, D. Stroud, Physical properties of macroscopically inhomogeneous media, in *Solid State Physics*, vol. 46, ed. by H. Ehrenreich, D. Turnbull (Academic Press, Boston, 1992), pp. 146–269
3. D.J. Bergman, The dielectric properties of composite materials—a problem in classical physics. Phys. Rep. **43**, 378 (1978)
4. M.P. Marder, *Condensed Matter Physics*, 2nd edn. (Wiley, Hoboken, 2010)
5. J.D. Joannopoulos, P.R. Vilenueve, S. Fan, *Photonic Crystals* (Princeton University Press, Princeton, 1995)
6. K. Sakoda, *Optical Properties of Photonic Crystals* (Springer, Berlin, 2001)
7. P.N. Favennec, *Photonic Crystals: Towards Nanoscale Photonic Devices* (Springer, Berlin, 2005)
8. A.R. McGurn, *Survey of Semiconductor Physics,* ed. by W. Boer (Wiley, New York, 2002)
9. Q. Gong, X. Hu, *Photonic Crystals: Principles and Applications* (CRC Press, Hoboken, 2013)
10. A.R. McGurn, Kerr nonlinear layered photonic crystal, in *Proceedings of SPIE 10112* (2017). https://doi.org/10.1117/12.2250040
11. A.R. McGurn, *Nonlinear Optics of Photonic Crystals and Meta-Materials* (Claypool & Morgan, San Rafael, CA, 2015)
12. K. Yee, Numerical solutions of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**(3), 302 (1966)
13. S.D. Gedney, *Introduction to the Finite-Difference Time-Domain Method for Soluting Maxwell's Equations* (Claypoo & Morgan, San Rafael, CA, 2011)
14. A. Taflove, *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House, Boston, 1995)
15. X.-Q. Sheng, W. Song, *Essentials of Computational Electromagnetics* (IEEE Wiley, Singapore, 2012)
16. W.C. Gibson, *The Method of Moments in Electromagnetics* (Chapman and Hall/CRC, Boca Raton, 2008)
17. H.M. El Misilmani, K.Y. Kabalan, M.Y. Abou-Shahine, M. Al-Husseini, A method of moment approach in solving boundary value problems. J. Electromagn. Anal. Appl. **7**, 61 (2015)
18. S. Humphries, *Finite-Element Methods for Electromagnetics* (Field Precission LLC, Albuquerque, 2010, CRC Press, Boca Raton, 1997)

Chapter 3
Photonic Crystals

In this chapter, the basic properties of photonic crystals are reviewed, along with some of the points of technological application of photonic crystals. New developments of photonic crystal technology are constantly being made and new technological applications for the design of photonic crystal based devices are frequently discovered. Due to this the present chapter is not meant to be a comprehensive review but is an introduction for students. Only an outline of the basic points needed to understand the principles of photonic crystal technology is made with some indications of where to pick up in the rather large literature on the subject.

Photonic crystals have important optical applications as they allow for the control of the propagation of light through space [1–5]. They do this using technology that affords for low losses in the transportation or confinement of optical energy. In their basic concept photonic crystals are periodic arrays of dielectric materials and they function in their interactions with light in a similar manner to how semiconductors interact with their conduction and valence band electrons. As with semiconductors the importance of the periodicity of the photonic crystal is in the development of a band structure for the dispersion relation of light [4, 5]. The band structure is a series of stop and pass frequency bands. Light at stop band frequencies cannot propagate within the system only light at pass band frequencies can move through the system. Depending on the applications, the periodicity can be one-, two-, or three- dimensional [1–5].

One-Dimensional Photonic Crystals

One-dimensional photonic crystals are periodic layerings. These can be of used in various mirror and optical coating applications, i.e., laser mirrors and mirrors for the generation of second harmonics being examples. Since the periodicity is the source of the reflection, low loss materials can be used in the design of such devices to offer an enhanced technology for mirror and mirror coating designs. The use of dielectric layerings is also found in nature where an example is seen in the cuticle of

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_3

insects that exhibit a metallic appearance in spite of the absence of metals in their design. This also is a feature observed in the feathers of some species of birds.

Two-Dimensional Photonic Crystals

Two-dimensional photonic crystals have applications in laser designs and in optical circuits [1–5]. Examples are surface emitting lasers, general resonance cavities, sensors, and antennas. For these applications, the system is formed as a periodic patterning written into a dielectric slab. The pattern is introduced in the plane of the surfaces of the slab as a periodicity of the dielectric properties. For such slabs light is confined within the slab by the dielectric mismatch at the slab surfaces. This is a slab waveguide effect. In addition, light with a component of motion parallel to the slab surfaces experiences a band structure in its propagation characteristics arising from the periodicity of the slab dielectric. This is the photonic crystal component in the interaction of the slab with light.

Both two-dimensional photonic crystals, with translational symmetry along one axis of space, and photonic crystal slabs are found to have electromagnetic modes that exhibit polarization dependent properties [1–5]. The electromagnetic solutions of these systems separate into modes with electric fields parallel to the plane of the periodic patterning and modes with magnetic fields parallel to the plane of the periodic patterning. Each of these polarizations exhibits its own distinct series of pass and stop bands in its electromagnetic dispersion relations, and the band structure is sensitive to the nature of the patterning. A patterning that supports regions of large stop bands for one polarization may be found to be significantly less effective in stop band formation for the other polarizations. Factors such as whether a photonic crystal slab is composed as a dielectric waveguide slab containing a periodic array of vacuum holes or a vacuum background containing dielectric disks are important in determining the band gaps found for the two different polarizations of electromagnetic waves.

An early problem in the study of two-dimensional and slab photonic crystals was to find dielectric patterns exhibiting overlapping stop bands for the modes of the two different polarizations [1–5]. This provides for a complete absence of propagation of electromagnetic modes at frequencies in the regions of the common stop bands. The first solutions of this problem were found in systems based on the triangle lattice dielectric pattern [1–5]. Since then, throughout many investigations, the triangle lattice has remained a commonly used pattern in the design of two-dimensional photonic crystal devices. Designs employing this lattice type have been developed to exhibit strong stop band gaps for all polarizations of light transported in the system. In addition to the triangle lattice, other types of patterning have found important applications that will be discussed later.

Three-Dimensional Photonic Crystals

Three dimensional systems have been of interests for circuits and for the design of materials with controlled emission characteristics as well as in antenna design [1–5]. These types of photonic crystals, however, are the most difficult to make for laboratory studies. For such structures to exhibit useful design characteristics it is often needed for them to exhibit periodicity on a mesoscopic scale. This requires the user of a variety of highly developed and precision techniques which operate at the level of nanoscales.

In three dimensional photonic crystals the periodicity is experienced by the light in all three directions of space [1–5]. This can lead to the possibility of a three-dimensional band structure in which a single frequency of light exhibits a stop band for propagation in all directions of space. A variety of such three-dimensional photonic crystals have been studied, exhibiting complete stop bands in all directions space. The earliest system found with a complete stop band was based on a diamond lattice of dielectric spheres and since then lattices with complete stop bands have been found in most three-dimensional Bravais lattice types.

The presence of a three-dimensional stop band offers the opportunity for the complete confinement of light by the photonic crystal. One application of this involves the Purcell effect [1–5]. An atom in an excited state can be suspended in that excited state if the photon emitted during its decay has a frequency within the stop band of the photonic crystal. In this situation the photon cannot propagate away from the atom. Seen in the context of the Fermi golden rule for transitions, there is no density of states to receive the light emission. Similarly, as another Fermi golden rule effect, the photonic crystal can have enhancements of photonic densities of states that will increase the rate of atomic decay. A number of interesting device applications can be achieved based on increases or decreases of the photonic densities of states in the band structure of the three-dimensional photonic crystal [1–5].

Photonic Crystal Fiber Waveguides and Lasers

Additional important systems for consideration are photonic crystal fibers [6, 7]. These are optical fibers with enhanced properties developed through the application of photonic crystal technology in their designs. In particular, the focus of the photonic crystal technology is on the improvement of the ability of fiber optical systems to confine and guide the transmission of light along the fiber. This results is an increase of energy efficiency of these systems from their standards in the absence of photonic crystal based designs.

Photonic crystal technology is introduced into optical fibers in the form of a cladding with designed properties developed through and tailored by the application of photonic crystal concepts [6, 7]. In this regard, claddings as confining mechanisms of light are developed for two different types of fiber based technologies. These include fibers designed to transfer signals and fibers designed to operate as laser amplifiers. Each of these two applications involves different types of photonic crystal cladding designs.

The cladding of an optical fiber is an outer shell on the fiber that exhibits different dielectric properties from the inner fiber core. In its simplest form the cladding aids in confining light transported by the fiber to propagate along the fiber without radiative losses. It does this through the change developed by the cladding in the dielectric properties of the fiber, going from the center of the fiber to its outer radius. The idea of cladding and cladding technology was developed in the design of optical fiber long before the introduction of the ideas of photonic crystals as a design element of cladding technology.

With the introduction of photonic crystal technology into cladding design, however, the combined technologies offer both novel design features of claddings and enhancement of some of the designs based on pre-photonic crystal ideas. Cladding technology as enhanced by photonic crystals forms the basis of two general technological applications that are of importance to telecommunications [6, 7]. In fibers designed for signal transmission the photonic crystal technology is used to create a cladding with effective dielectric properties that reduce radiative losses. This is a traditional cladding technique and the photonic crystal only acts to create a cladding medium with a desired dielectric.

In fiber lasers the photonic crystal stop band technology is employed to introduce a cladding with a stop band structure for light propagating near the perpendicular plane to the fiber axis. The presence of a stop band at the lasing frequency aids in the confinement of the light, keeping it subject to the lasing action within the fiber cavity [6, 7].

In this chapter, a discussion of the band structure and basic properties of a photonic crystal obtained within the context of the plane wave expansion methods and other analytical approaches will be made. Only brief mentions of results from computer simulation techniques [1–7] are given as these do not as readily reveal the physics of photonic crystals. Computer simulation techniques have already been treated in the Chapter on mathematical preliminaries and will not be gone over here. In this regard, some of the analytical techniques of Wannier functions Wannier functions and difference equation techniques will be discussed. These offer an interesting treatment of impurity modes in photonic crystals and give insights into methods for the treatment and understanding of band structure effects. They will be developed in the following as they apply to discussions of impurities, photonic crystal slab lasers, and photonic crystal waveguides [1–7].

3.1 Plane Wave Expansion Methods for the Determination of Photonic Crystal Band Structures

In this section a discussion is given of the plane wave method applied for the determination of the electromagnetic modes and dispersion relation of a photonic crystal [1–5]. The plane wave expansion is a commonly used method for determining the electromagnetic solutions of photonic crystals and has been used in the study of a variety of one-, two-, and three-dimensional photonic crystals [1–7]. In its application the electromagnetic fields and functional form of the periodic dielectric are expressed in terms of a Fourier expansion in plane waves. These expansions are then used to reduce the differential forms of the wave equations, generated from the Maxwell equations, to matrix eigenvalue problems. The matrix eigenvalue problems determine the Fourier coefficients of the plane wave expansion of the electromagnetic modes, relating them to the eigen-frequencies making up the dispersion relation of the system.

Discussions of a Truncated Two-Dimensional Photonic Crystal

As an interesting case that is related to an experimentally studied system the photonic band structure of a truncated two-dimensional periodic dielectric medium is discussed [8, 9]. This consists of an array of parallel axis dielectric cylinders forming a two-dimensional periodic array. The array of cylinders is located between two perfect conducting parallel plates that are perpendicular to the cylinder axes. A schematic of the configuration is provided in Fig. 3.1. As defined in the figure, the system is periodic in the $x_1 - x_2$ plane and the cylinder axes are parallel to the x_3-axis.

In the following discussions, the band structure calculations for the electromagnetic modes propagating between the two perfect conducting plates of the system in Fig. 3.1 are presented for dielectric cylinders arrayed in a square lattice [8, 9]. The square lattice of the array is defined so that the locations of the dielectric cylinder axes are at the lattice sites of the square lattice, and between the parallel plates and surrounding the dielectric cylinders is a vacuum background. The electromagnetic modes are confined in the region between the parallel plates where they exhibit a series of pass and stop bands due to the periodicity of the array of dielectric cylinders.

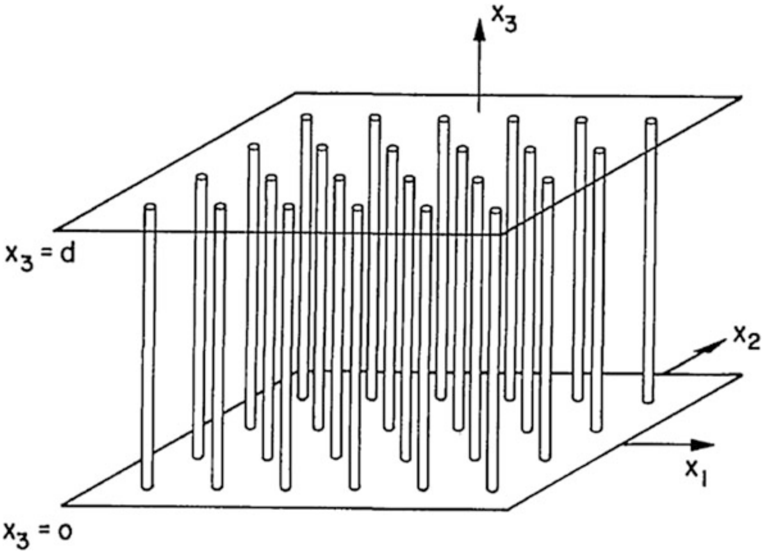


Fig. 3.1 Schematic figure of the photonic crystal square lattice array of dielectric cylinders located between two parallel perfect conducting plates [8]. Reproduced with permission from [8]. Copyright 1993 Optical Society of America

In these considerations the positions of the lattice sites of the square lattice in the $x_1 - x_2$ plane are given by two-dimensional position vectors of the form

$$\vec{r}(l_1, l_2) = l_1 a\hat{x}_1 + l_2 a\hat{x}_2. \tag{3.1}$$

Here a is the nearest neighbor separation between lattice points, l_1 and l_2 are integers, and $\hat{x}_1$ and $\hat{x}_2$ are unit vectors along x_1 and x_2 axes in the plane of the lattice. The dielectric cylinders within the $x_1 - x_2$ plane do not overlap one another so that the radii of the cylinders, R, satisfy the condition $2R < a$. Along the x_3 direction the separation between the perfect conducting plates is d. The perfect conductivity of the plates sets restrictive boundary conditions on the propagation of the solutions along the x_3 axis. As a result of the conditions arising from the geometry the modal solutions of the Maxwell equations are functions of x_1, x_2, and x_3.

Fourier Representation of the System Properties

In the plane wave method the dielectric properties of the system and the solutions obtained from the Maxwell equations are expressed as Fourier series sums of plane waves. As a consequence, it is necessary to consider the plane wave expansion of functions of the general form

$$f(x_1, x_2, x_3) \tag{3.2}$$

subject to the conditions set by the slab geometry of the problem outlined above. In the following, discussions are given of the plane wave expansions of (3.2).

First a treatment is made of the function considered as a general function of x_1 and x_2 in the $x_1 - x_2$ plane of the periodic lattice. This is followed by a specification to periodic functions of x_1 and x_2 in the $x_1 - x_2$ plane having the same periodicity as the lattice of cylinders. After these, considerations are given of the restrictions set on the Fourier series of the function due to the finite extent of the problem along the x_3 axis and of the resulting Fourier expansion including the x_3 variable. These results allow for the re-expression of the modal solution problems of the truncated system from the differential forms of the Maxwell equations to sets of matrix eigenvalue problems which are easily accessible to numerical methods.

As a first point the Fourier expansion of general functions defined in the $x_1 - x_2$ plane is given. In these discussions the x_3 variable is suppressed or ignored. For functions in the $x_1 - x_2$ plane, the notation in (3.1) extends to treat a general position in the $x_1 - x_2$ plane by writing the position vector

$$\vec{r}_{||} = x_1\hat{x}_1 + x_2\hat{x}_2 \tag{3.3}$$

for the point (x_1, x_2).

In terms of the position variable of (3.3) there are two types of functions that need to be treated. These included general functions of $\vec{r}_{||}$ such as those in (3.2), rewritten in the form $f(\vec{r}_{||}, x_3)$, and functions which are periodic functions of $\vec{r}_{||}$

having the periodicity of the lattice. As an example of the last case, the general form of the two-dimensional periodic dielectric function for the cylinder array between the plates is a function

$$\varepsilon\left(\vec{r}_{||}\right) \tag{3.4a}$$

with the periodicity property that

$$\varepsilon\left(\vec{r}_{||}\right) = \varepsilon\left(\vec{r}_{||} + \vec{r}(l_1, l_2)\right). \tag{3.4b}$$

In order to write the Fourier series of general functions of $\vec{r}_{||} = x_1\hat{x}_1 + x_2\hat{x}_2$ found in the treatment of the photonic crystal, the wave vectors of the form

$$\vec{k}_{||} = k_1\hat{x}_1 + k_2\hat{x}_2 \tag{3.5}$$

for the Fourier plane wave expansion have specific restrictions set on their coefficients k_1 and k_2. These arise by applying periodic boundary conditions over an $Na \times Na$ region of the $x_1 - x_2$ plane and taking the limit $N \to \infty$. This is a standard type of boundary condition applied in the treatment of the transport properties of many different physical systems.

Applying periodic boundary conditions on the parallel plate system so that it is periodic over an $Na \times Na$ region of the $x_1 - x_2$ plane reduces the wave vector continuum in (3.5) to the discrete set given by

$$\vec{k}_{||} = \frac{2\pi}{Na} n\hat{x}_1 + \frac{2\pi}{Na} m\hat{x}_2. \tag{3.6}$$

where n and m are integers. The general form of a plane wave which satisfies periodic boundary conditions over the $Na \times Na$ region of the parallel plate photonic crystal in the $x_1 - x_2$ plane is, consequently, written as

$$\exp\left[i\left(\frac{2\pi}{Na} n\hat{x}_1 + \frac{2\pi}{Na} m\hat{x}_2\right) \cdot (x_1\hat{x}_1 + x_2\hat{x}_2)\right]. \tag{3.7}$$

From the plane wave form in (3.7) a basis set is generated for the expansion of the electromagnetic modes with periodic boundary conditions over the $x_1 - x_2$ plane of the photonic crystal. It follows that a general function has the plane wave expansion [8, 9]

$$f\left(\vec{r}_{||}\right) = \sum_{n,m} F_{n,m} \exp\left[i\left(\frac{2\pi}{Na} n\hat{x}_1 + \frac{2\pi}{Na} m\hat{x}_2\right) \cdot (x_1\hat{x}_1 + x_2\hat{x}_2)\right]. \tag{3.8}$$

Here $F_{n,m}$ are a set of expansion coefficients, and it is seen directly from (3.8) that

$$f\left(\vec{r}_{||}\right) = f\left(\vec{r}_{||} + \vec{r}(l_1 N, l_2 N)\right) \tag{3.9}$$

where l_1 and l_2 are integers. Equation (3.9) demonstrates the periodicity over an $Na \times Na$ region of the $x_1 - x_2$ plane.

In the case that the function in the $x_1 - x_2$ plane is periodic in the lattice it follows that

$$f\left(\vec{r}_{||}\right) = f\left(\vec{r}_{||} + \vec{r}(l_1, l_2)\right) \tag{3.10}$$

where l_1 and l_2 are integers. This provides a further restriction on the Fourier series sum over wave vector space so that

$$f\left(\vec{r}_{||}\right) = \sum_{n,m} F^P_{n,m} \exp\left[i\left(\frac{2\pi}{a} n\hat{x}_1 + \frac{2\pi}{a} m\hat{x}_2\right) * (x_1\hat{x}_1 + x_2\hat{x}_2)\right]. \tag{3.11}$$

Here $F^P_{n,m}$ are a set of expansion coefficients for the case in which $f\left(\vec{r}_{||}\right)$ is a periodic function in the $x_1 - x_2$ plane with the periodicity of the lattice. Now it is obtained directly from (3.11) that [8, 9]

$$f\left(\vec{r}_{||}\right) = f\left(\vec{r}_{||} + \vec{r}(l_1, l_2)\right) \tag{3.12}$$

where l_1 and l_2 are integers. Equation (3.12) demonstrates the periodicity of $f\left(\vec{r}_{||}\right)$ over the lattice over the region of the $x_1 - x_2$ plane.

From the discussions given above it follows that general functions in the $x_1 - x_2$ plane are expressed in a plane wave basis with wave vectors of the form

$$\vec{k}_{||} = \frac{2\pi}{Na} n\hat{x}_1 + \frac{2\pi}{Na} m\hat{x}_2. \tag{3.13}$$

for n and m integers. On the other hand, functions that are periodic over the lattice defined in the $x_1 - x_2$ plane are expressed in a plane wave basis with wave vectors given by

$$\vec{G}_{||} = \frac{2\pi}{a} n\hat{x}_1 + \frac{2\pi}{a} m\hat{x}_2 \tag{3.14}$$

for n and m integers.

From these results, it is found that the set of wave vectors in (3.14) is a subset of the set of wave vectors defined in (3.13). In this regard, it is further interesting to note that a general wave vector of the form $\vec{k}_{||} = k_1\hat{x}_1 + k_2\hat{x}_2$ given in (3.13) is related to a wave vector in the first Brillouin zone of wave vector space, defined by

$\vec{k}'_{||} = k'_1\hat{x}_1 + k'_2\hat{x}_2$ with $|k'_1|, |k'_2| < \frac{\pi}{a}$. For these two wave vectors, $\vec{k}$ and $\vec{k}'$, there is always a vector $\vec{G}_{||} = G_1\hat{x}_1 + G_2\hat{x}_2$ of the form given in (3.14) such that

$$\vec{k}_{||} = \vec{k}'_{||} + \vec{G}_{||}. \tag{3.15}$$

Consequently, all wave vectors of the system are ultimately related to those within the first Brillouin zone.

This difference in the nature of the plane wave expansions of the two different types of functions in (3.8) and (3.11) becomes important in representing the solutions of the electromagnetic modes in wave vector space. Specifically, it will be seen later that the two types of wave vectors in (3.13) and (3.14) enter into giving a representation of the unique modal solutions and the dispersion relation of the modal solutions of the photonic crystal slab.

Next a consideration is given of the Fourier expansion including the x_3 variable. The x_3 variable has a complication in that the system is bounded by the two perfect conducting planes that have a separation of d along the x_3 axis. The boundary conditions related to the two perfect conducting plates are that the components of the electric field parallel to the plates are zero. Consequently, these components of the electric field have Fourier series in the x_3 variable which is based on a complete set of sine functions. These are defined between the two plates and are subject to the boundary conditions that the sine functions are zero at the location of the two perfect conducting planes positioned at $x_3 = 0$ and $x_3 = d$, respectively [8, 9].

The standard form of such a Fourier series for a function $g(x_3)$ defined between $x_3 = 0$ and $x_3 = d$ is

$$g(x_3) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{\pi n}{d} x_3\right). \tag{3.16}$$

The form in (3.16) is appropriate to represent functions between the plates which are zero at $x_3 = 0$ and $x_3 = d$.

Combining the result in (3.16) with those in (3.8) and (3.14) gives an expansion for a general function in x_1, x_2, and x_3 which is zero at $x_3 = 0$ and $x_3 = d$. The form of the function must be given by [7, 8]

$$F\left(\vec{r}_{||}, x_3\right) = \sum_{\vec{G}_{||}} \sum_{\vec{k}_{||}} \sum_{n=1}^{\infty} \tilde{F}_n\left(\vec{k}_{||} + \vec{G}_{||}\right) \exp\left[i\left(\vec{k}_{||} + \vec{G}_{||}\right) \cdot \vec{r}_{||}\right] \sin\left(\frac{n\pi}{d} x_3\right). \tag{3.17}$$

Here $\tilde{F}_n\left(\vec{k}_{||} + \vec{G}_{||}\right)$ are the Fourier coefficients, the sum on $\vec{k}_{||}$ is restricted to the first Brillouin zone, and the sum on $\vec{G}_{||}$ is over all $\vec{G}_{||}$ defined in (3.14).

Electrodynamic Equations

The wave equation for the propagation of the electromagnetic modes between the plates is obtained from the Maxwell equations. In the considerations of a system composed of a non-conducting dielectric media and in the absence of currents and net charges, these are given by

$$\nabla \cdot \vec{D} = 0 \tag{3.18a}$$

$$\nabla \cdot \vec{B} = 0 \tag{3.18b}$$

$$\nabla \times \vec{H} = \frac{1}{c}\frac{\partial \vec{D}}{\partial t} \tag{3.18c}$$

$$\nabla \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t} \tag{3.18d}$$

For linear optical media, in (3.18) the magnetic field and magnetic induction are related by $\vec{B} = \mu\vec{H} \approx \vec{H}$ and the electric field and displacement field are related by $\vec{D} = \varepsilon\vec{E}$ through the periodic dielectric function in (3.4a). The periodicity of the photonic crystal is seen to enter only through the dielectric function, and the magnetic induction fields are not significantly affect by the permeability of the media.

By taking the electric and magnetic fields of the photonic crystal modes to be harmonic waves represented by the forms [8, 9]

$$\vec{E}\left(\vec{r}_{||}, x_3|t\right) = \vec{E}\left(\vec{r}_{||}, x_3|\omega\right)\exp(-i\omega t) \tag{3.19a}$$

and

$$\vec{H}\left(\vec{r}_{||}, x_3|t\right) = \vec{H}\left(\vec{r}_{||}, x_3|\omega\right)\exp(-i\omega t), \tag{3.19b}$$

the time derivatives in (3.18c) and (3.18d) are replaced by $\frac{\partial}{\partial t} \to -i\omega$. It then follows, from taking the curl of Faraday's law and eliminating the magnetic induction by using Ampere's law, that the wave equation for a system with the periodic dielectric function given in (3.4a) is [8, 9]

$$\nabla \times \nabla \times \vec{E}\left(\vec{r}_{||}, x_3|\omega\right) = \varepsilon\left(\vec{r}_{||}\right)\frac{\omega^2}{c^2}\vec{E}\left(\vec{r}_{||}, x_3|\omega\right). \tag{3.20}$$

A similar equation for the magnetic induction is obtained by replacing the electric field in (3.20) by the magnetic induction. The focus in the following, however, is given to the electric field equations in (3.20). This is due to the fact that the boundary conditions at the parallel plates are for the electric fields.

Equations for the Modes of the Truncated Photonic Crystal

Equation (3.20) decomposes into a set of three differential equations which when written out are

$$\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}\left[-\frac{\partial^2 E_1}{\partial x_2^2}-\frac{\partial^2 E_1}{\partial x_3^2}+\frac{\partial^2 E_2}{\partial x_1 \partial x_2}+\frac{\partial^2 E_3}{\partial x_1 \partial x_3}\right]=\frac{\omega^2}{c^2} E_1. \tag{3.21a}$$

$$\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}\left[\frac{\partial^2 E_1}{\partial x_2 \partial x_1}-\frac{\partial^2 E_2}{\partial x_3^2}-\frac{\partial^2 E_2}{\partial x_1^2}+\frac{\partial^2 E_3}{\partial x_2 \partial x_3}\right]=\frac{\omega^2}{c^2} E_2. \tag{3.21b}$$

$$\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}\left[\frac{\partial^2 E_1}{\partial x_3 \partial x_1}+\frac{\partial^2 E_2}{\partial x_3 \partial x_2}-\frac{\partial^2 E_3}{\partial x_1^2}-\frac{\partial^2 E_3}{\partial x_2^3}\right]=\frac{\omega^2}{c^2} E_2. \tag{3.21c}$$

These are solved by converting them into matrix equations through the substitution of plane wave forms based on the discussions given earlier. The electric field components E_1 and E_2 are both zero at the upper and lower plates.

From (3.17) the general solutions the E_1 and E_2 components of the electric field are of the form

$$\begin{aligned} E_1\left(\vec{r}_{||}, x_3\right) &= \sum_{\vec{G}_{||}} \sum_{n=1}^{\infty} a_1^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad \times \exp\left[i\left(\vec{k}_{||}+\vec{G}_{||}\right) \cdot \vec{r}_{||}\right] \sin\left(\frac{n\pi}{d} x_3\right) \\ &= e^{i\vec{k}_{||}\cdot\vec{r}_{||}} \sum_{\vec{G}_{||}} \sum_{n=1}^{\infty} a_1^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad \times \exp\left[i\vec{G}_{||} \cdot \vec{r}_{||}\right] \sin\left(\frac{n\pi}{d} x_3\right) \end{aligned} \tag{3.22a}$$

and

$$\begin{aligned} E_2\left(\vec{r}_{||}, x_3\right) &= \sum_{\vec{G}_{||}} \sum_{n=1}^{\infty} a_2^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad \times \exp\left[i\left(\vec{k}_{||}+\vec{G}_{||}\right) \cdot \vec{r}_{||}\right] \sin\left(\frac{n\pi}{d} x_3\right) \\ &= e^{i\vec{k}_{||}\cdot\vec{r}_{||}} \sum_{\vec{G}_{||}} \sum_{n=1}^{\infty} a_2^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad \times \exp\left[i\vec{G}_{||} \cdot \vec{r}_{||}\right] \sin\left(\frac{n\pi}{d} x_3\right). \end{aligned} \tag{3.22b}$$

Due to the $\sin\left(\frac{n\pi}{d}x_3\right)$ factors these fields are zero on the perfect conducting plates located at $x_3 = 0$ and $x_3 = d$. In addition, the plane wave forms $\exp\left[i\left(\vec{k}_{||}+\vec{G}_{||}\right)\cdot\vec{r}_{||}\right]$ reduce the derivatives $\frac{\partial}{\partial x_1} \to i(k_1 + G_1)$ and $\frac{\partial}{\partial x_2} \to i(k_2 + G_2)$ to algebraic operators. This assists in the conversion of the differential wave equation to a matrix eigenvalue problem.

Substituting (3.22) into (3.21a) and (3.21b) it is found that the E_3 component must be of the form

$$\begin{aligned} E_3\left(\vec{r}_{||}, x_3\right) &= i\sum_{\vec{G}_{||}}\sum_{n=0}^{\infty}\gamma_n a_3^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad\times\ \exp\left[i\left(\vec{k}_{||}+\vec{G}_{||}\right)\cdot\vec{r}_{||}\right]\cos\left(\frac{n\pi}{d}x_3\right) \\ &= e^{i\vec{k}_{||}\cdot\vec{r}_{||}} i\sum_{\vec{G}_{||}}\sum_{n=0}^{\infty}\gamma_n a_3^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right) \\ &\quad\times\ \exp\left[i\vec{G}_{||}\cdot\vec{r}_{||}\right]\cos\left(\frac{n\pi}{d}x_3\right). \end{aligned} \tag{3.23a}$$

where

$$\gamma_n = \begin{cases} 1/2 & n = 0 \\ 1 & n \geq 1 \end{cases}. \tag{3.23b}$$

Notice in (3.23) that the x_3 dependence of E_3 introduces a $\cos\left(\frac{n\pi}{d}x_3\right)$ factor into (3.23). It arises from the need to match the x_3 dependence of the $\frac{\partial E_3}{\partial x_3}$ derivatives in (3.22a) and (3.22b) with that in (3.22a) and (3.22b) arising from the terms involving the E_1 and E_2 components. It also is needed in (3.21c) to match the x_3 dependence of the $\frac{\partial E_1}{\partial x_3}$ and $\frac{\partial E_2}{\partial x_3}$ derivatives with that in (3.23) arising from the E_3 terms.

An addition point to notice about the fields in (3.22) and (3.23) is that they are all of the general form [8, 9]

$$e^{i\vec{k}_{||}\cdot\vec{r}_{||}} U_{\vec{k}_{||},n}\left(\vec{r}_{||}, x_3\right) \tag{3.24}$$

involving a plane wave propagating within the $x_1 - x_2$ plane multiplying a function which is periodic in the photonic crystal lattice in the $x_1 - x_2$ plane. The periodic part of (3.24) is seen to depend both on the wave vector, $\vec{k}_{||}$, of the plane wave and the band index n associated with the eigenfrequency of the particular solution under consideration. In the x_3 direction there is no periodicity in the system but the fields satisfies the correct electromagnetic boundary conditions at the perfect conducting plates.

In an earlier chapter it was noted that wave functions in a periodic lattice are of the general form of a plane wave times a function which has the periodicity of the lattice. The results in (3.22) through (3.24) are consistent with this. They account for the periodicity of the system in the $x_1 - x_2$ plane and also take into account the absence of periodicity along the x_3-axis. Now some of the specific details of the solutions are studied for a number of case of interest for applications.

Example Solutions

The $n = 0$ term in the (3.23) expansion for E_3 is a constant independent of the x_3 variable. This term represents the case in which $E_1 = 0$ and $E_2 = 0$, i.e., only the E_3 component of the field is nonzero. For this limit the boundary conditions at the parallel perfect conducting plates are trivially satisfied, and the resulting solution of the system considering this mode is identical with that of a photonic crystal composed of infinite parallel dielectric cylinders. Substituting (3.22) and (3.23) into (3.21) and collecting the $n = 0$ results in the matrix eigenvalue problem [8, 9].

$$\sum_{\vec{G}'_{||}} \hat{\kappa}\left(\vec{G}_{||} - \vec{G}'_{||}\right)\left(\vec{k}_{||} + \vec{G}'_{||}\right)^2 a_3^{(0)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) = \frac{\omega^2}{c^2} a_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.25}$$

where

$$\frac{1}{\varepsilon\left(\vec{r}_{||}\right)} = \sum_{\vec{G}_{||}} \hat{\kappa}\left(\vec{G}_{||}\right) \exp\left[i\vec{G}_{||} \cdot \vec{r}_{||}\right]. \tag{3.26}$$

From (3.25) and (3.26) it is seen that only $a_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right)$ coefficients are involved in the matrix equation. This is due to the fact that none of the $\frac{\partial E_1}{\partial x_3}$ and $\frac{\partial E_2}{\partial x_3}$ derivatives of the $\sin\left(\frac{n\pi}{d} x_3\right)$ factors are constants, independent of x_3. As a consequence only the $n = 0$ terms are coupled together by the differential equations in (3.21) and the fields of the modal solutions depend only on x_1 and x_2.

The matrix in (3.25) is not symmetric but can be made to be symmetric. Defining [8, 9]

$$a_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right) = \frac{b_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right)}{\left|\vec{k}_{||} + \vec{G}_{||}\right|}, \tag{3.27}$$

and substituting into (3.25) gives

$$\sum_{\vec{G}'_{||}} \left|\vec{k}_{||} + \vec{G}_{||}\right| \hat{\kappa}\left(\vec{G}_{||} - \vec{G}'_{||}\right) \left|\vec{k}'_{||} + \vec{G}'_{||}\right| b_3^{(0)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) = \frac{\omega^2}{c^2} b_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right). \tag{3.28}$$

This is an eigenvalue problem involving a symmetric matrix, for the eigenvalues $\left\{\frac{\omega^2}{c^2}\right\}$ and their corresponding eigenvector sets $\left\{b_3^{(0)}\left(\vec{k}_{||} + \vec{G}_{||}\right)\right\}$, that determines the electromagnetic modes of the photonic crystal. The solutions of (3.28) are obtained applying standard numerical methods, and from (3.22), (3.23), and (3.27) yield the wave functions of the electromagnetic modes in position space.

For the numerical treatment of the problem in (3.28) $\hat{\kappa}\left(\vec{G}_{||}\right)$ as defined in (3.26) must be determined for the array of cylinders. From (3.26) it follows that [8, 9]

$$\begin{aligned}\hat{\kappa}\left(\vec{G}_{||}\right) &= \frac{1}{A}\int d\vec{r}_{||}\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}\exp\left[-i\vec{G}_{||}\cdot\vec{r}_{||}\right] \\ &= \frac{1}{A}\int d\vec{r}_{||}\left[\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}-1\right]\exp\left[-i\vec{G}_{||}\cdot\vec{r}_{||}\right]+\delta_{\vec{G}_{||},0}\end{aligned} \tag{3.29}$$

where the integral in (3.29) is over the entire photonic crystal and A is the area of the photonic crystal in the $x_1 - x_2$ plane. The area of the $x_1 - x_2$ plane can be broken into a set of identical smallest area units, a_c, containing a single lattice site of the photonic crystal lattice. (An example of such a division is to break the plane up into little squares centered about each cylinder so that the area A is obtained as a sum of the little squares.) For this division of the plane it then follows that $A = Na_c$, for N the number of lattice sites, relates A and a_c. With this notation (3.29) becomes

$$\hat{\kappa}\left(\vec{G}_{||}\right) = \frac{1}{a_c}\int\limits_{a_c} d\vec{r}_{||}\left[\frac{1}{\varepsilon\left(\vec{r}_{||}\right)}-1\right]\exp\left[-i\vec{G}_{||}\cdot\vec{r}_{||}\right]+\delta_{\vec{G}_{||},0} \tag{3.30}$$

where the integral is over a single a_c centered at the origin of the $x_1 - x_2$ plane, and the periodicity of the integrand has been used.

The integral in (3.30) is evaluated with the help of the identity [8, 9]

$$\exp\left(-iG_{||}r_{||}\cos\theta\right) = \sum_{m=-\infty}^{\infty}(-i)^m J_m\left(G_{||}r_{||}\right)e^{-im\theta} \tag{3.31}$$

where $\vec{G}_{||}\cdot\vec{r}_{||} = G_{||}r_{||}\cos(\theta)$. Substituting into (3.30) and doing the θ integral, it follows that

$$\begin{aligned}\hat{\kappa}\left(\vec{G}_{||}\right) &= \frac{1}{a_c}\int\limits_{a_c} d\vec{r}_{||}\left[\frac{1}{\varepsilon\left(r_{||}\right)}-1\right]\sum_{m=-\infty}^{\infty}(-i)^m J_m\left(G_{||}r_{||}\right)e^{-im\theta}+\delta_{G_{||},0} \\ &= \left[\frac{1}{\varepsilon_c}-1\right]\frac{1}{a_c}\int\limits_0^R r_{||}dr_{||}J_0\left(G_{||}r_{||}\right)+\delta_{G_{||},0}\end{aligned} \tag{3.32}$$

where ε_c is the dielectric constant of the cylinders of the photonic crystal. In evaluating (3.32), use has been made of the fact that cylinders of the photonic crystal have circular cross sections of radii R and that $\varepsilon\left(r_{||}\right) = 1$ outside of the cylinders.

Applying the Bessel function identity[8, 9]

$$\frac{d}{dx}[xJ_1(x)] = xJ_0(x) \tag{3.33}$$

on the far right of (3.32) gives the result

$$\hat{\kappa}\left(\vec{G}_{||} = 0\right) = \frac{1}{\varepsilon_c}f + (1 - f), \tag{3.34a}$$

$$\hat{\kappa}\left(\vec{G}_{||} \neq 0\right) = \left[\frac{1}{\varepsilon_c} - 1\right]f\frac{2J_1\left(G_{||}R\right)}{G_{||}R} \tag{3.34b}$$

Here in (3.34) $f = \pi R^2/a_c$ is the filling fraction of the cylinders in the plane of the photonic crystal plates. The result in (3.34), known as the form factor, is made for a square lattice but it is easily generalized to handle any two-dimensional photonic crystal array. Following a similar method, the form factor of three-dimensional photonic crystal of dielectric sphere is obtained in a closed form involving spherical Bessel functions.

Applying the form factor in (3.34) in (3.28) the photonic band structure of the $n = 0$ modes of the photonic crystal are evaluated. As mentioned earlier, these are not only modes of the truncated photonic crystal but they are modes of an array of non-truncated infinite cylinders on a two-dimensional lattice. The results for this system will now be discussed for a realization of the model that has been of experimental interest.

Experimental System for $n = 0$

In Fig. 3.2 the dispersion relation of the square lattice photonic crystal array is presented. The plot is made for $\frac{\omega a}{2\pi c}$ versus $\vec{k}_{||}$ where a is the nearest neighbor separation of the lattice sites of the square lattice and c is the speed of light. In the plot $n = 0$ so there is no dependence on the plate separation d, and the filling fraction was taken as $f = 0.4488$ in order to match experimental data published in [10]. The dielectric constant of the dielectric cylinders is $\varepsilon_c = 9$.

The inset in the figure shows the smallest square area in $\vec{k}_{||}$ wave vector space which contains all of the unique eigenvalue-eigenvector solutions for the photonic crystal of cylinders in a square lattice array. Outside of this area in wave vector space are only repeats of the solutions contained within the square inset.

At the center of the square in wave vector space is the point labeled $\bar{\Gamma}$. This is the origin of wave vector space, corresponding to $\vec{k}_{||} = (0, 0)$. The other labeled points at $\bar{M}$ and $\bar{X}$ are at the boundaries of the square inset. These points are a standard Group theory representation and are used to reference the dispersion relation plots presented below the inset. The dispersion relation is then presented along the lines $\bar{\Gamma}\bar{X}, \bar{X}\bar{M}$, and $\bar{M}\bar{\Gamma}$ in wave vector space. This is represented as a triangular path within the inset.

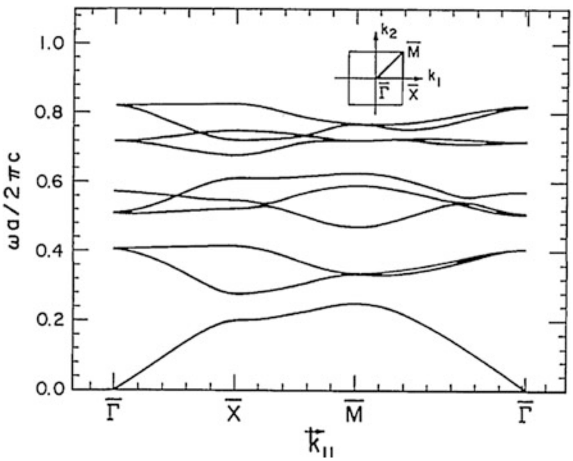


Fig. 3.2 Dispersion relation of the square lattice photonic crystal array [8]. The plot is of $\frac{\omega a}{2\pi c}$ versus $\vec{k}_{||}$ where a is the nearest neighbor separation of the lattice sites of the square lattice and c is the speed of light. In the plot $n = 0$, and the filling fraction taken as $f = 0.4488$ matches experimental data published in [10]. The dielectric constant of the dielectric cylinders is $\varepsilon_c = 9$. Reproduced with permission from [8]. Copyright 1993 Optical Society of America

Results for $n > 0$

For general $n \geq 1$ the matrix eigenvalue problem is obtained from (3.21), (3.22), and (3.34). Upon substituting these various forms, the infinite dimensional problem for the set of eigenvalues $\left\{\frac{\omega^2}{c^2}\right\}$ and their corresponding sets of eigenvectors $\left\{a_1^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right), a_2^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right), a_3^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right)\right\}$ takes the form [8, 9]

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||}-\vec{G}'_{||}\right)\left\{A_{11}a_1^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)+A_{12}a_2^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)+A_{13}a_3^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)\right\} \\ = \frac{\omega^2}{c^2}a_1^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right), \tag{3.35a}$$

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||}-\vec{G}'_{||}\right)\left\{A_{21}a_1^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)+A_{22}a_2^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)+A_{23}a_3^{(n)}\left(\vec{k}_{||}+\vec{G}'_{||}\right)\right\} \\ = \frac{\omega^2}{c^2}a_2^{(n)}\left(\vec{k}_{||}+\vec{G}_{||}\right), \tag{3.35b}$$

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right)\left\{A_{31}a_1^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) + A_{32}a_2^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) + A_{33}a_3^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right)\right\}$$
$$= \frac{\omega^2}{c^2} a_3^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.35c}$$

where

$$A_{11} = \left(k_2 + G'_2\right)^2 + \left(\frac{n\pi}{d}\right)^2, \tag{3.36a}$$

$$A_{12} = A_{21} = -\left(k_1 + G'_1\right)\left(k_2 + G'_2\right), \tag{3.36b}$$

$$A_{13} = A_{31} = \left(k_1 + G'_1\right)\left(\frac{n\pi}{d}\right), \tag{3.36c}$$

$$A_{22} = \left(k_1 + G'_1\right)^2 + \left(\frac{n\pi}{d}\right)^2, \tag{3.36d}$$

$$A_{23} = A_{32} = \left(k_2 + G'_2\right)\left(\frac{n\pi}{d}\right), \tag{3.36e}$$

and

$$A_{33} = \left(\vec{k}_{||} + \vec{G}'_{||}\right)^2. \tag{3.36f}$$

The matrix problem in (3.35) and (3.36) now involves all three components $a_1^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), a_2^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right)$, and $a_3^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right)$ of the electromagnetic fields in (3.22) and (3.23). As a consequence, the modes have non-zero electric fields with components in the $x_1 - x_2$ plane between the two perfect conducting plates. At the plates, however, $E_1 = 0$ and $E_2 = 0$, accounting for the dependence on d of all of the fields.

Large *d* Limit

Some further simplification of (3.35) and (3.36) can be made in the limit that d becomes large. For d approaching infinity these equations reduce to [8, 9]

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right)\left\{\left(k_2 + G'_2\right)^2 a_1^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) - \left(k_1 + G'_1\right)\left(k_2 + G'_2\right) a_2^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right)\right\}$$
$$= \frac{\omega^2}{c^2} a_1^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.37a}$$

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right)\left\{-\left(k_1 + G'_2\right)\left(k_2 + G'_2\right) a_1^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) + \left(k_1 + G'_1\right)^2 a_2^{(n)}\left(\vec{k}_{||} + \vec{G}'_1\right)\right\}$$
$$= \frac{\omega^2}{c^2} a_2^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.37b}$$

and

$$\sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right)\left(\vec{k}_{||} + \vec{G}'_{||}\right)^2 a_3^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) = \frac{\omega^2}{c^2} a_3^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right) \tag{3.37c}$$

For this system it is found that while (3.37a) and (3.37b) only involve the E_1 and E_2 field components, the third E_3 component is determined solely by (3.37c). The problem reduces to two coupled equations and one single equation. Looking further at (3.37) it is noted that these equations for the E_1, E_2, and E_3 field components are independent of n so that all of the solutions are degenerate in n.

The solutions of (3.37) have an interesting relationship to the modal solutions of a two-dimensional photonic crystal composed as an array of parallel axis infinite dielectric cylinders. The complete solutions of the electromagnetic modes of a two-dimensional photonic crystal composed as an array of parallel axis infinite dielectric cylinders separate into distinct polarizations of the fields parallel to the cylinder axes. These are modes with the electric fields parallel to the cylinder axes and modes with magnetic fields parallel to the cylinder axes. It can be shown that the equations in (3.37a) and (3.37b) are the same as those describing the E_1 and E_2 field components for a photonic crystal composed of infinite parallel dielectric cylinders with the magnetic field polarized along the axes of the dielectric cylinders. In addition the matrix equation in (3.37c) is seen to be the same as that in (3.25) for the $n = 0$ mode. Earlier it was also shown that (3.25) and its subsequently modified form in (3.27) and (3.28) described modes which are identical with those of a photonic crystal composed of infinite parallel dielectric cylinders with the electric field polarized along the axes of the cylinders. Consequently, the modal solutions of (3.37a) and (3.37b) along with those from (3.25), (3.27), and (3.28) constitute a complete set of solutions for a two-dimensional photonic crystal composed as an array of parallel axis infinite dielectric cylinders. This makes good sense as in the $d \to \infty$ limit the perfect conducting plate model becomes a two-dimensional photonic crystal composed as an array of parallel axis infinite dielectric cylinders [8, 9].

The $d \to \infty$ solutions and their relation to the two dimensional photonic crystal of infinite cylinders is very interesting [8, 9]. Many of the ideas and discussions of properties of photonic crystal are initially presented in the context of such two-dimensional photonic crystal models. It appears as a basis of discussion in many publications [8, 9].

Slab Solutions

Systems encountered experimentally, however, are generally of the form of photonic crystal slabs. Within the plane of the slab the light propagation is subject to a periodic patterning, but at the slab surfaces the dielectric mismatches between the slab and the regions outside the slab act to confine the light to the body of the slab. In this sense the truncated model studied here is instructive as it illustrates the greater complexity of experimental systems. These contain modes related to the two-dimensional systems of infinite cylinders as well as modes related to the higher order $n \geq 1$ solutions which are now to be discussed with a numerical example.

Continuing with the treatment of the dispersion relations of the truncated solutions of the photonic crystal, the discussions now turn to the $n \geq 1$ solutions of the system considered in Fig. 3.2 in which its $n = 0$ modes were presented. In Fig. 3.3a–c results are shown for the dispersion relations of the $n = 1$ modal solutions. These are plotted presenting the lowest ten bands of the dispersion relations of the truncated photonic crystal, respectively, for the cases $d/a = 0.5, d/a = 1.0$, and $d/a = 5.0$.

In Fig. 3.3 it is seen that the lines of the dispersion relation of the system are flatter and have larger frequency stop bands regions as the ratio d/a is decreased. This is in contrast to the $n = 0$ solutions in Fig. 3.2 which are independent of the ratio d/a. It is a general property of photonic crystal slab systems that the sets of higher $n \geq 1$ modes in the photonic crystal slab display a strong dependence on the slab thickness.

Limit that *d*/*a* Goes to Zero

In the extreme limit that $d/a \to 0$, it follows from (3.35) that [8, 9]

$$\left(\frac{n\pi}{d}\right)^2 \sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right) a_1^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) = \frac{\omega^2}{c^2} a_1^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.38a}$$

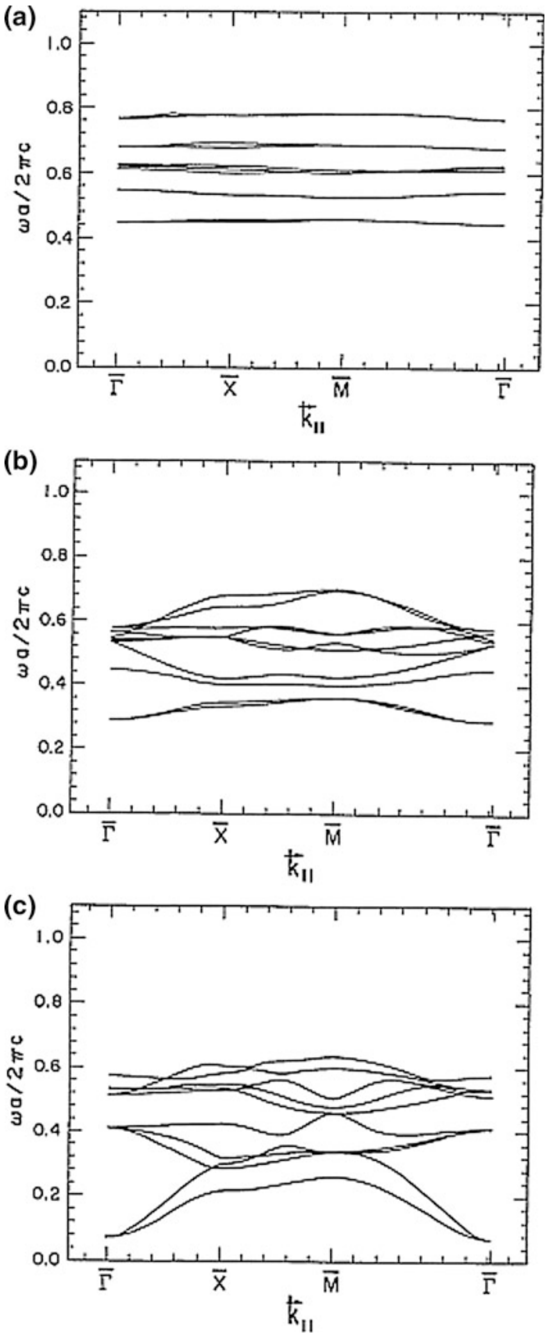
$$\left(\frac{n\pi}{d}\right)^2 \sum_{\vec{G}'_{||}} \kappa\left(\vec{G}_{||} - \vec{G}'_{||}\right) a_2^{(n)}\left(\vec{k}_{||} + \vec{G}'_{||}\right) = \frac{\omega^2}{c^2} a_2^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right), \tag{3.38b}$$

and $a_3^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right) = 0$. The eigenfunctions of the matrix problem in (3.38) are shown to be of the form of plane waves used to describe periodic functions of the lattice. Specifically, it is found that in term of the set of wave vectors $\{\vec{G}_{||}\}$,

$$a_i^{(n)}\left(\vec{k}_{||} + \vec{G}_{||}\right) = \exp\left(-i\vec{G}_{||} \cdot \vec{r}_{||}\right), \tag{3.39}$$

where $i = 1, 2$ generates a solution of the eigenvalue problem. These eigenfunctions are seen not to depend on the wave vectors $\vec{k}_{||}$ so that the eigenvalue problem is highly degenerate.

Fig. 3.3 Dispersion relation of the $n = 1$ modes of the truncated square lattice photonic crystal in Fig. 3.2. The filling fraction is $f = 0.4488$ and the dielectric constant of the dielectric cylinders is $\varepsilon_c = 9$. Results are shown for **a** $\frac{d}{a} = 0.5$, **b** $\frac{d}{a} = 1.0$, and **c** $\frac{d}{a} = 5.0$ [8]. Reproduced with permission from [8]. Copyright 1993 Optical Society of America



Upon substituting (3.39) into (3.38) the corresponding eigenvalues of (3.38) are found to be given by [8, 9]

$$\left(\frac{\omega d}{\pi n c}\right)^2 = \frac{1}{\varepsilon\left(\vec{r}_{||}\right)}, \tag{3.40}$$

which for the present problem has two values. This strong dependence of the eigenvalues on the position dependent dielectric function of the photonic crystal is an indication that the eigenfunctions of the photonic crystal corresponding to these eigenvalues are localized in space either within the dielectric cylinders or to the regions outside of the cylinders. For the specific model that is treated in Fig. 3.3, the two eigenvalues are [8, 9].

$$\frac{\omega a}{2\pi c} = \frac{n}{2\sqrt{\varepsilon_c}}\frac{a}{d}, \frac{n}{2}\frac{a}{d}. \tag{3.41}$$

These eigenvalues are observed numerical in the discussions of the square lattice system.

Solutions for *n* = 2

In Fig. 3.4a–c results for $\frac{\omega a}{2\pi c}$ versus wave vector are presented for $n = 2$ with, respectively, $\frac{d}{a} = 0.5, 1.0$, and 5.0. The figures show the ten lowest frequency bands of the system. As the separation of the plates is increased from $\frac{d}{a} = 0.5$ the band structure changes from an array of flat, well separated pass bands, to a dispersion relation in the $\frac{d}{a} = 5.0$ plot that contains no stop bands. Comparing the $\frac{d}{a} = 1.0$ band structure of the $n = 1$ and $n = 2$ plots in Figs. 3.3b and 3.4b, it is seen that the $n = 2$ system has flatter bands with more well defined regions of stop band.

In general it is found for an increasing n that the dispersion relations for a given plate separation tend to have flatter dispersion relations with larger stop band regions, and the dispersion relation becomes more compressed in frequency.

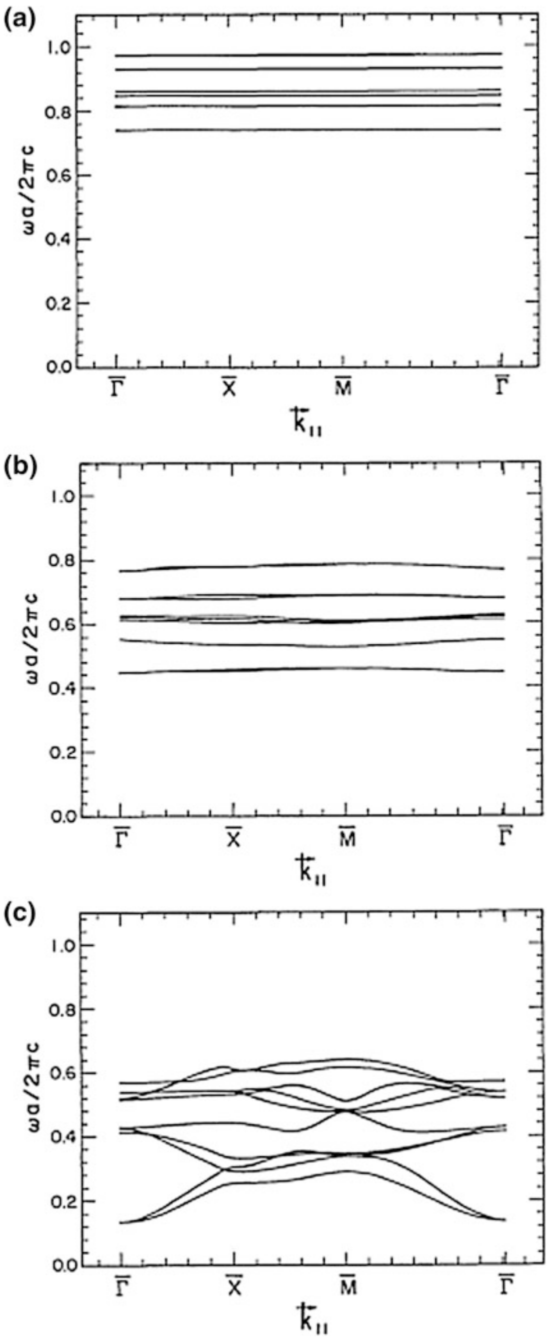
Triangle Lattice Truncated Slab

As a final band structure consideration some results for the triangle lattice are presented. As mentioned earlier, the triangle lattice was the first two-dimensional lattice shown to have a complete band gap for all polarizations of the solutions of the infinite cylinder photonic crystal.

In Fig. 3.5 the solutions of the triangle lattice photonic crystal for the case of the system of infinite cylinders is shown. The case with the electric field polarized parallel to the cylinder axes is presented in Fig. 3.5a. This corresponds to the $n = 0$ modes of the truncated triangle lattice photonic crystal. The case with the magnetic field polarized parallel to the cylinder axes is shown in Fig. 3.5b. As mentioned earlier the results in Fig. 3.5b are obtained as the modes of the $\frac{d}{a} \rightarrow \infty$ limit of the triangle lattice version of the truncated systems discussed earlier.

For these results, the dielectric constant of the cylinders was taken as $\varepsilon_{cylinder} = 1$ in a background medium of dielectric constant $\varepsilon_{background} = 13$, and the filling

Fig. 3.4 Photonic band structure of the truncated square lattice photonic crystal in Figs. 3.2 and 3.3 with $n = 2$ and for: **a** $\frac{d}{a} = 0.5$, **b** $\frac{d}{a} = 1.0$, and **c** $\frac{d}{a} = 5.0$ [8–10]. Reproduced with permission from [8]. Copyright 1993 Optical Society of America



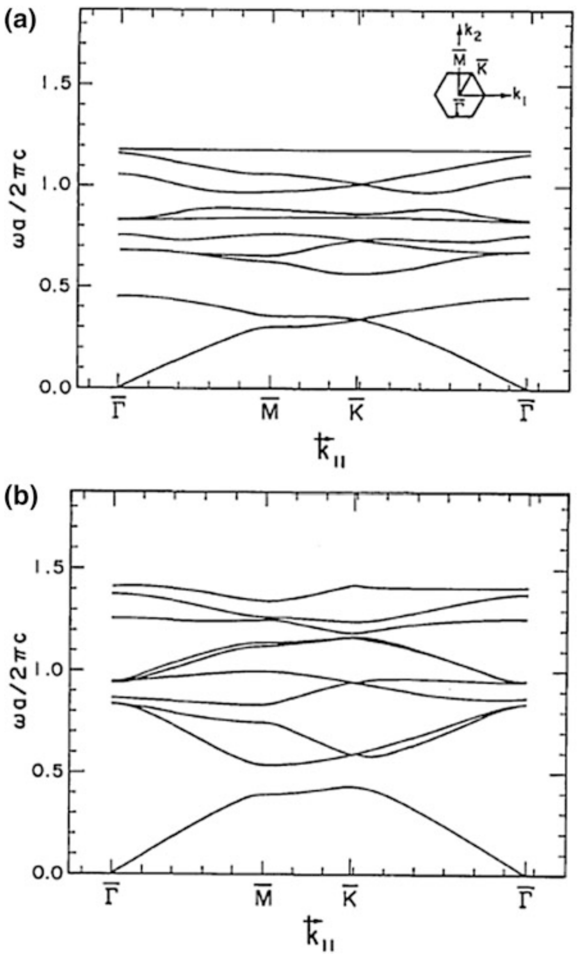


Fig. 3.5 Triangle lattice photonic crystal results: **a** band structure for the electric field parallel to the cylinder axes, corresponding to the $n = 0$ modes of the truncated triangle lattice photonic crystal, and **b** band structure for the magnetic field parallel to the cylinder axes [8]. Reproduced with permission from [8]. Copyright 1993 Optical Society of America

fraction of the cylinders in the lattice was taken as $f = 0.8358$. In both plots the cylinders are vacuum and are surrounded by a dielectric medium. This requires a small adjustment in the formulas given earlier.

As with the discussions of the square lattice system, an inset of the smallest area in wave vector space presenting the complete unique solutions of the system is shown in Fig. 3.5a. For the triangle lattice photonic crystal the smallest area is found to be a hexagon. Points of reference are indicated on the inset and these are used in the plots of the dispersion relations in wave vector space. The dispersion relations shown in Fig. 3.5 are plotted along lines between these inset points in wave vector space.

Notice in the plots of Fig. 3.5 that both polarization share a common stop band near $\frac{\omega a}{2\pi c} \approx 0.5$. In particular, the common stop band for the two modes is found in the region $0.450 < \frac{\omega a}{2\pi c} < 0.536$. For each of the two different modes the common

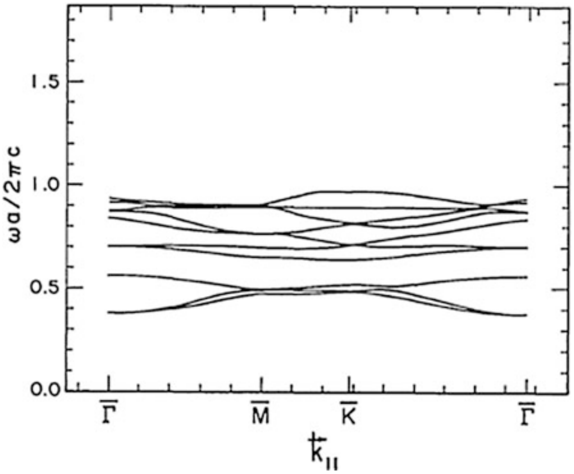


Fig. 3.6 Photonic band structure for the $n = 1$ modes of the truncated triangle lattice photonic crystal of vacuum cylinders with $\varepsilon_{cylinder} = 1$ surrounded by a dielectric background with $\varepsilon_{background} = 13$. The cylinders have a filling fraction $f = 0.8358$. The vacuum cylinders are surrounded by the dielectric background and the plate separation is $\frac{d}{a} = 1.0$ [8]. Reproduced with permission from [8]. Copyright 1993 Optical Society of America

stop band corresponds to their lowest frequency stop band. This is very convenient as the lowest frequency stop bands tend to be the largest stop bands and large stop bands are useful in many types of technological applications.

$n = 1$ Modes of Triangle Lattice Truncated Slab

Continuing with the triangle lattice photonic crystal, results are presented in Fig. 3.6 for the lowest ten $n = 1$ modes of the truncated triangle lattice photonic crystal. These are the first set of modes that exhibit a dependence on $\frac{d}{a}$. The modes presented in Fig. 3.6 are computed for the cylinders of dielectric constants $\varepsilon_{cylinder} = 1$ in a background medium with a dielectric constant $\varepsilon_{background} = 13$ and the filling fraction of the cylinders $f = 0.8358$ used in Fig. 3.5. In addition, the results in Fig. 3.6 are computed for $\frac{d}{a} = 1.0$.

In the limit $\frac{d}{a} \to \infty$ the truncated triangle lattice photonic crystal again is found to have a simple flat band structure. For the system in Fig. 3.6 two limiting values of the frequency are given as [8, 9].

$$\frac{\omega d}{2\pi^2 nc} \to 0.04414 \,\text{and}\, 0.15915. \tag{3.42}$$

3.2 Green's Function Method for Impurity Modes in Photonic Crystals

In this section the methods of Green's functions is developed for the investigation of single site impurities in two-dimensional photonic crystals [9, 12–14]. The Green's function method is a direct extension of the plane wave expansion method for finding the wave functions and the dispersion relation of photonic crystals. It is a flexible formulation which can be modified to handle a variety of problems involving clusters and finite or infinite ordered patterns of impurity features in photonic crystals.

An advantage of the method is in its generation of closed form expressions for the electromagnetic properties of the various impurity systems. For clusters of impurities, the Green's function method allows, through the applications of Group theory techniques, for the development of the modal solutions of the impurities to be classified in terms of the various irreducible represents of the symmetry groups of the impurities. This provides for a qualitative understanding of the modal frequency spectrum associated with the impurities which is not as directly evident from, for example, computer simulation studies.

Single site impurity problems have a long history in theoretical physics where they have been focused upon in the study of many different types of physical systems [13, 14]. Some of the more prominent of these include the treatment of electrons, magnon, and phonon scattering and bound states in impure crystalline materials. Both temperature independent and temperature dependent treatments have been given for these various physical systems.

Photonic crystal impurities are generally much easier to treat than impurities in the general many-body systems mentioned earlier [13, 14]. In photonic crystals the dielectric properties of the photonic crystal and the arrangement of the impurity materials within the photonic crystal is engineered into its basic design [9]. This is not the case with electron, magnon, and phonon systems [13, 14]. In these naturally occurring materials the impurity potentials that the electrons, magnon, and phonon modes encounter within the materials are generally poorly known or are part of the focus for determination in obtaining the solution of the impurity problem. This is not the case in photonic crystals where all the dielectric constants and the geometry of the photonic crystal pattern are known. For these systems the nature of the excitations are the focus of the problem. In addition, the focus of most of the impurity calculations in photonic crystals generally ignores the insignificant temperature dependence of the modal solutions of the system.

Computer simulation methods applied to photonic crystal impurity systems are an alternative to Green's functions approaches [1–5]. These have been described separately elsewhere in this text, and only require a direct implementation of the methodology of the simulations for the study of impurity problems. The methods available generally fall into two classifications as those based on the finite-difference time-domain techniques or the method of moments. Both approached require extensive computer algorithm designing techniques for the accuracy and efficiency of computations, and they have often been taken as the preferred approaches in the study of photonic crystal systems. In the following, the focus will be on the Green's function method, and the reader is referred to the literature for details of the implementation of the simulation approaches to impurities in photonic crystals.

Green's Function Formulation: Two-Dimensional Systems

In the following the Green's function treatment of the impurity problem will be formulated for a two-dimensional photonic crystal. The single site impurity problem will be the focus of the treatment followed by suggestions and literature references for applications to clusters. Some numerical results for the single site impurity in the two-dimensional photonic crystal will be presented and discussed.

This will be followed by discussions of the single impurity problem in a one-dimensional photonic crystal array of dielectric slabs. In this problem a single slab impurity is introduced into a finite array of a one-dimensional photonic crystal. A study will be made of the transmission of the array with a focus on the modification of the transmission properties, from those of the array in the absence of impurities, introduced into the system by the single impurity slab.

For the consideration of the impurity problem a two-dimensional photonic crystal of infinite parallel axes dielectric cylinders in vacuum is treated. The cylinders are ordered on a two-dimensional lattice, and an impurity is created in the photonic crystal by replacing one of the dielectric cylinders of the photonic crystal with an impurity cylinder. The impurity cylinder is of identical geometry to that of the cylinders forming the photonic crystal array, but it is of a different dielectric medium from that of the cylinders of the pure photonic crystal array.

The dielectric function describing the photonic crystal with an impurity has the general form [5, 9]

$$\varepsilon\left(\vec{r}_{||}\right) = \varepsilon_0\left(\vec{r}_{||}\right) + \delta\varepsilon\left(\vec{r}_{||}\right) \tag{3.43}$$

where $\vec{r}_{||} = x_1\hat{x}_1 + x_2\hat{x}_2$ is a vector in the $x_1 - x_2$ plane of the periodicity of the photonic crystal, and the dielectric function of the system is translationally invariant along the x_3-axis. The total dielectric function is composed of a periodic part from the pure photonic crystal and a part from the impurity material that has been introduced into the system.

The periodic dielectric function of the pure photonic crystal in (3.43) is the part.

$$\varepsilon_0\left(\vec{r}_{||}\right) = \varepsilon_0\left(\vec{r}_{||} + \vec{T}_{||}\right). \tag{3.44}$$

In (3.44) $\vec{T}_{||}$ is a translation vector in the plane of the two-dimensional lattice which translates the lattice into itself. Equation (3.44) then displays the periodicity of the photonic crystal in the absence of the impurity. The contribution to the dielectric function in (3.43) represented by

$$\delta\varepsilon\left(\vec{r}_{||}\right) \tag{3.45}$$

is the deviation from $\varepsilon_0\left(\vec{r}_{||}\right)$ due to the presence of impurity media. It is translationally invariant along the x_3-direction but is localized within the $x_1 - x_2$ plane.

As in the treatment of the pure photonic crystal the impurity mode solutions separate into modes with polarization of the electric and magnetic fields parallel to the cylinder axes of the system. For simplicity only modes of the impurity problem that are polarized with their electric fields parallel to the x_3-axis are considered in the following. The modes with magnetic fields parallel to the x_3-axis have a similar but mathematically more involved treatment to that given in the following considerations of the electric field polarization problem.

For modes with the electric field polarized along the x_3-axis, the fields have the general form

$$\vec{E}\left(\vec{r}_{||};t\right) = \left(0,0,E_3\left(\vec{r}_{||}|\omega\right)\right)e^{-i\omega t}. \tag{3.46}$$

The magnetic components of these modes are then shown from the Maxwell equations to be in the $x_1 - x_2$ plane. Working from the forms of Maxwell equations in the absence of currents and net charge densities, the field amplitude $E_3\left(\vec{r}_{||}|\omega\right)$ is obtained as a solution of a Helmholtz equation given by [5, 9]

$$\begin{aligned}&\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\varepsilon\left(\vec{r}_{||}\right)\frac{\omega^2}{c^2}\right)E_3\left(\vec{r}_{||}|\omega\right)\\&=\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\left[\varepsilon_0\left(\vec{r}_{||}\right)+\delta\varepsilon\left(\vec{r}_{||}\right)\right]\frac{\omega^2}{c^2}\right)E_3\left(\vec{r}_{||}|\omega\right)=0.\end{aligned} \tag{3.47}$$

Equation (3.47) can be rewritten into a format resembling an inhomogeneous differential equation with the following form [5, 9]

$$\begin{aligned}&\left[\frac{1}{\varepsilon_0\left(\vec{r}_{||}\right)}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)+\frac{\omega^2}{c^2}\right]E_3\left(\vec{r}_{||}|\omega\right)\\&=-\frac{\omega^2}{c^2}\frac{\delta\varepsilon\left(\vec{r}_{||}\right)}{\varepsilon_0\left(\vec{r}_{||}\right)}E_3\left(\vec{r}_{||}|\omega\right)\end{aligned} \tag{3.48}$$

This is a standard type of differential equation problem encountered in the study of bound states and scattering problems in electrodynamics and quantum mechanics. It has a standard methodology for its solution in terms of a Green's functions approach. This will now be outlined.

In discussing the Green's function approach to the solutions of (3.48). It is necessary to develop the idea of Green's functions of the operator on the left of (3.48). This requires a discussion of the expansions of the fields and the delta function operators in terms of the eigenvalues and eignvectors of the operator on the left of (3.48). This is now presented followed by the formulation of the solution of (3.48).

Solving the Green's Function and Inhomogeneous Equations

In terms of the left side of (3.48), the differential equation eigenvalue problem for the eigenvalues of the two dimensional photonic crystal in the absence of an impurity is [5, 8, 9].

$$\left[\frac{1}{\varepsilon_0\left(\vec{r}_{||}\right)}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)+\frac{\omega^2}{c^2}\right]E_3\left(\vec{r}_{||}|\omega\right)=0. \tag{3.49}$$

This can be rewritten into the standard general form

$$\frac{1}{\varepsilon_0\left(\vec{r}_{||}\right)}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)=-\lambda_n\left(\vec{k}_{||}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right) \tag{3.50}$$

for the set of eigenvalues $\left\{\lambda_n\left(\vec{k}_{||}\right)\right\}$ giving the modal eigenfrequencies $\frac{\omega^2}{c^2}$ and the set of eigenvectors $\left\{\psi_{n,\vec{k}_{||}}\left(\vec{r}_{||}\right)\right\}$ for the corresponding modal wave functions for $E_3\left(\vec{r}_{||}|\omega\right)$.

The problem in (3.49) was treated earlier in (3.25), (3.27), and (3.28). Specifically, it was shown there that the $n=0$ modes of the truncated two-dimensional lattice problem where identical to the modes of the non-truncated two-dimensional photonic crystal composed of infinite parallel axes dielectric cylinders. The matrix eigenvalue problem obtained from (3.49) and (3.50) is given by [5, 8, 9]

$$\sum_{\vec{G}'_{||}}M\left(\vec{k}_{||}+\vec{G}_{||},\vec{k}_{||}+\vec{G}'_{||}\right)C_n\left(\vec{k}_{||}+\vec{G}'_{||}\right)=\lambda_n\left(\vec{k}_{||}\right)C_n\left(\vec{k}_{||}+\vec{G}_{||}\right) \tag{3.51}$$

where

$$M\left(\vec{k}_{||}+\vec{G}_{||},\vec{k}_{||}+\vec{G}'_{||}\right)=\left|\vec{k}_{||}+\vec{G}_{||}\right|\hat{\kappa}\left(\vec{G}_{||}-\vec{G}'_{||}\right)\left|\vec{k}_{||}+\vec{G}'_{||}\right|. \tag{3.52}$$

(Note that the problem in (3.51) an (3.52) is essentially the same as that considered in (3.27) and (3.28).) In terms of the solutions of (3.51) and (3.52) it then follows that the eigenvectors of the differential equations in (3.50) are [5, 8, 9].

$$\psi_{n,\vec{k}_{||}}\left(\vec{r}_{||}\right)=\frac{1}{2\pi}\sum_{\vec{G}_{||}}\frac{C_n\left(\vec{k}_{||}+\vec{G}_{||}\right)}{\left|\vec{k}_{||}+\vec{G}_{||}\right|}e^{i\left(\vec{k}_{||}+\vec{G}_{||}\right)\cdot\vec{r}_{||}}. \tag{3.53}$$

The numerical solution of (3.51) and (3.52) generate $\left\{\lambda_n\left(\vec{k}_{||}\right)\right\}$ and the corresponding set $\left\{C_n\left(\vec{k}_{||}+\vec{G}_{||}\right)\right\}$ of Fourier coefficients for their respective modal wave functions. From these the spatial wave functions are generate using (3.53). Once this has been accomplished and the orthogonality relations of the spatial modes determined, general wave functions of the system and expressions for the spatial delta function can be written in terms of the modal solutions and their eigenvalues.

To obtain the solution of the single impurity problem in the photonic crystal it is necessary to have available a complete set of orthonormal modes of the photonic crystal. These are used to expand and investigate the general electromagnetic solutions of the photonic crystal and to obtain an expression for the Green's functions of the operator on the left of (3.48). In particular, the Green's function of the problem in (3.48), $G\left(\vec{r}_{||};\vec{r}'_{||}|\omega\right)$, is defined as the solution of [5, 8, 9]

$$\left[\frac{1}{\varepsilon_0\left(\vec{r}_{||}\right)}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)+\frac{\omega^2}{c^2}\right]G\left(\vec{r}_{||};\vec{r}'_{||}|\omega\right)=-\frac{1}{\varepsilon_0\left(\vec{r}_{||}\right)}\delta\left(\vec{r}_{||}-\vec{r}'_{||}\right) \tag{3.54}$$

and is expressed in terms of the complete orthonormal solutions of (3.50).

The orthogonality properties of the eigenfunction solutions of the differential equations in (3.50) can be obtained, starting with a consideration of the orthogonality properties of the solutions of the matrix eignenvalue problem in (3.51) and (3.52). From (3.51) and (3.52) it is found that the sets of coefficients $\left\{C\left(\vec{k}_{||}+\vec{G}_{||}\right)\right\}$ that are solutions for the matrix eigenvectors in (3.51) can be chosen to be real and normalized so that they satisfy the two conditions [5, 8, 9]

$$\sum_{\vec{G}_{||}}C_n\left(\vec{k}_{||}+\vec{G}_{||}\right)C_{n'}\left(\vec{k}_{||}+\vec{G}_{||}\right)=\delta_{n,n'} \tag{3.55}$$

and

$$\sum_{n}C_n\left(\vec{k}_{||}+\vec{G}_{||}\right)C_n\left(\vec{k}_{||}+\vec{G}'_{||}\right)=\delta_{\vec{G}_{||},\vec{G}'_{||}}. \tag{3.56}$$

These two relations are expressions defining the properties of the modal solutions for a fixed $\vec{k}_{||}$ in the first Brillouin zone. In particular, for fixed $\vec{k}_{||}$ there are multiple eigenvector solutions, each corresponding to one of the set of eigenvalues $\left\{\lambda_n\left(\vec{k}_{||}\right)\right\}$, such that solutions for different $\{n\}$ are orthonormal to one another. This follows from (3.55). In addition, the conditions obtained in (3.56) are a statement that the set of solutions for $\{n\}$ are complete and through linear combinations accurately represent general solutions of $\vec{k}_{||}$.

From (3.53) it is seen that the orthogonality relations of the set $\left\{\psi_{n\vec{k}_{||}}\right\}$, for a given $\vec{k}_{||}$ in the first Brillouin zone, are intimately connect to those of the $\left\{C\left(\vec{k}_{||}+\vec{G}_{||}\right)\right\}$ given in (3.55) and (3.56). To obtain these properties consider (3.50) rewritten in the form [5, 8, 9]

$$\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)=-\lambda_n\left(\vec{k}_{||}\right)\varepsilon_0\left(\vec{r}_{||}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right). \tag{3.57}$$

From (3.57) it then follows that [5, 8, 9]

$$\begin{aligned}&\int d^2r_{||}\psi^*_{m\vec{q}_{||}}\left(\vec{r}_{||}\right)\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)\\&\quad=-\lambda_n\left(\vec{k}_{||}\right)\int d^2r_{||}\varepsilon_0\left(\vec{r}_{||}\right)\psi^*_{m\vec{q}_{||}}\left(\vec{r}_{||}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right).\end{aligned} \tag{3.58}$$

where the integration is over the entire x_1-x_2 plane of the photonic crystal. Both $\vec{k}_{||}$ and $\vec{q}_{||}$ in (3.58) are fixed to be in the first Brillouin zone. This contains the complete array of unique solutions of the eigenfunctions of (3.53) and (3.57). The integral on the right side of (3.58) is of the classic form of an orthogonality, inner product, relationship. It only remains to evaluate the left side of the equation to completely determine the nature of the orthogonality.

Applying (3.53) in the left side of (3.58) and using the orthogonality relation in (3.55) it is seen that $\left\{\psi_{n\vec{k}_{||}}\right\}$ and $\left\{\psi_{m\vec{q}_{||}}\right\}$ satisfy the following relationships

$$\int d^2r_{||}\varepsilon_0\left(\vec{r}_{||}\right)\psi^*_{m\vec{q}_{||}}\left(\vec{r}_{||}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)=\frac{1}{\lambda_n\left(\vec{k}_{||}\right)}\delta\left(\vec{q}_{||}-\vec{k}_{||}\right)\delta_{m,n}, \tag{3.59}$$

and

$$\sum_n\int\limits_{BZ}d^2k_{||}\lambda_n\left(\vec{k}_{||}\right)\psi^*_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)\psi_{n\vec{k}_{||}}\left(\vec{r}\,'_{||}\right)=\frac{1}{\varepsilon_0\left(\vec{x}_{||}\right)}\delta\left(\vec{r}_{||}-\vec{r}\,'_{||}\right) \tag{3.60}$$

The first relationship in (3.59) describes the nature of the orthogonality between spatial modes that are states of different $\vec{k}_{||}$ and $\vec{q}_{||}$ and different n and m. The second relationship in (3.60) describes the completeness properties of modes of $\vec{k}_{||}$ and n as they are used to represent spatially dependent solutions of the photonic crystal system. The set $\left\{\psi_{n\vec{k}_{||}}\right\}$ is seen to give an accurate representation of general spatial solutions for the electromagnetic fields of the system, considering all wave vectors and dispersive bands of the modal solutions.

As a consequence of (3.59) and (3.60), the general form of the photonic crystal electric field polarized parallel to the x_3-axis is expressed in term of the set $\left\{\psi_{n\vec{k}_{||}}\right\}$ by [5, 8, 9]

$$E_3(\vec{r}_{||}|\omega) = \sum_n \int\limits_{BZ} \frac{d^2k_{||}}{(2\pi)^2} a_{n\vec{k}_{||}} \psi_{n\vec{k}_{||}}(\vec{r}_{||}). \tag{3.61}$$

Here the integral involves wave vectors for the complete set of modal solutions in the first Brillouin zone, the sum runs over the various dispersive bands of the photonic crystal band structure, and $\left\{a_{n\vec{k}_{||}}\right\}$ are a set of expansion coefficients expressing the linear combination of the modal solutions. This gives a general form of the solutions for the fields in the photonic crystal.

In the following discussions the general expression in (3.61) will be used to obtain a solution of the impurity problem defined in (3.48). For these considerations, the focus of the impurity problem is now to determine the $\left\{a_{n\vec{k}_{||}}\right\}$ so as to generate the solution of the inhomogeneous differential equation in (3.48). This is possible due to the completeness of the set of modal solutions $\left\{\psi_{n\vec{k}_{||}}\right\}$ in their representation of general spatial functions of the system.

Substituting the form in (3.61) into (3.48) and applying (3.57) it is obtained that [8, 9].

$$\sum_{n''} \int\limits_{BZ} \frac{d^2k'_{||}}{(2\pi)^2} \left\{-\lambda_{n'}\left(\vec{k}'_{||}\right) + \frac{\omega^2}{c^2}\right\} a_{n'\vec{k}'_{||}} \psi_{n'\vec{k}'_{||}}(\vec{r}_{||}) = -\frac{\omega^2}{c^2} \frac{\delta\varepsilon(\vec{r}_{||})}{\varepsilon_0(\vec{r}_{||})} E_3(\vec{r}_{||}|\omega). \tag{3.62}$$

Multiplying (3.62) by $\varepsilon(\vec{r}_{||})\psi^*_{n\vec{k}_{||}}(\vec{r}_{||})$ and integrating over $\vec{r}_{||}$, it follows from the orthogonality relation in (3.59) that the solutions for $\left\{a_{n\vec{k}_{||}}\right\}$ for the electromagnetic modal bound state on the impurity are obtained as

$$a_{n\vec{k}_{||}} = (2\pi)^2 \frac{\omega^2}{c^2} \frac{\lambda_n\left(\vec{k}_{||}\right)}{\lambda_n\left(\vec{k}_{||}\right) - \frac{\omega^2}{c^2}} \int d^2r_{||} \psi^*_{n\vec{k}_{||}} \delta\varepsilon(\vec{r}_{||}) E_3(\vec{r}_{||}|\omega). \tag{3.63}$$

Upon applying these results for $\left\{a_{n\vec{k}_{||}}\right\}$ in (3.61) a homogeneous integral equation is generated for the single site impurity problem. In this manner, the bound state solutions are given as the solutions of

$$E_3(\vec{r}_{||}|\omega) = \frac{\omega^2}{c^2} \int d^2r'_{||} G\left(\vec{r}_{||}, \vec{r}'_{||}|\omega\right) \delta\varepsilon\left(\vec{r}'_{||}\right) E_3\left(\vec{r}'_{||}|\omega\right) \tag{3.64}$$

where

$$G\left(\vec{r}_{||};\vec{r}'_{||}|\omega\right)=\sum_{n}\int_{BZ}d^{2}k_{||}\frac{\psi_{n\vec{k}_{||}}\left(\vec{r}_{||}\right)\lambda_{n}\left(\vec{k}_{||}\right)\psi^{*}_{n\vec{k}_{||}}\left(\vec{r}'_{||}\right)}{\lambda_{n}\left(\vec{k}_{||}\right)-\frac{\omega^{2}}{c^{2}}}. \tag{3.65}$$

It is seen upon substitution that (3.65) is the solution of the Green's function problem in (3.54) and further that applying the Green's function to the inhomogeneous problem in (3.48) reproduces the result in (3.64) and (3.65).

Bound State Modes

The bound state solutions of (3.64) and (3.65) are the impurity modes of the single site problem. These have frequencies that are outside of the pass band of the pure photonic crystals and have wave functions that are localized about the impurity materials introduced into the photonic crystal. There are a number of methods for obtaining the solutions of (3.64) and a number of simplifications that can be made in the treatment of the integral equations. These are essentially based on the nature of the photonic crystal system which offers some fundamental simplifications in its treatment that are not present in the treatment of electronic and vibrational impurities in crystalline, chemical, media. To conclude the single impurity discussions, these are now addressed.

If a function $f\left(\vec{r}_{||}|\omega\right)$ is defined such that [8, 9]

$$f\left(\vec{r}_{||}|\omega\right)=E_{3}\left(\vec{r}_{||}|\omega\right)\quad\text{in the region of non-zero }\delta\varepsilon\left(\vec{r}_{||}\right) \tag{3.66a}$$

$$=0\quad\text{outside the region of non-zero }\delta\varepsilon\left(\vec{r}_{||}\right), \tag{3.66b}$$

then it follows from (3.64) that

$$f\left(\vec{r}_{||}|\omega\right)=\frac{\omega^{2}}{c^{2}}\int d^{2}r'_{||}G\left(\vec{r}_{||},\vec{r}'_{||}|\omega\right)\delta\varepsilon\left(\vec{r}'_{||}\right)f\left(\vec{r}'_{||}|\omega\right). \tag{3.67}$$

Equation (3.67) is seen to relate the electric fields inside the region of non-zero $\delta\varepsilon\left(\vec{r}_{||}\right)$ back into themselves. It represents a well defined Fredholm integral equation of the first kind which is solved for the fields, $f\left(\vec{r}_{||}|\omega\right)$.

Once the $f\left(\vec{r}_{||}|\omega\right)$ fields are obtained, the general $E_{3}\left(\vec{r}_{||}|\omega\right)$ fields throughout all of space are obtained in terms of the $f\left(\vec{r}_{||}|\omega\right)$ solutions. In this way it follows from (3.64) and (3.66) that for all $\vec{r}_{||}$ [8, 9]

$$E_{3}\left(\vec{r}_{||}|\omega\right)=\frac{\omega^{2}}{c^{2}}\int d^{2}r'_{||}G\left(\vec{r}_{||},\vec{r}'_{||}|\omega\right)\delta\varepsilon\left(\vec{r}'_{||}\right)f\left(\vec{r}'_{||}|\omega\right). \tag{3.68}$$

The difference between (3.67) and (3.68) is seen in the need in (3.68) to evaluate the Green's function in (3.65) for all $\vec{r}_{||}$ rather than to the restricted region in which $\delta\varepsilon(\vec{r}_{||})$ is non-zero.

Numerical Evaluation of Impurity Equation
The numerical solutions of (3.67) can be obtained using two different approaches. In the first approach the integral equation is rewritten in terms of a matrix problem and the solvability conditions of the matrix equation are investigated. Within an application of Gauss-Legendre quadrature in two-dimensions the integral equation in (3.67) takes the matrix form

$$f\left(\vec{r}_{||}(i,j)|\omega\right) - g\frac{\omega^2}{c^2}\sum_{n,m} w_n w_m G\left(\vec{r}_{||}(i,j), \vec{r}_{||}(n,m)|\omega\right)\delta\varepsilon\left(\vec{r}_{||}(n,m)_{||}\right) f\left(\vec{r}_{||}(n,m)|\omega\right) \\ = 0. \tag{3.69}$$

Here $\{w_n\}$ are the weights for the Gauss-Legendre quadrature, $\vec{r}_{||}(n,m)$ are the points in the $x_1 - x_2$ plane that are used in forming the sums of the Gauss-Legendre quadrature in two-dimensions, and g is a constant related to the Gauss-Legendre quadrature through the measure of the region of integration.

Equation (3.69) is a linear homogeneous matrix equation for the field variables $\left\{f\left(\vec{r}_{||}(i,j)|\omega\right)\right\}$. Its solvability condition is that the determinate of the matrix on the left of (3.69) is zero. Evaluating the determinate of the matrix in (3.69) generates a nonlinear equation for the set of frequencies $\left\{\frac{\omega^2}{c^2}\right\}$ corresponding to the modal solutions of (3.67) and (3.69). For bound states of the system these must occur outside the pass band regions of the pure photonic crystal. Otherwise, the solutions represent resonant scattering modes that occur within the pass band of the photonic crystal.

As an example, consider the case of an impurity introduced into the photonic crystal by the replacement of a cylinder centered at the origin of the photonic crystal lattice by an impurity cylinder. The cylinders of the photonic crystal are formed of a medium with dielectric constant ε_c, and the replacement cylinder has the same geometry as the photonic crystal cylinders but is made of a dielectric medium of dielectric constant ε_c'.

In this case the points of the Gauss-Legendre quadrature are

$$\vec{r}_{||}(i,j) = \left(r_i \cos\theta_j, r_i \sin\theta_j\right) \tag{3.70}$$

where

$$r_i = \frac{R(q_i + 1)}{2} \tag{3.71a}$$

and

$$\theta_j = \pi\left(q_j + 1\right). \tag{3.71b}$$

In (3.71a) R is the radius of the dielectric cylinders of the photonic crystal so that r_i involves an integration on the impurity along the radian variable and in (3.71b) θ_j represents an integration over the angular variables on the impurity. In both (3.71), $-1 \le q_i \le 1$ for $i = 1, 2, 3, \ldots, P$ are Gauss-Legendre position points corresponding to the quadrature weights $\{w_i\}$.

For the single cylinder replacement integration proposed here, $g = \frac{\pi R^2}{4}$ in (3.69), and

$$\delta\varepsilon\left(\vec{r}_{||}\right) = \varepsilon_c' - \varepsilon_c \tag{3.72}$$

over the region $\left|\vec{r}_{||}\right| \le R$ of the cylinder of the photonic crystal that is being replaced. With these specifications, the formulae in (3.69) through (3.72) combine to provide a complete treatment of the single cylinder replacement problem.

As an additional side point, it should be noted here that the introduction of more general types of impurities is obtained through a straightforward generalization of the earlier formulae and of the numerical treatments of the single impurity problem discussed in the following. In particular, many types of dielectric changes, $\delta\varepsilon\left(\vec{r}_{||}\right)$, in the system can be inscribed within a cylinder of radius R. Such changes require only a simple modification of the above generated formulation.

The formulation for the replacement cylinder impurity is now applied for comparison with experimental results from an impurity system involving the removal of a single dielectric cylinder from the photonic crystal. Recent experiments on this system are found to be in reasonably good agreement with the results from (3.69) through (3.72).

Experimental Study

The square lattice photonic crystal composed of cylinders of dielectric constant $\varepsilon_c = 9$ surrounded by vacuum has been studied experimentally for the case of the system with filling fraction $f = 0.4488$. Both the band structure of the pure photonic crystal and the impurity modes for an impurity created in the photonic crystal through the removal of a single cylinder have been treated experimentally as well as in the theory present here.

Earlier, in Fig. 3.2 theoretical results for the band structure of the pure system have been presented. They are in good agreement qualitatively and reasonable quantitative agreement with the experimentally measured band structure for which the reader is referred to [9]. Upon the removal of a cylinder from the photonic crystal, it is found experimentally that an impurity mode appears in the system with a frequency in the second lowest stop band of the dispersion relation in Fig. 3.2.

From the band structure theory of the pure photonic crystal the second lowest stop band has an upper edge at $\frac{\omega a}{2\pi c} = 0.470$ (11.1 GHz) and a lower edge at

$\frac{\omega a}{2\pi c} = 0.413$ (9.76 GHz). From the determinant theory in (3.69) through (3.72) an impurity mode is found in the second lowest stop band at $\frac{\omega a}{2\pi c} = 0.45$ (10.63 GHz). In the experimental treatment, upon removal of a cylinder from the system a bound mode occurs in the experimental system at 11.2 GHz. This is within the second stop band of the measured results. The discrepancies between the theory and experiment may be due to the fact that the experiments were done on a system with finite length cylinders and on a finite array of the cylinders. Variations in the geometry and uniformity of the composite medium forming the cylinders may also enter into a more accurate comparison of the two approaches.

Second Approach to Impurity Problem

A second approach to finding the impurity mode frequencies from (3.67) and their fields from (3.68) can be made by reducing the integral equation in (3.67) to an eigenvalue problem. In particular, such a reduction occurs for impurities of the form

$$\begin{aligned} \delta\varepsilon\left(\vec{r}_{||}\right) &= \delta\varepsilon_0 \quad \text{for some region } S \text{ in the } x_1 - x_2 \text{plane} \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{3.73}$$

where $\delta\varepsilon_0$ a constant. For this case the integral equation in (3.67) can be rewritten into the form [5, 9]

$$f\left(\vec{r}_{||}|\omega\right) = \delta\varepsilon_0 \int_S d^2\vec{r}\,'_{||} \frac{\omega^2}{c^2} G\left(\vec{r}_{||}, \vec{r}\,'_{||}|\omega\right) f\left(\vec{r}\,'_{||}|\omega\right) \tag{3.74}$$

where S indicates an integration over the region in (3.73) for which $\delta\varepsilon_0$ is non-zero.

For a fixed value of ω (3.74) represents an integral equation eigenvalue problem for the set of eigenvalues $\{\delta\varepsilon_0\}$, giving the changes in the dielectric function needed to support a bound impurity mode about the region S. The corresponding wave functions of the bound state modes in S are given by the set $\left\{f\left(\vec{r}_{||}|\omega\right)\right\}$ associated with each of the $\{\delta\varepsilon_0\}$ eigenvalues for the wave functions of the impurity modes with frequency ω. The general fields in all of space are then obtained from (3.68) which relates them to the set $\left\{f\left(\vec{r}_{||}|\omega\right)\right\}$ through the integral transform

$$E_3\left(\vec{r}_{||}|\omega\right) = \delta\varepsilon_0 \int_S d^2\vec{r}\,'_{||} \frac{\omega^2}{c^2} G\left(\vec{r}_{||}, \vec{r}\,'_{||}|\omega\right) f\left(\vec{r}\,'_{||}|\omega\right) \tag{3.75}$$

where $\vec{r}_{||}$ is no longer restricted to the region S.

General Systems

The two methods in the preceding discussions are easily generalized to treat impurity geometries that are more involved than that of the single impurity problem. These include finite clusters of impurities, infinite sets of impurities introduced into the system to form waveguides, and infinite set of impurities to create two- and three-dimensional super-arrays.

A nice feature of the two treatments is that Group theory techniques, which have found standard applications in the study of impurity clusters in magnon, phonon, and electron systems, are directly extended to photonic crystal problems [13, 14]. These allow for the classification of the impurity frequency and wave functions by the irreducible representations of the symmetry Group of the impurity geometries [12]. This type of classification is not as nicely and quickly made in treatments based on alternative numerical simulation studies.

The application of Group theory methods also can be made to the theory photonic crystal waveguides and infinite two- and three- dimensional replacement arrays [5, 12]. These type of infinite arrays of site changes of the pure photonic crystal can exhibit resonant modes within the pass bands of the photonic crystal as well as modes bound to the infinite array with frequencies in the stop bands of the pure photonic crystal. The bound stop band modes are the modes of interest in many applications.

For a waveguide the bound stop band modes exhibit fields that are concentrated within and localized to the waveguide channel formed by the infinite array of site changes from the pure photonic crystal. The guided wave solutions are extended along the length of the waveguide channel and only propagate parallel to the channel. For such modes the translational symmetries of the systems are easily related to the band structure of the guided waves that propagate along the wave guide channel. This allows for a classification of the solutions in terms of the wave vectors of the irreducible translation group of the material forming the channel.

Two- and three-dimensional arrays of replacements in pure photonic crystals also have resonant and bound state modes depending on whether or not the modal solutions are within pass or stop band of the pure system, respectively. The stop band modes of the impurity propagate along the array of impurities and are similar to impurity bands observed in some types of impurity semiconductors. They are classified by the irreducible representations of the symmetry groups of the impurity array and may be of interest for the frequency transitions that they allow in the arrayed systems [5, 12].

To conclude the treatment of impurity modes it is of interest to consider a simple model which illustrates many of the features of the single site impurity problem in complex two- and three-dimensional photonic crystals. This is the problem of a one-dimensional layered photonic crystal into which an impurity layer is introduced. It is now discussed in the remainder of this section.

Impurity in One-Dimensional Layered Media

The one-dimensional photonic crystal with a single slab impurity not only illustrates many of the general properties of single site impurity modes introduced in two- and three-dimensional photonic crystals arrays, but it can be also used as a simple example of an approach to the impurity problem known as the super-cell approach. The super-cell method is a calculational approach for studying the properties of impurities in photonic crystals. In the approach, a pure photonic crystal is considered as being composed as a collection of identical large units or cells, i.e., these are the so-call super-cell units. An impurity is introduced into each of the super-cell units of the pure photonic crystal. The result is a new type of photonic crystal with a new periodicity.

The periodicity of the new photonic crystal is that of the impurities added to each super-cell. As the super-cells are made larger, the band structure of the new type of photonic crystal develops narrow frequency bands arising from the impurities added to each super-cell. As the separations between the impurities increases the bands arising from them flatten and give approximations for the impurity frequency of the original pure photonic crystal with an added single site impurity. The determination of the impurity mode frequency of the single site impurity in this manner is a consequence of the weakening in the super-cell system of the interaction of the impurity modes between neighboring super-cells of the photonic crystal as their separations increase.

The lattice of the photonic crystal can be considered as formed by the periodic repetition in space of a smallest spatial unit. In one-dimensions the unit is a segment of the line perpendicular to the interfaces of the layers forming the array. In two- and three-dimensional lattices the unit is a smallest repeat area or volume, respectively. In each of the one-, two-, and three-dimensional photonic crystals the super-cell is based on the same idea as the smallest repeat units in these respective systems. The super-cell units, however, are much larger that the smallest repeat unit of the system. As with the smallest repeat units, the super-cells again generate the entire photonic crystal lattice when they are repeated throughout space.

One-Dimensional Model

For the model of a one-dimensional photonic crystal, a photonic crystal composed as a periodic array of slabs of dielectric constant ε_a and vacuum are considered. The slabs are taken to be of width a and light of frequency ω is taken to travel in the system moving perpendicular to the slab interfaces.

Two impurity models based on this photonic crystal are considered [15]. In a first model, one of the slabs of dielectric constant ε_a is replace by a slab of dielectric constant ε_b. This is a model treating the impurity system as formed by slab replacement. In a second model one of the slabs of dielectric ε_a and width a has its width changed to xa where $0 \leq x \leq 3$ and the widths of its two surrounding vacuum slabs are changed to $\frac{3-x}{2}a$. This is a model which treats the change in the widths of a dielectric slab and its two vacuum neighbors. It does this in such a manner that going from the pure photonic crystal to the impurity model the total width of the three changed slabs remains $3a$.

The first model treated is the analogy in, for example, a two-dimensional photonic crystal of an impurity introduced into the system by replacing a dielectric cylinder with one of the same geometry but of a different dielectric material [15]. The second model is the analogy in a two-dimension photonic crystal array of dielectric cylinders of introducing an impurity by replacing a cylinder of the photonic crystal by a cylinder with a large or smaller radius than that of the cylinders of the original photonic crystal.

In the one-dimensional photonic crystal the super-cell solution is obtained as a product of transfer matrices [15]. These matrices relate the fields at one surface of a dielectric slab of the photonic crystal array to those at the other surface of the slab. If x_l is the position of the left surface of the dielectric slab and x_l is the position of the right surface of the dielectric slab, then the fields in the vacuum to the left of the slab are of the form

$$E_l(x) = E_{2R}e^{i\omega(x/c-t)} + E_{2L}e^{-i\omega(x/c+t)} \tag{3.76a}$$

and those to the right of the slab are of the form

$$E_r(x) = E_{1R}e^{i\omega(x/c-t)} + E_{1L}e^{-i\omega(x/c+t)}. \tag{3.76b}$$

In (3.76a) and (3.76b) the total fields within the vacuum are both seen to be composed of components propagating to the right and to the left. The coefficient of the right and left propagating components in (3.76a) and (3.76b) are shown to be related to one another through a transfer matrix $\overset{\leftrightarrow}{T}(x_l, x_r)$. This matrix is obtained from matching the boundary conditions at the slab surfaces on the fields in (3.76) and those within the dielectric slab. In this way it is found that [15].

$$\begin{vmatrix} E_{2R} \\ E_{2L} \end{vmatrix} = \overset{\leftrightarrow}{T}(x_l, x_r)\begin{vmatrix} E_{1R} \\ E_{1L} \end{vmatrix}. \tag{3.77}$$

Through the successive application of the transfer matrix the properties of the super-cell system can be investigated. To obtain the transmission properties of a large system composed of many super-cells, the finite but large portion of photonic crystal is surrounded by vacuum with incident and reflected wave boundary conditions applied on one side and transmitted wave boundary conditions applied on the opposite side of the finite portion. The impurities within the super-cells are introduced into the model by applying, in the course of the transfer process, an appropriate transfer matrix, $\overset{\leftrightarrow}{T}_i(x_l, x_r)$, relating the fields to the left and right of the impurity dielectric slab.

In this way the pass band structure of the impurity system is studied by determining the frequency characteristics of the light transmitted through the impurity photonic crystal. Light is transmitted through the system, composed of multiple super-cells containing single impurities, at the slightly renormalized pass band frequencies of the pure photonic crystal. The renormalization of the pass band frequencies comes from the introduction of impurities into the super-cells and the interactions of the impurities with the pass band modes that are present in the system in the absence of impurities. In addition, some new narrow frequency pass bands are observed arising from the weak interaction between impurity modes in each of the super-cells of the system in the presence of the impurities. These narrow frequency transmission bands are at the frequencies of the bound state modes of the single impurity problem of the photonic crystal [15].

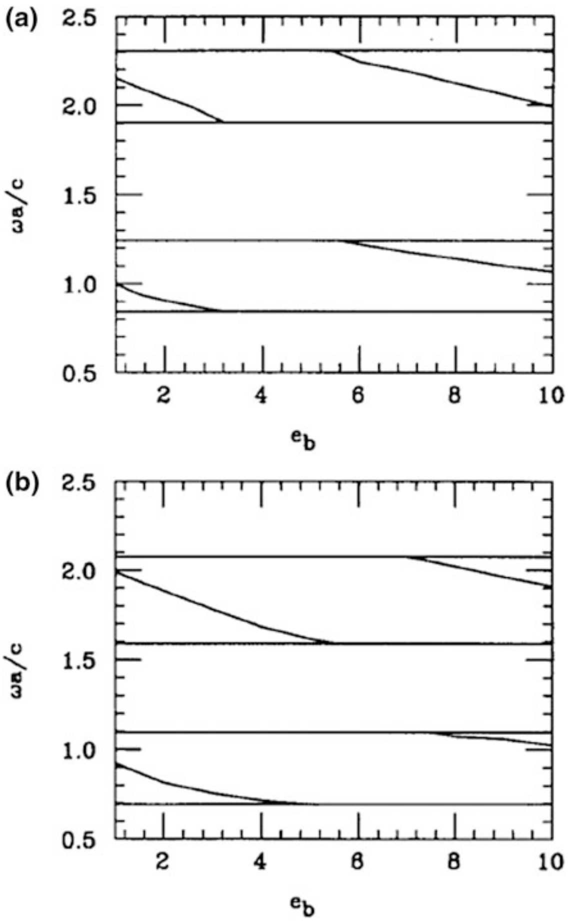


Fig. 3.7 Plot of impurity mode frequency, ω, as a function of the dielectric ε_b of the impurity. The horizontal lines indicate the edges of the stop bands and the impurity frequencies are plotted only within the stop bands. Results are presented for $\varepsilon_a = 4$ (top figure) and $\varepsilon_a = 6$ (bottom figure) [15]. Reproduced with permission from [15]. Copyright 1995 Elsiever

Some Results in One-Dimension

In Fig. 3.7 results are shown from a transmission study of the slab replacement problem in photonic crystals with $\varepsilon_a = 4$ and $\varepsilon_a = 6$. (This is the first type of replacement in which the medium of one of the dielectric slabs of the photonic crystal is changed from ε_a to ε_b.) Plots are presented for the bound state impurity mode frequencies in the stop band of the pure photonic crystal. The figures plot the results for the bound state frequency, ω, of the impurity modes within the stop bands of the photonic crystals versus the dielectric constant, ε_b, of the impurity replacement slab. The impurity slab is of the same thickness as the ε_a slabs of the photonic crystal and is only different in its dielectric constant.

In the figure the horizontal lines denote the frequency edges of the stop and pass bands of the photonic crystal. The stop bands are the frequency regions containing the plots of lines of ω versus ε_b, while the pass band frequency regions have nothing plotted within them. In this way for $\varepsilon_a = 4$ the stop bands are located between $0.843 \leq \frac{\omega a}{c} \leq 1.239$ and $1.903 \leq \frac{\omega a}{c} \leq 2.306$, and for $\varepsilon_a = 6$ the stop bands are located between $0.696 \leq \frac{\omega a}{c} \leq 1.094$ and $1.591 \leq \frac{\omega a}{c} \leq 2.075$. The plots presented within these regions are the bound state modes that are localized about the impurity medium.

As functions of increasing ε_b the impurity modes are found to enter the stop band at its upper edge and to decrease in frequency until they pass out of the stop band at its lower edge. This occurs a number of times in each of the two stop bands studied as ε_b is increased. In this way a series of periodically recurring impurity levels with increasing ε_b is created within the system. In addition, in both $\varepsilon_a = 4$ and $\varepsilon_a = 6$ systems there are regions of ε_b in which impurity modes are absent from one or both of the stop bands.

Modifications that occur within the pass band regions are scattering resonances of the impurity system. They have not been shown in the figure as they are of less technological interest than the impurity bound states. The resonant modes are dynamical processes which can favor the localization of the wave function in the vicinity of the impurity. Such modes, however, eventually move away from the impurity and travel off to infinity. These modes are never completely localized to the region of the impurity medium and are not further discussed in the following.

In Fig. 3.8 results are shown from a transmission study of the second type of slab replacement problem in photonic crystals with $\varepsilon_a = 4$ and $\varepsilon_a = 6$. (This is the second type of replacement in which the dielectric slab replacement is of the same dielectric medium as that of the photonic crystal slabs but the replacement slab is of a different width. In addition, the replacement is done in such a way that the replaced dielectric slab and it two nearest neighbor slabs retain their total length $3a$ in the photonic crystal.) Plots are presented for the bound state impurity mode frequencies in the stop band of the pure photonic crystal.

The figures plot the results for the bound state frequency, ω, of the impurity modes within the stop bands of the photonic crystals versus the width parameter, x, of the impurity replacement slab. Here the width of the replacement dielectric slab is xa and its two neighboring vacuum slabs are of length $(3 - x)a/2$. As in Fig. 3.7, the pass bands only contain modified resonant scattering states. These are of little technological interest and will not be discussed further [15].

In the figure the horizontal lines denote the frequency edges of the stop and pass bands of the photonic crystal and are at the same frequencies as those plotted in Fig. 3.7. The stop bands are the frequency regions containing the plots of lines of ω versus x, while the pass band frequency regions have nothing plotted within them. Plots presented within the stop band regions are the bound state modes that are localized about the impurity medium. They are the true bound and localized modes of the impurity photonic crystal.

As functions of increasing x the impurity mode frequencies are found to enter the stop band at its upper edge and to decrease in frequency until they pass out of the stop band at its lower edge. This occurs a number of times in each of the two stop bands studied as x is increased. In this way, a series of periodically recurring impurity levels with increasing x is created within the system. The shapes of some

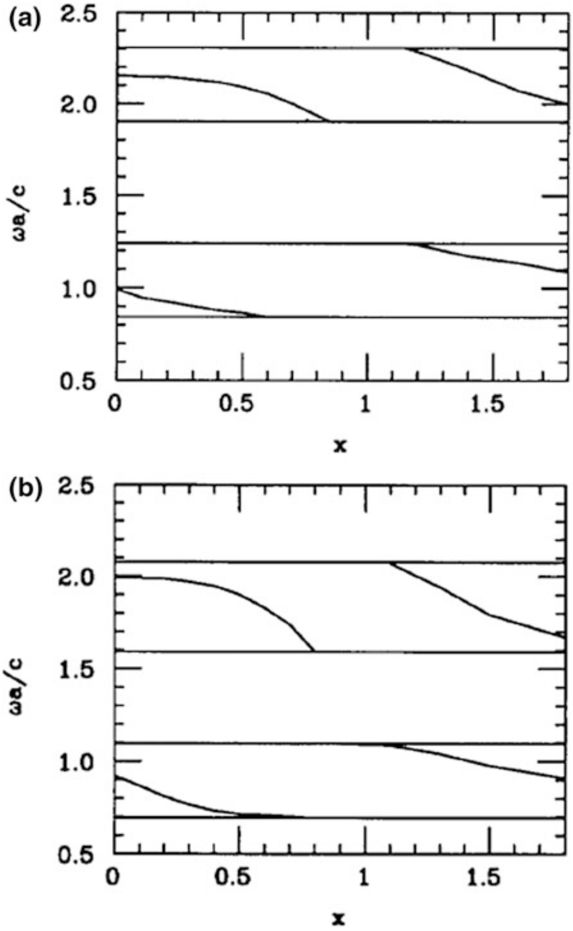


Fig. 3.8 Plot of impurity mode frequency, ω, as a function of the x of the impurity where the width of the dielectric slab is xa. The horizontal lines indicate the edges of the stop bands and the impurity frequencies are plotted only within the stop bands. Results are presented for $\varepsilon_a = 4$ (top figure) and $\varepsilon_a = 6$ (bottom figure) [15]. Reproduced with permission from [15]. Copyright 1995 Elsiever

of the bound mode curves in Fig. 3.8 as functions of x are seen to be convex, while the shape of the bound mode curves in Fig. 3.7 as functions of ε_b are concave. In addition, in both $\varepsilon_a = 4$ and $\varepsilon_a = 6$ system there are regions of x in which impurity modes are absent from one of both of the stop bands.

3.3 Method of Wannier Functions

An important analytical method for the treatment of localized impurities and waveguides in photonic crystals is the techniques of Wannier functions [5, 11, 16]. It is a method that was originally developed in the study of electron orbitals in metals and semiconductors. In those considerations, it has found many applications for the determination of electronic band structures and for the treatment of impurity modes in conductive materials. As with the earlier considered methods, introduced for the study of photonic crystals, a generalization of the Wannier function methods is often found to be useful in the treatment of photonic systems.

In the method of Wannier functions, the Block plane wave states of the electromagnetic modes in photonic crystals are used to create an alternative basis set of localized orthogonal functions. These localized functions, known as Wannier functions, can be used to represent the general solutions of the electrodynamics in the photonic crystal. The Wannier functions are particularly well suited in discussion of the properties of fields that are localized about impurities and within waveguides.

As has been seen in the earlier discussions, the periodic dielectric constant of the photonic crystal is formed by a repetition of a single dielectric unit throughout space. Consequently, these repeated units are related to one another by a translation vector, $\vec{T} = m_1\vec{a}_1 + m_2\vec{a}_2$, of the periodic lattice. (Here the reader is remained that the lattice translation vectors are linear combinations of the primitive lattice basis vectors with integer coefficients.) Similarly, the Wannier functions form a basis in which to describe the local behavior of the electromagnetic solutions within a basic repeat unit of the periodic dielectric and are designed so that the Block functions are expressed as a sum over the crystal lattice of localized Wannier functions. In this sum each Wannier function contribution is multiplied by a phase [5, 11, 16].

Whereas the Wannier functions emphasize the local behavior of the electromagnetic modes, the Block waves are modal solutions that emphasize the translational symmetry within the periodic lattice of the entire photonic crystal. In the following, the relationship between the Wannier and Block functions are discussed in terms of an analogy with the relationship between free space plane wave solutions and a delta function pulse of localized free space electromagnetic radiation fields.

Localized Wannier Functions

To this end, in the following, first the relationship in free space between plane waves and delta function pulses will be developed. Arguing by analogy, this will be used as a motivation for obtaining localized pulse-like Wannier functions from the Block wave modes of the periodic photonic crystal. All four of these types of functions can be used to represent general solutions in space.

The discussions begin by considering the plane wave forms of the electromagnetic modes in free space. In unbounded free space the modes of the electric field are plane waves given by [5].

$$E = \frac{1}{2\pi} e^{i\left(\vec{k}\cdot\vec{r} - \omega t\right)}. \tag{3.78}$$

Here for convenience the polarization of the field amplitude is ignored, i.e., only waves of unit amplitude and a single polarization are considered.

The plane waves in (3.78) are a complete orthogonal basis so that a general electromagnetic field can be written in terms of them as a Fourier transform. In this way the modes in (3.78) can be combined to express a highly localized electric pulse.

An extreme example of such a localized pulse is a delta-function pulse of radiation defined at $t = 0$. This is written in the form [5]

$$2\pi\delta(\vec{r}) = \frac{1}{2\pi}\int d\vec{k}e^{i\vec{k}\cdot\vec{r}} \tag{3.79}$$

and represents a highly localized pulse about the origin of coordinates in space.

These ideas can be developed further to generate a set of delta function pulses covering all of space. For example, the position of the pulse in (3.79) can be shifted to any location in the *x*–*y* plane by adding a phase to the integrand in (3.79) so that the integrand $e^{i\vec{k}\cdot\vec{r}}$ is replace by $e^{i\vec{k}\cdot\left(\vec{r}-\vec{R}\right)}$. With this new integrand, the position of the pulse in (3.79) is relocated to a position about $\vec{R}$.

An important idea arises from the relationship between extended modes in (3.78) and the highly localized modes in (3.79). The extended modes in (3.78) are a complete basis for studying the solutions of the system. The delta function modes of (3.79) that are generated from them by the Fourier transform are also a basis that can represent solutions of the system. Both can be used as a basis for expressing general solutions of the electrodynamics in free space. One is a set of functions that are extended throughout all space and the other is a set of functions that are highly localized in space.

A similar set of relationships to those in (3.78) and (3.79) are now developed between the Block wave basis of a periodic lattice and the localized Wannier function basis. The relationship between these two bases is developed similarly to the earlier development of the plane wave and delta function modes in free space. These are now discussed.

As with (3.79), a localized pulse of radiation in the periodic system can be generated as a Fourier transform of the extended Block modes in the periodic lattice of the photonic crystal. To understand this transformation, consider the Block wave modes of the photonic crystal to be of the form [5, 11, 16]

$$E(x,y) = b_{n,\vec{k}}(\vec{r})e^{-i\omega t} = e^{i\left[\vec{k}\cdot\vec{r}-\omega t\right]}u_{n,\vec{k}}(x,y) \tag{3.80}$$

where $u_{n,\vec{k}}(x,y)$ is the periodic function of the Block form, $\vec{k}$ is a wave vector in the first Brillouin zone, and n is a band index labeling the different frequency eigenmodes at fixed $\vec{k}$. Remember here that the band index is important in the periodic system and is used to distinguish the modal solutions of different frequencies corresponding to the same $\vec{k}$.

A localized function at each of the sites of the periodic lattice of the photonic crystal can now be created from (3.80). This is done employing the same techniques as those used to generate the delta function pulse in (3.79) from (3.78). In this way, localized functions of position are obtained from the Fourier transform of the Block wave basis and are given by [5, 16]

$$a_{n,(m_1,m_2)}(\vec{r}) = \frac{1}{\sqrt{N}} \sum_{\vec{k}} e^{-i\vec{k}\cdot\vec{R}} b_{n,\vec{k}}(\vec{r}), \tag{3.81}$$

where $\vec{R} = m_1\vec{a}_1 + m_2\vec{a}_2$ is a lattice translation vector and N is the number of lattice sites. The functions defined in (3.81) form a set of bases functions that are localized about the lattice sites, $\vec{r} = m_1\vec{a}_1 + m_2\vec{a}_2$. Notice that, because of the band structure of the dispersion relation, the bases functions are now associated with discrete sites of the lattice for modes of a particular band index n.

As an example of (3.81), consider $b_{n,\vec{k}}(\vec{r}) \propto e^{i\vec{k}\cdot\vec{r}}$. This has a uniform amplitude over the lattice so that the $a_{n,(m_1,m_2)}(\vec{r})$ generated from (3.81) is a pulse localized in the vicinity of $(\vec{r} - \vec{R}) \approx 0$. It is seen that the localization of the $a_{n,(m_1,m_2)}(\vec{r})$ as with (3.79) again arises from the constructive and destructive interference of the plane wave Block forms.

Properties of the Wannier Function Basis

Now consider some of the other properties of the Wannier basis functions defined in (3.81). These are important in their application to represent localized functions defined on the lattice.

The Wannier functions defined about different lattice sites are related to one another by a translation vector of the lattice, $\vec{T} = m_1\vec{a}_1 + m_2\vec{a}_2$. This follows from considering the function obtained from (3.80) and (3.81) which is given by [5, 16].

$$a_{n,(m_1,m_2)}(\vec{r}) = \frac{1}{\sqrt{N}} \sum_{\vec{k}} e^{i\vec{k}\cdot(\vec{r}-\vec{T})} u_{n,\vec{k}}(\vec{r}) = a_{n,(0,0)}(\vec{r} - \vec{T}). \tag{3.82}$$

Here the second equality is obtained by applying the translational invariance of $u_{\vec{k}}(x, y)$ within the periodic lattice. A consequence of (3.82) is that all of the Wannier functions on the lattice are related to $a_{n,(0,0)}(\vec{r})$ by a translation and, hence, are related to one another.

The set of all Wannier functions $\{a_{n,(m_1,m_2)}(\vec{r})\}$ generated in (3.81) then form a localized orthogonal basis of functions which are related to one another by translation vectors of the crystal lattice. This property will now be formally demonstrated using the definitions in (3.80) and (3.81).

The orthogonality of $\{a_{n,(m_1,m_2)}(\vec{r})\}$ arises from the symmetry and orthogonality properties of the Block waves in (3.80). To understand this consider the integral [5, 11, 16]

$$\begin{aligned}&\int \varepsilon(\vec{r})a^{*}_{n,(m_1,m_2)}(\vec{r})a_{n',\left(m'_1,m'_2\right)}(\vec{r})d\vec{r} = \int \varepsilon(\vec{r})a^{*}_{n,(0,0)}\left(\vec{r}-\vec{R}\right)a_{n',(0,0)}\left(\vec{r}-\vec{R}'\right)d\vec{r}\\ &= \frac{1}{N}\sum_{\vec{k},\vec{k}'} e^{i\left[\vec{k}\cdot\vec{R}-\vec{k}'\cdot\vec{R}'\right]}\int \varepsilon(\vec{r})\left(u^{*}_{n,\vec{k}}(\vec{r})e^{-i\vec{k}\cdot\vec{r}}\right)\left(u_{n',\vec{k}'}(\vec{r})e^{i\vec{k}\cdot\vec{r}}\right)d\vec{r}\\ &= \delta_{(m_1,m_2),\left(m'_i,m'_2\right)}\delta_{n,n'}\end{aligned} \tag{3.83}$$

where $\vec{R} = m_1\vec{a}_1 + m_2\vec{a}_2, \vec{R}' = m'_1\vec{a}_1 + m'_2\vec{a}_2$. Here in reducing the series of integrals in (3.83) to the product of Kronecker deltas on the far right of the equation, the translation and orthogonality properties of the Block waves are used.

Equation (3.83) is seen to be a statement of the orthogonality of the Wannier functions. As such it provides a basis for expanding the properties of the periodic system in states which are localized about the individual sites in the direct crystal lattice. Based on (3.83) an expansion of a general function, $\psi(\vec{r})$, of the system can be expressed in the form [5].

$$\psi(\vec{r}) = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} F\left(\vec{R}_j\right)a_{n,(0,0)}\left(\vec{r}-R_j\right). \tag{3.84}$$

Here $\vec{R}_j = m_{1,j}\vec{a}_1 + m_{2,j}\vec{a}_2$ where $\left(m_{1,j}, m_{2,j}\right)$ are the integers labeling the jth lattice site of the two-dimensional lattice, the sum is over all of the lattice sites, and $F\left(\vec{R}_j\right)$ are expansion coefficients.

Examples of the expansion in (3.84) are the Block waves of the system themselves. They are written in terms of the Wannier basis functions as

$$b_{n,\vec{k}}(\vec{r}) = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} e^{i\vec{k}\cdot\vec{R}_j}a_{n,(0,0)}\left(\vec{r}-R_j\right). \tag{3.85}$$

In (3.85), the Wannier functions $a_{n,(0,0)}(\vec{r})$ provide the localized variation of the Block wave about each lattice site. In addition, this local variation is multiplied at each site by a phase that changes from site to site along the lattice. As a results the overall translational symmetry of the Block wave solutions is maintained.

As shall be seen in the following considerations (3.84) and (3.85) are particularly useful results in the treatment of impurity problems. These problems are now undertaken in the following.

Impurity Problems

Consider the problem of a two-dimensional photonic crystal containing an impurity [11]. In the absence of an impurity the photonic crystal has modes that are the solutions of the Helmholtz equation of motion [5, 16]

$$\left[H_0 + V_p(\vec{r})\right]\psi(\vec{r}) = 0. \tag{3.86a}$$

where $H_0 = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$ describes the propagation of the waves in free space, $V_p(\vec{r}) = -n^2(x,y)\frac{\omega^2}{c^2}$ is the periodic potential of the photonic crystal, $\psi(\vec{r})$ is the electric field polarized perpendicular to the plane of propagation, ω is the frequency, and $n(x,y)$ is the periodic index of refraction of the system. Upon introducing a change in the periodic dielectric so that the $n^2(\vec{r}) \to n^2(\vec{r}) + \delta\varepsilon(\vec{r})$ replacement is made in (3.86a) one obtains the impurity equation [5, 16]

$$[H_0 + V(\vec{r})]\psi(\vec{r}) = \varepsilon(\vec{r})\frac{\omega^2}{c^2}\psi(\vec{r}). \tag{3.86b}$$

where $V(\vec{r}) = -\delta\varepsilon(\vec{r})\frac{\omega^2}{c^2}$ is the position dependent impurity potential arising from a change $\delta\varepsilon(\vec{r})$ of the dielectric constant from that of the pure photonic crystal, and the periodic potential has been written in terms of the permittivity and frequency.

To generate the solution of (3.86b), (3.84) is substituted into (3.86). Multiplying the resulting expression by $a^*_{n,(0,0)}\left(\vec{r} - \vec{R}_i\right)$, and integrating over $\vec{r}$ changes the differential equation problem into a set of difference equations. These difference equations relate the envelope functions $F\left(\vec{R}_j\right)$ defined at each lattice site to one another and are given by [5, 11, 16]

$$\sum_{j'}\left[h_0\left(\vec{R}_{j'}\right)F\left(\vec{R}_j - \vec{R}_{j'}\right) + V_{j,j'}F\left(\vec{R}_{j'}\right)\right] = \frac{\omega^2}{c^2}F\left(\vec{R}_j\right), \tag{3.87}$$

where

$$h_0\left(\vec{R}_j - \vec{R}_{j'}\right) = \int d\vec{r}a^*_{n,(0,0)}\left(\vec{r} - \vec{R}_j\right)H_0 a_{n,(0,0)}\left(\vec{r} - \vec{R}_{j'}\right), \tag{3.88}$$

and

$$V_{j,j'} = \int d\vec{r}a^*_{n,(0,0)}\left(\vec{r} - \vec{R}_j\right)V(\vec{r})a_{n,(0,0)}\left(\vec{r} - \vec{R}_{j'}\right). \tag{3.89}$$

The solutions of (3.87) for the envelope function $F\left(\vec{R}_j\right)$ of the Wannier function expansion in (3.84) offer an advantage over the study of the full impurity wave function solutions of (3.86). Equation (3.84) allows for the rapid changes in the full wave functions of the impurity modes to be accounted for by the localized Wannier functions. Consequently, the spatial dependence of the envelope functions in (3.84) tend to be slowly varying in space.

An important property of the envelope functions that will be useful in the following discussion is to note that for the case in which $V(\vec{r}) = 0$ the pure system limit is obtained. In this limit, (3.87) reduces to the form of a tight binding

Hamiltonian, and the envelope functions are plane waves. While the envelope functions involve a phase, the amplitude of the envelope functions are constant in space.

In the presence of an impurity, a treatment based on (3.84), written in term of the Wannier functions of the pure photonic crystal, greatly facilitates the problem. In particular, the localized nature of the Wannier functions and the localized nature of the impurity potential often combine, lead to $V_{j,j'}$ terms which quickly go to zero with increasing separation between the j, j' site. This limits the range of coupling between envelope sites in (3.87) and simplifies the process of obtaining the solution of the envelope function.

Reformulation as a Schrodinger Equation

A difficulty with (3.87), however, is that it is an algebraic equation that can still couple together many of the $F\left(\vec{R}_j\right)$ at different sites. In some cases that are of great interest, the algebraic difference equation can be further simplified by making an additional approximation. This reduces the problem of determining the envelope function defined in (3.84) to a standard type of Schrodinger differential equation problem [10, 15]. As a result, the slowly varying envelope function in (3.84) is obtained as a solution of the newly developed Schrodinger equations.

To see how the continuum limit is taken, noted that a simplification can be made to (3.87) by considering the first term on the left side of the equation. This term is given by $\sum_{j'} h_0\left(\vec{R}_{j'}\right) F\left(\vec{r}-\vec{R}_{j'}\right)$. Applying a Taylor series expansion in $\vec{r}$ to $F\left(\vec{r}-\vec{R}_{j'}\right)$ in this expression, it is found that [5, 11]

$$\sum_{j'} h_0\left(\vec{R}_{j'}\right) F\left(\vec{r}-\vec{R}_{j'}\right)=\sum_{j'} e^{-\vec{R}_{j'}\cdot\nabla} h_0\left(\vec{R}_{j'}\right) F(\vec{r}) \tag{3.90}$$

where ∇ is the gradient operator operating on the variables (x, y).

Equation (3.90) can now be used to develop the continuum limit for the first term on the left hand side of (3.87). In this limit, the difference sum between sites in (3.87) is to be replaced by a differential operator applied to $F(\vec{r})$. Consider this limit first for the simplified case in which $V(\vec{r})=0$.

To understand the replacement of the difference operator on the left hand side (3.87) by a differential operator, consider the case in which $V(\vec{r})=0$. In this limit (3.87) is solved by the $F\left(\vec{R}_j\right) \propto e^{-i\vec{k}\cdot\vec{R}_j}$ plane wave form, yielding an eigenvalue solution

$$\sum_{j'} h_0\left(\vec{R}_{j'}\right) e^{-i\vec{k}\cdot\vec{R}_{j'}}=\frac{\omega_0^2\left(\vec{k}\right)}{c^2}, \tag{3.91}$$

where $\omega_0\left(\vec{k}\right)$ is a frequency in the pure system.

In the continuum limit it is assumed that $F\left(\vec{R}_j\right) \propto e^{-i\vec{k}\cdot\vec{R}_j}$ becomes a general function of space $F(\vec{r}) \propto e^{-i\vec{k}\cdot\vec{r}}$. (It is also assumed that even in the presence of $V(\vec{r})$, the F 's are slowly varying functions over the periodic lattice.) In addition, in taking the continuum limit, the expression on the right side of (3.91) can be replaced by an operator so that the eigenvalue problem, rewritten in terms of (3.90), is expressed as [5]

$$\sum_{j'} h_0\left(\vec{R}_{j'}\right)e^{-\vec{R}_{j'}\cdot\nabla}F(\vec{r}) = \frac{\omega_0^2\left(\vec{k}\right)}{c^2}F(\vec{r}) \tag{3.92a}$$

where $F(\vec{r}) \propto e^{-i\vec{k}\cdot\vec{r}}$ is the plane wave eigenmode. Consequently, under the sum of these processes, the eigenvalues of the two problems are the same, and the continuum wave function correctly reduces to the discrete envelope function at the lattice sites.

These ideas can be taken one step further. The wave vectors in k-space can be approximately replaced with spatial differential operators by applying the quantum mechanical $\vec{k} \leftrightarrow -i\nabla$ correspondence. Under this replacement (3.92a) becomes.

$$\sum_{j'} h_0\left(\vec{R}_{j'}\right)e^{-\nabla\cdot\vec{R}_{j'}}F(\vec{r}) = \frac{\omega_0^2(-i\nabla)}{c^2}F(\vec{r}). \tag{3.92b}$$

Though (3.92b) is valid for the system in the absence of the impurity potential, it is also a reasonably good approximation in the case that the amplitude of $F(\vec{r})$ has a slow spatial variation.

Consequently, it follows from (3.90) that in the continuum limit

$$\sum_{j'} h_0\left(\vec{R}_{j'}\right)F\left(\vec{r}-\vec{R}_{j'}\right) = \frac{\omega_0^2(-i\nabla)}{c^2}F(\vec{r}). \tag{3.93}$$

From this result it is seen that, as a reasonable approximation for short range potentials, (3.87) and (3.93) then yield an equation for the envelope function of the form [5, 16].

$$\left[\frac{\omega_0^2(-i\nabla)}{c^2} + V(\vec{r})\right]F(\vec{r}) = \frac{\omega^2}{c^2}F(\vec{r}). \tag{3.94}$$

A condition on the approximations leading to (3.94) is that the continuous function $V(\vec{r})$ is an approximation for discrete couplings $V_{j,j'}$. Underlying this assumption is the ideas that $V_{j,j'}$ is a short ranged potential compared to the spatial variation of $F(\vec{r})$. In other words, $V_{j,j'} \approx V\left(\vec{R}_j\right)\delta_{j,j'}$ is a basic consideration in determining the continuum form of the impurity potential.

For impurity systems studied in nanophotonics, the modes of interest are bound states having frequencies located within the stop bands of the photonic crystals. The frequency terms in such systems, $\frac{\omega_0^2(\vec{k})}{c^2}$, can often be well represented as extrema of the upper or lower bound of the pass band of the pure system. Consequently, a parabolic approximation of the band by the form [5, 11, 16].

$$\frac{\omega_0^2\left(\vec{k}\right)}{c^2} = E_0 + \frac{\hbar^2}{2m}k^2 \tag{3.95a}$$

is useful for purposes of illustrating the stop band bound states. (Here E_0 and $\frac{\hbar^2}{2m}$ are to be treated as parameters of the parabolic fit to the band extrema. These are written in the notation of studies of the Schrodinger equations to make the impurity problem in (3.94) resemble the impurity problem for electron levels in semiconductors. In photonic crystal systems, however, the units of these two parameters are different from those of the electron problem.)

Applying this notation in (3.95a) it follows that $\frac{\hbar^2}{2m} > 0$ for an impurity state below the bottom of a pass band and $\frac{\hbar^2}{2m} < 0$ for an impurity state above the top of a pass band. In the parabolic approximation in (3.95a) applied to (3.92), the impurity equation (3.94) can be rewritten in the form [5, 16].

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) + E_0 - \frac{\omega^2}{c^2}\right]F(\vec{r}) = 0. \tag{3.95b}$$

This is in the standard form found in the treatment of impurity problems in semiconductor physics so that standard methods from the study of the impurity conductivity problem can be applied to its treatment.

An Example

A frequent choice of impurity potential is one of the form $V(\vec{r}) = -\delta\varepsilon(\vec{r})\frac{\omega^2}{c^2}$ where $\delta\varepsilon(\vec{r})$ represents the change in the photonic crystal dielectric constant arising from the presence of impurity media [5, 11]. Applying this potential in (3.95b) it is then found that the form of a Schrodinger equation impurity problem for the optical system becomes [5, 16].

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + E_0 - \frac{\omega^2}{c^2}\right]F(\vec{r}) = \delta\varepsilon(\vec{r})\frac{\omega^2}{c^2}F(\vec{r}). \tag{3.96}$$

The solution of (3.96) gives the envelope function, $F(\vec{r})$ corresponding to the modes at frequency, ω. If the modal frequency ω is located within the stop band of the pure photonic crystal system, the solutions are the bound states modes which are localized about the impurity media. In the following the focus will be on studying the bound state impurity modes with frequencies inside a stop band. Modes with frequencies within the pass bands of the photonic crystal are resonant modes. These are not of interest here.

The solution of (3.96) is obtained using standard Green function methods [5, 11–14]. In the Green's function approach, the Green's function, $G(\vec{r}, \omega)$, for (3.96) is determined as a solution of [17].

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + E_0 - \frac{\omega^2}{c^2}\right]G(\vec{r}, \omega) = \delta(\vec{r}). \tag{3.97}$$

In the determination of the bound state Green's function, the boundary conditions on the solutions are that the Green's function must vanish at infinite separation from the localized impurity media. Applying these conditions for the Green's function of the two-dimensional photonic crystal, it is found that [5, 17]

$$G(\vec{r}, \omega) = \frac{1}{2\pi}\frac{2m}{\hbar^2}K_0\left(\left[\frac{2m}{\hbar^2}\left(E_0 - \frac{\omega^2}{c^2}\right)\right]^{1/2} r\right), \tag{3.98}$$

where $K_0(x)$ is a modified Bessel function of the second kind.

Applying the Green's function in (3.98) to write the solution of (3.96), it follows that the formal the solution of (3.96) is given as a solution of the integral equation [5].

$$F(\vec{r}) = \frac{\omega^2}{c^2}\int d\vec{r}' G(\vec{r} - \vec{r}', \omega)\delta\varepsilon(\vec{r}')F(\vec{r}'). \tag{3.99}$$

That $F(\vec{r})$ obtained from (3.99) is a formal solution of (3.96) can be seen upon a direct substitution of (3.99) into (3.96) and with the application of the relationship in (3.97). The original problem of solving the differential equation in (3.96) is now transformed to a treatment of the solution of the integral equation in (3.99).

For a frequency $\frac{\omega^2}{c^2}$ chosen within the stop band of the photonic crystal, the solutions of the integral equation in (3.99) then yield the envelope functions of the bound state modes. This is very useful as the solution of the integral equation in (3.99) is often easier to generate than the direct solution of the original differential equation problem of (3.96). As an illustration of such a solution of (3.99) a discussion will now be given for the special case of a single site impurity formed in a photonic crystal by cylinder replacement.

Numerical Illustration

As an example, a single site impurity is treated for the case in which the impurity is formed by replacing one of the cylinders of the photonic crystal. This amounts to changing the dielectric constant of a single cylinder within the photonic crystal. The system considered will be a two-dimensional photonic crystal formed as an array of infinite dielectric cylinders, and the radii of the cylinders of the photonic crystal and impurity cylinder will be taken as R.

If the change in the dielectric constant of the replaced photonic crystal cylinder is denoted by $\delta\varepsilon_{00}$, then (3.99) for the impurity modes takes the form [5, 11, 16].

$$F(\vec{s}) = e \int_S d\vec{s}' K_0(|\vec{s} - \vec{s}'|) F(\vec{s}'). \tag{3.100}$$

In this equations the following notation has been adopted: $\vec{s} = \left[\left(2m/\hbar^2\right)\left(E_0 - \left(\omega^2/c^2\right)\right)\right]^{1/2}\vec{r}$, the eigenvalues are of the form $e = \frac{1}{2\pi}\frac{\left(\omega^2/c^2\right)\delta\varepsilon_{00}}{E_0 - \left(\omega^2/c^2\right)}$, and S is a circular region of integration which is of radius $s_0 = \left[\left(2m/\hbar^2\right)\left(E_0 - \left(\omega^2/c^2\right)\right)\right]^{1/2} R$ and which is centered about the origin of the photonic crystal lattice.

The bound state solutions are obtained from (3.100) in terms of the eigenvalue solutions for e as a function of $E_0, \hbar^2/2m$, and ω^2/c^2. Following the determination of the eigenvalue e and eigenvector $F(\vec{s})$, the relation [5, 11]

$$\delta\varepsilon_{00} = 2\pi\left(\frac{E_0}{\omega^2/c^2} - 1\right)e \tag{3.101}$$

then gives the values of $\delta\varepsilon_{00}$ which are necessary to support an impurity mode of frequency ω. In order to have a true bound state at the impurity site, the frequency must be chosen to be within the stop band of the photonic crystal. Otherwise, the solution represents a resonance of the system.

Results for e plotted versus s_0 are presented in Fig. 3.9. These are obtained from the numerical solution of (3.100) [5, 11]. The curves labeled *i* and *ii* are, respectively, the lowest and next to lowest eigenvalues of a series of eigenvalue solutions obtained from (3.100).

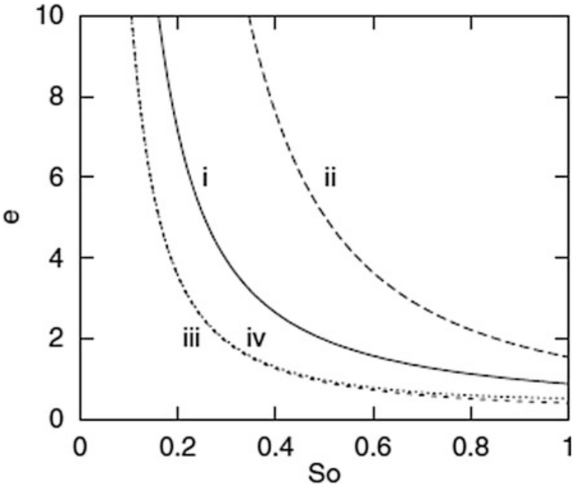


Fig. 3.9 Plot of e versus So taken from [11]. Reproduced with permission from [11]. Copyright 2005 IOP Publishing

The figure is taken from [11] and contains addition plots related to other considerations of the single site problem, For additional discussions of these other labeled curves, that are not related to the discussions presented here, the reader is referred to [11].

As an important point, however, it should be noted that the figure shows, for a fixed s_0, that there are a series of different e solutions to the integral equations. Consequently, there is a series of different $\delta\varepsilon_{00}$ solutions, each of which support their own bound impurity modes. These modes are each described by a related envelope function solutions, $F\left(\vec{R}\right)$.

3.4 Photonic Crystal Waveguides: Analytical Models

In this section, a simple analytical model will be used to illustrate some of the basic features found in photonic crystal waveguides, waveguides with impurities, and networks of waveguides [18]. These systems can be studied by means of computer simulations [1–5], and this is the approach taken for many engineering applications. However, the approach presented in this section is an analytical method that can be quickly and easily solved to demonstrate general properties of the photonic crystal waveguides. This provides a certain amount of pedagogical insight that is not always made available from a computer simulation study.

The approach presented here is based on a set of algebraic difference equations that can be generated applying the methods used in the formulation of (3.64). For the discussions, the difference equations are developed for a two-dimensional photonic crystal formed as a periodic array of infinite dielectric cylinders with axes parallel to the x_3-axis and which are arrayed periodically in the $x_1 - x_2$ plane. To simplify the consideration, the modes of interest in the system are restricted to those with components of the electric field polarized along the x_3-axis.

The Integral Equation Model

If a position dependent dielectric impurity of the form $\delta\varepsilon\left(\vec{r}_{||}\right)$ is added to the periodic dielectric function of the two-dimensional photonic crystal, it can be shown using the techniques leading to (3.64) that the electric field of the photonic crystal modes are solutions of the integral equation [18, 19].

$$E_3\left(\vec{r}_{||}|\omega\right) = \frac{\omega^2}{c^2}\int d^2r'_{||}G\left(\vec{r}_{||},\vec{r}'_{||}|\omega\right)\delta\varepsilon\left(\vec{r}'_{||}\right)E_3\left(\vec{r}'_{||}|\omega\right) \tag{3.102}$$

Here ω is the mode frequency and $G\left(\vec{r}_{||},\vec{r}'_{||}|\omega\right)$ is the electromagnetic Green's function of the two-dimensional photonic crystal in the absence of impurity materials.

For $\delta\varepsilon(\vec{r}_{||}) = \delta\varepsilon_0$ a constant over its spatial region of definition and zero otherwise, the integral equation in (3.102) becomes an eigenvalue problem for the eigenvector fields $E(\vec{r}_{||}|\omega)$ and eigenvalues proportional to $\delta\varepsilon_0$.

In the case that ω is within a pass band of the photonic crystal the Green's function is that of a propagating mode, but for the case in which the frequency is within a stop band of the photonic crystal the Green's function represents that of an exponential decaying mode. Consequently, solutions of (3.102) within a stop band are found to be localized states bound in the vicinity of the impurity media. States within the pass band are resonant scattering states and are not of interest here.

Single Impurity Bound States: A Difference Equation Approach

A simple problem that can be treated using (3.102) is that of a single site impurity with [18]

$$\delta\varepsilon(\vec{r}_{||}) = \delta\varepsilon_0 \,\text{for}\, |\vec{r}_{||}| < R_\varepsilon, \delta\varepsilon(\vec{r}_{||}) = 0\,\text{otherwise} \tag{3.103}$$

where $\delta\varepsilon_0$ is a constant and R_ε is the radius of a region over which $E_3(\vec{r}_{||}|\omega)$ is slowly varying in space. For this case (3.102) can be written as

$$E_3 = \gamma_s \alpha E_3 \tag{3.104}$$

where $\gamma = \delta\varepsilon_0 A_\varepsilon$ for the impurity cross sectional area A_ε, α is proportional to the Green's function integral over A_ε, and E_3 is the average field in A_ε.

If the mode frequency ω is chosen in the stop band of the pure photonic crystal the solutions will be a site impurity bound states, localized about the position of the impurity. On the other hand, if the mode frequency is in a pass band of the photonic crystal it will represent a scattering resonance which is extended throughout the photonic crystal.

The method outlined above is essentially the same as that used to study localized bound state impurity modes and scattering resonances in impurity semiconductors. In the photonic crystal, the impurity modes then must satisfy the algebraic equation

$$[1 - \gamma\alpha]E_3 = 0 \tag{3.105}$$

so that the values of $\gamma \propto \delta\varepsilon_0$ which support impurity modes are determined by

$$\gamma_s = \frac{1}{\alpha}. \tag{3.106}$$

For bound states, the Green's function $G(\vec{r}_{||}', \vec{r}_{||}'|\omega)$ is determined at a given stop band frequency ω and used to evaluate α for the particular pure photonic crystal being studied.

The treatment in (3.103) through (3.106) is the simplest type of photonic crystal impurity that can be treated. More complex impurities involve a number of impurity sites. In the following a generalization of these equations will be made to treat a waveguide formed as a linear array of impurities along the x_1-axis. The waveguide is formed by placing impurity media identically at each dielectric cylinder in a line of cylinders along the x_1-axis. For the treatment, a simple square lattice photonic crystal of lattice constant a and dielectric cylinders of radii $R < \frac{a}{2}$ will be taken as the pure photonic crystal.

Photonic Crystal Waveguides: Difference Equation Approach

The impurity media array representing the waveguide channel in the system is written in the form [18, 19]

$$\delta\varepsilon\left(\vec{r}_{||}\right) = \delta\varepsilon_0 \quad \text{for} \left|\vec{r}_{||} - na\hat{x}_1\right| < R_\varepsilon,$$
$$\delta\varepsilon\left(\vec{r}_{||}\right) = 0 \quad \text{otherwise} \tag{3.107}$$

where n runs over the integers. This then represents an addition of impurity $\delta\varepsilon_0$ to the pure photonic crystal in a region of radius R_ε centered at each lattice site na on the x_1-axis. The result is an infinite array of impurities along the x_1 -axis with symmetry under translation by $na\hat{x}_1$. This is the form of a one-dimensional waveguide formed by making a linear array of single site impurities of the type in (3.103).

For the impurity array in (3.107), the integral eigenvalue problem in (3.102) can be treated in the same approximation used in going from (3.102) to (3.104) for the single site impurity [18, 19]. In this way, the integral equation in (3.102) becomes a set of difference equations of the form

$$E_3(n) = \gamma[\alpha E_3(n) + \beta(E_3(n+1) + E_3(n-1))] \tag{3.108}$$

where n runs over the integers.

Here $E_3(n)$ is the electric field in the impurity media at the nth lattice site along the x_1-axis, $\gamma = \delta\varepsilon_0 A_\varepsilon$ for the impurity cross sectional area A_ε, α is proportional to the Green's function integral over A_ε of the impurity medium in the nth site, β is proportional to the Green's function integral over A_ε of the impurity medium in the (n + 1)th or in the (n − 1)th site, and A_ε is the cross sectional area of a single impurity site. Notice, that the coefficients in (3.108) are evaluated at the modal frequency ω which is taken to be in a stop band of the pure photonic crystal and that in the limit of a single site impurity (3.108) reduces to (3.104). In addition, the decay in the fields of the waveguide modes with separation from the waveguide channel is mediated by the exponential decay of the Green's function in (3.102) in the bulk of the photonic crystal.

As a simplification of the mathematics, it is assumed in (3.108) that the exponential spatial decay of the coupling between neighboring sites is quick enough that only nearest neighbor couplings are needed in the difference equation. This can always be arranged in a photonic crystal. Specifically, if the separation of the neighboring impurity sites forming the waveguide channel is sufficiently great the higher order couplings can be made very small. This may require nearest neighbor impurity sites to be separated by na where $n > 1$ is the nearest neighbor separation.

The difference equations in (3.108) can be solved by assuming an electric field of the form

$$E_3(n) = E_0 e^{inka}. \tag{3.109}$$

Upon substitution of (3.109) into (3.108) the difference equation eigenvalue problem reduces to

$$E_0 = \gamma[\alpha + 2\beta \cos(ka)]E_0 \tag{3.110}$$

so that the values of $\gamma \propto \delta\varepsilon_0$ which support guided modes bound to and traveling along the waveguide channel are determined by

$$\gamma = \frac{1}{\alpha + 2\beta \cos(ka)}. \tag{3.111}$$

For a given guided mode frequency ω located within the stop band of the pure photonic crystal, the Green's function $G\left(\vec{r}_{||}, \vec{r}^{\prime}_{||}|\omega\right)$ is determined and used to evaluate α and β for the particular pure photonic crystal being studied. The value of $\gamma \propto \delta\varepsilon_0$ which supports a guided mode of stop band frequency ω is then obtained from (3.111).

The waveguide problem treated in (3.108) through (3.111) is found to be closely related to the tight binding model in condensed matter physics [20]. This has been used to study electron systems in which the conduction is by electron hopping between atomic sites of the lattice. In this case the excitations are fermion. It has also been applied to boson system such as those found in the study of lattice vibrations. Both of these systems require a full quantum mechanical approach. The model in (3.108), however, is being treated strictly as a model of classical electrodynamics.

Photonic Crystal Waveguide Containing a Single Site Impurity

A generalization of the waveguide problem to one containing a single site impurity within the wave guide channel can be made by adjusting the set of difference equations in (3.108). The problem proposed is shown schematically in Fig. 3.10. For the treatment, a generalized model of the site impurity problem will now be solved which allows the waveguide channel to differ in the regions above and below the site impurity. This allows for a variety of configurations that could be of technological interest.

Consider then a problem in which $\gamma \propto \delta\varepsilon$ in the waveguide above the impurity site, $\gamma_0 \propto \delta\varepsilon_0$ in the waveguide below the impurity site, and at the impurity site $\gamma_1 \propto \delta\varepsilon_1$. The difference equations for the system are

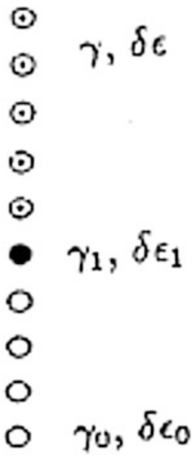


Fig. 3.10 The waveguide with a single site impurity. Only the sites of the waveguide channel are shown and the waveguide channel is formed by cylinder replacement. The system considered is a square lattice photonic crystal and the waveguide and impurity are formed by replacement of the photonic crystal cylinders in the row along the x-axis of the photonic crystal. Reproduced with permission from [18]. Copyright 1967 American Physical Society. Copyright 2002 American Physical Society

$$E_3(n) = \gamma_n[\alpha E_3(n) + \beta(E_3(n+1) + E_3(n-1))] \tag{3.112a}$$

for $|n| \geq 2$ where $\gamma_n = \gamma \propto \delta\varepsilon$ for $n \geq 2, \gamma_n = \gamma_0 \propto \delta\varepsilon_0$ for $n \leq -2$,

$$E_3(\pm 1) = \gamma_\pm[\alpha E_3(\pm 1) + \beta E_3(\pm 2)] + \gamma_1 E_3(0) \tag{3.112b}$$

where $\gamma_+ = \gamma, \gamma_- = \gamma_0$, and

$$E_3(0) = \gamma_1 \alpha E_3(0) + \gamma\beta E_3(1) + \gamma_0 \beta E_3(-1). \tag{3.112c}$$

Bound state solutions of the difference equations in (3.112) can exist. These are localized about the waveguide site at the origin and are of the form

$$E_3(n) = E_0 e^{-nqa} \tag{3.113a}$$

for $n \geq 1$,

$$E_3(0) = E_0', \tag{3.113b}$$

and

$$E_3(n) = E_0'' e^{nka} \tag{3.113c}$$

for $n \leq -1$.

Upon substitution of (3.113) into (3.112) it follows for ω within a stop band of the pure photonic crystal that

$$\gamma = \frac{1}{\alpha + 2\beta \cosh qa}, \tag{3.114a}$$

$$\gamma_0 = \frac{1}{\alpha + 2\beta \cosh ka}, \tag{3.114b}$$

and

$$\gamma_1 = \frac{1}{\alpha + \beta(e^{-qa} + e^{-ka})}. \tag{3.114c}$$

Here γ and γ_0 are fixed parameters of the waveguide channel and (3.114a) and (3.114b) are used to determine the values of q and k describing the spatial exponential decay of the waveguide fields in the channel. Once these decay parameters are determined, (3.114c) then determines the value of $\gamma_1 \propto \delta\varepsilon_1$ required in order to bind the impurity mode to be localized about the origin site in the waveguide channel.

Some interesting limits of the bound state problem can be obtained from a consideration of (3.114). In the limit that q, $k \to \infty$ it is found from (3.114a) and (3.114b) that

$$\gamma = \gamma_0 = 0 \tag{3.115a}$$

and from (3.114c) that

$$\gamma_1 = \gamma_s = \frac{1}{\alpha}. \tag{3.115b}$$

This is the limit of a single site impurity in the pure photonic crystal given in (3.103) through (3.106) and discussed earlier.

On the other hand, if only $k \to \infty$ then the problem reduces to an impurity site located at the end of a semi-infinite waveguide. In this case,

$$\gamma = \frac{1}{\alpha + 2\beta \cosh qa}, \tag{3.116a}$$

$$\gamma_0 = 0, \tag{3.116b}$$

and

$$\gamma_1 = \frac{1}{\alpha + \beta e^{-qa}}. \tag{3.116c}$$

For a frequency in the stop band of the photonic crystal, (3.116a) then gives q in terms of the waveguide parameter γ, and (3.116c) determines the value of the impurity parameter $\gamma_1 \propto \delta\varepsilon_1$ needed for a bound state localized on the waveguide impurity site.

If the waveguide channel on both sides of the channel impurity have the same dielectric constants then (3.114) become

$$\gamma = \gamma_0 = \frac{1}{\alpha + 2\beta \cosh qa}, \tag{3.117a}$$

and

$$\gamma_1 = \frac{1}{\alpha + 2\beta e^{-qa}}. \tag{3.117b}$$

For a weakly localized mode in which $q \approx 0$ these reduce to

$$\gamma = \gamma_0 = \frac{1}{\alpha + 2\beta}, \tag{3.118a}$$

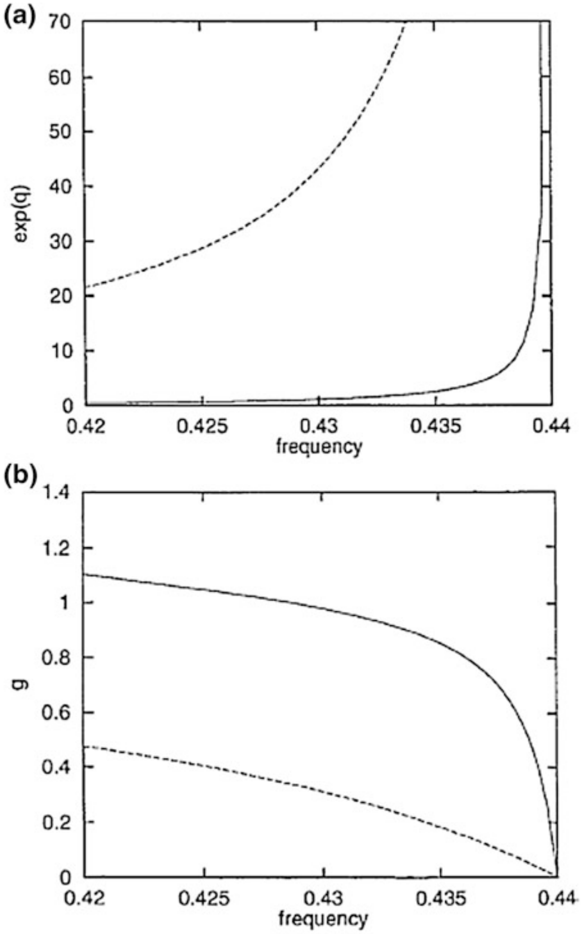
and

$$\gamma_1 = \frac{1}{\alpha + 2\beta(1 - qa)}. \tag{3.118b}$$

The solutions of the set of (3.117) for an example of a waveguide impurity are presented in Fig. 3.11. The pure system photonic crystal is the square lattice photonic crystal used for the band structure studies presented earlier in Figs. 3.2, 3.3, and 3.4 [18]. The plots are for: a) $\exp(q)$ versus frequency and b) $g = \frac{\gamma}{\gamma_s}$ versus frequency. Here γ_s is defined in (3.106) and γ_1 has been fixed at the value of γ_s evaluated at $\frac{\omega a}{2\pi c} = 0.440$. For the cases studied in the plots the media added to the cylinders of the waveguide channel had a square cross section [18]. The lengths of the square sides are indicated in the figure captions.

It is seen in the figures that as the frequency of the bound mode approaches that of a single site impurity of the photonic crystal at $\frac{\omega a}{2\pi c} = 0.440$, the channel dielectric parameters rapidly decay to zero. The divergence of the channel decay parameters is also approached as the waveguides are no longer present in this limit. Away from $\frac{\omega a}{2\pi c} = 0.440$ the channel dielectric needed to support the mode increases rapidly and the spatial decay of the impurity mode within the channel decreases.

Fig. 3.11 Plots of: a) $\exp(q)$ versus frequency, $\frac{\omega a}{2\pi c}$, and b) $g = \frac{\gamma}{\gamma_s}$ versus frequency, $\frac{\omega a}{2\pi c}$, for the $q = k$ system in (3.117). Here γ_s is defined in (3.106) and γ_1 has been fixed at the value of γ_s evaluated at $\frac{\omega a}{2\pi c} = 0.440$. The sides of the added media are $0.1a$ (lower curves) and $0.01a$ (upper curves). Reproduced with permission from [18]. Copyright 2002 American Physical Society



The difference equation theory developed in the preceding impurity and waveguide problems also lends itself to the treatment of photonic crystal circuits. These are more complex arrays of interconnecting waveguide networks which offer a wider variety of signal processing. As an example of these type of circuits, a waveguide coupler circuit will now be studied.

Waveguide Coupler

An example of a pair of coupled waveguides is shown in Fig. 3.12 [18]. Each of the two waveguided channels are formed into the shape of a U, one an upper upright U and the second a lower inverted U. Both waveguides are infinite in length, with the bottoms of the two U's coming into closest proximity in the horizontal region at the

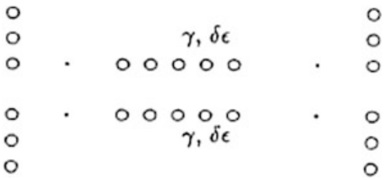


Fig. 3.12 Two coupled waveguides in the form of an upright U and an inverted U. A weak interaction between the two waveguides occurs along the length of the horizontal region of closes approach of the two waveguides. As with Fig. 3.10, only the sites of the waveguide channel are shown and the waveguide channel is formed by cylinder replacement. Reproduced with permission from [18]. Copyright 2002 American Physical Society

bottom of the two U. Here there is a weak interaction between the two channels which allows for the modes in each of the waveguide channels to interact with one another.

The system composing the two waveguide channels is described by the following set of difference equations, representing a generalization of the waveguide equation in (3.108) to the case of the two weakly coupled waveguides in Fig. 3.12: The vertical sides of the two U's are described by difference equations of the form

$$E_3(j,l) = \gamma[\alpha E_3(j,l) + \beta(E_3(j,l+1) + E_3(j,l-1))]. \tag{3.119a}$$

Here in the case of the upper U $j = 0$ or N and $l = d+1,\ d+2,\ d+3,\ \ldots$ where d is a positive integer giving the separation of the waveguide channels at their closest approach. In the case of the inverted lower U $j = 0$ or N for $l = -1,\ -2,\ -3, \ldots$. At the bottom, horizontal, region of the two U's the difference equations take the form

$$E_3(l,j) = \gamma[\alpha E_3(l,j) + \beta(E_3(l+1,j) + E_3(l-1,j)) + \delta E_3(l,j_0)] \tag{3.119b}$$

where $(j,j_0) = (d,0)$ or $(0,d)$ are the y-coordinates of the two different horizontal waveguide channels and $l = 1,\ 2,\ 3,\ \ldots,\ N-2,\ N-1$ marks off the x-coordinate along both horizontal channels. The joining conditions of the horizontal and vertical segments of the waveguides are provided by the four relations

$$E_3(0,d) = \gamma[\alpha E_3(0,d) + \beta(E_3(1,d) + E_3(0,d+1)) + \delta E_3(0,0)], \tag{3.119c}$$

$$E_3(0,0) = \gamma[\alpha E_3(0,0) + \beta(E_3(1,0) + E_3(0,-1)) + \delta E_3(0,d)], \tag{3.119d}$$

$$E_3(N,d) = \gamma[\alpha E_3(N,d) + \beta(E_3(N,d+1) + E_3(N-1,d)) + \delta E_3(N,0)], \tag{3.119e}$$

and

$$E_3(N,0) = \gamma[\alpha E_3(N,0) + \beta(E_3(N,-1) + E_3(N-1,0)) + \delta E_3(N,d)]. \quad (3.119f)$$

In (3.119a–3.119f) the weak interaction between the two different waveguide channels is given by δ. In the limit that $\delta = 0$ the two channels have no interactions between them and the two U's cease to interact with one another. The γ in (3.119) represent the waveguide parameter in (3.108) which characterizes the channels of each of the two waveguides.

The solution to (3.119) is obtained in the form

$$E_3(0,l) = a_0 e^{ipla} + b_0 e^{-pla}, \quad (3.120a)$$

where $l = d+1,\ d+2,\ d+3,\ldots;$

$$E_3(0,l) = c_0 e^{ipla} + d_0 e^{-ipla}, \quad (3.120b)$$

where $l = -1, -2, -3, \ldots;$

$$E_3(N,l) = re^{ip|l|a}, \quad (3.120c)$$

where $l = d+1,\ d+2,\ d+3,\ldots;$

$$E_3(N,l) = ue^{ip|l|a}, \quad (3.120d)$$

where $l = -1, -2, -3, \ldots;$

$$\begin{aligned} E_3(l.,d) = {} & [i\ \sin(qla)x + \cos(qla)x_1]e^{ikla} \\ & + [-i\ \sin(qla)y + \cos(qla)y_1]e^{-ikla}, \end{aligned} \quad (3.120e)$$

where $l = 0,\ 1,\ 2,\ 3, \ldots,\ N$; and

$$\begin{aligned} E_3(l.,0) = {} & [\cos(qla)x + i\ \sin(qla)x_1]e^{ikla} \\ & + [\cos(qla)y - i\ \sin(qla)y_1]e^{-ikla}, \end{aligned} \quad (3.120f)$$

where $l = 0,\ 1,\ 2,\ 3,\ \ldots,\ N$. In this solution b_0 and c_0 are, respectively, the amplitudes of the incident waves in the upper and lower waveguide channels, and the amplitudes r and u are, respectively, those of the transmitted waves in the upper and lower channels.

Upon substitution of (3.120) into (3.119) three conditions are found relating $\{k,\ p,\ q\}$ to the variables characterizing the waveguides and their interactions with one another. These are

$$\gamma = \frac{1}{\alpha + 2\beta \cos(pa)}, \tag{3.121a}$$

$$\gamma = \frac{1}{\alpha + 2\beta \cos(qa)\cos(ka)}, \tag{3.121b}$$

$$2\alpha \sin(qa)\sin(ka) = \delta, \tag{3.121c}$$

In addition, four equations are obtained which allow for the solution of reflected wave amplitudes $\{a_0, d_0\}$ and the transmitted wave amplitudes $\{r, u\}$ in terms of the incident wave amplitudes $\{b_0, c_0\}$.

For an illustration in Fig. 3.13 is plotted the transmission amplitudes in the low channel (solid line) and the upper channel (dashed line) for an incident wave incident on the coupler from the upper channel, i.e., $b_0 = 1.0$ and $c_0 = 0.0$. The plot is presented as a function of the coupler length N. For the further details of the study the reader is referred to the original paper.

The point of the results in Fig. 3.13 is that, due to the weak coupling between the channels, a mode sent into the coupler in one waveguide can emerge from the coupler in either of the waveguide channels or as partially transmitted in both channels. The type of transmission observed depends on the coupling strength between the waveguides and the length of the waveguide coupling region.

The preceding results have all been for systems composed of linear media. The theory can also be modified to treat Kerr nonlinear media [18, 21]. The presence of optical nonlinearity in the model greatly increases the types of excitations generated within the system.

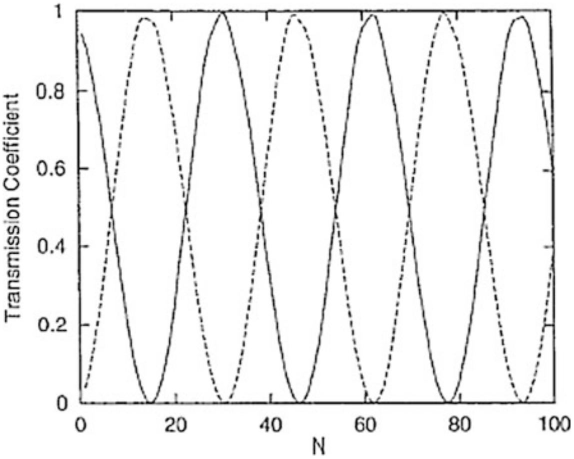


Fig. 3.13 Plot of the Transmission coefficients versus N for the waveguide coupler. Reproduced with permission from [18]. Copyright 2002 American Physical Society

Kerr Nonlinear Optical Media

In the case of Kerr nonlinear media, the electrical permittivity is dependent on the intensity of the applied electric field. This allows for the excitations of the system to become dependent on their field intensities. In addition, it allows new type of excitations to appear in the nonlinear system. These include bight, dark, and grey solitons [21, 22]. In discrete systems, these are often termed intrinsic localized modes.

Intrinsic localized modes can only exist in systems formed of nonlinear media. The localized modes are generated in these systems because the intrinsic localized mode field amplitudes induce changes in the system dielectric properties which in turn support the field amplitudes of the intrinsic localized modes. The development of the modes is then a completely self-consistent arrangement between the modes and the media supporting them. In the following it will be shown that photonic crystals composed of Kerr media support these new types of excitations.

An example of a single site Kerr impurity is straightaway developed from the form of the single site impurity introduced in (3.103). The Kerr impurity can be introduced into the problem by taking the change in dielectric due to the single site impurity represented in (3.103) to be given by the field dependent form

$$\delta\varepsilon\left(\vec{r}_{||}\right) = \delta\varepsilon_0\left[1+\lambda|E_3|^2\right] \quad \text{for } \left|\vec{r}_{||}\right| < R_\varepsilon, \delta\varepsilon\left(\vec{r}_{||}\right) = 0 \text{ otherwise} \tag{3.122}$$

where $\delta\varepsilon_0$ is a constant and R_ε is the radius of a region over which $E_3\left(\vec{r}_{||}|\omega\right)$ is slowly varying in space. In (3.122) the term $\lambda|E_3|^2$ represents the change in the dielectric due to the field $E_3\left(\vec{r}_{||}|\omega\right)$ at the impurity site. The form in (3.112) has been used to study various single impurity problems and, in a modification for multiple sites, various waveguide problems.

For the waveguide problem, the difference equation in (3.108) for the waveguide of linear optical media becomes, for the Kerr nonlinear system, a nonlinear difference equation of the form

$$E_3(n) = \gamma\left\{\begin{array}{l} \alpha\left[1+\lambda|E_3(n)|^2\right]E_3(n) \\ +\beta\left[1+\lambda|E_3(n+1)|^2\right]E_3(n+1) \\ +\beta\left[1+\lambda|E_3(n-1)|^2\right]E_3(n-1) \end{array}\right\}. \tag{3.123}$$

(Note that in the linear limit that $\lambda \to 0$, (3.123) is seen to reduce to (3.108). It can be shown that (3.123) for an infinite waveguide exhibits various bright, dark, and grey soliton-like intrinsic localized modes [21, 22] as new important excitations of the nonlinear system. For the conditions governing the existence of these modes and their detailed properties the reader is referred to the literature [21–23]. Here only a brief illustration of some of their basic properties is given by treating an interesting transmission problem involving intrinsic localized modes.

In the following an example of a barrier transmission problem for a Kerr media barrier contained within an otherwise linear media waveguide will be treated. This illustrates many of the interesting features of the renormalization of modes in the system due to the introduction of nonlinearity and of the existence of new intrinsic localized modes due to the introduction of nonlinearity into the system.

Waveguide Barrier of Kerr Optical Media

One of the problems treated using (3.108) and (3.123) is that of the scattering from a barrier of seven Kerr sites in and otherwise linear media waveguide [24]. The barrier is formed by replacing seven consecutive waveguide channel sites with Kerr media sites. An incident plane wave form is reflected and transmitted from the barrier.

The problem is solved by using the difference equations and the form of the transmitted wave to generate a recursive solution. In this way, the transmitted wave is traced back recursively through the barrier media and eventually emerges as an incident and reflected wave on the opposite side of the barrier media. Having related the incident, reflected and transmitted amplitudes to one another, the transmission and reflection coefficients are obtained from these.

In [24] the transmission and reflection coefficients were obtained and evaluated for a specific model. These results will now be presented and the interested reader referred to [24] for the details of the system. The results shown below illustrate a typical behavior of these type of barrier problems.

Figure 3.14 presents results for the barrier transmission coefficient versus $g = \gamma\alpha$, comparing linear and nonlinear limits of the Kerr barrier media. In each of the two plots reproduced here, the linear and nonlinear results are slightly shifted except in the region $0.75 < g < 0.8$. In the region $0.75 < g < 0.8$ a sharp peak of near perfect transmission is observed. These are the result of barrier transmission assisted by the excitation of an intrinsic localized mode. In these cases, the modes look like bright solitons. For further details of these studies the reader is referred to the original papers.

Aside from the intrinsic localized mode, the other barrier modes are present in both the linear and Kerr media systems. Only a renormalization is observed between the two systems.

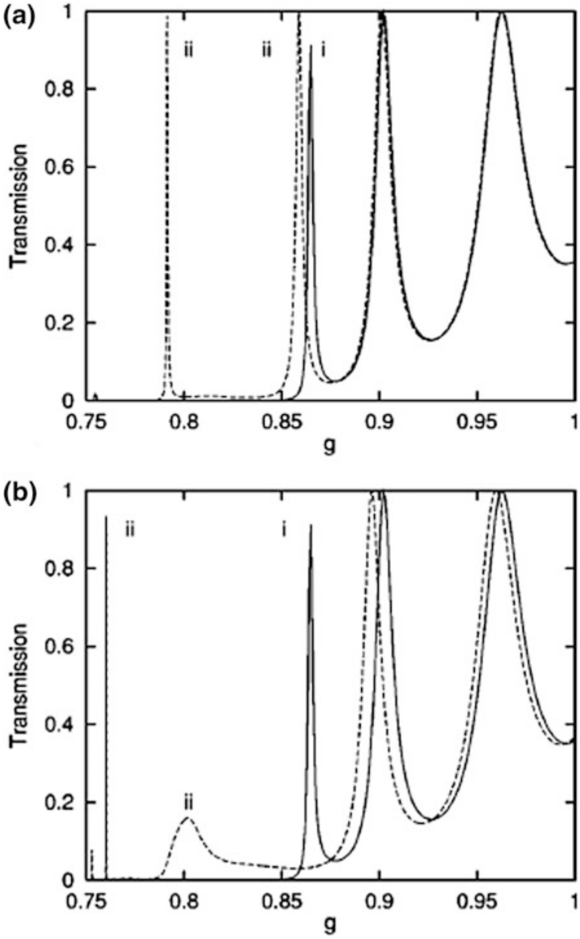


Fig. 3.14 Plots of the transmission coefficient versus $g = \gamma\alpha$ at $\frac{\omega a}{2\pi c} = 0.440$ at $ka = 2.5$ for a barrier of seven Kerr media sites. In **a** are results for the linear media limit of the barrier for $\lambda|E|^2 = 0.0$ (labeled i) and $\lambda|E|^2 = 0.00025$ (labeled ii) in terms related to the amplitude of the transmitted wave. In **b** are results for the linear media limit of the barrier for $\lambda|E|^2 = 0.0$ (labeled i) and $\lambda|E|^2 = 0.0015$ (labeled ii) in terms related to the amplitude of the transmitted wave. Reproduced with permission from [24]. Copyright 2004 American Physical Society

References

1. J.D. Joannopoulos, P.R. Vilenueve, S. Fan, *Photonic Crystals* (Princeton University Press, Princeton, 1995)
2. K. Sakoda, *Optical Properties of Photonic Crystals* (Springer, Berlin, 2001)
3. P.N. Favennec, *Photonic Crystals: Towards Nanoscale Photonic Devices* (Springer Verlag, Berlin, 2005)
4. A.R. McGurn, in *Survey of Semiconductor Physics,* ed. by W. Boer (Wiley, New York, 2002) Chapter 13
5. A.R. McGurn, *The Nonlinear Optics of Photonic Crystals and Meta-Materials* (Claypool & Morgan, San Refael, 2015)
6. P. Russell, Photonic crystal fibers. Science **229**, 358–362 (2003)
7. F. Poli, A. Cucinotta, S. Selleri, *Photonic Crystal Fibers: Properties and Applications* (Springer, Berlin, 2007)

8. A.A. Maradudin, A.R. McGurn, The photonic band structure of a truncated, two-dimensional, periodic medium. J. Opt. Soc. Am. B **10**, 307 (1993)
9. A.A. Maradudin, A.R. McGurn, Photonic band structures of two-dimensional dielectric media, in *Photonic Band Gaps and Localization*, ed. by C.M. Soukoulis, Nato ASI Series Series B: Physics vol. 308, (Plenum Press, New York, 1993), pp. 247–268
10. S.L. McCall, P.M. Platzmann, R. Dalichaouch, D. Smith, S. Schultz, Microwave propagation in two-dimensional dielectric lattices. Phys. Rev. Lett. **67**, 2017 (1991)
11. A.R. McGurn, Impurity mode techniques applied to the study of light sources. J. Phys. D App. Phys. **38**, 2338 (2005)
12. M.S. Dresselhaus, G. Dresselhaus, A. Jorio, *Group Theory: Applications to Condensed Matter Physics* (Springer, Berlin, 2007)
13. A. Gonis, *Green Functions for Ordered and Disordered Systems* (Noth-Holland, Amsterdam, 1992)
14. D.G. Duffy, *Green's Functions with Applications* (CRC Press, Boca Raton, 2015)
15. H.G. Algul, Y.F. Chang, Y. Zhong, A.R. McGurn, Impurity modes in 1-D periodic optical systems. Phys. B **205**, 19–23 (1995)
16. R.A. Smith, *Wave Mechanics of Crystalline Solids*, (Chapman and Hall, London, 1968) Chap. 11
17. G.B. Arfken, H.J. Weber, *Mathematical Methods for Physicists*, 6th edn. (Elsevier, Amsterdam, 2005)
18. A.R. McGurn, Photonic crystal circuits: localized modes and waveguide couplers. Phys. Rev. B **65**, 075406 (2002)
19. A.R. McGurn, Green's function theory for rows and periodic defect arrays in periodic band structures. Phys. Rev. B **53**, 7059 (1997)
20. H. Iach, H. Luth, *Solid-State Physics: An Introduction to Principles of Material Science* (Springer, Berlin, 2002)
21. A.R. McGurn, Intrinsic localized modes in nonlinear photonic crystal waveguides: dispersive modes. Phys. Lett. **A260**, 314 (1999)
22. A.R. McGurn, Intrinsic localized modes in nonlinear photonic crystal waveguides. Phys. Lett. A **251**, 322 (1999)
23. A.R. McGurn, *Nonlinear Optics of Photonic Crystals and Meta-Materials* (published by Morgan & Claypool for IOP Publishing as part of IOP Concise Physics, 2015) ISBN 978-1-6817-4107-9 (ebook), ISBN 978-1-6817-4043-0 (print) ISBN 978-1-6817-4235-9 (mobi), IOP Concise Physics ISSN 2053-2571 (online) ISSN 2054-7307 (print)
24. A.R. McGurn, G. Birkok, Transmission anomalies in Kerr media photonic crystal circuits: Intrinsic localized modes. Phys. Rev. B **69**, 235105 (2004)

Chapter 4
Plasmonics

Plasmonics is the study of surface electromagnetic waves at interfaces between two different media [1–7]. It is involved with how surface electromagnetic waves can be generated on an interface, how bulk electromagnetic modes interact with surface waves to couple in or out of the them, and how surface electromagnetic waves interaction with other types of excitations related to the surface. As shall be seen, surface electromagnetic waves are important as they can be used to study the structure and properties of a surface or to enhance some of the physical processes that take place on the surface. In this chapter, the basic properties of these excitations will be presented along with some discussions of possible technological applications that are proposed for them.

The focus will be on surface waves on planar surfaces [1–7], planar surfaces with localized structures on them [8], planar gratings [1–8], and randomly rough surfaces which are planar on average [1–7, 9, 10]. More exotic surfaces such as those on spheres, cylinders, etc. will not be a treated here though these can be very important in the explanation of interesting phenomena such as, for example, the optical glory and rainbow effects [10]. The focus will be on analytical methods as these tend to offer an insight into the physics of the basic surface wave mechanism more than computer simulation methods. Some references to simulations and experiments, however, will be provided but it is not meant to give a detailed review of these nor should the discussions of the analytical work be considered as an attempt at a detailed review of these methods [1–7].

In the discussions of the Maxwell equations at a planar interface between two media, a treatment is commonly given of the reflection and refraction of a wave incident on the interface. In most texts, however, it is often left out that there are other solutions of the Maxwell equations related to the interface. These are the surface electromagnetic waves known as surfaces plasmons or surface polaritons or surface plasmon-polaritons [1–11]. The surface electromagnetic waves are solutions that are bound to the interface and propagate parallel to the plane of the interface between the two different media. The waves are bound to the interface in the sense

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_4

that their field intensities decrease to zero at increasing perpendicular separation form the interface and their energy flow is parallel to the surface [1–11].

Because of the translation symmetry of the interface the surface wave solutions are completely distinct from the modes interacting with the interface form the bulk [1–7]. These two different types of solutions only mix when the translational symmetry of the surface is destroyed. When this occurs the existence of surface electromagnetic waves on the interface gives rise to many interesting effects in the diffuse scattering of radiation from the interface and are an essential element in understanding the scattering properties of the surface. In turn, interesting effects are also seen in the properties of the propagations of surface modes related to their interaction with bulk electromagnetic modes and these effects are also an essential element in the understanding of the surface electromagnetic waves on the interface.

Surface polaritons are the general form of surface wave solutions at the interface [1–7]. Often to simplify discussions the quasi-static limit of the surface polaritons are discussed. These are known as surface plasmons and their solutions are obtained by taking the speed of light to be infinite in the surface polariton solutions. Not all surfaces support surface plasmon-polariton excitations and there are specific sets of requirements on the dielectric properties of the two materials forming the interface for solutions of these types of modes to exist. In addition, surface imperfection and roughness couple the surface modes to bulk modes and along with dielectric losses from the media can lead to a finite lifetime or finite propagation distance along the interface for these modes. The topography of the surface also has interesting effects on the structure of the dispersion relation of plasmon-polaritons along the interface. Bulk and surface electromagnetic waves can also be coupled together through the use of prisms in the so-called Otto and Kretschmann couplings [1–7]. These are common methods for the experimental study of these excitations in many types of systems.

In the following the solutions for plasmon-polaritons on planar surfaces and thin films with planar surfaces will be treated. These will focus on the basic properties of cases involving simple examples of these systems. After this, discussions of effects on the plasmon-polariton from surface structure features on otherwise planar surfaces and from periodic surfaces are given. A focus will be on the band structures, frequencies of the bound modes at imperfections, and scattering into bulk modes. Next, effects in the scattering of light from rough surfaces and thin films with rough surfaces will be related to weak localization effects of the surface electromagnetic waves in these systems. Discussions of the nature of strong and weak localization in the disordered systems will be give so as to develop an understanding of the mechanisms contributing to Anderson localization effects [9, 10]. This is followed by a review of some of the technological applications of surface electromagnetic waves including: surface enhanced Raman scattering, enhanced transmission from plasmonic films, plasmonics of metamaterial surfaces, etc.

4.1 Surface Plasmon-Polaritons on a Planar Interface

In this section the wave function solutions of surface plasmon-polaritons propagating at the planar interface between two different media are discussed [1–7]. A general treatment is given, determining the form of the fields at and in the neighborhood of the surface between the media and the dispersion relations of the plasmon-polariton waves. These results are calculated and presented as functions of the dielectric properties of the two media supporting the surface wave. In addition, some numerical studies related to systems of technological interest are presented and discussed.

Not all planar surfaces support surface plasmon-polaritons because of restrictions from the electrodynamics [1–7]. As shall be seen later, there are a very specific set of conditions on the dielectric properties that the two media forming the interfaces must have in order for surface plasmon-polariton solutions to exist. These conditions also limit the form of the dispersion relations of the plasmon-polariton modes. Generally, however, surface plasmons-polaritons may exist along a great variety of interfaces between metals and dielectrics and between two different dielectrics. Both of these general types of surfaces (between metal-dielectric and dielectric-dielectric media) which support surface plasmon-polaritons exhibit qualitatively different forms of solutions [1–7]. These will now be studied.

Consider a planar interface between two media which is located at $x = 0$. One medium has dielectric constant $\varepsilon_{>}(\omega)$ in the region $x > 0$ and the second medium has dielectric constant $\varepsilon_{<}(\omega)$ in the region $x<0$ (see Fig. 4.1 for a schematic of the interface and the two media). A solution will be given for surface plasmon-polariton waves propagating along the interface in the z-direction. The solution is obtained from the Maxwell equations by applying surface wave boundary conditions at the interface to yield a modal wave having fields localized near the surface between the two media and vanishing at infinity [1–7].

The form of the Maxwell equations in the following considerations are written as

$$\nabla \cdot \vec{E} = 0, \tag{4.1a}$$

$$\nabla \cdot \vec{B} = 0, \tag{4.1b}$$

$$\nabla \times \vec{E} + \frac{1}{c}\frac{\partial \vec{B}}{\partial t} = 0, \tag{4.1c}$$

and

$$\nabla \times \vec{B} - \frac{\mu\varepsilon}{c}\frac{\partial \vec{E}}{\partial t} = 0. \tag{4.1d}$$

These are for the case in which there is no free charge in the system, and it is assumed in the calculation that no net free charge is treated within the problem. In

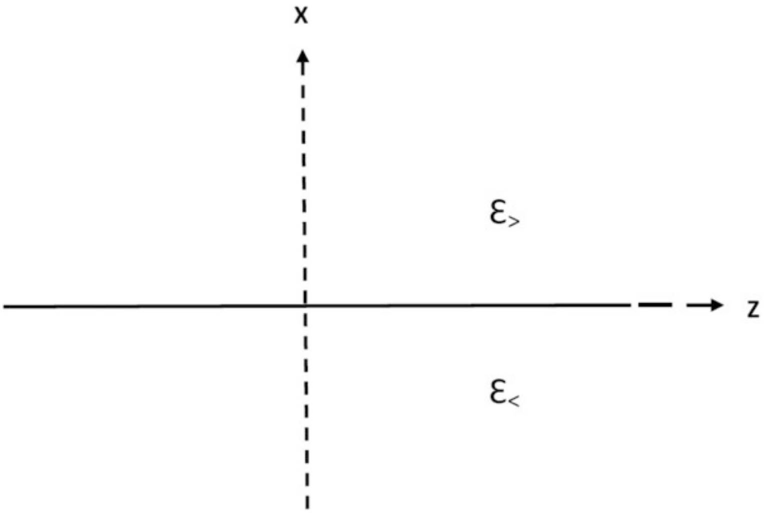


Fig. 4.1 Schematic of the planar interface between two semi-infinite bulk media that supports surface electromagnetic waves

addition, the current has been left off from Amperes law. This can be done by taking advantage of the ambiguity between the dielectric and current responses in a frequency dependent system. In the calculations presented later, the dielectric functions used are those which treat the total response of the system to frequency dependent fields as a dielectric response. Other formulations are available which treat the total response of the system as a current response or as a combination of dielectric and current responses. These last cases are not of interest here.

Under these conditions, from the two curl equations and Gauss's law the wave equation for the electric fields follows and is given by

$$\nabla \times \nabla \times \vec{E} + \frac{\mu\varepsilon}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} = -\nabla^2 \vec{E} + \frac{\mu\varepsilon}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} = 0 \tag{4.2}$$

with a similar wave equation obtained from (4.1b), (4.1c), and (4.1d) for the magnetic induction. These equations determine all of the bulk and surface waves in the system once the appropriate boundary conditions are applied for their solution.

The form of the solution of (4.2) for surface waves moving in the z-direction along the planar surface [1–7] between the two media are: for $x > 0$ and $\alpha_> > 0$

$$\vec{E}_>\left(\vec{r}, t\right) = \vec{E}^0_> \exp(-\alpha_> x)\exp[i(kz - \omega t)], \tag{4.3a}$$

and for $x < 0$ and $\alpha_< > 0$

$$\vec{E}_{<}\left(\vec{r},t\right) = \vec{E}^{0}_{<} \exp(\alpha_{<} x) \exp[i(kz - \omega t)]. \tag{4.3b}$$

Here from the wave equation in (4.2) applied in the regions above and below the interface

$$\alpha^2_{>} = k^2 - \frac{\mu\varepsilon_{>}}{c^2}\omega^2 \tag{4.4a}$$

and

$$\alpha^2_{<} = k^2 - \frac{\mu\varepsilon_{<}}{c^2}\omega^2, \tag{4.4b}$$

and the vector amplitudes of the waves are given by

$$\vec{E}^{0}_{>} = \left(E^{0}_{>,x}, E^{0}_{>,y}, E^{0}_{>,z}\right) \tag{4.5a}$$

and

$$\vec{E}^{0}_{<} = \left(E^{0}_{<,x}, E^{0}_{<,y}, E^{0}_{<,z}\right). \tag{4.5b}$$

In (4.3) the continuity of the electric field at the interface requires that the wave vector component parallel to the surface, k, is the same for the field solutions both above and below the interface [1–7]. In addition, for a wave solution to be bound to and localized on the interface, the electric field intensities must decay in directions perpendicularly away from the surface. The α's in (4.3) have been chosen accordingly for solutions localized about the interface.

Substituting (4.3) into Gauss's law it follows that the x and z components of the electric field are related to one another by

$$E^{0}_{>,x} = \frac{ik}{\alpha_{>}} E^{0}_{>,z}. \tag{4.6a}$$

$$E^{0}_{<,x} = -\frac{ik}{\alpha_{<}} E^{0}_{<,z}. \tag{4.6b}$$

Consequently, the general form of the surface plasmon-polarition along the interface for the fields above and below the surface are, respectively, given by

$$\vec{E}_{>}\left(\vec{r},t\right) = \left(\frac{ik}{\alpha_{>}} E^{0}_{>,z}, E^{0}_{>,y}, E^{0}_{>,z}\right) \exp(-\alpha_{>} x) \exp[i(kz - \omega t)] \tag{4.7a}$$

and

$$\vec{E}_<\left(\vec{r},t\right) = \left(-\frac{ik}{\alpha_<}E^0_{<,z}, E^0_{<,y}, E^0_{<,z}\right)\exp(\alpha_< x)\exp[i(kz-\omega t)] \tag{4.7b}$$

From Faraday's law the magnetic inductance corresponding to the electric fields in (4.7a) and (4.7b) are, respectively,

$$\vec{B}_>\left(\vec{r},t\right) = \frac{c}{\omega}\left(-kE^0_{>,y}, i\frac{\mu\varepsilon_>}{c^2}\frac{\omega^2}{\alpha_>}E^0_{>,z}, i\alpha_> E^0_{>,y}\right)\exp(-\alpha_> x)\exp[i(kz-\omega t)] \tag{4.8a}$$

and

$$\vec{B}_<\left(\vec{r},t\right) = \frac{c}{\omega}\left(-kE^0_{<,y}, -i\frac{\mu\varepsilon_<}{c^2}\frac{\omega^2}{\alpha_<}E^0_{<,z}, -i\alpha_< E^0_{<,y}\right)\exp(\alpha_< x)\exp[i(kz-\omega t)]. \tag{4.8b}$$

The fields in (4.7) and (4.8) are the general form of the surface wave modal solutions above and below the interface which now must be match with boundary conditions at the interface. This completes the determination of the fields and the dispersion relations of the modes.

The boundary conditions at the surface relate the fields of the solutions above and below the interface to one another. The conditions that hold at the interface are that the tangential components of the electric fields and the normal components of the electric displacement vectors are continuous. From the continuity of the electric fields parallel to the surface it follows that [5]

$$E^0_{>,y} = E^0_{<,y} \tag{4.9a}$$

and

$$E^0_{>,z} = E^0_{<,z}. \tag{4.9b}$$

From the continuity of the normal component of the displacement vector at the surface it follows that

$$\frac{\varepsilon_>(\omega)}{\varepsilon_<(\omega)} = -\frac{\alpha_>(\omega)}{\alpha_<(\omega)}. \tag{4.10}$$

Equation (4.9) completes the field solutions for both the electric fields and magnetic induction. Equation (4.10) provides a fundamental limitation on the system, yielding the conditions need for solutions to exist and the dispersion relation of the waves. It is interesting to note in (4.10) that surface waves only exist along the interface when $\varepsilon_>(\omega)$ and $\varepsilon_<(\omega)$ have opposite signs. The medium with the negative dielectric constant is, consequently, a reflective medium.

The continuity of the z component of the magnetic induction requires that [5]

$$(\alpha_> + \alpha_<)E^0_{>,y} = 0. \tag{4.11}$$

This along with (4.9a) requires that the y component of the electric field is zero so that the electric fields of the surface waves are in the x-z plane. The magnetic induction, on the other hand, is along the y-axis. The resulting surface plasmon-polariton, unlike waves in the bulk of the medium, is not a transverse wave.

The general solution for the wave functions and dispersion relations of surface plasmon-polaritons given in (4.1)–(4.11) above will now be studied for specific types of interface media. In one example an interface between a dielectric and a metal is treated. This is followed by a treatment of an interface between two different dielectric media. These two different types of media interfaces will be seen to exhibit distinctly different behaviors arising from differences in the frequency dependent dielectric functions of metals and insulators.

4.1.1 *Example of a Dielectric-Metal or Semiconductor Interface*

As an example of an important class of structures supporting surface plasmon-polaritons, considerations are given to the surface waves on a planar dielectric-metal or semiconductor interface [5]. For these interfaces, the dielectric of the metal and semiconductor systems will be treated within the context of the Drude model approximation. This gives a rough modeling of some of the general properties of these types of conductors. Consequently, both the metals and semiconductors studied here will be described by this same form of the dielectric response.

For these considerations a simple model is treated of an $x = 0$ interface separating a semi-infinite dielectric from a semi-infinite metal. The model consists of a non-conducting dielectric medium described by a frequency independent dielectric constant $\varepsilon > 0$ in the region $x > 0$. In the region $x < 0$ the medium is a metal with a frequency dependent dielectric constant.

In the following discussions, first the nature and origins of $\varepsilon_m(\omega)$ are explained. This is followed by a solution of the surface wave problem on the $x = 0$ interface.

The Nature of the Dielectric Response, $\varepsilon_m(\omega)$

The dielectric constant of the bulk metal is taken to be of a standard form given by [5, 7, 12, 13]

$$\varepsilon_m(\omega) = \varepsilon_\infty \left(1 - \frac{\omega_p^2}{\omega^2}\right). \tag{4.12}$$

Here ε_∞ is the dielectric response at infinite frequency and ω_p is the bulk plasma frequency. The bulk plasma frequency is the frequency at which collective modes of the conduction electrons (known as plasmons) occur in the bulk of the metal. The plasmons exist when $\varepsilon_m(\omega) = 0$ and are seen later to be closely related to the surface plasmon-polaritons.

The form of the metal dielectric constant in (4.12) is obtained from a simple model. In its most elementary form the model treats the conduction electrons of the metal as a gas of non-interacting freely moving classical particles. This is a very simplistic model but gives a functional form for the dielectric response that is the same as that found in more advanced treatments.

For an electric field applied in the z direction of an infinite bulk electron gas the equation of motion of the electrons in the free particle gas is [7, 12, 13]

$$m\frac{d^2z}{dt^2} = -eE \tag{4.13}$$

where $E(t) = E_0 \mathrm{e}^{-i\omega t}$ is the time-dependent applied electric field. From the solution of (4.13) for $z(t)$ a time-dependent polarization of the electron gas is obtained, having the form

$$P = -nez = -\frac{ne^2}{m\omega^2}E \tag{4.14}$$

where n is the electron carrier density and e is the positive fundamental unit of charge. The dielectric function resulting from (4.14) then follows as [13]

$$\varepsilon_m(\omega) = 1 + 4\pi\frac{P(\omega)}{E(\omega)} = 1 - \frac{\left(\omega_p^0\right)^2}{\omega^2}, \tag{4.15}$$

where $\omega_p^0 = \sqrt{\frac{4\pi ne^2}{m}}$ is the plasma frequency of the conduction electron gas.

In the result in (4.15) only the response of the gas of conduction electrons are accounted for, and other contributions need to be taken into account for a complete picture of the dielectric response of the metal. These additional contributions to the response of the metal arise from the positive ions forming the metal. Not only do the bound electrons of the positive ion background give a dielectric response in addition to that of the conduction electrons, but they also modulate the response of the condition electrons. The bound charge response of the positive ions is expected to be similar to the response of the dielectric media above the $x = 0$ interface, i.e., it gives a frequency independent contribution to the dielectric of the system. In addition, the contribution of the positive ions to the dielectric response modifies the plasma frequency of the electron gas in its response to the applied field.

The positive ion background in which the electrons move also contributes a dielectric response to the metal. The background response of the ions shall be denoted ε_P and approximated as frequency independent. It adds to the response of the conduction electrons in (4.15) to give the complete dielectric function of the metal. Making this addition to (4.15), after a little algebra the dielectric constant in (4.12) is obtained.

In (4.12) the combined frequency independent part of the dielectric response is given by [5, 7, 12, 13]

$$\varepsilon_\infty = \varepsilon_P + 1. \tag{4.16a}$$

In addition, the plasma frequency of the entire system is renormalized to have the form

$$\omega_P^2 = \frac{\left(\omega_P^0\right)^2}{\varepsilon_\infty}. \tag{4.16b}$$

This gives the frequency of the bulk plasma waves of the metal.

The resulting (4.12) provides a successful description of many of the features of the response of metals to frequency dependent applied electric fields. The description can be made to yield both qualitative and quantitative forms for the dielectric behavior of experimentally encountered systems or to serve as a curve fitting form for the dielectric constant data of metallic systems.

Some experimental data of typical values of the plasma frequency in metals, described by (4.16b), are listed in Table 4.1. These present a representative range of values found in metal systems and give an ideas of the energies associated with the dielectric response in (4.12) [5, 7, 12, 13].

The dielectric responses of the metal and dielectric are now used to obtain the surface plasmon-polariton modes along the $x = 0$ interface.

Surface Wave Dispersion Relation

The dispersion relation of the surface waves follows from substituting the dielectric constants for the dielectric, ε, and the metal, $\varepsilon_m(\omega) = \varepsilon_\infty\left(1 - \frac{\omega_p^2}{\omega^2}\right)$, into (4.10).

Table 4.1 Data for plasma frequencies in metals

Metal	Plasmon energy in eV
Li	7.12
Na	5.71
K	3.72
Mg	10.6
Al	15.3

Equation (4.10) then generates all of the modes at the interface studied as a function of frequency [5, 7, 12, 13].

Squaring both sides of (4.10) and using the relations in (4.4), it is found that [5, 7, 12–14]

$$\frac{\varepsilon^2}{\varepsilon_\infty^2\left(1-\frac{\omega_p^2}{\omega^2}\right)^2} = \frac{k^2 - \frac{\mu\varepsilon}{c^2}\omega^2}{k^2 - \frac{\mu\varepsilon_\infty}{c^2}\left(1-\frac{\omega_p^2}{\omega^2}\right)\omega^2}, \tag{4.17}$$

relating k^2 to ω^2 in terms of parameters characterizing the material forming the surface. Upon collecting the terms in k^2 to one side of the equation, it follows that

$$\left[\varepsilon_\infty\left(1-\frac{\omega_p^2}{\omega^2}\right)+\varepsilon\right]k^2 = \frac{\mu\varepsilon\varepsilon_\infty}{c^2}\left(\omega^2-\omega_p^2\right). \tag{4.18}$$

Equation (4.18) is then rewritten as a quadratic equation in ω^2 which is solved for the dispersion relation of the surface waves. After a little algebra, two solutions for ω^2 in terms of k^2 are obtained in the form

$$\omega_\pm^2 = \frac{1}{2\varepsilon_\infty}\left\{\varepsilon_\infty\omega_p^2 + \frac{\varepsilon+\varepsilon_\infty}{\mu\varepsilon}c^2k^2 \pm \left[\left(\varepsilon_\infty\omega_p^2 + \frac{\varepsilon+\varepsilon_\infty}{\mu\varepsilon}c^2k^2\right)^2 - 4\frac{\varepsilon_\infty^2}{\mu\varepsilon}\omega_p^2c^2k^2\right]^{1/2}\right\}, \tag{4.19}$$

exhibiting a dispersion for the surface electromagnetic modes at the interface for all values of positive and negative wave vector k.

To understand the nature of the modes described by (4.19) begin by looking at the limiting form of the dispersion relation as $k^2 \to \infty$ and as $k^2 \to 0$. In the limit that $k^2 \to \infty$, (4.19) gives [5, 7, 12–14]

$$\omega_+^2 \to \frac{\varepsilon+\varepsilon_\infty}{\mu\varepsilon\varepsilon_\infty}c^2k^2 \tag{4.20a}$$

and

$$\omega_-^2 \to \frac{\omega_p^2}{1+\frac{\varepsilon}{\varepsilon_\infty}}. \tag{4.20b}$$

In the limit that $k^2 \to 0$, (4.19) gives

$$\omega_+^2 \to \omega_p^2 \tag{4.21a}$$

and

$$\omega_-^2 \to 0. \tag{4.21b}$$

As is seen from these limits and is confirmed by the numerical results presented later, the ω_-^2 branch of the dispersion is the branch associated with the surface plasmon-polaritons waves propagating along the interface. It is bounded above by the limit in (4.20b) and approaches zero at zero wave vector, representing the greatest departure from the dispersion relations of the bulk modes. In addition, from (4.7) and (4.8) the wave functions are strongly localized around the interface between the two media, and the localization increases with increasing wave vector.

The other ω_+^2 branch is bounded from below by the limit in (4.21a) and as the wave vector becomes infinite ω_+^2 approached infinity, merging with the light line of the bulk modes. The ω_+^2 modes are, generally, light like and eventually merge with the bulk modes. In addition, because $\frac{\varepsilon_m(\omega)}{\varepsilon} > 0$ over the frequency range of the ω_+^2 modes they do not satisfy the condition in (4.10) required of surface waves to exist [5, 7, 12–14].

The plasmon-polariton branch of the dispersion relation is evaluated for a vacuum-InSb interface and presented in Fig. 4.2 as a plot of ω_- versus wave vector k [14]. [Here 14 is followed in approximating the response of n doped InSb by the Drude form in (4.12).] It is seen from the plot that all along the dispersion relation the plasmon-polariton branch falls below the $\omega(k) = ck$ light line of the bulk modes of light. As the wave vector k increases the frequency of the plasmon-polariton is found to rise quickly from zero and approach the $\omega_- = \frac{\omega_p}{\sqrt{1+\frac{\varepsilon}{\varepsilon_\infty}}}$ limit at $k \to \pm\infty$.

For small wave vectors near $k \approx 0$ the dispersion relation is a linear form [5, 7, 12–14]

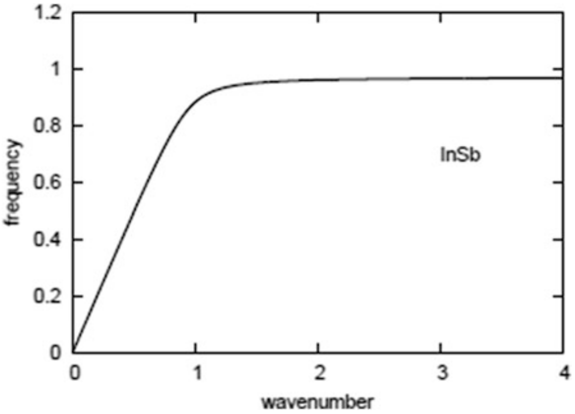


Fig. 4.2 The dispersion relation of the surface plasmon-polariton propagating along a planar vacuum-InSb n-doped semiconductor interface [5, 13]. Plotted is $\frac{\omega_-}{\omega_p}$ versus $\frac{ck}{\omega_p}$

$$\omega_{-} \approx \frac{1}{\sqrt{\mu\varepsilon}} ck \tag{4.22}$$

which is just below the light line $\omega(k) = ck$ of the dielectric. The modal wave functions in this limit look like perturbed bulk modes of light, but all of the surface waves along the surface plasmon-polartion branch have fields localized about the interface. In addition, the degree of field localization of the surface waves increases as k increases along the dispersion relation plot.

For all values of k, the surface wave dispersion relation is distinctly different from the dispersion of the bulk light modes. Consequently, the plasmon-polariton modes do not interact with the light modes, and are distinctly different excitations from the bulk light modes in the system.

This is no longer true if the translational symmetry of the interface is broken so that the excitations at the surface are not characterized as states of wave vector k. With the loss of translational symmetry the surface and bulk modes mix with one another and the surface plasmon-polaritons scatter from the surface and into bulk modes of the vacuum. This causes the surface waves to develop a finite life time in their propagation along the interface. The lifetime of the surfaces waves decreases with the increase of surface disorder and with the consequent larger mixing of surface and bulk excitations.

As the wave vector is increased from zero in Fig. 4.2, the surface plasmon-polariton branch of modes quickly approaches the $\omega_{-} = \frac{\omega_p}{\sqrt{1+\frac{1}{\varepsilon_\infty}}}$ limit. As a result, the frequencies of the plasmon-polariton modes are seen to be closely related to the frequency of the bulk plasmon excitations of the metal which occur at the plasma frequency ω_p.

Traveling along the branch of surface waves with increasing wave vector, the surface modes in the flat branch of excitations approaching the $\omega_{-} = \frac{\omega_p}{\sqrt{1+\frac{1}{\varepsilon_\infty}}}$ limit differ a great deal from the bulk propagating light modes. They look more like plasmon excitations and less like light modes as the dispersion bends farther and farther away from the light line and the modal fields are increasingly localized on the interface. The modes in this part of the surface plasmon-polariton dispersion relation are farther from the light line than the modes near $k \approx 0$. Consequently, they are more stable against scattering into bulk light modes which radiate away from the interface. These are the modes of most interest in the development of surface plasmon-polariton effects [5, 7, 12–14].

4.1.2 Example of a Dielectric-Dielectric Interface

Another example of an important class of surface plasmon-polaritons are those surface electromagnetic waves supported on a planar interface between two different semi-infinite dielectrics [5, 14]. These type of surface plasmon-polaritons are associated with dielectric resonances arising from the interaction of the electric fields with phonon excitations in one of the dielectric media. They are commonly found in systems in which one of the media (i.e., the resonant medium) is an ionic material.

In a simple treatment of the dielectric-dielectric planar interface, a non-conducting dielectric medium described by a frequency independent dielectric constant $\varepsilon > 0$ is in the region $x > 0$. In the region $x < 0$, however, the medium is taken to be a non-conducting dielectric with a frequency dependent dielectric function $\varepsilon_p(\omega)$ displaying frequency resonances. In the following, a discussion is given of the nature of dielectric systems with resonances and the nature of their dielectric functions. This is followed by a treatment of the plasmon-polaritons at the interface described earlier.

The Nature of the Frequency Dependent Response, $\varepsilon_p(\omega)$

The resonances are associated with the strong coupling of the electromagnetic fields to the phonon modes of the lattice that are commonly found in some types of dielectric materials. An example of such a material is an ionic crystal. In these systems the transverse optical phonon modes of the material represent polarization waves of counter vibrating positive and negative ions. The counter vibrating ions give rise to a time dependent polarization which couples with applied electric fields.

The standard form of the dielectric function in systems with strong coupling to the vibrational polarization modes is given by [7, 13]

$$\varepsilon_p(\omega) = \varepsilon_\infty \left(1 - \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2} \right). \tag{4.23}$$

Here ε_∞ is the dielectric response at infinite frequency, ω_{TO} is the frequency of the transverse phonon modes of the material, and ω_{LO} is the frequency of the longitudinal phonon modes of the material. The transverse optical phonons are modes involving counter moving positive and negative ions which travel in directions perpendicular to the wave vector of the transverse optical phonon. In the longitudinal optical modes the counter moving positive and negative ions travel in the direction parallel to the wave vector of the longitudinal optical phonon.

The form of the dielectric constant in (4.23) is obtained from a simple model considering an infrared electromagnetic wave interacting with an ionic crystal. An infrared electromagnetic wave applied to an ionic medium is a transverse wave with a wavelength that is slowly varying over typical inter-ionic separations in the crystal. The transverse nature of the electromagnetic wave means that it will most strongly couple to transverse polarization waves of the medium, and its long wavelength favors strong interactions only with the long wavelength polarization waves of the medium.

As a simplistic consideration, the essential features of the system can be described as a harmonically varying applied electric field interacting with the polarization of an ionic crystal. In this interaction the electric field is taken to couple only with the long wavelength transverse optical modes of the crystal, and the shorter wavelength modes of the ionic polarization are ignored. The coupling is through the electromagnetic interaction of the field with the counter propagating ions which form the time varying polarization of the optical mode.

A simple model for this interaction is a driven harmonic oscillator equation for the ionic polarization given by [7]

$$\frac{d^2P}{dt^2} + \omega_{TO}^2 P = \frac{ne^2}{\mu} E \tag{4.24}$$

where $E(t) = E_0 e^{-i\omega t}$ is the time-dependent applied electric field, and the ionic polarization vector of the transverse optical modes of the crystal is

$$P = -nez \tag{4.25}$$

Here n is the electron carrier density, e is the positive fundamental unit of charge, μ is the reduced mass of the two counter-propagating ions, and z is the relative displacement from equilibrium of the ions in the polarization. Notice that in this model the modes of all the counter moving ions in the system move in synchronization.

Using the definition of the polarization in (4.25), (4.24) can be rewritten in the form of an harmonic oscillator equation for z (the separation of the ions from equilibrium) subject to a forcing term from the applied electric field. This equation is [7, 13]

$$\mu \frac{d^2z}{dt^2} + \mu\omega_{TO}^2 z = -eE. \tag{4.26}$$

It is essentially the equation studied for the dielectric response of a metal but with an added harmonic oscillator term. The charge in the ionic dielectric is bound ionic charge subject to harmonic motion and not free charge as that found in the metallic dielectric function.

In addition, the polarization interacts with the electric field through the electromagnetic wave equations. To see the nature of the interaction, consider the wave equation for a wave in a general bulk dielectric medium of dielectric constant ε. It has the form

$$\nabla^2 \vec{E} - \frac{1}{c^2}\frac{\partial^2 \varepsilon \vec{E}}{\partial t^2} = 0. \tag{4.27a}$$

By applying the standard relations $\vec{D} = \varepsilon\vec{E} = E + 4\pi\vec{P}$ to (4.27a) it is rewritten into the form

$$\nabla^2 \vec{E} - \frac{1}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 \vec{P}}{\partial t^2}. \tag{4.27b}$$

This equations describes the field arising from the dynamics of the polarization and the polarization arising from the field. The nature of the dynamics of the coupled electric and polarization fields can now be determined from (4.24) and (4.27).

Upon substituting plane wave forms proportional to $e^{i\left(\vec{k}\cdot\vec{r}-\omega t\right)}$ for the field and polarization, from (4.27) it follows that

$$k^2\vec{E} = \omega^2\left(\vec{E}+4\pi\vec{P}\right). \tag{4.28}$$

In the long wavelength, $k \to 0$, limit (4.24) and (4.27) yield solutions for the electric field and polarization that are proportional to $e^{i\left(\vec{k}\cdot\vec{r}-\omega t\right)}$.

Substituting these plane wave forms in (4.24) and (4.28) results in the matrix equation for $E(\omega)$ and $P(\omega)$ given by [7, 14]

$$\begin{vmatrix} \omega^2 & 4\pi\omega^2 \\ ne^2/\mu & \omega^2-\omega_{TO}^2 \end{vmatrix} \begin{vmatrix} E \\ P \end{vmatrix} = 0 \tag{4.29}$$

where E and P are the plane wave amplitudes of the parallel or antiparallel electric field and polarization responses of the material.

The matrix equation is solved for $E(\omega)$ and $P(\omega)$ which are the amplitudes of the field and polarization, respectively. Setting the determinant of (4.29) equal to zero gives two eigenvalues. The first is

$$\omega^2 = 0 \tag{4.30a}$$

which represents a $k = 0$ bulk light mode of the general material with $E(\omega = 0) = \frac{\mu}{ne^2}\omega_{TO}^2 P(\omega = 0)$. A second eigenvalue

$$\omega^2 = \omega_{TO}^2 + \frac{4\pi ne^2}{\mu} \tag{4.30b}$$

represents a bulk plasmon-polariton mode with $E(\omega) = -4\pi P(\omega)$. Notice that the renomalized bulk mode of light in (4.30a) has parallel electric field and polarization vectors while the plasmon-polariton mode in (4.30b) has anti-parallel electric field and polarization vectors.

The solutions in (4.30) represent the $k = 0$ modes in the bulk dielectric medium. Later, as was the case with the bulk plasma waves in the problem of the dielectric-metal interface, these modes will be seen to be important in understanding the surface plasmon-polarition arising on the dielectric-dielectric interface. The next consideration is the determination of the general frequency dependent dielectric response of the bulk system.

From the (4.24) the dielectric response of the ionic medium to electromagnetic waves of frequency ω are obtained. Using standard expressions for the linear dielectric response, it follows that the dielectric function of the relative motion of the ions is given by [7, 12, 13]

$$\varepsilon_{po}(\omega) = 1 + 4\pi \frac{P(\omega)}{E(\omega)} = 1 + \frac{4\pi ne^2}{\mu\left(\omega_{TO}^2 - \omega^2\right)}. \tag{4.31}$$

This represents the resonant response of the ionic medium due to the movement of the positive and negative ions relative to one another. It is not, however, the only feature to the response of the ionic medium to an externally applied field. The ionic crystal contains additional dynamics processes than those related to the relative motion of the ions composing it.

Equation (4.31) only represents the response of the ions of the material in their harmonic motion relative to one another. This is not the complete response of the ionic crystal. Just as in the electron gas model, there are additional contributions from the bound electrons in the ions themselves. As in the problem of the dielectric function of the metal, the bound charge of each ion gives a constant frequency independent contribution to the dielectric of the material. This additional response adds to the dielectric in (4.31). Consequently, the dielectric function for the total response of the ionic crystal can be written in the form [7, 12, 13]

$$\varepsilon_p(\omega) = \varepsilon_\infty + \frac{4\pi ne^2}{\mu\left(\omega_{TO}^2 - \omega^2\right)}. \tag{4.32}$$

Here ε_∞ is the frequency independent dielectric of the total ionic material. It also is the $\omega \to \infty$ limit of the dielectric function.

A further simplification of (4.32) can be made, putting it into a more illuminating form. At zero frequency (4.32) becomes

$$\varepsilon_p(0) = \varepsilon_\infty + \frac{4\pi ne^2}{\mu\omega_{TO}^2} \tag{4.33}$$

so that (4.32) is rewritten as [7, 12, 13]

$$\varepsilon_p(\omega) = \varepsilon_\infty + [\varepsilon(0) - \varepsilon_\infty]\frac{\omega_{TO}^2}{\omega_{TO}^2 - \omega^2}. \tag{4.34}$$

From (4.34) and Gauss's law it follow that the longitudinal electromagnetic modes of the ionic crystal are obtained as the solution of [7, 12, 13]

$$\varepsilon_p(\omega_{TO}) = 0. \tag{4.35}$$

Solving (4.35) for ω_{TO}^2 and applying it to rewrite (4.34) gives the form of the dielectric function of the ionic material in (4.23). Equation (4.23) for the total response of the ionic crystal response is used in the later discussions describing the dielectric response of the medium below the $x = 0$ interface. The medium above the surface is assumed either to have no dielectric resonances or to have resonances at frequency other than those considered later.

The resulting dielectric function for the ionic material in (4.23) provides a successful description of many of the features of the response of ionic media to frequency dependent applied electric fields. The description can be made to yield both qualitative and quantitative forms for the dielectric behavior of experimentally encounter systems or to serve as a curve fitting form for the dielectric constant data of ionic systems.

Experimental data of typical values of the parameters in (4.23) for representative ionic materials are listed in Table 4.2 [7]. Some of these are used in the following discussions of the surface waves on the dielectric-dielectric interfaces.

Surface Wave Dispersion

The dielectric response in (4.23) is now used to obtain the surface plasmon-polariton modes along the $x = 0$ interface.

In the following example the dispersion relation of the surface plasmon-polariton at the interface is computed and the wave functions of the plasmon-polariton are discussed. A model is studied in which the dielectric medium in the region $x > 0$ is described by a frequency independent dielectric constant, while the medium in the region $x < 0$ is described by the resonant dielectric functional form of (4.23).

The dispersion relation of the surface waves follows from substituting the dielectric constants for the dielectric above the interface (described by ε, a frequency independent constant) and the dielectric below the interface (described by the resonant form $\varepsilon_d(\omega) = \varepsilon_\infty\left(1 - \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right)$) into (4.10). Equation (4.10) then generates all of the modes at the interface studied as a function of their frequency.

Squaring both sides of (4.10) and using the relations in (4.4), it is found that [13]

$$\frac{\varepsilon^2}{\varepsilon_\infty^2\left(1 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right)^2} = \frac{k^2 - \frac{\mu\varepsilon}{c^2}\omega^2}{k^2 - \frac{\mu\varepsilon_\infty}{c^2}\left(1 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right)\omega^2}, \tag{4.36}$$

relating k^2 to ω^2 in terms of parameters characterizing the material forming the surface. Upon collecting the terms in k^2 to one side of the equation, it follows that [13]

$$\left[\varepsilon_\infty\left(1 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right) + \varepsilon\right]k^2 = \frac{\mu\varepsilon\varepsilon_\infty}{c^2}\omega^2\left(1 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right) \tag{4.37}$$

Table 4.2 Some values of parameters in Equation 4.23

Material	ε_∞	ω_{TO} in 10^{13} s^{-1}	ω_{LO} in 10^{13} s^{-1}
NaCl	2.25	3.1	5.0
GaAs	10.9	5.1	5.5
Si	11.7	9.9	9.9
GaSb	14.4	4.3	4.6
LiF	1.9	5.8	12.0

Equation (4.37) is then rewritten as a quadratic equation in ω^2 which is solved for the dispersion relation of the surface waves. After a little algebra, two solution for ω^2 in terms of k^2 are obtained in the form

$$\omega_{\pm}^2 = \frac{1}{2\varepsilon_\infty}\left\{\varepsilon_\infty\omega_{LO}^2 + \frac{\varepsilon+\varepsilon_\infty}{\mu\varepsilon}c^2k^2 \pm \left[\left(\varepsilon_\infty\omega_{LO}^2 + \frac{\varepsilon+\varepsilon_\infty}{\mu\varepsilon}c^2k^2\right)^2 - 4\frac{\varepsilon_\infty}{\mu\varepsilon}\left(\varepsilon_\infty\omega_{LO}^2 + \varepsilon\omega_{TO}^2\right)c^2k^2\right]^{1/2}\right\}, \tag{4.38}$$

exhibiting a dispersion for the electromagnetic modes at the interface for values of positive and negative wave vector k.

The ω_-^2 solution in (4.38) is the branch of interest in the study of surface plasmon-polaritons on the dielectric-dielectric interface. From (4.7) and (4.8), it is seen to correspond to wave functions having fields localized at the interface. It also satisfies the condition in (4.10) which requires a negative ionic dielectric constant for there to be surface waves on the interface. With increasing wave vectors, the fields of the surface waves in these solutions become increasingly localized at the interface.

The ω_+^2 solution in (4.38) is not of interest in the discussion of plasmon-politons. It does not satisfy the conditions in (4.10) for the modes to exist as surface waves. As with the metal-vacuum ω_+^2 solutions, it will not be treated further.

In Fig. 4.3 a plot of ω_- versus k is made for a vacuum-InSb interface. (In this case the system is pure, undoped, InSb.) The modal dispersion relation looks similar to the surface plasmon-polariton dispersion relation at the vacuum metal interface except that there is a region of k over which the surface waves do not exist [13].

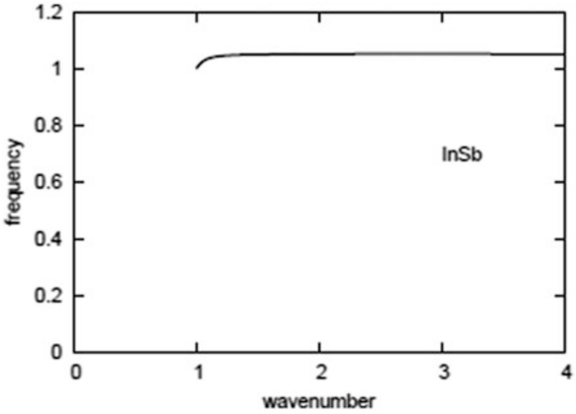


Fig. 4.3 Plot of $\frac{\omega_-}{\omega_{TO}}$ versus $\frac{ck}{\omega_{TO}}$ for surface plasmon-polaritons on a vacuum-InSb interface [14]. In this case the InSb is not doped [13]

Real solutions of (4.38) only can exist over the region of frequency $\omega_{TO} \leq \omega \leq \omega_{LO}$. This is the region over which the frequency dependent dielectric function is negative. In these considerations, it should be noted that the Lyddane-Sach-Teller relation [7],

$$\frac{\omega_{LO}^2}{\omega_{TO}^2} = \frac{\varepsilon(0)}{\varepsilon(\infty)}, \tag{4.39}$$

relates the longitudinal and transverse mode frequencies of a general ionic material to the frequency dependent dielectric function of the material at zero and infinite frequencies. The dielectric ratio on the left of (4.39) is, consequently, greater than 1.

The asymptotic value of the dispersion relation as $k \to \infty$ is [7, 13]

$$\omega_{-}(k \to \infty) = \left[\omega_{TO}^2 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{1 + \frac{\varepsilon}{\varepsilon_\infty}}\right]^{1/2}, \tag{4.40}$$

which is bounded above by ω_{LO}. In the ionic system, a lower bound on the wave vector arises from the condition that $\omega_{-} \geq \omega_{TO}$ in order that the dielectric function in (4.23) be negative. This lower limit on the wave vector is a new feature of the system arising from the dielectric form in (4.23). In the case of the metallic interface, the metallic dielectric function is negative for all frequencies less than ω_p. Consequently, in the metal-vacuum system solutions for surface waves exist at all wave vectors while the ionic-vacuum system there are gaps in the wave vector space which support surface waves.

4.1.3 Example of a Metallic Slab in Vacuum

As shall be seen later, many of the technologies of plasmonics involve systems which are based on thin film geometries [13–15]. Consequently, it is a useful example to consider the nature of the plasmon-polaritons on a slab of material in vacuum. This geometry exhibits a complexity of plasmon-polariton modes arising from the interaction of the surface electromagnetic modes on the two surfaces of the slab. As a simple example, illustrating many features found in the general variety of slab systems, a metallic slab in vacuum is treated.

The geometry of the slab is shown in the schematic drawing in Fig. 4.4. For the slab of thickness d, the slab surfaces are taken at $x = \frac{d}{2}$ and $-\frac{d}{2}$ so that the slab has reflection symmetry in the y-z plane. This is a facilitation in generating the modes propagating along the slab and which are found to have symmetric and anti-symmetric wave functions under reflection through the y-z plane. In the region $\frac{d}{2} \geq x \geq -\frac{d}{2}$ the metallic medium forming the slab has a dielectric constant

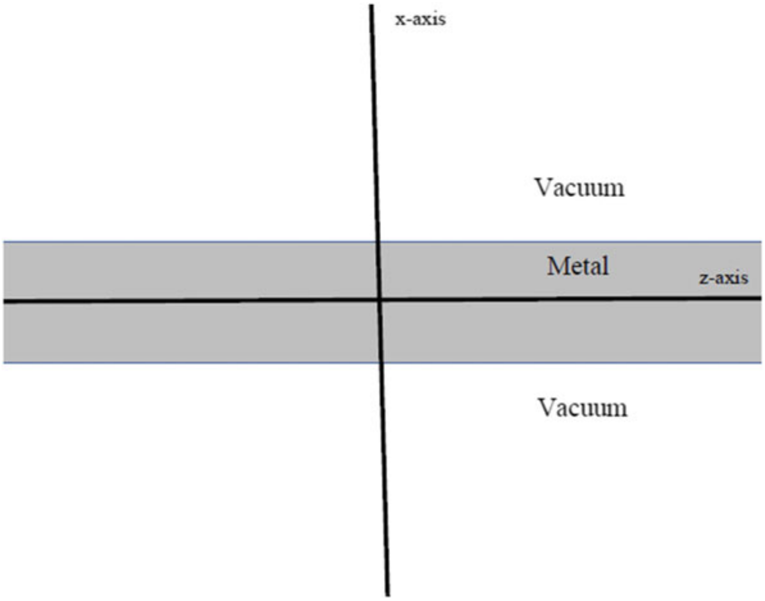


Fig. 4.4 Schematic of a metal slab surrounded by vacuum

$$\varepsilon_m(\omega) \tag{4.41}$$

given by (4.12). Outside of the slab is vacuum.

Within the slab the electric field solutions can be classified into solutions which have z-components that are symmetric about the y-z plane and solutions which have z-components that are anti-symmetric about the y-z-plane. This separation comes from the reflection symmetry of the slab about the $x = 0$ plane. Each of these modes are now separately treated starting with the symmetric modes [13].

Symmetric Solutions

For the symmetric modes propagating along the z-axis the wave function in the vacuum above the slab are from (4.7a) given by [13]

$$\vec{E}_{>}\left(\vec{r},t\right) = \left(\frac{ik}{\alpha}E^{0}_{>,z}, 0, E^{0}_{>,z}\right)\exp\left(-\alpha x\right)\exp\left[i(kz-\omega t)\right] \tag{4.42a}$$

Here the condition in (4.11) has been used to set the y-component of the electric field to zero. Below the slab, upon applying the same condition in (4.11) to set the y-component of the electric field to zero, the electric field is from (4.7b) given by

$$\vec{E}_{<}\left(\vec{r},t\right)=\left(-\frac{ik}{\alpha}E^{0}_{<,z},0,E^{0}_{<,z}\right)\,\exp\left(\alpha x\right)\,\exp\left[i(kz-\omega t)\right]. \tag{4.42b}$$

In both (4.42) for the vacuum above and below the slab

$$\alpha^{2}=k^{2}-\frac{1}{c^{2}}\omega^{2}. \tag{4.43}$$

The symmetric solutions within the slab are of the form

$$\vec{E}^{sym}_{in}\left(\vec{r},t\right)=\left(-\frac{ik}{p}E^{0}_{in,z}\,\sin\,px,0,E^{0}_{in,z}\cos\,px\right)\,\exp\left[i(kz-\omega t)\right], \tag{4.44a}$$

where

$$k^{2}+p^{2}=\varepsilon_{m}(\omega)\frac{\omega^{2}}{c^{2}} \tag{4.44b}$$

for the consideration of propagating modes within the slab. The forms in (4.42)–(4.44) are now matched at the upper and lower surfaces of the slab.

At the upper and lower surfaces the continuity of the component of electric field parallel to the slab surfaces requires that [13]

$$E^{0}_{in,z}\cos\frac{pd}{2}=E^{0}_{>,z}\mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.45a}$$

and

$$E^{0}_{in,z}\cos\frac{pd}{2}=E^{0}_{<,z}\mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.45b}$$

respectively. At the upper and lower surfaces the continuity of the component of electric displacement field normal to the slab surfaces sets the conditions

$$-\frac{\varepsilon_{m}(\omega)}{p}E^{0}_{in,z}\sin\frac{pd}{2}=\frac{1}{\alpha}E^{0}_{>,z}\mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.46a}$$

and

$$\frac{\varepsilon_{m}(\omega)}{p}E^{0}_{in,z}\sin\frac{pd}{2}=-\frac{1}{\alpha}E^{0}_{<,z}\mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.46b}$$

respectively.

Dividing (4.45a) by (4.46a) and (4.45b) by (4.46b) it follows that [13]

$$\varepsilon_m(\omega) = -\frac{p}{\alpha} \cot p\frac{d}{2}. \tag{4.47}$$

Equation (4.47) is the dispersion relation relating ω to k for the modes propagating along the slab.

Antisymmetric Solutions

For the antisymmetric modes propagating along the z-axis the wave function in the vacuum above the slab is from (4.7a) given by [13]

$$\vec{E}_>\left(\vec{r}, t\right) = \left(-\frac{ik}{\alpha} E^0_{>,z}, 0, -E^0_{>,z}\right) \exp\left(-\alpha x\right) \exp\left[i(kz - \omega t)\right] \tag{4.48a}$$

In (4.48a) the condition in (4.11) has, again, been used to set the y-component of the electric field to zero. Below the slab, upon applying the same condition in (4.11) to set the y-component of the electric field to zero, the electric field form in (4.7b) gives [13]

$$\vec{E}_<\left(\vec{r}, t\right) = \left(-\frac{ik}{\alpha} E^0_{<,z}, 0, E^0_{<,z}\right) \exp(\alpha x) \exp[i(kz - \omega t)]. \tag{4.48b}$$

In both (4.48)

$$\alpha^2 = k^2 - \frac{\omega^2}{c^2}. \tag{4.49}$$

The antisymmetric solutions within the slab are of the form

$$\vec{E}^{anti}_{in}\left(\vec{r}, t\right) = \left(-\frac{k}{p} E^0_{in,z} \cos px, 0, iE^0_{in,z} \sin px\right) \exp[i(kz - \omega t)] \tag{4.50a}$$

where

$$k^2 + p^2 = \varepsilon_m(\omega) \frac{\omega^2}{c^2} \tag{4.50b}$$

for propagating modes within the slab. The forms in (4.48)–(4.50) are now matched through the boundary conditions at the upper and lower surfaces of the slab.

At the upper and lower surfaces of the slab the continuity of the component of electric field parallel to the slab surfaces gives the conditions [13]

$$iE^0_{in,z} \sin\frac{pd}{2} = -E^0_{>,z} e^{-\alpha\frac{d}{2}}, \tag{4.51a}$$

and

$$iE^0_{in,z} \sin \frac{pd}{2} = -E^0_{<,z} \mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.51b}$$

respectively. At the upper and lower surfaces of the slab the continuity of the component of electric displacement field normal to the slab surfaces requires that [13]

$$\frac{\varepsilon_m(\omega)}{p} E^0_{in,z} \cos \frac{pd}{2} = \frac{i}{\alpha} E^0_{>,z} \mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.52a}$$

and

$$\frac{\varepsilon_m(\omega)}{p} E^0_{in,z} \cos \frac{pd}{2} = \frac{i}{\alpha} E^0_{<,z} \mathrm{e}^{-\alpha\frac{d}{2}}, \tag{4.52b}$$

respectively.

Dividing (4.51a) by (4.52a) and (4.51b) by (4.52b) it follows that [13]

$$\varepsilon_m(\omega) = \frac{p}{\alpha} \tan p \frac{d}{2}. \tag{4.53}$$

The solutions of (4.53) set the dispersion relation (ω as a function of k) of the antisymmetric modes.

Exponential Solutions within the Slab

The solutions presented earlier for the symmetric and antisymmetric modes represent one of two possible cases in the study of the slab modes. The solutions presented are based on forms for the fields within the slab involving sine and cosines in the x coordinate. In addition to these solutions, there are also solutions for the slab modes based on fields within the slab described by exponential forms in the x coordinate. The solutions for this case follow the same steps as the solutions presented above and will not be considered in detail here. The results for these modes, however, are summarized in the following.

The solutions of the slab modes involving exponents of the x coordinate have fields within the slab that for symmetric modes are given by

$$\vec{E}^{sym}_{in}\left(\vec{r}, t\right) = \left(-\frac{ik}{p_1} E^0_{in,z} \sinh p_1 x, 0, E^0_{in,z} \cosh p_1 x\right) \exp[i(kz - \omega t)]. \tag{4.54a}$$

Matching the boundary conditions at the slab surfaces with the vacuum fields in (4.42) gives an equation for the dispersion relations of the symmetric modes that is of the form [13]

$$\varepsilon_m(\omega) = -\frac{p_1}{\alpha} \coth p_1 \frac{d}{2} \tag{4.54b}$$

where $p_1^2 = k^2 - \varepsilon_m(\omega)\frac{\omega^2}{c^2}$.

The solutions of the slab modes involving exponents of the x coordinate have fields within the slab that for antisymmetric modes are given by [13]

$$\vec{E}_{in}^{asym}\left(\vec{r}, t\right) = \left(-\frac{ik}{p_1} E_{in,z}^0 \cosh p_1 x, 0, E_{in,z}^0 \sinh p_1 x\right) \exp[i(kz - \omega t)]. \tag{4.55a}$$

Matching the boundary conditions at the slab surfaces with the vacuum fields in (4.48) gives an equation for the dispersion relations of the antisymmetric modes that is of the form [13]

$$\varepsilon_m(\omega) = -\frac{p_1}{\alpha} \tanh p_1 \frac{d}{2} \tag{4.55b}$$

where $p_1^2 = k^2 - \varepsilon_m(\omega)\frac{\omega^2}{c^2}$.

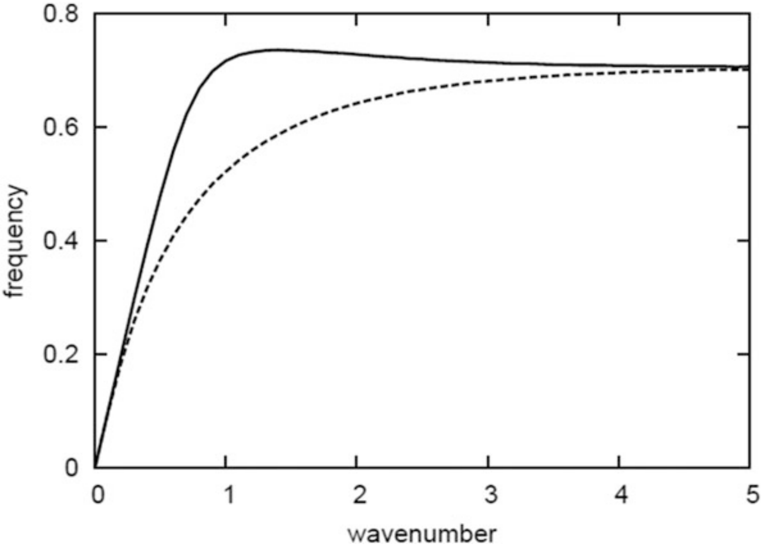


Fig. 4.5 A plot of the frequency, $\frac{\omega}{\omega_p}$, as a function of the wavenumber, $\frac{kc}{\omega_p}$, is presented for a slab with a dielectric constant given by $\varepsilon_m(\omega) = 1 - \frac{\omega_p^2}{\omega^2}$ and which is surrounded by vacuum [15]. In the plot the upper curve is a solution from (4.55b) and the lower curve is a solution for (4.54b). Both curves are plotted for $\frac{\omega_p d}{c} = 1.0$

Numerical Example

As an example of the modes of the slab, in Fig. 4.5 a plot of ω as a function of k is presented for a slab with a dielectric function of the form $\varepsilon_m(\omega) = 1 - \frac{\omega_p^2}{\omega^2}$ [15]. The results are for the modes in (4.54)–(4.55) characterized by exponential behavior within the slabs.

In the plot the modes are set by the parameters $\frac{\omega}{\omega_p}, \frac{ck}{\omega_p}$, and $\frac{\omega_p d}{c}$ and are seen to occur in pairs at a given $\frac{ck}{\omega_p}$. For the curves presented in the plots the thickness of the slab was chosen such that $\frac{\omega_p d}{c} = 1.0$. For these solutions, the modes in (4.54) are the low frequency solutions and those in (4.55) are the high frequency modes. It is found that as the slab thickness increases the frequencies of the two $\frac{\omega_p d}{c}$ solutions approach one another.

4.2 Surface Plasmon-Polariton Modes for Shape Resonances, Gratings, and Light Scattering from Rough Surfaces

Another important class of interfaces in the consideration of electromagnetic surface waves includes interfaces with surface bumps, surface gratings, and surface roughness [13, 14]. The presence of a bump on an otherwise planar interface gives rise to the scattering of plasmon-polaritons from the bump and to the possibility of bound states becoming localized in the region of the bump. These two phenomena show up in the optics of both the light scattered from the surface and in the light propagating on the surface. Bumps on interfaces enter into a number of important technological considerations based on optical surface properties of the surface shape resonances associated with the surface features.

Surface roughness also has important effects on the optical properties exhibited at interfaces, and these include effects on both bulk and surface electromagnetic waves. Surface roughness is a type of imperfection to perfectly planar surfaces that is always present to some degree in experimental systems. Nevertheless, it is also responsible for important physical effects that in themselves can be of significant importance in technology.

Rough interfaces are divided into two categories: surface gratings and random surface roughness. Gratings are interfaces that exhibit periodicity in their spatial properties. This periodicity leads to diffractive effects in the light scattered from the surface and to a band structure associated with the surface electromagnetic waves that are supported along the interface.

Random rough surfaces, on the other hand, have surface profiles which exhibit random disorder. This type of disorder leads to a general diffuse scattering of light form the surface and to lifetime effects in the propagation of surfaces waves along the interface. In both random rough and grating systems bulk light and surface electromagnetic waves can couple to one another.

In the following surfaces supporting bumps, gratings, and random surface roughness will be considered.

4.2.1 Shape Resonances

In the following the electromagnetic modes bound to features on an otherwise planar interface between two media are considered for surfaces supporting plasmon-polaritons [14]. Specifically, a localized imperfection on an otherwise planar surface can be found to bind electromagnetic modes in its neighborhood on the surface. The surface electromagnetic modes are attached to the imperfection site and require specific conditions in the shape of the surface features and dielectric properties of the media forming the interface for their existence on the surface. They are termed surface shape resonances and in their conception are similar to electronic p- or n-bound impurity modes that are found in impurity doped semiconductors. The excitation of surface shape resonance are observed in the surface scattering of light and particles from the supporting surfaces.

The treatment given here focuses on plasmon surface shape resonances bound at an impurity on an otherwise planar vacuum-metal interface. This provides the simplification of involving calculations made for the quasi-static modes of the system in which the retardation effects are not taken into account. The solutions are obtained from the Laplace equations, but nonetheless illustrate many of the properties of the system treated outside the quasi-static limit and in the context of the full wave equation. For the treatment of the polariton surface shape resonances, obtained as solutions of the full electromagnetic wave equations, aside from some brief comments made later, the reader is referred to the literature.

As an additional simplification of the presentation, the surface will be treated as a one-dimensional surface described by a surface profile function of the form [14]

$$z = \xi(x). \tag{4.56}$$

Here (4.56) relates the z-coordinate of the surface to the position x on the x-axis, and the surface is considered to be translationally invariant along the y-direction. The function $\xi(x)$ describes a localized feature on the surface which is centered on $x = 0$ and vanishes quickly away from $x = 0$.

In the following a treatment will be made for the modal solutions bound to the localized surface feature and having wave functions that depend only on x and z. Solutions for wave functions with y dependence exist, but these are for modes bound to the localized feature and propagating along the y direction. These modes are interesting but will not be considered here.

The surface shape resonance modes are obtained for the surface defined in (4.56) in terms of the solutions of the Laplace equations for the interface between vacuum and a dielectric medium described by the frequency dependent dielectric constant $\varepsilon(\omega)$. Though the considerations are based on the $\frac{\omega}{c} \to 0$ approximation, the dynamics of the system still enters the problem through $\varepsilon(\omega)$ in the region about this limit. For the considerations presented in the following the region $z > \xi(x)$ contains vacuum and the region $z < \xi(x)$ contains the $\varepsilon(\omega)$ dielectric.

Consider a surface between the regions of vacuum and metal of dielectric constant $\varepsilon(\omega)$ that has a localized ridge upon it described by (4.56). The solutions of the electromagnetic potential in these two regions are determined in the quasi-static limit from the Laplace equations [14]

$$\nabla^2 \phi^{>}(x, z|\omega) = 0 \text{ for } z > \xi(x) \tag{4.57a}$$

and

$$\nabla^2 \phi^{<}(x, z|\omega) = 0 \text{ for } z > \xi(x). \tag{4.57b}$$

The solutions of (4.57) that are of interest describe potentials which are localized on the interface and decrease to zero at $z \to \pm\infty$. In addition to being localized on the interface, the potentials must also be localized on the surface feature to which they are bound.

Solutions with these properties have the general form given by the Fourier transforms involving plane waves on the interface which decay exponentially in amplitude as $z \to \pm\infty$. In particular,

$$\phi^{>}(x, z|\omega) = \int \frac{dk}{2\pi} A(k\omega) \mathrm{e}^{i(kx - |k|z)} \tag{4.58a}$$

describes the potential above the surface, and

$$\phi^{<}(x, z|\omega) = \int \frac{dk}{2\pi} A^{<}(k\omega) \mathrm{e}^{i(kx - |k|z)} \tag{4.58b}$$

describes the potential below the surface [14].

At the interface the forms in (4.58) are matched by the boundary conditions [14]

$$\phi^{>}(x, z|\omega)_{z=\xi(x)} = \phi^{<}(x, z|\omega)_{z=\xi(x)} \tag{4.59a}$$

and

$$\hat{n} \cdot \nabla \phi^{>}(x, z|\omega)_{z=\xi(x)} = \varepsilon(\omega) \hat{n} \cdot \phi^{<}(x, z|\omega)_{z=\xi(x)}. \tag{4.59b}$$

Applying these boundary conditions to (4.58) yields, after some algebra, an homogeneous integral equation for the amplitude $A(q\omega)$ in the Fourier transform in (4.58a). The integral equation has the form [15]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}A(q\omega) = \int \frac{dp}{2\pi} J(|q|-|p||q-p)\left[1-\frac{q}{|q|}\frac{p}{|p|}\right]|p|A(p\omega) \tag{4.60}$$

where the Kernel involves the surface roughness

$$J(a|b) = \int dx \mathrm{e}^{ibx}\frac{\mathrm{e}^{a\xi(x)-1}}{a}. \tag{4.61}$$

For the case in which the surface profile function $\xi(x)$ is an even function of x, it follows from symmetry considerations that

$$\phi^{>}(-x,z|\omega) = \pm\phi^{>}(x,z|\omega) \tag{4.62}$$

i.e., $\phi^{>\,or<}(x)$ is an even or odd function of x. Consequently, in (4.58) these symmetry considerations carry over to the coefficients of the Fourier transform for $\phi^{>}(x)$ so that

$$A(-q\omega) = \pm A(q\omega). \tag{4.63}$$

Following some algebra and the use of these symmetries (4.60) becomes [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}A(q\omega) = \pm\frac{1}{\pi}\int_0^\infty dpJ(q-p|q+p)pA(p\omega) \tag{4.64}$$

The homogeneous integral equation in (4.64) can be solved by treating it as an eigenvalue problem. The eigenvalue problem of interest has the form [14]

$$\frac{1}{\pi}\int_0^\infty dpJ(q-p|q+p)pA_s(p\omega) = \lambda_s A_s(q\omega) \tag{4.65}$$

and determines the set of eigenvalues $\{\lambda_s\}$ and the corresponding eigenvectors $\{A_s(q\omega)\}$ for a fixed frequency ω and surface profile function $\xi(x)$. In terms of the eigenvalues the surface shape resonances occur at frequencies determined by the conditions [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1} = \pm\lambda_s. \tag{4.66}$$

In the case of a metal with a dielectric constant of the form

$$\varepsilon(\omega) = \varepsilon_\infty\left(1-\frac{\omega_p^2}{\omega^2}\right), \tag{4.67}$$

the condition in (4.66) gives surface shape resonances with frequencies

$$\omega = \left[\frac{\varepsilon_\infty(1 \mp \lambda_s)}{1 + \varepsilon_\infty \pm \lambda_s(1 - \varepsilon_\infty)}\right]^{1/2} \omega_p. \tag{4.68}$$

For an insulator with a dielectric constant given by [14]

$$\varepsilon(\omega) = \varepsilon_\infty \left(1 + \frac{\omega_{LO}^2 - \omega_{TO}^2}{\omega_{TO}^2 - \omega^2}\right), \tag{4.69}$$

the condition in (4.66) gives surface shape resonances with frequencies

$$\omega^2 = \omega_{TO}^2 + \frac{(1 \mp \lambda_s)\varepsilon_\infty\left(\omega_{LO}^2 - \omega_{TO}^2\right)}{1 \pm \lambda_s + \varepsilon_\infty(1 \mp \lambda_s)}. \tag{4.70}$$

The surface shape resonances correspond to the real values of ω obtained from either (4.68) or (4.70) in terms of the eigenvalues of the (4.65), and the expressions for the $A(q\omega)$ related to these ω are computed as the solutions of the integral equation in (4.65). For the numerical solution of (4.65) the integral in (4.65) can be converted from an integral equation eigenvalue problem into a matrix eigenvalue problem using quadrature methods. These results then constitute the complete solution of the surface shape resonance problem, giving both the frequencies and the wave functions of the system.

For real dielectric functions, the surface shape resonance problem computed in the quasi-static limit generally yields real values of ω. These are stable modes which do not decay in time. This is not the case with the solutions obtained by including retardation effects. Upon the reintroduction of retardation effects into the considerations for the full unrestricted solutions of the Maxwell equations, the surface shape resonance modes exhibit decay due to surface scattering and the radiation of electromagnetic fields from the surfaces. These effects contribute to give surface shape resonance solutions with finite lifetimes and are a manifestation of the loss of translational symmetry on the surface.

Illustrative Example

As an example of the quasi-static surface shape resonances, consider the surface shape resonances of a small perturbation on the planar interface with a Lorentzian profile given by [14]

$$\xi(x) = \frac{AR^2}{x^2 + R^2}. \tag{4.71}$$

In the case that A is positive (4.71) represents a surface with a ridge, and in the case that A is negative (4.71) represents a surface groove. For the limit of a small perturbation from the planar interface, A in the Lorentzian form is the small parameter characterizing the amplitude of the ridge or groove on the surface.

Under these conditions of the small amplitude limit, (4.61) becomes [14]

$$\begin{aligned} J(q-p|q+p) &= \int dx \mathrm{e}^{-i(q+p)x} \frac{\mathrm{e}^{(q-p)\xi(x)}-1}{q-p} \\ &\approx \int dx \mathrm{e}^{-i(q+p)x} \zeta(x) \end{aligned} \tag{4.72}$$

and for the particular profile function form in (4.71)

$$\begin{aligned} J(q-p|q+p) &= \int dx \mathrm{e}^{-i(q+p)x} \frac{AR^2}{x^2+R^2}. \\ &= \pi A\, \mathrm{Re}^{-R}(q+p) \end{aligned} \tag{4.73}$$

The integral equation eigenvalue problem for the surface shape resonances given in (4.65) then becomes in the approximation for $J(q-p|q+p)$ in (4.73)

$$A\mathrm{Re}^{-Rq} \int\limits_0^\infty dp \mathrm{e}^{-Rp} p A_s(p\omega) = \lambda_s A_s(q\omega) \tag{4.74}$$

Multiplying both sides of (4.74) by $q\mathrm{e}^{-Rq}$ and integrating over q reduces (4.74) to the algebraic eigenvalue problem [14]

$$\left[\lambda_s - \frac{1}{4}\frac{A}{R}\right] \int\limits_0^\infty dp \mathrm{e}^{-Rp} p A_s(p) = 0 \tag{4.75}$$

so that the eigenvalue is

$$\lambda_s = \frac{1}{4}\frac{A}{R}. \tag{4.76}$$

From (4.76) it should be noted that positive λ_s are found for surface ridges (i.e., $\frac{A}{R} > 0$) and negative λ_s are found for surface grooves (i.e., $\frac{A}{R} < 0$). This is then a distinguishing characteristic between these two types of surface features. Both types of feature will be seen to support bound surface shape resonances [14].

In terms of the vacuum-metal interface (4.76) and (4.68) give solutions for surface shape modes with frequencies obtained from the form [14]

$$\omega = \left[\frac{\varepsilon_\infty\left(1 \mp \frac{A}{4R}\right)}{1+\varepsilon_\infty \pm \frac{A}{4R}(1-\varepsilon_\infty)}\right]^{1/2} \omega_p. \tag{4.77}$$

The frequencies of the surface shape resonances in (4.77) depend not only on the even or odd parity of the potentials of the modes bound to the ridge or groove. The nature of the frequencies found also depend on whether or not the feature is a ridge or a groove on the surface.

For the case that $\varepsilon_\infty = 1$ and the surface feature is a ridge (i.e., $\frac{A}{R} > 0$) it is seen that the surface shape resonance potentials having even parity in x are at lower frequencies than the flat surface plasmon modes. For the case that $\varepsilon_\infty = 1$ and the surface feature is a groove (i.e., $\frac{A}{R} < 0$) it is seen that the surface shape resonance potentials having even parity in x are at higher frequencies than the flat surface plasmon modes [14]. The situation is reversed in the cases of the surface shape resonant potentials having odd parity in x [14].

4.2.2 Scattering from Gratings

In the following the electromagnetic modes bound to a periodic interface between two media are considered for surfaces supporting surface electromagnetic waves [14, 15]. As in the case of the solutions of the dynamics of the surface shape resonances, the treatment focuses on plasmon modes on a vacuum-metal interface. This provides the simplification of involving calculations made in the quasi-static limit, which are based on solutions of the Laplace equation and ignores the retardation effects from the full Maxwell equation treatment of the polariton problem. At the end of the presentation of the quasi-static treatment, however, some qualitative discussions of the more general problem of surface polariton solutions on a periodic surfaces will be given. Nevertheless, for a detailed study of the surface polariton limit, the reader is again referred to the literature [1–6, 14, 15].

To simplify the presentation, the surface will be treated as one-dimensional in the sense that it is described by a surface profile function of the form [14]

$$z = \xi(x), \tag{4.78a}$$

and it is translationally invariant along the y-direction. The surface profile function in (4.78a) then relates the z-coordinate of the surface to the position x on the x- axis. It is taken to be a periodic function $\xi(x)$ such that [14]

$$\xi(x) = \xi(x + na) \tag{4.78b}$$

for n an integer, and in (4.78b) a represents the smallest repeat distance of the surface along the x-axis. The periodic surface profile function is chosen such that the average of the surface profile over the x-y plane gives [14]

$$\langle \xi(x) \rangle = 0. \tag{4.78c}$$

The plasmon modes are obtained for the surface defined in (4.78) in terms of the solutions of the Laplace equations for the interface between vacuum above the surface and a metallic medium described by the frequency dependent dielectric constant $\varepsilon(\omega)$ below the surface. Since the considerations are based on the quasi-static, $\frac{\omega}{c} \to 0$, approximation, the dynamics of the system enters the problem through the frequency dependence of $\varepsilon(\omega)$. This is an essential point as only at nonzero frequencies can metals exhibit the wider range of dielectric properties needed in the discussions to follow. Specifically, in the later considerations $\varepsilon(\omega)$ is required to be negative for surface wave solutions to exist on the interface.

Consider a surface of the form of a periodic grating described by (4.78) which is the interface between a semi-infinite region of vacuum and a semi-infinite region of metal of dielectric constant $\varepsilon(\omega)$. In the regions above and below the surface, the solutions of the electromagnetic potentials will be determined in the quasi-static limit from the Laplace equations [14]

$$\nabla^2 \phi^> (x, z|\omega) = 0 \quad \text{for } z > \xi(x) \text{ above the surface} \tag{4.79a}$$

and

$$\nabla^2 \phi^< (x, z|\omega) = 0 \quad \text{for } z < \xi(x) \text{ below the surface,} \tag{4.79b}$$

respectively. For a simplification in the treatment of the surface modes, (4.79) have been written for potentials which are only functions of x and z Consequently, here and in the following only surface modes propagating along the interface in the x-direction are treated and all y-motion is suppressed. The solutions obtained under these conditions exhibit the largest effects from the periodic interface, displaying the essential interesting features of surface waves on a periodic interface.

Modes that propagate along the y-direction are also of interest for technological applications and their development is a straightforward generalization of the presentation given here for x-propagating modes. The development of these modes will be left to the reader to work out.

The surface wave solutions of (4.79) that are of interest here describe potentials which are localized on the interface and decrease to zero at $z \to \pm\infty$. Solutions with these properties have the general form of a Fourier transforms involving plane waves on the interface which decay exponentially in amplitude as $z \to \pm\infty$. In particular, the appropriate Fourier transforms are [14]

$$\phi^> (x, z|\omega) \int \frac{dk}{2\pi} A(k\omega) \mathrm{e}^{i(kx - |k|z)} \tag{4.80a}$$

describing the potential above the surface, and

$$\phi^< (x, z|\omega) \int \frac{dk}{2\pi} A^< (k\omega) \mathrm{e}^{i(kx + |k|z)} \tag{4.80b}$$

describing the potential below the surface.

From matching the boundary conditions on the Fourier transforms at the interface, it is found that $A(k\omega)$ in (4.80a) is determined by the integral equation [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}A(q\omega)=\int\frac{dp}{2\pi}J(|q|-|p||q-p)\left[1-\frac{q}{|q|}\frac{p}{|p|}\right]|p|A(p\omega) \tag{4.81}$$

where the Kernel of the integral equation involves the function

$$J(a|b)=\int dx\mathrm{e}^{-ibx}\frac{\mathrm{e}^{a\xi(x)}-1}{a}. \tag{4.82a}$$

From (4.81) and (4.82a) it is seen that the information about the surface profile is introduced into (4.81) on the right of the equation through the function $J(a|b)$ while the dielectric properties of the surface materials enter only on the left of the equation. This has an interesting consequence in the limit of a weakly rough surface profile. In the limit of a weakly rough grating profile function

$$J(a|b)\approx\int dx\mathrm{e}^{-ibx}\xi(x), \tag{4.82b}$$

so that for this limit of a rough surface it is then found that $\varepsilon(\omega)\approx-1$ for (4.81) to have a solution. This condition is a restriction on the range of frequencies of the solutions in this limit.

The periodic symmetry of the surface profile function requires that the solutions of (4.79) and (4.80) be of the form [14]

$$\phi^{>}(x,z|\omega)=\mathrm{e}^{ikx}U_k^{>}(x,z|\omega) \tag{4.83a}$$

and

$$\phi^{<}(x,z|\omega)=\mathrm{e}^{ikx}U_k^{<}(x,z|\omega) \tag{4.83b}$$

where $U_k^{>}(x+na,z|\omega)=U_k^{>}(x,z|\omega)$ and $U_k^{<}(x+na,z|\omega)=U_k^{<}(x,z|\omega)$ for n an integer. This is a fundamental restriction which was discussed earlier in the treatments of the wave functions of periodic systems. The details presented at that time were for quite general systems with periodic equations so that the reader is referred to the Chapter on Photonic crystals for more details about the functional forms in (4.80). In the following the forms in (4.83b) will be used to obtain solutions for the wave functions of the surface waves.

The solution of (4.83a) from (4.81) will be the focus of considerations in the following considerations. The wave function form in (4.83a) contains the essential elements needed for the determination of the dispersion relation of the surface plasmons on the periodic interface. For (4.80a) to have the form in (4.83a), $A(q\omega)$ must be given by

$$A(q\omega) = 2\pi \sum_{m=-\infty}^{\infty} A_m(k\omega)\delta(q - k_m) \tag{4.84}$$

where

$$k_m = k + \frac{2\pi m}{a}. \tag{4.85}$$

This can be seen from a simple substitution of the form in (4.84) into the integral for $\phi^{>}(x, z|\omega)$.

Upon substituting (4.84) into (4.80a) it follows that [14]

$$\phi^{>}(x, z|\omega) = \mathrm{e}^{ikx} \sum_{m=-\infty}^{\infty} A_m(k\omega)\mathrm{e}^{i\frac{2\pi m}{a}x}\mathrm{e}^{-\left|k+\frac{2\pi m}{a}\right|z} \tag{4.86a}$$

where

$$U_k^{>}(x, z|\omega) = \sum_{m=-\infty}^{\infty} A_m(k\omega)\mathrm{e}^{i\frac{2\pi m}{a}x}\mathrm{e}^{-\left|k+\frac{2\pi m}{a}\right|z} \tag{4.86b}$$

has the required periodicity

$$U_k^{>}(x, z|\omega) = U_k^{>}(x + na, z|\omega) \tag{4.87}$$

for n and integer.

To obtain the solutions for $\{A_m(kw)\}$, (4.84) is substituted into (4.81). This gives an homogeneous integral equations for the required $\{A_m(kw)\}$, in terms of the surface profile function and the dielectric properties of the materials forming the interface. The substitution gives, [14]

$$\begin{aligned} &\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1} 2\pi \sum_{m=-\infty}^{\infty} A_m(k\omega)\delta(q - k_m) = \int dpJ(|q| - |p||q - p)\left[1 - \frac{q}{|q|}\frac{p}{|p|}\right]|p| \\ &\quad \times \sum_{m=-\infty}^{\infty} A_m(k\omega)\delta(p - k_m) = \sum_{m=-\infty}^{\infty} \int dx\mathrm{e}^{-i(q-k_m)x} \frac{\mathrm{e}^{(|q|-|k_m|)\xi(x)} - 1}{|q| - |k_m|}\left(|k_m| - \frac{q}{|q|}k_m\right)A_m(k\omega) \end{aligned} \tag{4.88}$$

where on the far right the definition of $J(a|b)$ in (4.82) has been used. The integral on the far right of (4.88) can be rewritten through the application of the identity

$$\int_{-\infty}^{\infty} dxf(x) = \sum_{l=-\infty}^{\infty} \int_{0}^{a} dxf(x + la), \tag{4.89}$$

so that (4.88) becomes [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}2\pi\sum_{m=-\infty}^{\infty}A_m(k\omega)\delta(q-k_m)=\sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\int_0^a dx\mathrm{e}^{-i(q-k_m)la}$$
$$\mathrm{e}^{-i(q-k_m)x}\frac{\mathrm{e}^{(|q|-|k_m|)\xi(x)}-1}{|q|-|k_m|}\left(|k_n|-\frac{q}{|q|}k_m\right)A_m(k\omega) \tag{4.90}$$

where the periodicity of the surface profile function, $\xi(x+la)=\xi(x)$, has been used.

The integral equation in (4.90) can be further reduced by applying the Fourier identity

$$\sum_{l=-\infty}^{\infty}\mathrm{e}^{-i(q-k_m)la}=\frac{2\pi}{a}\sum_{n=-\infty}^{\infty}\delta(q-k_{m+n}). \tag{4.91}$$

Applying the identity and after some algebra gives.

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}2\pi\sum_{m=-\infty}^{\infty}A_m(k\omega)\delta(q-k_m)$$
$$=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\frac{2\pi}{a}\delta(q-k_{n+m})\int_0^a dx\mathrm{e}^{-i(q-k_m)x}\frac{\mathrm{e}^{(|q|-|k_m|)\xi(x)}-1}{|q|-|k_m|}\left(|k_m|-\frac{q}{|q|}k_m\right)A_m(k\omega). \tag{4.92}$$

This, in turn, is rewritten into the form

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}2\pi\sum_{m=-\infty}^{\infty}A_m(k\omega)\delta(q-k_m)$$
$$=\sum_{m=-\infty}^{\infty}2\pi\delta(q-k_m)\sum_{n=-\infty}^{\infty}\int_0^a dx\mathrm{e}^{-i(q-k_n)x}\frac{\mathrm{e}^{(|q|-|k_n|)\xi(x)}-1}{|q|-|k_n|}\left(|k_n|-\frac{q}{|q|}k_n\right)A_n(k\omega). \tag{4.93}$$

Equating the delta function coefficients on both sides of (4.93) gives a matrix equation for the fields above the surface. In particular, the coefficients $\{A_m(k\omega)\}$ are found to satisfy [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}A_m(k\omega)=\sum_n M_{m,n}(k)A_n(k\omega). \tag{4.94}$$

for the matrix

$$M_{m,n}(k) = \frac{|k_m||k_n| - k_m k_n}{|k_m|(|k_m| - |k_n|)} \frac{1}{a} \int_0^a dx \mathrm{e}^{-i\frac{2\pi}{a}(m-n)x} \left[\mathrm{e}^{i(|k_m|-|k_n|)\xi(x)} - 1 \right]. \tag{4.95}$$

From (4.95) it is seen that the diagonal components of the matrix are zero. This has consequences in the form of the solution of (4.94). In addition, other restrictions on the form of the matrix arise from symmetry considerations. These shall now be addressed and followed by the solution of (4.94) obtained by reducing the problem to an eigenvalue problem.

The matrix equation in (4.94) is seen to exhibit symmetries which are an aid in computing its solutions. These arise from the periodicity of the grating profile function contributing to the properties observed in the nature of the electromagnetic modes in both position and wave vector space. The symmetries in position space have already been discussed so that the focus in the following will be on the symmetries in wave vector space. While the symmetry of the system in position space set the form of the wave functions of the electromagnetic modes, in wave vector space the symmetry of the system sets the nature of the dispersion relation of the modes.

From (4.84) it is seen that for a fixed integer l,

$$\begin{aligned} A(q\omega) &= 2\pi \sum_{m=-\infty}^{\infty} A_{m+l}(k\omega)\delta(q - k_{m+l}) \\ &= 2\pi \sum_{m=-\infty}^{\infty} A_m\left(\left(k + \frac{2\pi l}{a}\right)\omega\right)\delta(q - k_m). \end{aligned} \tag{4.96}$$

This sets the requirement that the $\{A_m(k\omega)\}$ satisfy the identity

$$A_m\left(\left(k + \frac{2\pi l}{a}\right)\omega\right) = A_{m+l}(k\omega). \tag{4.97}$$

In a similar manner by taking $k \longrightarrow k + \frac{2\pi l}{a}$ the matrix in (4.93) is found to have the property that

$$M_{m,n}\left(k + \frac{2\pi l}{a}\right) = M_{m+l,n+l}(k). \tag{4.98}$$

Consequently, using these relations in (4.96) it follows that [14]

$$\frac{\varepsilon(\omega) + 1}{\varepsilon(\omega) - 1} A_m\left(\left(k + \frac{2\pi l}{a}\right)\omega\right) = \sum_n M_{m,n}\left(k + \frac{2\pi l}{a}\right) A_n\left(\left(k + \frac{2\pi l}{a}\right)\omega\right), \tag{4.99a}$$

or

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1}A_{m+l}(k\omega) = \sum_n M_{m+l,n+l}(k)A_{n+1}(k\omega). \tag{4.99b}$$

Equation (4.99b) again reduces to the original matrix problem in (4.94). As a result of this reduction it follows that all of the solutions for $\{A_m(k\omega)\}$ are distinct within a region of wave vector space that is of length $\frac{2\pi}{a}$ in wave vector space. Outside such an interval the $\{A_m(k\omega)\}$ solutions just repeat themselves.

The solutions of (4.94) are now obtained by solving the matrix eigenvalue problem defined by [14]

$$\sum_n M_{m,n}(k)A_n^s(k\omega) = \lambda_s A_m^s(k\omega). \tag{4.100}$$

This gives the set of distinct eigenvectors $\left\{A_m^s(k\omega)\right\}$ corresponding to the set of eigenvalues $\{\lambda_s\}$. In treating these solutions the symmetry properties of $M_{m,n}(k)$ in wave vector space as well as in the matrix form offer a great simplification.

From the earlier discussions, it is seen from the periodic properties that the problem in (4.100) need only be solved within an interval of wave vector space that is of length $\frac{2\pi}{a}$. Consequently, the focus is usually set for obtaining solutions on the interval $0 < k < \frac{2\pi}{a}$ or $\frac{\pi}{a} < k < \frac{\pi}{a}$. The last form is usually referred to as the first Brillouin zone of the system and is a common choice for the presentation of results. From the symmetry of the matrix with its transpose it also follows that the eigenvalues occur in $\pm\lambda_s$ pairs. Once these eigenvalue pairs are determined the dispersion of the system is given by [14]

$$\frac{\varepsilon(\omega)+1}{\varepsilon(\omega)-1} = \pm\lambda_s. \tag{4.101}$$

Illustrative Example

As an example of the formulation, consider the limit of a weakly rough grating. A simple yet interesting case to treat is a periodic profile given by

$$\xi(x) = A\cos\left(\frac{2\pi}{a}x\right) \tag{4.102}$$

where A is the small grating amplitude parameter. This illustrates the basic properties of periodic profiles while simplifying the Fourier integrals involved in the matrix in (4.95) and, consequently, the form of the matrix in the eigenvalue problem.

From (4.95), in the weak grating limit, it is found that the $M_{n,m}(k)$ matrix function of the eigenvalue problem for the sinusoidal profile takes the form [14]

$$\begin{aligned} M_{m,n}(k) &= \frac{|k_m||k_n| - k_m k_n}{|k_m|(|k_m| - |k_n|)}\frac{1}{a}\int\limits_0^a dx\mathrm{e}^{-i\frac{2\pi}{a}(m-n)x}\left|\mathrm{e}^{i(|k_m|-|k_n|)\xi(z)} - 1\right| \\ &\approx \frac{|k_m||k_n| - k_m k_n}{|k_m|(|k_m| - |k_n|)}\frac{1}{a}\int\limits_0^a dex^{-i\frac{2\pi}{a}(m-n)x}\left[(|k_m| - |k_n|)A\cos\left(\frac{2\pi}{a}x\right)\right]. \end{aligned} \tag{4.103}$$

Upon evaluating the integral in (4.103) the $M_{n,m}(k)$ matrix becomes a tri-diagonal matrix given by

$$M_{m,m=0}, \tag{4.104a}$$

$$M_{m,m+1} = \frac{|k_m||k_{m+1}| - k_m k_{m+1}}{|k_m|}A, \tag{4.104b}$$

and

$$M_{m,m-1} = \frac{|k_m||k_{m-1}| - k_m k_{m-1}}{|k_m|}A. \tag{4.104c}$$

This is a sparse matrix which offers a great simplification from the original general form of the matrix. The resulting tri-diagonal matrix eigenvalue problem is an extensively studied problem with many existent routines available for its treatment.

In a crude approximation of the tri-diagonal matrix eigenvalue problem in (4.104) consider the truncated 3 × 3 eigenvalue problem obtained from $M_{m,m}$ by taking $n = -1, 0, 1$ and $m = -1, 0, 1$. The solutions of the eigenvalues are given by [14]

$$\lambda_s = \pm A\sqrt{\frac{a^2}{|k_0||k_1|} + \frac{b^2}{|k_0||k_{-1}|}} \tag{4.105}$$

where

$$a = |k_0||k_1| - k_0 k_1 \tag{4.106a}$$

and

$$b = |k_0||k_{-1}| - k_0 k_{-1}. \tag{4.106b}$$

The resulting eigenvalues are again restricted to the region $-\frac{\pi}{a} < k < \frac{\pi}{a}$ and are seen to occur in $\pm\lambda_s$ pairs.

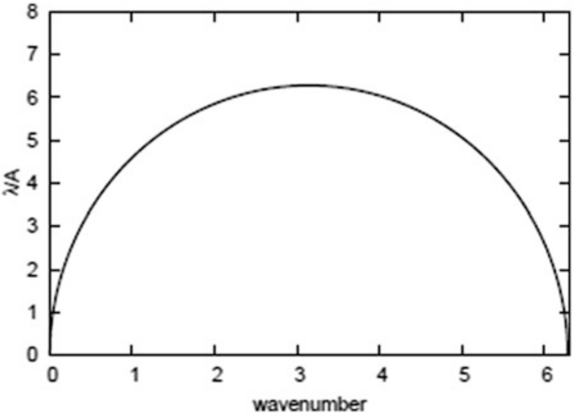


Fig. 4.6 Plot of $\frac{\lambda_s}{A} = \sqrt{\frac{a^2}{|k_0||k_1|} + \frac{b^2}{|k_0||k_{-1}|}}$ versus wave vector, *ka*, for wave vectors in a unit cell of the reciprocal lattice [14]. In the special case presented here only a single eigenvalue is obtained for the perturbation limit of a small amplitude grating

As an example, in Fig. 4.6 a plot is presented for $\frac{\lambda_s}{A} = \pm\sqrt{\frac{a^2}{|k_0||k_1|} + \frac{b^2}{|k_0||k_{-1}|}}$ versus wave vector over a unit cell of the reciprocal lattice. The $\lambda_s(k)$ are seen to be periodic in k with the property that $\lambda_s(k) = \lambda_s(-k)$. Introducing the values of $\pm\lambda_s$ into (4.101) provides an approximation to the dispersion relation of some of the modes within the quasi-static limit. The results following from Fig. 4.6 are qualitative similar to the results for the largest eigenvalues obtained in [14] from an evaluation of the matrix eigenvalue problem using a 38 × 38 matrix. The reader is referred to the literature for the details of a more precise numerical evaluation of the matrix eigenvalue problem and for the resulting surface shape resonance frequencies.

Remarks on General System: Retardation Effects

The above discussions were for the quasi-static, $c \to \infty$, limit of the grating problem. For modes in this limit there are no radiative losses from the grating surface, and the solutions of the system separate distinctly into surface and bulk modes. Upon the reintroduction of the retarded limit into the study of the electromagnetic modes of the grating, the possibility arises of radiation of the surface waves away from the interface by scattering into bulk modes of the system. These new surface solutions, which radiate away from the interface, are known as leaky waves [14].

In the following, discussions will be presented to develop a qualitative understanding of the nature of the bound and leaky surface waves of the grating. Bound modes are surface waves that propagate along the interface without scattering into bulk modes of the system, and due to the grating periodicity they exhibit a frequency band structure. For lossy dielectric media bound modes may decay due to Joule heating, but they do not radiate away from the interface. Leaky modes are modes that can decay due to dielectric losses, but they definitely decay through scattering transitions into bulk radiative states propagating away from the surface. The conditions leading to these two different types of solutions on the grating will now be discussed in terms of the frequency band structure of the surface waves and its relationship to the dispersion relation for light in the bulk [14].

To qualitatively understand the existence of leaky waves, the example of plasmon-polaritons at a vaccum-metal interface is considered. This offers a simple illustration of how the difference between bound and leaky surface wave solutions arises upon the reintroduction of retardation effects. In addition, the treatment given here can be easily extended to surface waves on other vacuum-dielectric gratings.

A focus in the discussions is on the presentation of the dispersion relation of the grating surface waves in the first Brillouin zone. As mentioned earlier this region of wave vector space contains all of the unique solutions of the periodic system, and the solutions with wave vectors outside this region of k-space are duplicates of those in the first Brillouin zone. The restriction of the unique solutions of the problem to the first Brillouin zone arises solely due to the periodic symmetry of the grating and is present for both the quasi-static modes obtained from the Laplace equations and for the full solutions of the Maxwell equations.

In addition, for the discussion presented in the following, the material parameters typical of the planar surface solutions presented in Fig. 4.2 will be used to characterize the materials separated by the grating interfaces. This provides for a simple illustration of the principles involved in defining the deference between and existence of surface waves and leaky waves at grating interfaces [14].

Consider a vacuum-metal grating with a surface profile function of period a along the x-direction and which is translationally invariant along the y-direction. Specifically,

$$z = \xi(x) = \xi(x + na), \tag{4.107}$$

for an integer n, gives the z-coordinate of the interface in terms of the position along the x-axis. Due to the periodicity of (4.107) the solutions for the electromagnetic waves of the system propagating along the x-direction are uniquely represented by their positions in the first Brillouin zone. This restricts the modes propagating in the x-direction with wave vector k to wave vectors within the region

$$-\frac{\pi}{a} \leq k \leq \frac{\pi}{a}. \tag{4.108}$$

In addition to the important condition in (4.108) on the modes in k-space there are restrictions on the stability of the solutions found in the first Brillouin zone. These restrictions arise from the introduction into the system of retardation effects and are now addressed.

An additional important restriction on the modes in the first Brillouin zone comes from examining a plot in the first Brillouin zone of the so-called light line. The light line for the vacuum-metal interface is defined by the condition that

$$\omega = ck. \tag{4.109}$$

For the quasi-static limit in which $c \to \infty$ it is seen that the light line vanishes as a boundary of the dispersion relations in the first Brillouin zone. For this case the classification of k-solutions into those with frequencies above and those with frequencies below the light line does not exist as all solutions in the quasi-static limit are below the light line. This, however, is not the case with the full Maxwell equation solutions that include retardation effects. As shall be seen the k-solutions with retardation effects are divided into distinct sets of solutions with frequencies above and below the light line.

The classification of surface wave solutions into those with frequencies above the light line and those with frequencies below the light line is very important as it provides a fundamental restriction on the forms of the surface waves. If a surface wave solution has a frequency above the light line, it can decay into linear combinations of electromagnetic modes on the light line. These light line modes are the bulk vacuum electromagnetic solutions of the vacuum-metal system. During the breakup process the frequency and wave vector of the surface wave solution must be conserved as it breaks up into combinations of the lower frequency bulk modes. This is facilitated by the linear dispersion in (4.109) of frequency to wave vector which allows for sums of bulk modes with the same frequency and wave vectors as those of the surface solutions in the region above the light line [14].

A similar reasoning for the surface modes with frequencies below the light line can be given. For surface wave solutions with frequencies below the light line, no combinations of light line modes are available to accommodate the decay of the surface waves. As with the polariton modes on the planar vacuum-metal surface, the surface waves at frequencies below the light line are stable. In the case of the periodic profile, however, a series of stop and pass band arise as well as the presentation of all of the surface wave solutions within the first Brillouin zone. This acts as a complication in the treatment of the system [14].

In the limit that the periodic profile function in (4.107) vanishes the system approaches a flat surface and the surface waves of the grating reduce to the surface waves on the flat surface. It is illustrative of the point to consider the first Brillouin zone representation of the surface wave solution of the periodic system as they approach their flat surface limit of the periodic system. This involves essentially folding up the flat surface dispersion relation into the first Brillouin zone. Specifically, the surface wave dispersion relation on the flat surface given in Fig. 4.2 by $\omega(k)$ is replaced in this presentation by the dispersion relation $\omega_{BZ}(k_{BZ})$ in the first Brillouin zone defined by [14]

$$\omega_B(k_{BZ}) = \omega\left(k - n\frac{2\pi}{a}\right). \tag{4.110}$$

Here n is an integer chosen such that $k_{BZ} = k - n\frac{2\pi}{a}$ is a wave vector in the first Brillouin zone for the system of smallest repetition distance a.

In Fig. 4.7 results are shown for the flat surface dispersion relation represented as in (4.110) within the first Brillouin zone of a system with smallest repetition distance a. For a comparison the light line has been shown for the separation of modes into surface waves below the light line and leaky modes above the light line. As a weak periodicity of smallest repeat distance a is introduced into the modes of

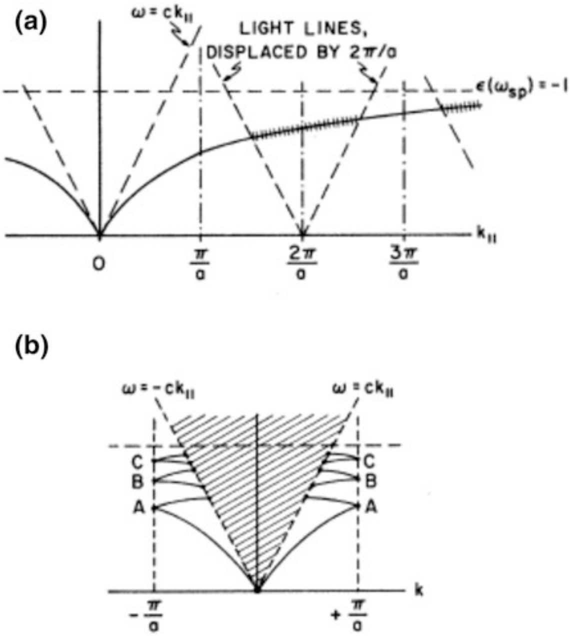

(a)
$\omega = ck_{\parallel}$
LIGHT LINES, DISPLACED BY $2\pi/a$
$\epsilon(\omega_{sp}) = -1$
$k_{\parallel}$
O
$\frac{\pi}{a}$
$\frac{2\pi}{a}$
$\frac{3\pi}{a}$
(b)
$\omega = -ck_{\parallel}$
$\omega = ck_{\parallel}$
C
B
A
C
B
A
k
$-\frac{\pi}{a}$
$+\frac{\pi}{a}$

Fig. 4.7 Plot of the folded flat surface dispersion relation into the first Brillouin zone of an imaginary periodic surface lattice with smallest repeat distance *a* [14]. In **a** the flat-surface dispersion relation is drawn in the extended-zone scheme. In **b** the non-radiating portions of the flat surface dispersion have been translated into the first Brillouin zone. This gives the reduced zone representation of the dispersion relation upon introducing the periodic grating. The cross-hatched regions become unstable with respect to radiation into the vacuum upon the introduction of the periodic grating. Reproduced with permission from [14]. Copyright 1981 American Physical Society

the plot small gaps are introduced at the edges of the Brillouin zone and at the edges of the light line. These gaps increase with increasing amplitude of the grating profile function and further leaky modes develop as the modal solutions pass through the light line limit.

4.2.3 Scattering from Rough Surfaces

Surface plasmons are often found to provide an important mechanism in the scattering of light from randomly rough surfaces [5, 9–11]. This is generally true for interfaces satisfying the conditions to support surface waves. In particular, at such surfaces the components of the diffuse scattering of light from the randomly rough surface are to a certain degree mediated through the virtual excitation of surfaces electromagnetic waves. These observations can be continued to the treatment of the scattering and transmission of light at and through a slab of materials supporting surface electromagnetic waves along its randomly rough interfaces [5, 9–11].

Light incident on a randomly rough surface is coupled by the roughness to the surface electromagnetic waves on the interface. In the absence of surface roughness the translational invariance of a planar interface requires that the bulk modes refracted and reflected by the interface are separate, distinct, modes from the surface waves. These distinct solutions are modes of fixed wave vector component parallel to the interface. Upon the introduction of surface roughness, however, the states of the system are no longer states of fixed wave vector components in the interface. The wave vector states are mixed by the loss of translational symmetry and this allows the incident wave to excite surface waves along the interface.

In this scheme the diffuse scattering from the rough surface proceeds as follows: Light is incident on the interface and part of it is directly reflected or transmitted through the interface. An additional part of the incident light is coupled through the rough interface and scattered into surface electromagnetic waves. These surface electromagnetic waves propagate along the interface, but, because of the roughness of the interface, they are eventually scattered into bulk modes that are reflected or transmitted through the surface. The higher order scattering processes in this scenario are all mediated by the surface electromagnetic waves.

This process has some interesting consequences as it allows the incident light to probe the nature of the surface waves on the rough interface. The wave function probe relates the wave functions of the surface waves to prominent features in the diffusely scattered light considered as a function of the scattering angle.

It is known that surface waves traveling on a randomly rough interface are Anderson localized by the disorder. The phase coherent scattering as the surface wave travels along the rough interface causes scattering from different parts of the surface to add not only in amplitude but in a phase interference sum. As a result, the waves in one- or two dimensional motion through a random medium can be shown to be localized, bound, states within the medium [6]. They are not modes extended throughout the entire system and they do not propagate through the system.

These considerations are for an infinite one- or two-dimensional medium. If a finite one- or two-dimensional piece of material is treated, it is possible that the length of the one- or two-dimensional medium over which the wave functions are localized could exceed the lengths of the finite media. In this case the localization effects would be less evident in the properties related to the wave functions of the excitations in one- or two-dimension.

In the case of scattering from a randomly rough interface, an anomaly associated with the Anderson localization of the surfaces waves is observed in the diffuse scattering of radiation from the interface. Scattered light moving in directions opposite to the light incident on the interface exhibits an enhancement peak in the intensity of diffusely scattered radiation considered as a function of the scattering angle. This is an enhanced retroreflectance and the height and width of the intensity peak of diffusely scattered light as a function of the scattering angle are directly related to the length scale of the region over which the localized wave functions of the surface waves are bound and confined on the interface [9].

Technically, in most cases the retroreflectance enhancement is the result of weak Anderson localization [9]. Weak Anderson localization is a phase coherence in the interaction of the surface excitations with the surface roughness as they propagate along the interface. This interaction leads to an enhanced cross section for the surface waves to be scattered in directions along the interface that are opposite to that in which they originally traveled before the scattering. The enhancement in the scattering cross section of the surface waves for directions opposite that in which the wave is originally moving is due to the phase interference of the scattered wave with itself. It is not found in the scattering of classical mechanical particles that do not carry phase information [9].

As the surface wave scattering increases and the dielectric losses of the system decrease to zero, weak Anderson localization drives the surface wave system to a phase transition resulting in strong Anderson localization of the surface waves. Strong Anderson localization is the case in which the intensity of the backscattering of the surface waves transforms the system into completely localized surface wave functions bound and confined to finite regions of the rough interface. Both weak and strong localizations of surface waves lead to the enhanced retroreflectance in the light scattered from the interface and into the bulk [9].

The transformation from weak to strong localization is a true second order phase transition, exhibiting the common properties of second order transitions observed in many systems, e.g., phase transitions in magnetic, ferroelectrics, liquid helium, superconductors, etc. The ordering in systems that undergo second order transitions begins to appear in the system before the transition to the fully ordered state actually occurs. It develops as long range fluctuations contained within finite regions in which the system locally adopts the appearance of the ordered system. Outside of the region of the correlated fluctuation the system ordering breaks down so that there is a typical correlation length past which the system appears more disordered than ordered. Weak Anderson localization is the exhibition of localized fluctuations in the system before the transition to the fully strong Anderson localized state [9].

In three dimensional systems, the Anderson localization transition, with increasing system disorder, is present as a second order transition. In these systems as the disorder in the media is increased the transition from weak to strong Anderson localization occurs when a particular degree of disorder is achieved within the system [9, 10].

In a famous result, it has been shown that in one- and two-dimensional disordered systems an arbitrarily small disorder leads the wave functions of the system to be strongly localized [6]. This, however, is true in the absence of losses in the media of the system or in the absence of thermal effects which can cause hopping between localized states of the system. In the optical systems considered later it is the dielectric losses of the media and the scattering of surface modes into bulk modes due to the surface disorder that causes the system to exhibit weak rather than strong localization [9, 10].

The effects of weak localization on rough surfaces is best studied for weakly rough surfaces in which the wavelength of the scattered light is much larger than the heights and depths of the surface profile and is of order of the typical distance between neighboring peaks and depths along the rough surface. This is the region in which the enhanced retroreflectance of the interfaces arises solely from weak localization effects [9].

For surfaces with larger height distributions and close neighboring peaks and valleys other physical effects contribute to the creation of an enhanced retroreflectance peak in the diffuse scattering from the surfaces and these effects can contribute to or overwhelm the enhancement effects that may be present due to weak localization. These other effects are commonly known as shadowing effects [9]. On strongly rough surfaces the light incident on the surface can cast shadows on the surface just as trees cast shadows on the ground in the daytime. Away from the retroreflection direction an observer of the surface sees a dimed landscape which is a mixture of bright patches of reflected sunlight and dark shadow patches. In the retroreflection direction, however the shadows are covered up behind the peaks along the rough surface and the observer sees only a full bright surface with no shadows. In this way the diffusely scattered light from the surface exhibits a retroreflection peak in the plot of the diffuse scattering as a function of angle above the surface.

Shadowing is an important effect in astronomy where it is referred to as the opposition effect [9]. When the Earth is between the Sun and a reflecting astronomical object a maximum is observed in the intensity of the light reflected from the object to an observer on Earth. This enhancement is even observed in the light scattered from interplanetary dust [9].

The important application of shadow casting effects is not limited only to astronomy. It is also significant in the scattering of light from many systems of application in engineering and condensed matter physics. It enters into the design of paints and coating, applications to radar, and to the study scattering in surface studies [9].

In the following an outline of analytical calculations of the retroreflectance due to the weak Anderson localization of plasmons on a metal-vacuum interface is given. The calculation shows many of the theoretical features that are found in systems exhibiting retroreflection and weak Anderson localization. In the treatment the surfaces will be weakly rough so that shadowing effects are excluded from consideration [9].

To make things simple, the disorder of the system is taken as a one-dimensional disorder. This provides the basis for an analytical approach revealing in a straightforward way much of the physics involved in the scattering process. As an additional point, experimental studies have been made on the one-dimensionally rough surfaces treated here [9]. Using photoresist methods and the generation of speckled light with the appropriate statistical properties, the one-dimensional random surfaces treated theoretically here have been made and are found to validate the calculations presented later.

Weak Localization in Diffuse Scattering: The Model

The one-dimensional rough surface is described by a continuous surface profile which on average is the x-z plane, with the z-axis normal to the average plane of the surface. For the case of one-dimensional disorder, the surface is taken to be translationally invariant parallel to the y-axis but is described by as statistical random continuous function along the x-axis.

A schematic of the surface and the scattering geometry is shown in Fig. 4.8. The vacuum-metal interface is described by a surface profile function of the form [9]

$$z_{surf} = \xi(x) \tag{4.111}$$

relating the z-coordinate of the surface to the coordinate on the x-axis. In the region above the surface (i.e., for $z > \xi(x)$) contains vacuum and the region below the surface (i.e., for $z < \xi(x)$) is filled with metal.

For this scattering geometry, an electromagnetic plane wave propagating in the x-z plane is incident from the vacuum onto the rough metal surface. It is reflected by the disordered interface, exhibiting both specular and diffuse components of reflection. The focus in the following is to relate the diffusely reflected light to the nature of the surface plasmons supported at the interface.

The semi-finite region of metal and the scattering surface are described by a complex dielectric function $\varepsilon_m(\omega) = \varepsilon_{m,1}(\omega) + i\varepsilon_{m,2}(\omega)$. For the scattering of visible light a silver surface is studied with the properties $|\varepsilon_{m,2}(\omega)| \ll |\varepsilon_{m,1}(\omega)|$. This assures that the scattering from the surface reflects features of weak Anderson

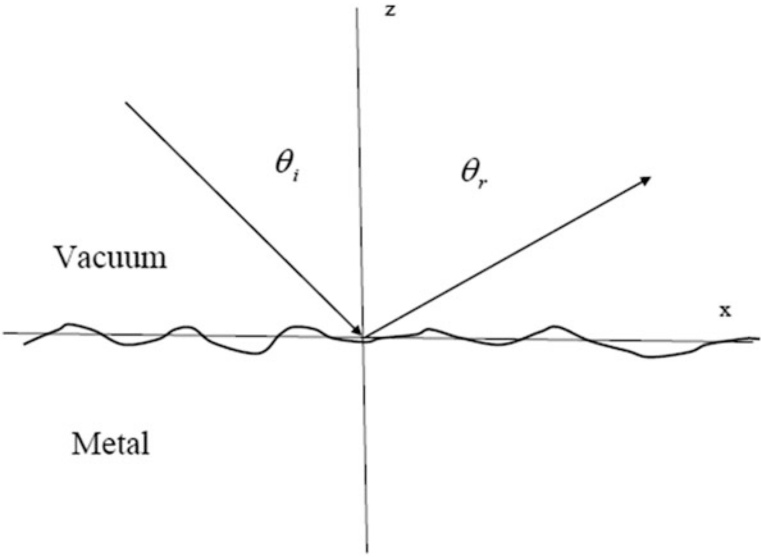


Fig. 4.8 Schematic of the scattering of an incident plane wave of light incident from vacuum on a randomly rough vaccum-metal surface [9]

localization rather than those of strong Anderson localization. In addition, it is seen later that $\varepsilon_{m,1}(\omega) < -1$ is required for surface polaritons to exist on the vacuum-metal interface [9].

For the treatment presented later, the function $\xi(x)$ is a continuous random function. In calculations, typically, $\xi(x)$ is chosen as one of a set of Gaussian random function, $\{\xi(x)\}$, with well defined statistical properties. This facilitates the formulation of the problem in an analytic Green's function approach which can be developed in terms of Feynman diagrammatics.

Once the set $\{\xi(x)\}$ is chosen, the physical properties of the system are formally determined in terms of a selected $\xi(x)$ surface from the set. At the end of the calculation an average is performed over the entire set $\{\xi(x)\}$. This is taken to give the average response of the random surface in terms of the statistical characterization of the set of surfaces. The procedure of obtaining the physical properties of the system by averaging over the set $\{\xi(x)\}$ can be shown to give the same result as averaging the physical properties over the entire length of one realization, $\xi(x)$, of the surface [9].

Statistical Properties of the Surface: Gaussian Random Surfaces

For the following discussions the set of surface profiles $\{\xi(x)\}$ is statistically characterized by [9]

$$\langle \xi(x) \rangle = 0 \tag{4.112a}$$

and

$$\langle \xi(x)\xi(x') \rangle = \sigma^2 \exp\left(-\frac{|x-x'|^2}{a^2}\right). \tag{4.112b}$$

In (4.112) $\langle\rangle$ indicates an average over the set of random profile functions, $\sigma^2 = \left\langle \xi^2(x) \right\rangle$ and a is the correlation length of the surface roughness. The correlation length gives an indication of the length along the surface over which a point on the surface profile is dependent on the values of its neighboring profile points.

The system is Gaussian random which means that the higher order correlation functions of $\xi(x)$ are written in terms of all possible pairing of the pair correlations in (4.112b). Due to (4.112a), products of odd numbers of $\xi(x)$ are zero and only even products are nonzero. As an example, the correlation function for a product of four surface profile functions is expressed as three term giving all of the possible pairings of the surface profile functions in the system. It is expressed as [9]

$$\begin{aligned}\langle \xi(x_1)\xi(x_2)\xi(x_3)\xi(x_4) \rangle = &\langle \xi(x_1)\xi(x_2) \rangle \langle \xi(x_3)\xi(x_4) \rangle \\ &+ \langle \xi(x_1)\xi(x_3) \rangle \langle \xi(x_2)\xi(x_4) \rangle \\ &+ \langle \xi(x_1)\xi(x_4) \rangle \langle \xi(x_2)\xi(x_3) \rangle.\end{aligned} \tag{4.113}$$

The importance of the Gaussian random statistics is that it allows from the development of a Wick's theorem for the scattering in the system. This is helpful as the resulting treatment of the scattering solutions can be written as an expansion in Feynman diagrams.

Scattering Calculation: A Green's Function Approach

A treatment of the scattering from the earlier described one-dimensionally rough surface is given, taking the *x*-*z* plane to be the plane of incidence of the incident light. The translational symmetry of the surface in the *y*-direction requires that both the incident and scattered waves propagate in the plane of incidence. In addition, for this scattering geometry the light can be treated as being composed of two different polarization components, a p-polarized component and/or an s-polarized component. The p-polarization component of light has its magnetic field polarized perpendicular to the plane of incident, while the s-polarization component of light has its electric field polarized perpendicular to the plane of incidence.

Due to the translational symmetry of the surface parallel to the *y*-axis the scattered light retains the polarization of the incident wave. The scattering of the two different polarization components do not mix, but the scattering of the incident p-polarized wave components are only into p polarized waves and the scattering of the incident s-polarized wave components are only into s polarized waves.

The separation into two polarization components is very important. It is found that incident light with p-polarization couples to the surfaces plasmons through the surface disorder. This is the only polarization that couples to the surface waves, and the coupling to the surface waves allows its scattering dynamics to reveal much more about the nature of the disordered surface than do scattering solutions that do not couple to the surface waves. The other s-polarized component of light does not couple to the surface plasmons through the roughness of the interface. Unlike the p-polarization the s-polarization does not exhibit any of the interesting scattering effects arising from interactions with surface plasmons. Consequently, only the p-polarization will be treated in the following.

In this case of p-polarized light, the magnetic field above the surface is polarized perpendicular to the plane of incidence and is of the form [9]

$$\vec{H}(x,z,t) = \big(0, H_y(x,z), 0\big)\mathrm{e}^{-i\omega t} \tag{4.114}$$

where

$$H_y(x,z,t) = \mathrm{e}^{ik_0x - i\alpha_0(k_0\omega)z} + \int \frac{dq}{2\pi} R(q|k_0)\mathrm{e}^{iqx + i\alpha_0(q\omega_0)z}. \tag{4.115}$$

Here the first term on the right in (4.115) is the incident wave on the interface and the second term on the right is the reflected wave, composed of both the specular and diffuse components of scattered light. In addition,

$$\alpha_0(q\omega) = \left[\frac{\omega^2}{c^2} - q^2\right]^{1/2}, \quad q^2 < \frac{\omega^2}{c^2} \tag{4.116a}$$

and

$$\alpha_0(q\omega) = i\left[q^2 - \frac{\omega^2}{c^2}\right]^{1/2}, \quad q^2 > \frac{\omega^2}{c^2} \tag{4.116b}$$

where for the incident wave $\alpha_0(k_0\omega)$ is real.

Equations (4.114)–(4.116) represent the standard form for an incident wave being diffusely scattered above a random rough surface. The object in the following is to determine $R(q|k_0)$ in terms of the dielectric properties and surface disorder of the interface. This is done by applying the boundary conditions at the random surface.

The form in (4.115) is used to match boundary conditions at the surface of the random interface. For this treatment care must be taken in applying random surface boundary conditions to the form in (4.115). Such an application may introduce additional assumptions into the treatment of the system. While the fields of the form in (4.115) are valid in the region $z > \xi(x)_{\max}$ where $\xi(x)_{\max}$ is the maximum of the surface profile function, it is only in a limited sense that the form in (4.115) is also valid within the entire region $z > \xi(x)$. In particular, the assumption that (4.115) can be continued to the region $z > \xi(x)$ may not be true if the roughness of the surface is to strong.

The assumption that (4.115) is the true form of the solution for the entire region $z > \xi(x)$ is commonly made in the study of surface physics of disordered interfaces [9]. It is known as the Rayleigh hypothesis, and it is only accurate for weakly rough surfaces. Its accuracy has been tested against results from computer simulation studies and some analytical works, and it is generally known to be a good assumption in the case that [9]

$$\frac{\sigma}{a} < 0.1. \tag{4.117}$$

Here (4.117) is characterizing the surface disorder in terms of the parameters for the Gaussian random surfaces in (4.112) and (4.113). The condition in (4.117) is also consistent with the requirement that shadowing effects are absent from the random rough surface so that the validity of the Rayleigh hypothesis is a very good assumption in the calculations that follow.

Matching the boundary conditions at $z = \xi(x)$, Brown et al. [11], have shown that in the weak roughness limit the reflection amplitude in (4.115) is expressed as a series in terms of the surface profile. This is done using an integral form for the refracted fields below the surface which is treated using the Rayleigh hypothesis, Green' theorem, and the extinction theorem to match the fields in (4.115) above the surface. For the details of this development, the reader is referred to the literature [11].

For the discussions here it is only important to understand the details of the fields above the rough interface and the physical insight that their representations give regarding the processes contributing to the diffuse scattering of radiation from the interface.

From the Brown et al. [11] formulation the reflection amplitude of light of frequency ω is expressed as a series of scattering and multiple scattering events from the rough interface. It has the standard form [9, 11]

$$R(q|k_0) = 2\pi\delta(q - k_0)R_0(k_0) - 2iG_0(q)T(q|k_0)G_0(k_0\alpha(k_0)). \tag{4.118}$$

In (4.118) the first term on the right of the equation is the Fresnel coefficient for the reflection of p-polarized light from a flat interface. It is given by

$$R_0(k) = \frac{\varepsilon_m(\omega)\alpha_0(k\omega) - \alpha(k\omega)}{\varepsilon_m(\omega)\alpha_0(k\omega) + \alpha(k\omega)} \tag{4.119}$$

where

$$\alpha(k\omega) = \left[\varepsilon_m(\omega)\frac{\omega^2}{c^2} - k^2\right]^{1/2} \tag{4.120}$$

with $\text{Re}\alpha(k\omega) > 0$ and $\text{Im}\alpha(k\omega > 0)$, and $\alpha_0(k\omega)$ is from (4.116). This describes the specular reflection of the incident fields from the surface. The remaining term on the right represents the scattering and multiple scattering from the rough interface which will be the focus of the discussion of plasmon effects in the diffuse scattering of light form the interface.

The second terms on the right in (4.118) represent the scattering and multiple scattering terms of light from the interface disorder. These processes gives rise to diffuse scattering from the rough surface and to the renormalization of the specular reflection as light is scattered from the specular to the diffuse components with increasing surface disorder.

In the second term, the Green's function for the propagation of surface polaritons of frequency ω on the flat interface is [9, 11]

$$G_0(k) = \frac{i\varepsilon_m(\omega)}{\varepsilon_m(\omega)\alpha_0(k\omega) + \alpha(k\omega)}, \tag{4.121a}$$

and the T-matrix describing the scattering of a surface polaritons as it moves along the interface is given by the form [9, 11]

$$T(p|k) = V(p|k) + \int \frac{dq}{2\pi} V(p|q)G_0(q)T(q|k), \tag{4.121b}$$

Here, in the T-matrix defined in (4.121b), the surface plasmon scattering potential for a plasmon of frequency ω is given by

$$V(p|k) = \frac{\varepsilon_m(\omega) - 1}{\varepsilon_m^2(\omega)} \hat{\xi}(p-k)[\varepsilon_m(\omega)pk - \alpha(p\omega)\alpha(k\omega)], \tag{4.121c}$$

where

$$\hat{\xi}(p) = \int dx \, \exp[-ipx]. \tag{4.121d}$$

The potential in (4.121c) is correct to leading order in the surface roughness. Higher order terms which do not qualitatively affect the later discussions have been neglected in (4.121c).

The T-matrix in (4.121b) describes the repeated scattering of surface plasmons as they propagate along the rough interface. It is of the standard form of a T-matrix describing motion through a disordered media. The flat surface plasmon moves along the surface and, in the leading term of the T-matrix series, it is scattered once by the surface roughness. This is the first term in the scattering series of the T-matrix. In the second term of the series the flat surface plasmon is scattered twice by the random surface. In the nth term of the series the flat surface plasmon is scattered n times by the surface, etc. For a weakly rough surfaces the terms of successively higher number of scatterings should be successively decreasing with the increasing number of scattering events. All of the scattering information about surface plasmon propagation is eventually contained within the sum of the T-matrix series. Consequently, the reflection coefficient in (4.115) is expressed in terms of the set of scattering processes for plasmon propagation along the interface contained within the T-matrix.

Another important function for characterizing the propagation of surface plasmons along the random interface is the surface plasmon Green's function. It contains the entire single plasmon response for propagation along the interface and like the reflection amplitude is also expressed in terms of the T-matrix for scattering along the surface. The two response functions, the scattering amplitude and the surface plasmon Green's functions, will be shown in the following to be closely related to one another.

In terms of the T-matrix series in (4.121b) the Green's function for the propagation of surface plasmons on the rough surface is [9, 11]

$$G(p|k) = 2\pi\delta(p-k)G_0(k) - G_0(p)T(p|k)G_0(k). \tag{4.122}$$

This is the standard Dyson equation result for the full Green's function of the random system in terms of the Green's function in the absence of scattering and the scattering potential arising from the interaction of the surface waves with surface roughness. The Green's function in (4.122) explains the response of the surface electromagnetic waves on the random interface as a sequence of scattering interactions with the surface roughness, relating an inputted stimuli to the rough surface plasmons to their outputted response from the system.

From (4.118) and (4.122) a simple relationship is found between the $R(q|k_0)$ and $G(q|k_0)$ responses of the system. Specifically, it is follows from these equations that [9, 11]

$$R(q|k_0) = 2\pi\delta(q - k_0)[R_0(k_0) - 2iG(k_0)\alpha_0(k)] + 2iG(q|k_0)\alpha(k_0). \tag{4.123}$$

The differential reflection coefficient for the scattering of radiation from the rough surface is computed from (4.115). It is related to the angular distribution in the far field limit of the ratio of the scattered to the incident light power flow above the surface. This ratio is computed as the quotient of the Poynting vector of the scattered radiation obtained from the second term on the right side of (4.115) divided by the Poynting vector of the incident radiation from the first term on the right side of (4.115). In particular, the ratio of these Poynting vectors averaged over the surface roughness is

$$\frac{|S_{refl}|}{|S_{inc}|} = \frac{1}{L} \int\limits_{q^2 < \frac{\omega^2}{c^2}} \frac{dq}{2\pi} \frac{\alpha_0(q)}{\alpha_0(k_0)} \left\langle |R(q|k_0)|^2 \right\rangle, \tag{4.124}$$

where $q = \frac{\omega}{c} \sin\,\theta_r, k_0 = \frac{\omega}{c} \sin\,\theta_i$ in terms of the angles θ_r and θ_i of the reflected and incident radiation, respectively, and ω is the frequency of the elastically scattered light.

The integral in (4.124) can be rewritten as an integration in the scattering angle θ_r of the reflected radiation to obtain

$$\frac{|S_{refl}|}{|S_{inc}|} = \int d\theta_r \frac{\partial R}{\partial \theta_r}. \tag{4.125}$$

Here the differential reflection coefficient for the diffuse scattering is given by [9, 11]

$$\frac{\partial R}{\partial \theta_r} = \frac{\omega}{2\pi c} \frac{\cos^2\,\theta_r}{\cos\,\theta_i} \frac{\left\langle |R(q|k_0)|^2 \right\rangle}{L}. \tag{4.126}$$

In (4.126) it is seen that the diffuse reflection from the surface is expressed directly in terms of the reflection amplitude of the scattering form in (4.115).

Applying (4.123) to the (4.126) for differential reflection coefficient of the diffuse reflection of radiation from the surface yields a form rewritten in terms of the Green's function for the surface plasmons on the rough interface. Performing this, (4.126) then becomes [9]

$$\frac{\partial R}{\partial \theta_r} = \frac{1}{L}\frac{2}{\pi}\left(\frac{\omega}{c}\right)^3 \cos^2\,\theta_r\,\cos\,\theta_i \left\langle |G(q|k_0)|^2 \right\rangle_{diffuse}. \tag{4.127}$$

This relates the differential reflection coefficient for the diffusely scattered radiation directly to the Green's function response of the plasmons on the rough interface. A study of the Green's function for the propagation of surface waves along the interface contains all of the physics of the diffuse scattering of the p-polarization of electromagnetic waves from the random interface. The study of the surface plasmon Green's function is next addressed.

Surface Plasmon Green's Function

The function of interest in (4.127)

$$\left\langle |G(q|k_0)|^2 \right\rangle \tag{4.128}$$

is technically known as a two-particle Green's function. It involves the product of two single particle plasmon Green's functions along the interface and contains much more information about the response of the system of plasmons to external stimuli and to each other than is contained in the single particle Green's functions. In many body theories the two particle Green's functions generally are responsible for the transport functions of the system, e.g., the electrical conductivity, the magnetic and electric susceptibilities, etc. Single particle Green's functions such as those in (4.122) do not describe the transport properties of systems but are closely related to the dispersion relations and lifetime of the excitations in many body systems. The single particle Green's functions are obtained by directly averaging (4.122) over the interface, probing the effects of the system on a single plasmon propagating at the interface. These are first discussed followed by the study of the two-particle Green's functions.

Averaging (4.122) over the random surface, the single plasmon Green's function for a surface plasmon of frequency ω is found to be given by [9]

$$\langle G(q|k)\rangle = 2\pi\delta(q-k)G(k) \tag{4.129}$$

where

$$G(k) \cong \frac{C_1(\omega)}{k - K_{SP}(\omega) - i\Delta_{Total}(\omega)} - \frac{C_1(\omega)}{k + K_{SP}(\omega) + i\Delta_{Total}(\omega)}, \tag{4.130a}$$

$$C_1(\omega) = \frac{\varepsilon_{m,1}(\omega)\sqrt{-\varepsilon_{m,1}(\omega)}}{1-\varepsilon_{m,1}^2(\omega)}, \tag{4.130b}$$

$$K_{SP}(\omega) = \frac{\omega}{c}\left[\frac{\varepsilon_{m,1}(\omega)}{\varepsilon_{m,1}(\omega)+1}\right]^{1/2}, \tag{4.130c}$$

$$\Delta_{Total}(\omega) = \Delta_{\varepsilon}(\omega) + \Delta_{SP}(\omega), \tag{4.130d}$$

$$\Delta_{\varepsilon}(\omega) = \frac{\varepsilon_{m,2}(\omega)K_{sp}}{2\varepsilon_{m,1}(\omega)\left[\varepsilon_{m,1}(\omega)+1\right]}, \tag{4.130e}$$

and

$$\Delta_{SP}(\omega) \cong 2\sqrt{\pi}a\sigma^2 C_1^2(\omega)\left[\frac{\varepsilon_{m,1}(\omega)-1}{\varepsilon_{m,1}(\omega)}\right]^2 K_{SP}^4(\omega)\exp\left(-a^2K_{SP}^2(\omega)\right) \tag{4.130f}$$

Equation (4.130a) is the standard from of a single particle Green's function with poles in the complex frequency wave vector plane representing the dispersive properties of the surfaces plasmon-polaritons.

In the Green's function of (4.130a), $K_{SP}(\omega)$ is the surface plasmon-polariton wave vector for a surface plasmon-polarition of frequency ω. It is expressed in terms of $\varepsilon_{m,1}(\omega)$ in (4.130c). In addition, $\Delta_{Total}(\omega)$ in (4.130a) is the imaginary part of the Green's function pole. It represents the decay of the plasmon-polarition as it propagates along the interface. From (4.130d), $\Delta_{Total}(\omega)$ is seen to include the lowest order losses coming from the surface roughness scattering and the dielectric losses to the $k \approx \pm K_{SP}(\omega)$ surface plasmon-polariton poles from the materials forming the interface. The lowest order contribution in the surface roughness losses is given by $\Delta_{SP}(\omega)$ and enters the pole structure of $G(k)$ in (4.130a) through $\Delta_{Total}(\omega) \approx \Delta_{\varepsilon}(\omega) + \Delta_{SP}(\omega)$ in (4.130d). The $\Delta_{\varepsilon}(\omega)$ contribution to $\Delta_{Total}(\omega)$ in (4.130d) is given in (4.130e) and represents Joule losses in the system from the imaginary part of the dielectric constant of the metal. This is solely a materials property. The residue of the Green's function pole, $C_1(\omega)$, is given in (4.130b).

The functions in (4.129) and (4.130) are seen to be of importance in determining the dispersive properties of the single plasmon modes along the interface. Specifically, the pole singularities of $G(k)$ give the frequency versus wave vector renormalized dispersion relation for the surface plasmons. From the imaginary parts of the frequency poles of $G(k)$ the life time of the rough surface plasmons before they are scattered away into other modes of the system is obtained. In addition, the imaginary parts of the wave vector poles of $G(k)$ give the propagation distance of the plasmons on the rough surface before they are scattered away into other modes of the system. From (4.130a) the length of propagation along the surface of the plasmons is $\frac{1}{\Delta_{Total}}$ so that the plasmons decay in space by the exponential form $\mathrm{e}^{-\Delta_{Total}x}$.

In order to compute the diffuse scattering of radiation of frequency ω from the rough interface, it is necessary to study the two particle Green's function for the propagation of surface polaritons of frequency ω on the rough interface. The specific form of the two particle Green's function that is of interest in the calculation of the diffuse scattering is given by the surface averaged function given by $\langle G(q|k)G^*(p|k)\rangle$. This two particle Green's function is expressed in terms of a series expansion in scattering and multiple scattering processes of the plasmon with the rough interface. The series is an expansion which is reminiscent of the T-matrix treatment of the averaged single particle Green's function in (4.122) and (4.129).

The two-particle Green's function $\langle G(q|k)G^{*}(p|k)\rangle$ is computed from (4.122) by taking the product of $G(q|k)$ and $G^{*}(p|k)$ given by the expansion in (4.122). The product of the left hand sides of (4.122) give $G(q|k)G^{*}(p|k)$, and the product of the right hand sides of (4.122) result in a complicated series involving the T-matrix scattering terms. Both sides of the resulting products are averaged, with the left hand side giving $\langle G(q|k)G^{*}(p|k)\rangle$ and the righthand side giving an expression in the average scattering from the surface. The averages of the righthand side over the rough interface is made by applying the properties of the Gaussian random surface averages in (4.112) and (4.113).

Performing this averaging and following some algebra a Bethe-Salpeter integral equation for the two particle Green's function is obtain. The specific form of the Bethe-Salpeter equation for the two particle Green's function is [9]

$$\begin{aligned}\langle G(q|k)G^{*}(p|k)\rangle &= 2\pi\delta(q-k)2\pi\delta(p-k)|G(k)|^{2}\\ &\quad + G(q)G^{*}(p)\int\frac{dr}{2\pi}\int\frac{ds}{2\pi}\langle\Gamma(qr|ps)\rangle\langle G(r|k)G^{*}(s|k)\rangle.\end{aligned} \tag{4.131}$$

The equation in (4.131) expresses the desired two-particle Green's function $\langle G(q|k)G^{*}(p|k)\rangle$ as a series sum of scattering and multiple scattering process with the random surface. The irreducible vertex function, $\langle\Gamma(qr|ps)\rangle$, contains all of the fundamental scattering process contributing to the propagation of the two interacting plasmons as they move along the interface.

The first term on the righthand side of (4.131) describes two plasmons moving along the system without interacting with one another. The higher order processes in the second term on the righthand side describe higher order correlated scatterings of the two surface plasmons by the surface roughness. Details of the processes that enter into $\langle\Gamma(qr|ps)\rangle$ will be discussed later, but first some remarks on the general form of $\langle G(q|k)G^{*}(p|k)\rangle$ and $\langle\Gamma(qr|ps)\rangle$ must be made. These lead to a simplification of the treatment of (4.131).

In (4.131) the average on the left side of the equality is given by the general form

$$\langle G(q|k)G^{*}(p|k)\rangle = 2\pi\delta(q-p)G_{1}(p|k). \tag{4.132}$$

Here averaging the product $G(q|k)G^{*}(p|k)$ over the surface randomness restores translational invariance along the average interface, resulting in the delta-function on the right side. The restoration of translational invariance along the average interface also applies to the vertex function in (4.132) which has the general form

$$\langle\Gamma(qr|ps)\rangle = 2\pi\delta(q-r-p+s)\Gamma_{0}(qr|ps), \tag{4.133}$$

again reflecting the restored translational symmetry of the averaged system. The extraction of the delta-functions that arise upon the surface averaging contributes a great simplification in the form of (4.131) and results in a more tractable integral equation for the determination of the two-particle Green's functions.

Placing the delta-function forms in (4.132) and (4.133) into (4.131) results in an expression for $G_1(p|k)$ in terms of $\Gamma_0(qr|ps)$. Specifically, the function $G_1(p|k)$ is obtained as a solution of the integral equations

$$G_1(q|k) = 2\pi\delta(q-k)|G(k)|^2 + |G(q)|^2 \int \frac{dr}{2\pi} \Gamma_0(qr|qr) G_1(r|k). \tag{4.134}$$

In this integral equation for $G_1(q|k)$ the first term on the right is the only delta-function in the equation. It represents two plasmons propagating in the system without interacting with one another through correlated scattering with the surface roughness. It does not contribute to diffuse scattering in the system. All of the diffuse scattering contributions are contributed from the second term on the right hand sider of (4.134).

It is seen from (4.127) that the two-particle Green's function $\left\langle |G(q|k_0)|^2 \right\rangle$ is directly related to the diffuse scattering. The relevant Green's function is obtained from (4.132) and (4.134) to be of the form

$$\langle G(q|k) G^*(q|k) \rangle = L G_1(q|k). \tag{4.135}$$

where the relationship between the length of the scattering surface L and the delta-function has been used, i.e.,

$$L = 2\pi\delta(0) = \int dx \tag{4.136}$$

Applying (4.127) and (4.135) it then follows that the differential reflection coefficient for the diffuse scattering is [9]

$$\frac{\partial R}{\partial \theta_r} = \frac{2}{\pi}\left(\frac{\omega}{c}\right)^3 \cos^2\theta_r \cos\theta_i G_1(q|k_0)_{diffuse}, \tag{4.137}$$

where $q = \frac{\omega}{c}\sin\theta_r$ and $k_0 = \frac{\omega}{c}\sin\theta_i$ satisfy the conditions of elastic diffuse scattering, i.e., $q \neq k_0$.

For a weakly rough surface the leading order scattering term in the irreducible vertex function is

$$\langle \Gamma(qr|ps) \rangle \approx \langle V(q|r) V^*(p|s) \rangle = 2\pi\delta(q-r-p+s) V_0(qr|ps), \tag{4.138a}$$

where.

$$V_0(qr|ps) = \sqrt{\pi}a\sigma^2\left|\frac{\varepsilon_m(\omega)-1}{\varepsilon_m^2(\omega)}\right|^2[\varepsilon_m(\omega)qr-\alpha(q\omega)\alpha(r\omega)][\varepsilon_m(\omega)ps-\alpha(p\omega)\alpha(s\omega)]^*\mathrm{e}^{-\frac{1}{4}a^2(q-r)^2}. \tag{4.138b}$$

This is the lowest order term in the scattering interaction between the two surface plasmons with the rough surface. The scattering described by it represents the correlated scattering of the two plasmons from the same region of the random interface.

For the evaluation of (4.134) at this level of approximation

$$\Gamma_0(qr|qr) \approx V_0(qr|qr), \tag{4.139}$$

and the full solution of (4.134) for the interaction in (4.139) is generated iteratively. Upon iterating (4.134) a series is developed of the form [9]

$$\begin{aligned} G_1(q|k) = {} & 2\pi\delta(q-k)|G(k)|^2 + |G(q)|^2\Gamma_0(qk|qk)|G(k)|^2 \\ & + |G(q)|^2\int\frac{dr}{2\pi}\Gamma_0(qr|qr)|G(r)|^2\Gamma_0(rk\backslash rk)|G(k)|^2 \\ & + |G(q)|^2\int\frac{dr}{2\pi}\Gamma_0(qr|qr)|G(r)|^2\int\frac{dr'}{2\pi}\Gamma_0(rr'|rr')|G(r')|^2\Gamma_0(r'k|r'k)|G(k)|^2 \\ & + \ldots \end{aligned} \tag{4.140}$$

where here the first four terms of the iterative interaction are displayed.

The iterated series in (4.140) is represented in Feynman diagrams as the sum over a set of so-called ladder diagrams shown in Fig. 4.9. The two solid parallel horizontal lines in the figure represent the Green's functions of two propagating polaritons and the successive scatterings of the system are described by the vertical

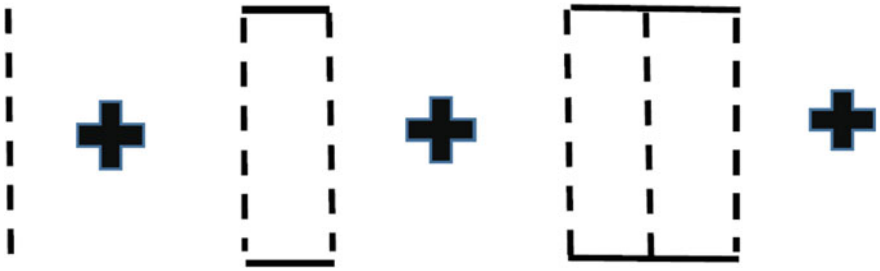


Fig. 4.9 Ladder diagram contributions to the irreducible vertex [9]

dashed lines. The surface waves propagate freely between consecutive correlated surface scatterings of the two plasmons. For weak scattering each term in the series is weaker than the previous terms in the series and the higher order terms in the series correspond to processes with increasing amounts of surface scatterings [9].

The ladder diagram approximation is a commonly applied approximation in the discussion of transport properties of many-body systems expressed in terms of two-particle Green's functions [9]. As an example, the transport properties of electrical and thermal conductivities are related, respectively, to the two-particle electron Green's functions and the two-particle phonon Green's functions of electron and insulating lattice systems. Additionally, the two-particle Green's functions also enter into the scattering of neutrons from magnetic systems and in the magnetic susceptibilities exhibited by magnetic materials.

In these theories the ladder approximation is responsible for the leading order behavior of the conductivities and susceptibilities and scatterings displayed in these various systems. The results from the ladder diagram treatment are also found in elementary discussions based on kinetic theory. In the scattering of light from the randomly rough surface, the ladder diagram approximation contributes to a general diffused scattering above the surface. The resulting diffuse scattering of light is continuous and slowly varying in the scattering angle above the surface [9].

To evaluate the integral equation of (4.134) and (4.140) it is useful to apply the pole approximation in evaluating the integrals over r, r', etc. In the pole approximation it is shown that the dominant contribution to the integrand comes from the overlapping poles of the Green's function product forms (i.e., the $|G(r)|^2$) within the various integrands of the iterative series in (4.140). In the limit of weak scattering, the overlap of the Green's function product poles, for the purposes of doing the integrals in (4.134) and (4.140), is described as [9]

$$|G(r)|^2 \cong \frac{\pi C_1^2}{\Delta_{Total}}[\delta(r - K_{SP}) + \delta(r + K_{SP})]. \tag{4.141}$$

The form in (4.141) is essentially a residue times the sum of delta-functions at the singularities of the surface plasmon Green's function. With each iteration of the series in (4.134) and (4.140) a new integral is generated involving a new integration variable r'' with a new $|G(r'')|^2$ which again dominates the integral in r''. Using (4.141) in the evaluations of the integrals of the iterative series generated in (4.140) reduces the series of terms involving multiple integrals to an algebraic series of algebraic terms.

Applying the pole approximation in (4.141) to (4.140) and summing the resulting algebraic series gives the following result for the two-particle Green's function for the diffuse scattering within the ladder approximation [9]

$$\left\langle |G(q|k_0)|^2 \right\rangle = L\{2\pi\delta(q-k_0)|G(k_0)|^2 + |G(q)|^2|G(k_0)|^2 \\ \times \left[K(q|k_0) + C_1^2/\left[2\Delta_{Total}\left(1-(\Delta_{SP}/\Delta_{Total})^2\right)\right]\right. \\ \times (K(q|K_{SP})K(K_{SP}|k_0) + K(q|-K_{SP})K(-K_{SP}|k_0) \\ + (\Delta_{SP}/\Delta_{Total})[K(q|K_{SP})K(-K_{SP}|k_0) + K(q|-K_{SP})K(K_{SP}|k_0)])]\} \tag{4.142}$$

Here [9]

$$K(q|k) = \sqrt{\pi}a\sigma^2\left|\frac{\varepsilon_m(\omega)-1}{\varepsilon_m^2(\omega)}\right|^2|\varepsilon_m(\omega)qk - \alpha(q\omega)\alpha(k\omega)|^2\exp\left[-\frac{1}{4}a^2(q-k)^2\right]. \tag{4.143}$$

Then from (4.127) and (4.137) the differential reflection coefficient for the diffuse scattering within the ladder approximation is obtained as [9].

$$\left(\frac{\partial R}{\partial\theta_r}\right)_L = \frac{2}{\pi}\left(\frac{\omega}{c}\right)^3\cos^2\theta_r\cos\theta_i|G(q)|^2|G(k_0)|^2 \\ \times \left[K(q|k_0) + C_1^2/\left[2\Delta_{Total}\left(1-(\Delta_{SP}/\Delta_{Total})^2\right)\right]\right. \\ \times (K(q|K_{SP})K(K_{SP}|k_0) + K(q|-K_{SP})K(-K_{SP}|k_0) \\ + (\Delta_{SP}/\Delta_{Total})[K(q|K_{SP})K(-K_{SP}|k_0) + K(q|-K_{SP})K(K_{SP}|k_0)])]\}. \tag{4.144}$$

Equation (4.144) provides for a general diffuse scattering above the rough interface, but while it treats important contributions to the surface scattering it leaves off some other very important contributions. By its nature the ladder approximation does not include coherent processes. It is a ballistic approach to the transport problem [9].

Some of the phase coherent processes that are left off in the ladder approximation contribute to weak Anderson localization effects that can be prominent in the scattering from randomly rough interfaces. In the following these types of phase coherent scattering processes are now included in the determination of the differential reflection coefficient from the rough interface. The total diffuse scattering is then given as a sum of scattering terms from the ladder approximation added with the phase coherent scattering which is now treated.

Phase coherent processes that contribute to weak localization effects are the contributions from diagrams know as maximally crossed diagrams. These types of diagrams are shown in Fig. 4.10, and represent processes in which the scattering of the two plasmons encounter the surface roughness scattering in reversed order. The diagrams arise naturally in the pairings arising in (4.113).

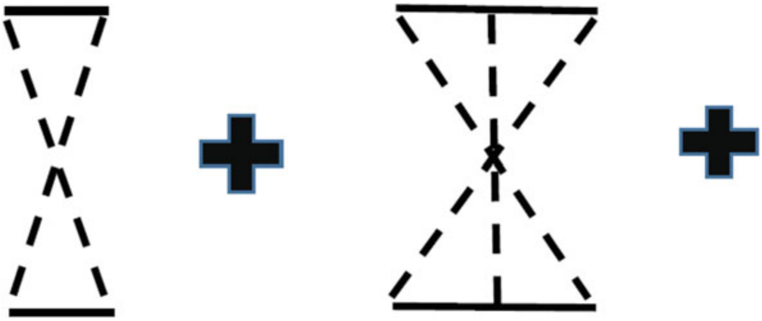


Fig. 4.10 Maximally crossed diagram contribution to the irreducible vertex [9]

The maximally crossed contributions and their localization and weak localization effects in transport processes have been studied for a variety of electron, phonon, and photon systems. A good explanation and interpretation of the how the phase coherent effects enter in these diagrams has been offer by Bergmann and can be found in [6]. In the following only the first diagram of the series is discussed. For surfaces with dielectric losses this diagram contributes most of the weak localization effects.

The contribution to the irreducible vertex function from the first diagram in Fig. 4.10 is [9]

$$\langle \Gamma_{crossed}(qk|qk)\rangle = \int \frac{ds}{2\pi} \int \frac{dt}{2\pi} \langle V(q|s)V^*(t|k)\rangle G(s)G^*(t)\langle V(s|k)V^*(q|t)\rangle. \tag{4.145}$$

The two plasmons on the surface are seen to encounter the rough surface, scattering from the surface roughness in reverse order. The average over the surface disorder, again, restores the translational symmetry to the theory so that, as with (4.135), the vertex contribution in (4.145) takes the form

$$\langle \Gamma_{crossed}(qk|qk)\rangle = L\Gamma_{crossed,0}(qk|qk) \tag{4.146}$$

with $L = 2\pi\delta(0) = \int dx$. Consequently, it follows from (4.145) that [9]

$$\Gamma_{crossed,0} = \int \frac{dt}{2\pi} V_0(q, q+k-t|t,k)G(q+k-t)G^*(t)V_0(q+k-t,k|q,t). \tag{4.147}$$

In the evaluation of (4.147) the dominant contribution to the integrand comes from the poles of the Green's function product, $G(q+k-t)G(t)$. For these integrals, as with the approximation in (4.141) used to evaluate the integrals of the ladder diagrams, an approximation for $G(q+k-t)G(t)$ can be written as

$$G(s)G^*(r-s) \cong \frac{4\pi C_1^2 \Delta_{Total}}{r^2 + 4\Delta_{Total}^2}[\delta(s-K_{SP}) + \delta(s+K_{SP})]. \tag{4.148}$$

Using (4.148) in the evaluation of (4.147) and following some algebra gives [9]

$$\Gamma_{crossed,0}(qk|qk) \cong \frac{4C_1^2 \Delta_{Total}}{(q+k)^2 + 4\Delta_{Total}^2} K(k|K_{SP})K(k|-K_{SP}) \tag{4.149}$$

where $K(r|t)$ is defined in (4.143).

The irreducible vertex contribution in (4.149) has an interesting functional form, showing the effects of weak localization. Notice that (4.149) is basically a Lorentzian function of $q+k$ and gives a maximum contribution to the scattering at $q=-k$. This maximum shows up in the system as an enhancement of backscattering in the propagation of the surface plasmons and will be seen later to lead to enhanced retroreflection in the diffuse scattering of light from the surface.

Before discussing these effects and the total diffuse scattering from the surface, some remarks will be made about the effect of including the complete set of maximally crossed diagrams in Fig. 4.10. These will be seen to give a renormalization of the retroreflection from the surface but do not change the qualitative nature of the diffuse scattering from the rough surface.

The irreducible vertex including all of the maximally crossed contributions in Fig. 4.10 has a similar form to that of the first crossed diagram in (4.146). In particular, due to the restoration of translation symmetry upon surface averaging, the sum of the maximally crossed terms takes the form

$$\left\langle \Gamma_{MCrossed}^{T}(qk|qk) \right\rangle = L\Gamma_{MCrossed}(qk|qk). \tag{4.150}$$

The sum of diagrams in Fig. 4.10 is done in [9] where it is shown that

$$\Gamma_{MCrossed}(qk|qk) = \frac{4C_1^2 \Delta_{Total}}{(q+k)^2 + 4\Delta_{Total}^2} \frac{1}{1-(\Delta_{SP}/\Delta_{Total})^2}[A(k)+B(k)]. \tag{4.151}$$

Here

$$A(k) = K(k|K_{SP})K(k|-K_{SP}) \tag{4.152a}$$

and

$$B(k) = \frac{1}{2}\frac{\Delta_{SP}}{\Delta_{Total}}\left(K^2(k|K_{SP}) + K^2(k|-K_{SP})\right) \tag{4.152b}$$

where $K(r|t)$ is defined in (4.143).

Using the irreducible vertex in (4.151) the contributions of the maximally crossed diagrams to the diffuse scattering cross section for the random surface is obtained. Specifically, form (4.118), (4.123), (4.125), (4.131), (4.150), and (4.151) these contributions are [9]

$$\left(\frac{\partial R}{\partial \theta_r}\right)_{MCrossed} = \frac{2}{\pi}\left(\frac{\omega}{c}\right)^3 \cos^3\theta_r \cos\theta_i |G(q)|^2 |G(k_0)|^2 \\ \frac{4C_1^2\Delta_{Total}}{(q+k_0)^2+4\Delta_{Total}^2}\frac{1}{1-(\Delta_{SP}/\Delta_{Total})^2}[A(k)+B(k)]. \tag{4.153}$$

The total contribution to the diffuse scattering from the rough surface is then obtained from (4.144) and (4.153) as [9]

$$\left(\frac{\partial R}{\partial \theta_r}\right) = \left(\frac{\partial R}{\partial \theta_r}\right)_{Ladder} + \left(\frac{\partial R}{\partial \theta_r}\right)_{MCrossed}. \tag{4.154}$$

In the following some examples of the scattering from (4.154) for weakly rough surfaces are given. These illustrate the typical weak localization effects found in the scattering from weakly rough surfaces that support surface plasmons.

Illustrative Example

The total diffuse scattering exhibited by (4.154) has been computed for visible light at 4579 Å scattering from a randomly rough vacuum-silver surface. Results for the differential reflection coefficient of light incident from vacuum at an incident angle of 20° is plotted in Fig. 4.11 as a function of the diffuse scattering angle. The

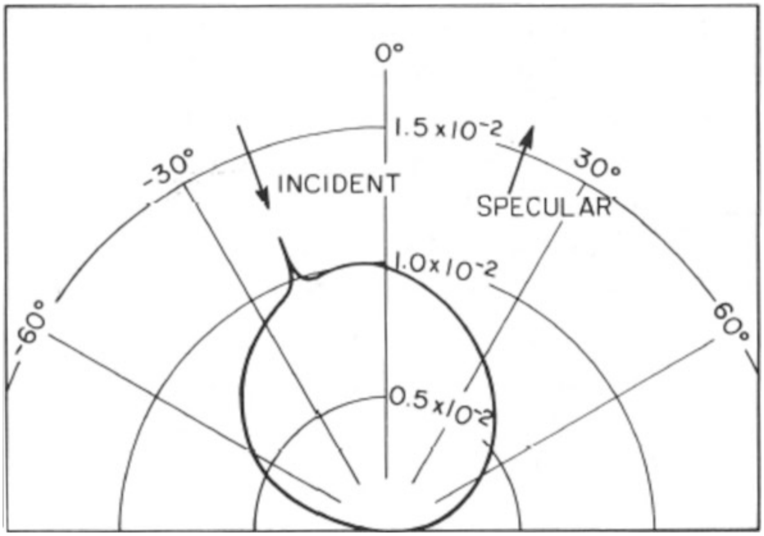


Fig. 4.11 Enhanced backscattering from the analytic theory [9]. Reproduced with permission from [9]. Copyright 1990 Elsevier

surface roughness for the plot is Gaussian random and characterized by the parameters $a = 1000$ Å, $\sigma = 50$ Å [9, 12].

In the plot a general diffuse scattering is observed at all scattering angles above the surface. This comes from the contribution of the ladder diagrams. In addition, a sharp peak in the diffuse scattering is observed in the light diffusely reflected opposite to the incident plane wave direction. This is the enhanced backscattering or refroreflectance of light. It comes from the contribution of the maximally crossed diagrams and is significant only over a range of a couple of degrees around the backscattering direction.

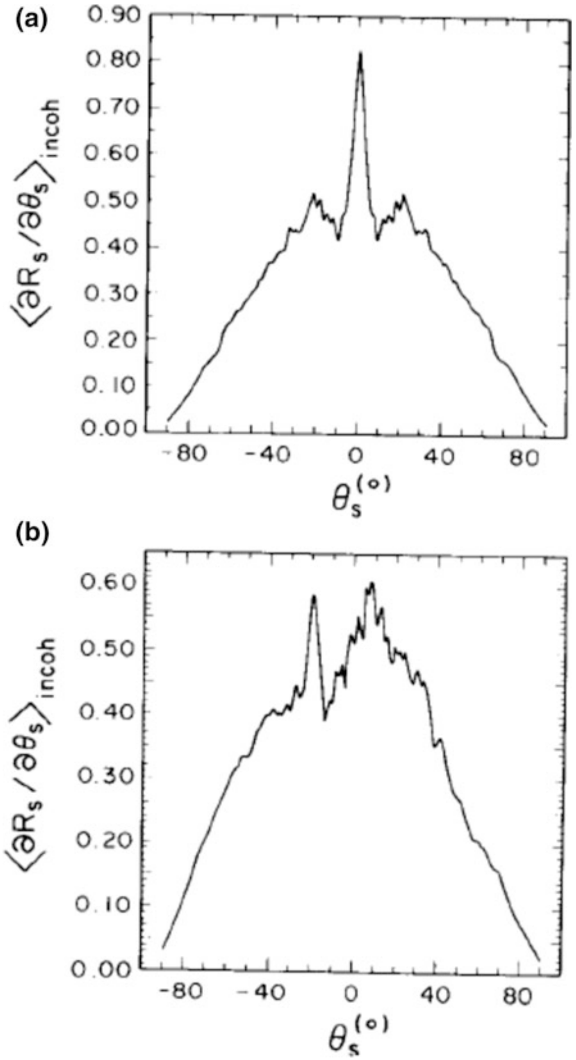


Fig. 4.12 Computer simulation study of enhanced backscattering [16]. In **a** is normal incident radiation and in **b** is incident radiation at 20°. In both cased the backscattering enhancement is opposite the incident radiation. Reproduced with permission from [16]. Copyright 1990 Academic Press, Inc

The width of the backscattering peak is related to the distance that the surface plasmon travels along the interface before it is reradiated into the vacuum or is dissipated into dielectric losses. It is generally found that the backscattering enhancement peak becomes more pronounced as the incident angle approaches normal incidence, being greatest for normal incidence [9].

The theoretical results presented in Fig. 4.11 have been observed experimentally. A number of experimental results confirming enhanced backscattering are found in [16, 17]. In addition, computer simulation studies have also confirmed the enhancement effects from a number of different surfaces with different degrees of weak roughness.

As an example of a simulation study, in Fig. 4.12 results for the differential reflection coefficient of light scattered from a vaccum-silver surface are presented [16]. The study is made for 6127 Å light incident on a Gaussian random silver surface characterized by $a = 2$ μm and $\sigma = 1.2$ μm. Figure 4.2a shows results at normal incidence while Fig. 4.2b presents results at an incident angle of 20°. A large enhanced backscattering peak is observed in both plots.

Again, prominent backscattering peaks, which arise from phase coherence, are observed in the diffuse scattering. The peaks from the simulation are seen to also exhibit side lobes around the backscattering peak. These are diffraction effects associated with the phase coherence of the backscattering peak.

A similar treatment to that given in the preceding discussions for a vacuum-metal interface can be made for the scattering and transmission of light by thin films with rough surfaces. In this case not only are enhancement peaks found in the diffusely scattered radiation from the rough surfaces of thin films, but enhancements are also observed in the diffusely transmitted light through the thin films. The enhancements in the diffuse transmission of light are also due to the effect of weak localization of the surface plasmons on the rough surfaces of thin films. These effects from thin films are now addressed in the following.

4.2.4 Surface Plasmon-Polariton Modes for Light Scattering from Thin Films with Rough Surfaces

In the following a problem related to the diffuse scattering of light from a rough surface is treated. This is the problem of light scattering from the rough surface of a thin film supporting surface plasmons-polaritons. For this system, the weak localization of surface plasmon-polaritons on the thin film shows up in the light diffusely transmitted through the film as well as in the light diffusely reflected from the thin film. Just as in the earlier treated case of light incident on the randomly rough silver surface the diffuse scattering of the light reflected from the thin film exhibits a retroreflectance enhancement peak for diffusely backscattered light into the incident beam and propagating opposite the incident beam. In addition, the diffusely transmitted light traveling away from the thin film in a direction opposite to that of the specularly reflected light from the thin film is found to have an enhancement peak in its differential transmission coefficient [10]. This enhancement is shown to arise from the weak localization of the surface plasmon-polaritons on the film surfaces.

A simple model for the reflection and transmission of light from a thin film with a randomly rough surface [10] is shown schematically in Fig. 4.13. As with the earlier discussions of the scattering from the rough vacuum-silver surface, to facilitate the discussions the surface roughness is again treated as one-dimensional roughness. The thin film is translationally invariant along the y-direction but is disordered along the x-direction. The disorder is characterized by a Gaussian random profile function $z_{surface} = \xi(x)$ relating the coordinates (x, z) of the surface. For this surface geometry the x-z plane is taken as the plane of incident of the light in the system, and, again, due to these symmetry conditions both the incident and scattered light travels in the plane of incidence [10].

Under these conditions, a thin film of average thickness d is considered with vacuum in the regions $z > \xi(x)$ and $z < -d$. In the region $-d < z < \xi(x)$ is a metal film characterized by the frequency dependent dielectric constant $\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega)$ where $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$ are, respectively, the real and imaginary parts of the dielectric constant [10].

For an additional simplification of the treatment, only the upper surface of the thin film is taken to be randomly rough. This does not qualitatively affect the scattering and transmission from the random systems but only facilitates the statistical averaging over the disorder in the system. For the scattering geometry of Fig. 4.13, the light in the system is incident on the thin film from the region $z > \xi(x)$. Its scattering components are then diffusely reflected into the region $z > \xi(x)$ and diffusely transmitted into the region $z < -d$.

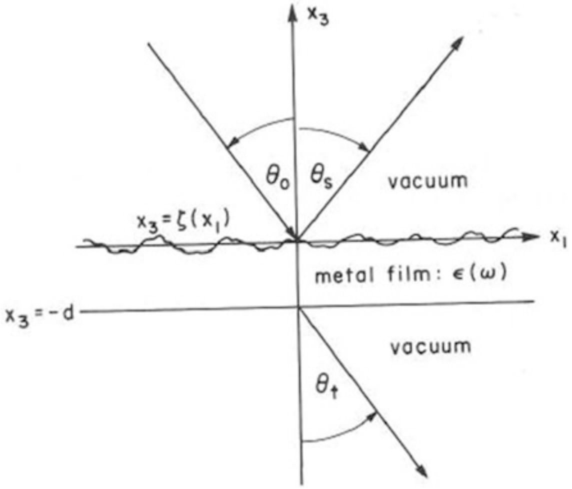


Fig. 4.13 Schematic of the scattering-transmission for a thin film [10]. The figure is from the original paper in [10], and $x = x_1$, $y = x_2$, and $z = x_3$ relates the coordinate notation use here to that used in [10]. Reproduced with permission from [10]. Copyright 1989 Elsevier

The statistical properties of the Gaussian random surface profile function $\xi(x)$ of the upper surface of the thin film are the same as those considered earlier in the case of the scattering of light from a rough surface. Specifically, $\langle\xi(x)\rangle = 0$ and $\langle\xi(x)\xi(x')\rangle = \sigma^2 \exp\left(-|x-x'|^2/a^2\right)$, where the angular brackets denote an average over the ensemble of realizations of the surface profile functions. The higher order correlations of the surface profile functions are again broken down into sums of terms involving products of pair correlations by treating all pairwise combinations [10].

The Fields and their Relation to the Differential Reflection and Transmission Coefficients

As with the scattering from a one-dimensionally rough surface which is disordered in the x-direction, in order for incident light traveling in the x-z plane of incident to couple to the surface plasmon-polaritons of the thin film it must be p-polarized. P-polarized light incident in the x-z plane has its magnetic field vector perpendicular to the plane of incidence, being of the general form $\vec{H}(\vec{r},t) = (0, H_y(x,z|\omega), 0)\mathrm{e}^{-i\omega t}$. The other s-polarization component of light, with the electric field perpendicular to the plane of incidence, does not couple to the surface plasmon-polaritons on the random interface and is not of interest here [10].

For the scattering geometry in Fig. 4.13, the solution in the region $z > \xi(x)$ is given by [10]

$$H_y(x,z|\omega) = \exp[i(kx - i\alpha_0(k)z)] + \int \frac{dq}{2\pi} R(q|k)\exp[iqx + i\alpha_0(q)z], \tag{4.155a}$$

where the first term on the right is the incident plane wave and the second term on the right contains the specular and diffusely scattered light from the thin film. The wave vector of the incident wave, $k = \frac{\omega}{c}\sin\theta_0$, is related to the angle of incidence θ_0 in the usual way. In the region $-d < z < \xi(x)$ the form of the solution

$$H_y(x,z|\omega) = \int \frac{dq}{2\pi}\exp[iqx]\{B(q|k)\exp[-i\alpha(q)z] + C(q|k)\exp[i\alpha(q)z]\}, \tag{4.155b}$$

gives a mixture of upward and downward propagating components of light within the thin film. The solution for the transmitted light from the thin film in the region $z < -d$ has the form

$$H_y(x,z|\omega) = \int \frac{dq}{2\pi} T(q|k)\exp[iqx - i\alpha_0(q)z]. \tag{4.155c}$$

It represents a mixture of the specularly and diffusely transmitted light through the thin film and propagating away from it. In the forms for the general solutions listed in (4.155)

$$\alpha_0(q) = \left[\frac{\omega^2}{c^2} - q^2\right]^{1/2}, \text{with } \mathrm{Re}\alpha_0(q) \text{ and } \mathrm{Im}\alpha_0(q) > 0, \tag{4.156a}$$

and

$$\alpha(q) = \left[\varepsilon(\omega)\frac{\omega^2}{c^2} - q^2\right]^{1/2}, \text{with } \mathrm{Re}\alpha(q) > 0 \text{ and } \mathrm{Im}\alpha(q) > 0. \tag{4.156b}$$

The coefficients $B(q|k), C(q|k), R(q|k)$, and $T(q|k)$ for the complete solution of the scattering of the p-polarization component are obtained by matching the boundary conditions at the two surfaces of the thin film.

In matching the boundary conditions the Rayleigh hypothesis for weakly rough surfaces is again applied. This is the assumption that the form in (4.155a) is valid for $z > \xi(x)$ and not just for $z > \xi(x)_{maximum}$. Likewise the form in (4.155b) is assumed to be valid for $-d < z < \xi(x)$ and not just for $-d < z < \xi(x)_{minimum}$. For weakly rough surfaces of the type considered here, which have relatively smooth profiles, the assumptions of the Rayleigh hypothesis are known to be effective. They are not found to affect the results presented in the following which are consistent with experiment and computer simulation studies [10].

The differential reflection and transmission coefficients are expressed in terms of the solutions for the reflection amplitude, $R(q|k)$, in (4.155a) and the transmission amplitude, $T(q|k)$, in (4.155c). The coefficients of reflection and transmission, as with the randomly rough surface scattering problem, are obtained by computing and comparing the Poynting vectors of incident, reflected, and transmitted light from the film surfaces. From these Poynting vectors the differential reflection coefficient for reflectance from the rough surface of the thin film is

$$\frac{\partial R}{\partial \theta_s} = \frac{1}{L}\frac{\omega}{2\pi c}\frac{\cos^2\theta_s}{\cos\theta_0}\left\langle |R(q|k)|^2 \right\rangle, \tag{4.157}$$

where $q = \frac{\omega}{c}\sin\theta_s$ and $k = \frac{\omega}{c}\sin\theta_0$. Similarly, the differential transmission coefficient for the transmission through the rough surface of the thin film is found to be given by

$$\frac{\partial T}{\partial \theta_t} = \frac{1}{L}\frac{\omega}{2\pi c}\frac{\cos^2\theta_t}{\cos\theta_0}\left\langle |T(q|k)|^2 \right\rangle, \tag{4.158}$$

where in this case $q = \frac{\omega}{c}\sin\theta_t$ and $k = \frac{\omega}{c}\sin\theta_0$. In (4.157) and (4.158) L is the length of the scattering surface and θ_0, θ_s, and θ_t are, respectively, the incident, diffuse reflection, and diffuse transmission angles measure as in Fig. 4.13.

Formulation of the Green's Function Approach

Upon matching the boundary conditions and following some algebra, T-matrix equations are developed for the reflection and transmission amplitudes in (4.155a) and (4.155c). In the resulting formulation the mean thickness of the thin film, d, is assumed to be small but large enough such that $|\exp[i\alpha(q)d]| \ll 1$. Consequently, in the theory presented in the following all of the work is based on retaining terms of the lowest nonzero order in $\exp[i\alpha(q)d]$. This leads to an analytically tractable theory exhibiting the interesting enhancement effects in both the differential refection and transmission for the thin film. It is accurate for weak roughness and for thin films satisfying $|\exp[i\alpha(q)d] \ll 1|$.

Specifically, under these conditions the reflection amplitude takes the form [10

$$R(q|k) = 2\pi\delta(q-k)R_0(k) - 2iG_0(q)T_R(q|k)G_0(k)\alpha_0(k). \tag{4.159a}$$

Here the first term on the right side of (4.159a) gives the specular scattering from the flat surface, and the second term represents the rough surface scattering which renormalizes the specular reflection from the thin film and also contributes to the general diffuse scattering contributions in the region above the thin film.

In the same way the transmission amplitude for the thin film with surface roughness takes the form

$$T(q|k) = 2\pi\delta(q-k)T_0(k) - 2iG_0(q)T_T(q|k)G_0(k)\alpha_0(k). \tag{4.159b}$$

The first term on the right of (4.159b) gives the transmission for the thin film with flat surfaces. The second term on the right represents the rough surface scattering which renormalizes the flat surface transmission and contributes to a general diffuse transmission into the region below the thin film.

In (4.159)

$$G_0(k) = \frac{i\varepsilon(\omega)}{\varepsilon(\omega)\alpha_0(k) + \alpha(k)} \tag{4.160a}$$

is the surface plasmon-polariton Green's function for the thin film with flat surfaces,

$$R_0(k) = \frac{\varepsilon(\omega)\alpha_0(k) - \alpha(k)}{\varepsilon(\omega)\alpha_0(k) + \alpha(k)} \tag{4.160b}$$

is the Fresnel reflection coefficient for the thin film with smooth surfaces, and

$$T_0(k) = \frac{4\varepsilon(\omega)\alpha_0(k)\alpha(k)}{[\varepsilon(\omega)\alpha_0(k) + \alpha(k)]^2}\exp\{-i[\alpha_0(k) - \alpha(k)]d\} \tag{4.160c}$$

is the smooth surface transmission coefficient for the thin film. Equations (4.160b) and (4.160c) enter the specular terms for the reflection and transmission of the smooth surface thin film while (4.160a) describes the propagation response of the surface electromagnetic on the smooth surface thin film.

The T-matrices in (4.159) contain the scattering corrections to the reflection and transmission amplitudes due to the randomness of the upper surface of the thin film.

From the boundary conditions and the thin film approximations they are found to be of the forms [10]

$$T_R(q|k) = V_R(q|k) + \int \frac{dp}{2\pi} V_R(q|p) G_0(p) T_R(p|k) \tag{4.161a}$$

and

$$T_T(q|k) = V_T(q|k) + \int \frac{dp}{2\pi} V_T(q|p) G_0(p) T_R(p|k). \tag{4.161b}$$

In these T-matrix forms there are two different scattering potentials given by

$$V_R(q|k) = \frac{\varepsilon(\omega) - 1}{\varepsilon^2(\omega)} \xi(q-k)[\varepsilon(\omega)qk - \alpha(q)\alpha(k)] \tag{4.162a}$$

and

$$V_T(q|k) = 2\frac{\varepsilon(\omega) - 1}{\varepsilon(\omega)} \xi(q-k) \frac{qk + \alpha_0(q)\alpha(k)}{\varepsilon(\omega)\alpha_0(q) + \alpha(q)} \alpha(q) \exp\{-i[\alpha_0(q) - \alpha(q)]d\}. \tag{4.162b}$$

In (4.162)

$$\xi(Q) = \int dx \exp(-iQx)\xi(x) \tag{4.163}$$

is the Fourier transform of the rough surface profile function, and the potentials in (4.162) are correct to lowest order in the weak surface roughness limit.

The T-matrix $T_R(q|k)$ describes the scattering of the surface plasmon-polaritons along the rough surface of the thin film. It relates the surface plasmon-polariton Green's functions for the rough surface thin film to the smooth surface thin film Green's functions. In terms of the T-matrix the rough surface Green's function is given by

$$G(q|k) = 2\pi\delta(q-k)G_0(k) + G_0(q)T_R(q|k)G_0(k). \tag{4.164}$$

The first term on the right in (4.164) is the Green's function for the surface plasmon-polariton on a thin film with a smooth surface. The second term includes the multiple scattering interactions on the rough surface. From the T-matrix in (4.161a) it is found to be composed of a sum of multiple scattering processes on the rough surface.

The first term in the sum of $T_R(q|k)$ gives a single scattering of the surface plasmon-polariton as it passes along the surface, while the nth term in the sum of $T_R(q|k)$ describes n successive scattering of the surface plasmon-polariton as it passes along the rough surface. In this way, (4.164) represents the total dynamics of the single plasmon-polariton along the rough surface of the thin film.

From (4.157), (4.159), and (4.164) the differential reflection coefficient for the diffuse reflection can be written in terms of the Green's function for the plasmon-polariton propagation on the rough surface thin film [10]

$$\frac{\partial R}{\partial \theta_s} = \frac{4}{L} \frac{\omega^3}{2\pi c^3} \cos^2 \theta_s \cos \theta_0 \left\langle |G(q|k)|^2 \right\rangle. \tag{4.165}$$

The diffuse reflection in (4.157) arises as a coupling of the incident light, through the rough surface, to surface plasmon-polaritons propagating along the interface. This is followed by the scattering of the plasmon-polaritons along the rough interface until they ultimately couple, by the surface roughness, out of the surface and into bulk modes radiating away from the random interface. In particular, the average of the Green's function product in (4.165) is a two-particle Green's function. It represents the propagation of two surface plasmon-polaritons in their correlated scattering along the random surface during the intermediate processes between the incident and scattered light.

At the level of approximation considered here, the evaluation of (4.165) for the diffusely reflected light from the thin film with rough surfaces yields essentially the same result as that obtained for the diffusely reflected light from the rough surface of a semi-infinite metal. This was treated earlier where the two-particle surface plasmon-polariton Green's function was calculated in terms of sums of ladder and maximally crossed diagrams and will not be further pursued here. The focus in the following will be on the diffuse transmission of light through the thin film with a randomly rough surface. Again the diffuse transmission will be shown to be directly related to the two-particle surface plasmon-polariton Green's function which will be evaluated in the context of the ladder and maximally crossed diagrams studied earlier.

The $T_T(q|k)$ T-matrix in (4.161b) for the transmission processes is complicated by the need of the surface scattered light to be transported through the thin film and exit the lower smooth surface of the thin film. Transmission through the thin film leads to a significant reduction of the intensity of the scattered light. Due to this, most of the multiple scattering in $T_T(q|k)$ is contained within the $T_R(q|k)$ term entering into the integrand of the integral on the right of (4.161b).

Both terms entering $T_T(q|k)$ on the right of (4.161b) contain only a single $V_T(q|k)$ scattering term. This potential involves transmission through the thin film and is generally small due to the decay of the fields as the light passes through the thin film. Terms with multiple $V_T(q|k)$ scattering terms would be significantly reduce from those contain a single $V_T(q|k)$ scattering term. Consequently, they are not included in the scattering approximation presented here.

From (4.158), (4.159b), (4.161b), and (4.164) the differential transmission coefficient for the diffusely transmitted light through the thin film with the randomly rough surface is obtained in terms of the rough surface plasmon-polariton Green's function. Specifically, the differential transmission coefficient for the diffuse transmission can be written in terms of the Green's function for the plasmon-polariton propagation on the rough surface thin film [10]

$$\frac{\partial T}{\partial \theta_t} = \frac{4}{L}\frac{\omega^3}{2\pi c^3}\cos^2\theta_t \cos\theta_0 \left| G_0\left[\frac{\omega}{c}\sin\theta_t\right]\right|^2 \left\langle |F(q|k)|^2 \right\rangle, \tag{4.166}$$

where $F(q|k)$ is expressed in terms of $G(q|k)$ by

$$F(q|k) = \int \frac{dp}{2\pi} V_T(q|p)G(p|k). \tag{4.167}$$

The diffuse transmission in (4.166) arises as a coupling of the incident light, through the rough surface, to surface plasmon-polaritons propagating along the interface. This is followed by the scattering of the plasmon-polaritons along the rough interface until they ultimately couple, by the surface roughness, out of the surface and into thin film modes radiating away from the random interface. Eventually these thin film modes pass thorough the smooth lower surface of the then film and radiate into the bulk media below the thin film. In particular, the average of the $F(q|k)$ product in (4.166) is related to the two-particle Green's function of the surface plasmon-polaritons. It represents the propagation of two surface plasmon-polaritons in their correlated scattering along the random surface during the intermediate processes between the incident and scattered light.

The form of $F(q|k)$ explicitly introduces the $V_T(q|k)$ scattering potential into the calculation of the diffuse transmission. The $V_T(q|k)$ scattering potential is smaller than the $V_R(q|k)$ scattering potential for scattering along the rough surface as it includes the effects of transmission through the thin film as part of its interaction. It accounts for the fact that the diffuse transmission of light through the thin film with surface roughness is a small effect. Higher order terms in the $V_T(q|k)$ scattering, which are left off in the present theory, would be significantly less than those given in (4.166).

In the evaluation of (4.166) for the differential transmission coefficient, it is necessary to determine the two-particle scattering function given by

$$\left\langle |F(q|k)|^2 \right\rangle. \tag{4.168}$$

The averaging process in (4.168) is complicated by the presence of the two different scattering potentials,$V_R(q|k)$ and $V_T(q|k)$, combined with the pairwise correlations of the surface profile functions arising from the Gaussian random surface statistics. In particular, as a result of this, four pairwise contractions of the scattering potentials are introduced into the calculation upon performing the average in (4.168). These potential pairing in the averaging processes are

$$V_1(qr|ps) = \langle V_T(q|p)V_T^*(r|s)\rangle, \tag{4.169a}$$

$$V_2(qr|ps) = \langle V_R(q|p)V_R^*(r|s)\rangle, \tag{4.169b}$$

$$V_3(qr|ps) = \langle V_T(q|p)V_R^*(r|s)\rangle, \tag{4.169c}$$

and

$$V_4(qr|ps) = \langle V_R(q|p)V_T^*(r|s)\rangle, \tag{4.169d}$$

The four pairing in (4.169) complicated the summation of the ladder and maximally crossed diagrams that were shown in the treatment of rough surface scattering to contribute the dominant effects.

Diagrammatic Solution

As was shown earlier in the considerations of rough surface scattering, the ladder diagrams are responsible for a general diffuse reflection from the rough vacuum-metal interface. The same is true in the thin film problem for a general diffuse reflection above the thin film. In addition, however, for the thin film problem the ladder diagrams also contribute to a general diffuse transmission into the region below the thin film. The ladder diagrams represent the successive interaction of two surface plasmon-polaritons in a correlated scattering involving closely located portions of the scattering surface. The ladder diagram approach to perturbative scattering from a weakly disorder system typically is used to consider to lowest order the transport processes of a randomly disordered media.

In Fig. 4.14 the relevant ladder and the four scattering potentials in (4.168) and (4.169) are shown for the determination of $\left\langle |F(q|k)|^2 \right\rangle$ in the context of the ladder approximation. These contribute the dominant effects, in the limit of weak surface roughness, for the diffuse transmission of light at general transmission angles through the rough surfaced thin film. Summing the ladder contributions gives a contribution to $\left\langle |F(q|k)|^2 \right\rangle$ of the form [10]

$$\left\langle |F(q|k)|^2 \right\rangle_L \approx LF_L(q|k)|G_0(k)|^2, \tag{4.170a}$$

where $F_L(q|k)$ is written as

$$F_L(q|k) = V_a(q|k) + \frac{C_1^2}{2\Delta_{Total}} \frac{1}{1 - \frac{\Delta_{SP}^2}{\Delta_{Total}^2}} [A(q|k) + B(q|k)]. \tag{4.170b}$$

In (4.170) L is the length of the surface in the x-direction, $K_{SP}(\omega) = \frac{\omega}{c}\sqrt{\frac{\varepsilon_1(\omega)}{\varepsilon_1(\omega)+1}}$, $C_1 = \frac{\varepsilon(\omega)\sqrt{-\varepsilon_1(\omega)}}{1-\varepsilon_1^2(\omega)}$, $\Delta_\varepsilon(\omega)$ is the decay rate of the surface plasmon-polariton due to the imaginary part of the dielectric constant of the metal film, $\Delta_{SP}(\omega) \ll \Delta_{\varepsilon}(\omega)$ is the decay rate of the surface plasmon-polariton due to scattering from the surface roughness, and $\Delta_{Total}(\omega) = \Delta_{\varepsilon}(\omega) + \Delta_{SP}(\omega)$. [Note: That the surface plasmon-polariton decay rate from the imaginary part of the dielectric and the surface scattering for the thin film treated here are the same as for the semi-infinite metal treated in (4.130).] In addition, $A(q|k)$ and $B(q|k)$ in (4.170) are [10]

$$A(q|k) = V_a(q|K_{SP})\left[V_b(K_{SP}|k) + \frac{\Delta_{SP}}{\Delta_{Total}} V_b(-K_{SP}|k)\right] \tag{4.171a}$$

and

$$B(q|k) = V_a(q|-K_{SP})\left[V_b(-K_{SP}|k) + \frac{\Delta_{SP}}{\Delta_{Total}} V_b(K_{SP}|k)\right]. \tag{4.171b}$$

with the functions $V_a(q|k)$ and $V_b(q|k)$ expressed as

$$\begin{aligned} V_a(q|p) = 4\sqrt{\pi}a\sigma^2 \left|\frac{\varepsilon(\omega)-1}{\varepsilon(\omega)}\right|^2 |\alpha(q)|^2 \frac{|qp+\alpha_0(q)\alpha(p)|^2}{|\varepsilon(\omega)\alpha_0(q)+\alpha(q)|^2} \\ \times \exp\{2\mathrm{Im}[\alpha_0(q)-\alpha(q)]d\} \exp\left[-a^2\frac{(q-p)^2}{4}\right]. \end{aligned} \tag{4.172a}$$

and

$$V_b(q|p) = \sqrt{\pi}a\sigma^2 \left|\frac{\varepsilon(\omega)-1}{\varepsilon^2(\omega)}\right|^2 |\varepsilon(\omega)qp - \alpha(q)\alpha(p)|^2 \exp\left[-a^2\frac{(q-p)^2}{4}\right]. \tag{4.172b}$$

In the discussions to follow the ladder contribution in (4.170) will be found to give a slowly varying contribution to the differential transmission coefficient of the diffuse transmission below the thin film. Its intensity distribution in space is quite similar to the diffuse differential reflection coefficient above the thin film, but it is of a somewhat reduced intensity. The reduced intensity is due to the decay in the electromagnetic fields as they penetrate into the thin film.

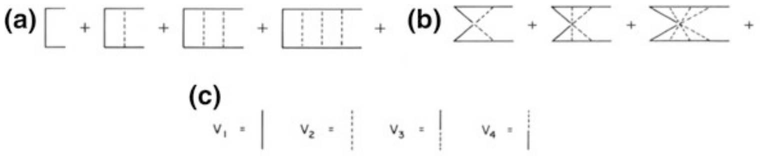


Fig. 4.14 Diagramatics for: **a** the sum of ladder diagrams, **b** the sum of maximally crossed diagrams, and **c** the four different scattering lines used in (**a**) and (**b**) [10]. Reproduced with permission from [10]. Copyright 1989 Elsevier

In the earlier treatment of the scattering from a rough metal surface, the maximally crossed diagrams were shown to be responsible for the weak localization effects in the diffuse reflection of light from the rough interface. Similarly, in the following discussions of scattering from the thin film with rough surfaces, maximally crossed diagrams are now shown to be the source of weak localization effects in the reflection and transmission of light by the thin film.

These diagrams, as with the ladder diagrams, represent the successive interaction of two surface plasmon-polaritons in a correlated scattering involving closely located portions of the scattering surface. However, in the case of the maximally crossed diagrams the closely located portions of the scattering surface are encountered by each of the two plasmon-polaritons in a reversed scattering sequence from the other. This can be seen in Fig. 4.14 which shows the maximally crossed diagrams contributing to the multiple scatterings of the two plasmon-polaritons on the random surface of the thin film. A consequence of the reverse sequence to the scattering of the two plasmon-polaritons is that a phase memory is preserved in the scattering amplitude. The resulting phase memory is known as weak localization and contributes an enhanced backscattering to the response of the system which eventually shows up as enhancement peaks in the differential reflection and transmission coefficients of the thin film.

Results for Diffuse Reflection

For the scattering of light incident on a rough metal surface the weak localization effects were observed as an enhancement in the diffusely scattered light in the direction opposite that of the incident beam. This is the enhanced retroreflectance peak in the differential reflection coefficient for the light scattered by the rough surface. The width of the retroreflectance peak in the differential reflection coefficient was shown to be related to the propagation distance of the surface plasmon-polarition before its decays into other modes.

In the thin film system studied here, the same retroreflectance peak is found in the differential reflection coefficient of the light reflected by the thin film. At the level of approximation of the thin film discussion present here, the differential reflection is described by the same formulae as that for the rough scattering surface, and the width of the retroreflectance peak is again related to the decay length for plasmon-polariton motion along the surface of the thin film. The identity of the result for the thin film with the surface scattering result of the rough vacuum-metal interface is due to the very weak transmission of light though the thin film considered in the following treatment.

Results for Diffuse Transmission

Only the weak localization effects in the diffuse transmission of light through the thin film are treated in the following. These effects also arise from the weak localization of the surface plasmon-polaritons on the rough surface of the thin film. They result in an enhancement peak in the diffuse transmission below the thin film that has the same width as that of the differential reflection coefficient of the diffusely scattered light above the thin film. Both peaks are related to the decay length of the surface plasmon-polariton on the rough interface, and this accounts for their same widths.

In Fig. 4.14 the relevant maximally crossed diagrams and the four scattering potentials in (4.168) and (4.169) are shown for the determination of $\left\langle |F(q|k)|^2 \right\rangle$ in the context of the approximation involving maximally crossed diagram. These contribute the dominant weak localization effects, in the limit of weak surface roughness, for the diffuse transmission of light through the rough surfaced thin film. The diagrams contribute as a sum of scattering sequences representing the phase coherent scatterings as the plasmon-polaritons propagate along the rough interface [10].

Summing the maximally crossed contributions in Fig. 4.14 gives a contribution to $\left\langle |F(q|k)|^2 \right\rangle$ for the phase coherent, weak localization effects, that has the form

$$\left\langle |F(q|k)|^2 \right\rangle_{MC} \approx LF_{MC}(q|k)|G_0(k)|^2, \tag{4.173a}$$

where $F_{MC}(q|k)$ is written as

$$F_{MC}(q|k) = \frac{2C_1^2 \Delta_{Total}}{(q+k)^2 + \Delta_{Total}^2} \frac{1}{1 - \frac{\Delta_{SP}^2}{\Delta_{Total}^2}} [A_{MC}(q|k) + B_{MC}(q|k)]. \tag{4.173b}$$

In these L is again the length of the surface along the x-axis and $K_{SP}(\omega)$, C_1, $\Delta_\varepsilon(\omega)$, $\Delta_{SP}(\omega) \ll \Delta_\varepsilon(\omega)$, and $\Delta_{Total}(\omega) = \Delta_\varepsilon(\omega) + \Delta_{SP}(\omega)$ have been defined below (4.170). The functions $A_{MC}(q|k)$ and $B_{MC}(q|k)$ in (4.173) are found to be given by the forms [10]

$$A_{MC}(q|k) = K(-k|K_{SP})K^*(k|K_{SP}) + K(-k| - K_{SP})K^*(k| - K_{SP}) \tag{4.174a}$$

and

$$B_{MC}(q|k) = \frac{\Delta_{SP}}{\Delta_{Total}} [K(-k|K_{SP})K^*(k| - K_{SP}) + K(-k| - K_{SP})K^*(k|K_{SP})]. \tag{4.174b}$$

In (4.174)

$$K(r|s) = 2\sqrt{\pi} a\sigma^2 \left| \frac{\varepsilon(\omega) - 1}{\varepsilon(\omega)} \right|^2 \frac{\alpha(r)}{\varepsilon^*(\omega)} \frac{[rs + \alpha_0(r)\alpha(s)][\varepsilon(\omega)rs - \alpha(r)\alpha(s)]^*}{[\varepsilon(\omega)\alpha_0(r) + \alpha(r)]} \times \exp\{-i[\alpha_0(r) - \alpha(r)]d\} \exp\left[-a^2 \frac{(r-s)^2}{4}\right]. \tag{4.175}$$

The contribution to the maximally crossed scattering in (4.173)–(4.175) yield results for the weak localization contributions to $\left\langle |F(q|k)|^2 \right\rangle$ that are valid in the limit that

$$\Delta_{SP}(\omega) \ll \Delta_{Total}(\omega). \tag{4.176}$$

This means that the dielectric losses in the system dominate the losses from scattering by the surface roughness into other surface and radiative electromagnetic modes. Consequently, only the lowest order maximally crossed diagrams in Fig. 4.14 contribute significantly to the enhancement peak in the differential transmission coefficient of the diffusely transmitted light.

The peaked enhancement effect in the diffuse transmission enters through (4.173). From (4.173) the enhanced transmission occurs when

$$q \approx -k. \tag{4.177}$$

This is the condition for a maximum in to exist in $\left\langle |F(q|k)|^2 \right\rangle_{MC} \approx LF_{MC}(q|k)|G_0(k)|^2$ and $F_{MC}(q|k)$. It arises specifically from the Lorentzian form

$$\frac{2C_1^2\Delta_{Total}}{(q+k)^2 + \Delta_{Total}^2} \tag{4.178}$$

in (4.173b) and contributes to both functions defined in (4.173). The width of the Lorentzian peak is seen in (4.173b) to be set by Δ_{Total} which also gives the decay length for surface plasmon-polarition propagation along the rough interface of the thin film.

The differential transmission coefficient for the diffuse transmission of light though the thin film is now obtained as a sum of the contributions from the ladder diagrams and the maximally crossed diagrams. It follows from (4.166), (4.170), and (4.173) that the differential transmission coefficient is given in terms of $F_L(q|k)$ and $F_{MC}(q|k)$ by [10]

$$\frac{\partial T}{\partial \theta_t} = \frac{4}{L}\frac{\omega^3}{2\pi c^3}\cos^2\theta_t \cos\theta_0 \left| G_0\left[\frac{\omega}{c}\sin\theta_t\right]^2 \right| \{F_L(q|k) + F_{MC}(q|k)\}|G_0(k)|^2, \tag{4.179}$$

where $q = \frac{\omega}{c}\sin\theta_t$ and $k = \frac{\omega}{c}\sin\theta_0$ for the incident angle θ_0 and transmission angle θ_t.

The contribution of the ladder diagrams, $F_L(q|k)$, is a smooth slowly varying function of the transmission angle θ_t. It represents a general background at all transmission scattering angles below the thin film. From (4.173), (4.179), and the Lorentz form in (4.178) it is seen that the maximally crossed diagram terms,

$F_{MC}(q|k)$, in the differential transmission coefficient contribute a single narrow peak to the transmission coefficient which is center about $q = -k$ and is of a width set by Δ_{Total}. This peak represents transmitted light in the region below the thin film moving in a direction directly opposite the motion of the specularly reflected light in the region above the thin film. The peak in the differential transmission coefficient comes from the weak localization of the surface waves traveling on the thin film. Unlike the scattering from the ladder diagrams terms, it is a phase coherent scattering process.

Illustrative Example

In Fig. 4.15 (4.179) is evaluated for the differential reflection and transmission coefficients for a thin silver film in vacuum. Light of optical wavelength $\lambda = 4579$ Å is incident on the metal film with an angle of incidence of $\theta_0 = 20°$. The dielectric constant of silver at this wavelength is of the form $\varepsilon(\omega) - 7.5 + i0.24$. For the plots, the roughness on the upper surface was taken to be Gaussian randomness characterized by $\sigma = 50$ and $a = 1000$ Å. Two plots are shown one for a film of thickness $d = 800$ Å and one for a film thickness of $d = 1000$ Å.

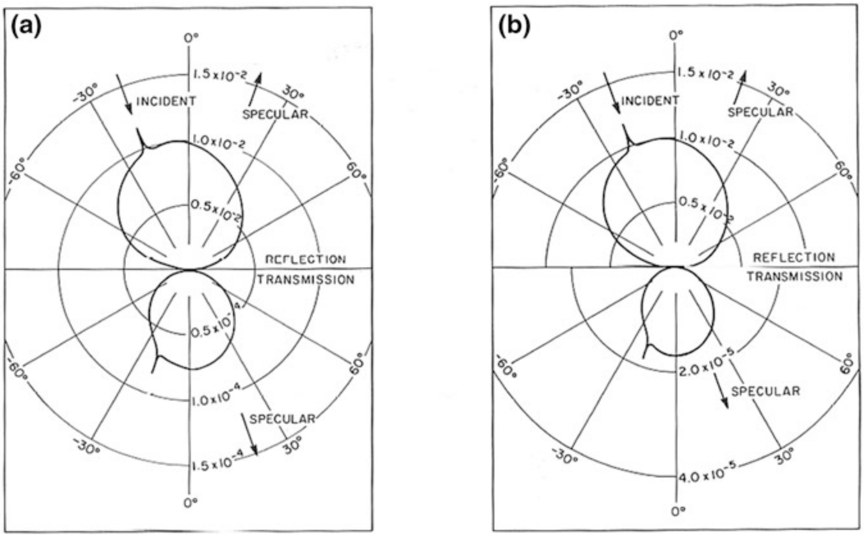


Fig. 4.15 Differential reflection and transmission coefficients for the diffusely reflected and transmitted light through a thin metal film with a rough upper surface [10]. Reproduced with permission from [10]. Copyright 1989 Elsevier

4.2.5 Speckle Correlations in the Reflection and Transmission of Light Through a Thin Film

Another aspect of the scattering and transmission of light through a thin film that is influenced by surface plasmon-polaritons is the speckle correlations in the diffusely reflected and transmitted light from the rough surfaces of the thin film [18–26]. In the following a discussion of these effects is presented in the context of the model just treated for the scattering of light from a thin film. This is preceded by an introductory discussion of the speckle phenomenon, some of its general features, and its technological applications.

Speckle is commonly observed in the scattering of laser light of a single wavelength from a rough surface [20–26]. It is the grainy pattern of dark and bright intensities found in the light viewed from over the entire scattering surface, with the features of the bright and dark patches being dependent on the wavelength of the light and the roughness of the scattering surface. A common example of the effect is seen in the reflected light from laser pointers used at conference talks.

When the laser beam is scattered from a room wall or a presentation screen the viewer sees a fine mixture of bright and dark grains in the illuminated region of the pointer. As more wavelengths of light are introduced into the scattering, the overlap of the patterning from the different wavelengths tends to washout the granular pattern. This is why speckle is observed in the dot of coherent light from a laser pointer and not so much from a focused beam of while light.

The bulk features of the granular mixture in the speckle pattern can be grossly explained in a simple theoretical approximation based on a random walk addition of the phasors of light scattered from the rough surface [18–26]. This approach treats the scattering from the surface as a simple one encounter scattering (i.e., single scattering) from the surface. The scattered light from different points of the surface is represented by amplitudes and phases which vary from point to point along the random surface. These components of light add to produce a net field at the point of observation.

By making some reasonable simple assumptions on the statistical properties of the amplitudes and phases of the different components of scattered light arriving at the point of observation, the distribution of grain intensities within the speckle pattern is generated. In general, the assumption of a random walk addition of the phasors of light at the observation point explains many of the features found in the intensity distribution within the speckle pattern. This portrays light from each point on the surfaces as being part of a random walk contribution to the total field at the point of observation. In this manner, the distribution of intensities observed within the spackle pattern is shown to be distributed in a Poisson distribution.

A Simple Theory of Basic Speckle Phenomena

Goodman [18, 19] was the first to apply the random walk treatment to the study of the intensity distributions within the speckle pattern. The Poisson distribution of speckle intensities was shown to be of the general form [18, 19]

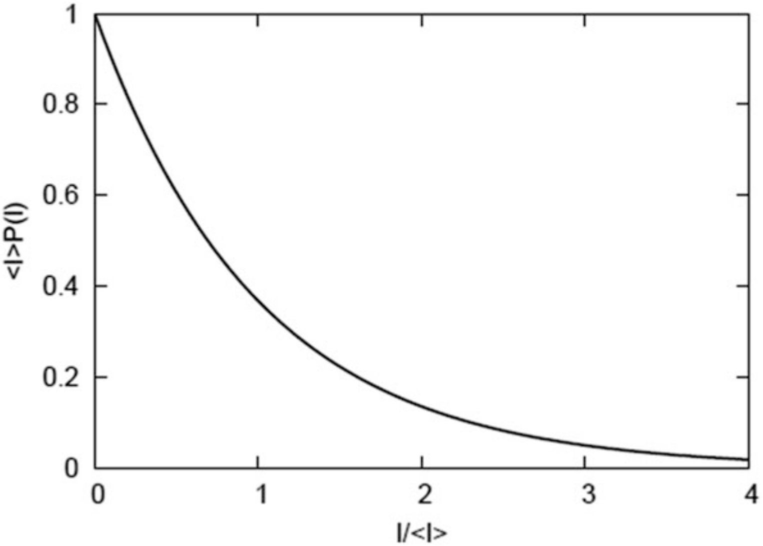


Fig. 4.16 Plots of $P(I)$ versus the speckle intensity, $\frac{I}{\langle I\rangle}$, from the random walk treatment of the distribution of speckle intensities [18, 19]

$$P(I) = \frac{1}{\langle I\rangle}\exp\left(-\frac{I}{\langle I\rangle}\right), \tag{4.180}$$

where $\langle I\rangle$ is the average intensity of the speckle pattern. In Fig. 4.16 the distribution in (4.180) for $P(I)$ versus the speckle intensity, I is plotted. The most probable intensity is zeros and the probability decreases uniformly as the intensity increases.

The Poisson distribution offers an explanation of the relative occurrences of the bright and dark intensities. It does not, however, explain correlations between the dark and bright patches or their relative sizes. These properties depend on the detailed geometry of the rough scattering surfaces and on the wavelength of the coherent radiations illuminating the surface. In the case that the rough surface supports surface plasmon-polaritons, it is also expected that the scattering of incident light into and out of surface plasmon-polaritons should be seen in the features of the speckle patterning from the surface [18, 19].

A More Complete Approach with Applications: Correlations

In the following discussions a more detailed approach to the treatment of speckle patterns will relate the correlations of the bright and dark grains in the speckle pattern to the propagation of surface plasmon-polaritons along the rough surface. The focus of the presentation will be on surfaces that support surface plasmon-polaritons and on how the presence of these excitations in the system influence the features of the speckle pattern [18, 19].

This approach treats the scattering of light as a multiple scattering (multiple interaction) of the light from the rough surface [20–26]. The multiple scattering events, which manifest themselves as the scattering into and out of surface plasmon-polaritons propagating along the surface, then introduce correlations within the speckle pattern. In this way the distribution of the speckles within the speckle pattern are found to be dependent in part on the properties of the surfaces waves found at the interface. The correlation of features from the surface waves within the pattern is determined from a study of the intensity-intensity correlation functions of light diffusely scattered at the rough surface.

The speckle features of a rough surface are the foundation of a number of technological techniques that are important both in the study of surfaces and in the study of the scattering from volume disordered systems. Examples of applications include: speckle metrology, speckle photography, holographic interferometry, electronic speckle pattern interferometry, speckle imagining, dynamic speckle, and biospeckle [27, 28].

Many of these techniques are implemented for the determination of surface roughness, surface deformations, or effects related to material flows. In general, the measurements in these studies may involve either static systems or systems which change in time. Information in such applications of speckle is obtained on the systems being studied from the changes in the speckle pattern in time or through the change in the interference of a scattered speckle field from a rough surface as it interacts with a reference beam. In these last mentioned processes the reference beam and the light sent to scatter from the rough surface often originate from the same light source.

The techniques just mentioned are based on the application of speckle technology, focused on effects which do not account for the correlations in the speckle pattern due to the presence of surface wave effects. These correlations are expected to introduce a new feature into the applications of speckle as a technology for the characterization of surfaces. In addition, they are of interest in providing a fundamental understanding of the interaction of surface electromagnetic waves with the scattered fields generated at the surface.

In the next subsection, the model for the scattering of light from a thin film with a rough surface is reformulated to treat the correlations arising in the speckle pattern due to the presence of surface plasmon-polaritons. Similar Green's function techniques to those applied to the study of rough surface reflection and transmission coefficients form the basis of the calculations. As with the reflection and transmission coefficients the weak localization of surface plasmon-polaritons are shown to have an important influence on the speckle pattern generated from the thin film with surface disorder.

The Model and Definition of Speckle Correlation Function

The model treated is the same as that studied in the discussions of the reflection and transmission coefficients of light from a thin metal film with surface roughness [20, 21]. In the following, this is briefly summarized, after which a discussion of the theoretical considerations for the speckle correlations in the diffusely scattered radiation from the film is presented.

In the model of the thin film the system is composed of vacuum in the region $x_3 > \xi(x_1)$, a metal film characterized by an isotropic, complex, frequency-dependent dielectric function $\varepsilon(\omega)$ for $-d < x_3 < \xi(x_1)$, and vacuum for $x_3 < -d$. (A schematic representation of this and the associated scattering geometry is shown in Fig. 4.17.) The rough upper surface of the thin film has a one-dimensional disorder characterized by a profile function $\xi(x_1)$ that is a single valued Gaussian random process. For simplicity only one surface is consider to be randomly rough. This restriction to one surface does not affect the basic nature of the qualitative results obtained for the system in the following treatment [20, 21].

The statistical properties of the surface profile functions are, specifically, characterized by the surface averages [20, 21]

$$\langle \xi(x_1) \rangle = 0, \tag{4.181a}$$

$$\left\langle \xi(x_1)\xi(x_1') \right\rangle = \sigma^2 \exp(-|x_1 - x_1'|^2/a^2) \tag{4.181b}$$

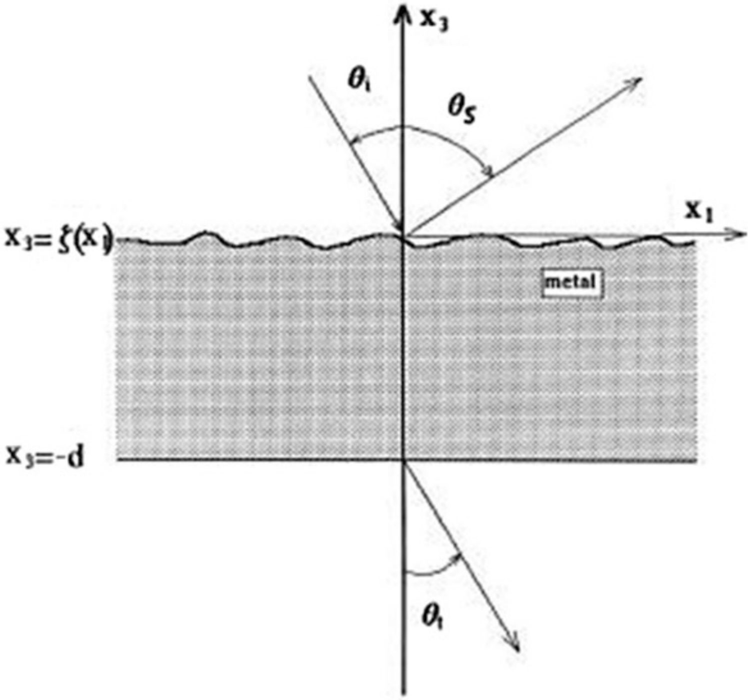


Fig. 4.17 Schematic of the thin film and its scattering geometry [20, 21]. Reproduced with permission from [20]. Copyright 1989 Elsevier

where $\langle\rangle$ indicates an average of the profile over the rough surface. In the statistical characterization provided by (4.181), σ characterizes the root mean squared height distribution about the x_1-axis and a sets the correlation length of the surface profile along the x_1-axis.

As with the previous discussions of surface profiles with Gaussian random disorder, all higher order correlations of $\xi(x_1)$'s are expressed as products of the correlation functions in (4.181). In particular, in the case of a product of four surface profile functions

$$\begin{aligned}\langle\xi(x_1)\xi(x_1'')\xi(x_1')\xi(x_1''')\rangle &= \langle\xi(x_1)\xi(x_1')\rangle\langle\xi(x_1'')\xi(x_1''')\rangle \\ &+ \langle\xi(x_1)\xi(x_1'')\rangle\langle\xi(x_1')\xi(x_1''')\rangle + \langle\xi(x_1)\xi(x_1''')\rangle\langle\xi(x_1')\xi(x_1'')\rangle.\end{aligned} \tag{4.182}$$

Correlations involving a product of an odd number of surface profile functions are zero. This is due to the result in (4.181a).

As with the reflection and transmission problem treated earlier, the incident and reflected electromagnetic planes in the region $x_3 > \xi(x_1)$ are taken to be p-polarized. A formal solutions for the form of the fields in this regions is then given by [20, 21]

$$\begin{aligned}H_2^I(x_1, x_3|\omega) &= \exp[i(kx_1 - i\alpha_0(k,\omega)x_3)] \\ &+ \int_{-\infty}^{\infty} \frac{dq}{2\pi} R(q|k)\exp[iqx_1 + i\alpha_0(q,\omega)x_3],\end{aligned} \tag{4.183a}$$

where $\alpha_0(q,\omega) = \left[\left(\frac{\omega}{c}\right)^2 - q^2\right]^{1/2}$, $\mathrm{Re}\alpha_0(q,\omega) > 0$, and $\mathrm{Im}\alpha_0(q,\omega) > 0$. In terms of the incident angle, θ_i, and the scattering angle, θ_s, in Fig. 4.17 the components of wave vector parallel to the x_1-axis in (4.183a) are $k = \frac{\omega}{c}\sin\theta_i$ and $q = \frac{\omega}{c}\sin\theta_s$.

The p-polarized fields in (4.183a) are of the form of incident and reflected plane waves that couple to the surface electromagnetic waves on the thin film. Due to the one-dimensional nature of the surface roughness, the p-polarization of the fields is maintained during the propagation of light throughout the thin film system. This is true for the incident, reflected, transmitted and bulk wave of the thin film. The other s-polarized plane wave solutions do not excite the surface electromagnetic waves that are of interest in the following discussions. Consequently, the case of an incident s-polarized plane wave interacting with the thin film will not be considered here.

Corresponding to the solution of the form in (4.183a), in the region $-d < x_3 < \xi(x_1)$ the form of the solutions of the p-polarized fields are

$$H_2^{(II)}(x_1, x_3|\omega) = \int_{-\infty}^{\infty} \frac{dq}{2\pi} \exp(iqx_1) \times \left[A(q|k)\mathrm{e}^{i\alpha(q,\omega)x_3} + B(q|k)\mathrm{e}^{-i\alpha(q,\omega)x_3}\right] \tag{4.183b}$$

where $\alpha(q, \omega) = \left[\varepsilon(\omega)\left(\frac{\omega}{c}\right)^2 - q^2\right]^{1/2}$, $\mathrm{Re}\alpha(q, \omega) > 0$ and $\mathrm{Im}\alpha(q, \omega) > 0$. Finally, in the region $x_3 < -d$ the transmitted wave solutions for the p-polarization are given by the form

$$H_2^{III}(x_1, x_3|\omega) = \int_{-\infty}^{\infty} \frac{dq}{2\pi} T(q|k) \exp[iqx_1 - i\alpha_0(q, \omega)(x_3 + d)], \tag{4.183c}$$

where $q = \frac{\omega}{c}\sin\theta_t$ in terms of the transmission angle θ_t in Fig. 4.17.

The nature of the propagation of the p-polarized fields interacting with the thin film are obtained by matching the solutions in (4.183), using the electromagnetic boundary conditions. These conditions involves the continuity of the electric and magnetic fields at the interfaces of the thin film.

The formal solutions in (4.183) for the three regions are all matched by the electromagnetic boundary conditions to obtain a set of equations which are then solved for the amplitudes $R(q|k)$, $A(q|k)$, $B(q|k)$, and $T(q|k)$. The results for these amplitudes are expressed in terms of the surface profile functions, the various wave vectors, and the dielectric constant of the film. Of particular interest are the coefficients $R(q|k)$ and $T(q|k)$ as these are the scattering amplitudes from the film of the reflected and transmitted fields, respectively. These are used to determine the speckle pattern generated by the interaction of light with the thin film.

In this way, the scattering amplitude $R(q|k)$ for the reflection of light form the surface of the thin film is determined to be given by [20, 21]

$$R(q|k) = -2\pi\delta(q - k) - 2iG_R(q|k)\alpha_0(k, \omega) \tag{4.184}$$

where $G_R(q|k)$ is the Green's function for the propagation of surface plasmon-polaritons along the rough surface of the thin film. For weak surface roughness the electromagnetic surface wave Green's function is obtained as a solution of the Dyson equation

$$G_R(q|k) = 2\pi\delta(q - k)G_0(k, \omega) + G_0(q, \omega) \int_{-\infty}^{\infty} \frac{dp}{2\pi} V_R(q|p)G_R(p|k) \tag{4.185}$$

where

$$G_0(k,\omega) = \frac{i\Lambda_+(k,\omega)}{\alpha_0(k,\omega)D_-(k,\omega)} \tag{4.186}$$

is the Green's function for the propagation of surface plasmon-polaritons along the thin film in the absence of surface roughness. In (4.186) the parameters entering into the smooth surface Green's function are given by

$$D_\pm(k,\omega) = \Lambda_\pm \pm \frac{\alpha(k,\omega)\Lambda_-(k,\omega)}{\varepsilon(\omega)\alpha_0(k,\omega)}, \tag{4.187a}$$

$$\Lambda_\pm = f_-(k,\omega) \pm f_+(k,\omega), \tag{4.187b}$$

and

$$f_\pm(k,\omega) = \frac{1}{2}\left[1 \pm \varepsilon(\omega)\frac{\alpha_0(k,\omega)}{\alpha(k,\omega)}\right]\exp[\mp i\alpha(k,\omega)d]. \tag{4.187c}$$

The scattering interactions of light with the rough surface in (4.184) arise from the second term on the right side of (4.185). These terms involve the scattering potential $V_R(q|k)$ which depends on the surface roughness profile function. To lowest order in the surface roughness profile function $\xi(x_1)$ the scattering potential of the surface roughness entering into the Dyson equation in (4.185) is

$$V_R(q|k) = v_R(q|k)\hat{\xi}(q-k), \tag{4.188a}$$

where

$$\hat{\xi}(p) = \int_{-\infty}^{\infty} dx_1 \exp(-px_1)\xi(x_1) \tag{4.188b}$$

is the Fourier transform of the surface roughness profile function, and

$$v_R(q|k) = \frac{\varepsilon(\omega)-1}{\varepsilon^2(\omega)} \times \left[\varepsilon(\omega)qk - \frac{\alpha(q,\omega)\alpha(k,\omega)\Lambda_-(q,\omega)\Lambda_-(k,\omega)}{\Lambda_+(q,\omega)\Lambda_+(k,\omega)}\right]. \tag{4.188c}$$

Equations (4.184)–(4.188) formulate a complete description of the single and multiple scattering processes entering into the reflection of light from the thin film. Single scattering processes involve terms in (4.184) in which only a single factor of $V_R(q|k)$ enters, while multiple scattering processes involve terms containing multiple factors of $V_R(q|k)$. The multiple scattering terms account for processes in which the incident light is scattered by the surface roughness into surface plasmon-polaritons which then propagate along the thin film as surface electromagnetic waves.

As the surface electromagnetic waves move along the rough surface they scatter from the surface roughness. This renormalizes their dispersive properties and amplitudes. Eventually the surface electromagnetic waves are scattered away from the surface and into bulk radiating plane waves. This creates the reflected wave generated by (4.183)–(4.188).

In terms of $R(q|k)$ the intensity, $I(q|k)$, of the reflected light from the rough surface of the thin film is characterized by [20, 21]

$$I(q|k) = \frac{1}{L}\frac{\omega}{c}\frac{\alpha_0(q,\omega)}{\alpha_0(k,\omega)}|R(q|k)|^2, \tag{4.189}$$

where L is the length of the rough surface along the x_1 -axis, and $|k| < \frac{\omega}{c}$ and $|q| < \frac{\omega}{c}$ are required for the incident and reflected waves to propagate to and from the thin film. The intensity in (4.189) then represents the scattering process of a single plane wave incident from the vacuum above the thin film in Fig. 4.17 to a plane wave reflected from the surface into the vacuum above the thin film in Fig. 4.17.

In the following, discussions will be presented on the use of the intensity defined in (4.189) to study the statistical features of the speckle generated by the scattering of light from the thin film. This will involve statistical analysis on the level of single and multiple incident plane waves as they interact with and are scattered from the thin film.

Speckle Correlation Function

In order to characterize the statistics of the bright and dark patches of the speckle patterns generated by the interaction of the plane waves of incident light with the thin film, a measure of the variation of the scattered field intensity relative to the mean intensity of the scattered fields is introduced. Representing the variation of the scattered field about its mean value by $\Delta I(q|k)$, for a single incident plane $\Delta I(q|k)$ is defined by

$$\Delta I(q|k) = I(q|k) - \langle I(q|k)\rangle \tag{4.190}$$

Here $\langle\rangle$ is an average over the scattering surface.

For two different incident plane waves a correlation function based on (4.190) for the speckle pattern created by the two plane waves can be introduced. The correlation function is define so as to provide a measure of the similarity of the two different speckles arising from each of the plane waves that are incident on the thin film. The correlation function for this measure of the speckle pattern is defined as [20, 21]

$$C(q,k|q',k') = \langle \Delta I(q|k)\Delta I(q'|k')\rangle = \langle I(q|k)I(q'|k')\rangle - \langle I(q|k)\rangle\langle I(q'|k')\rangle, \tag{4.191}$$

where $|k|, |k'| < \frac{\omega}{c}$ and $|q|, |q'| < \frac{\omega}{c}$ are required for the incident and reflected waves, respectively, to propagate to and from the rough surface in the region of vacuum above the thin film.

Defined in this way, high values of $C(q, k|q', k')$ indicate that the speckle pattern generated by one of the plane waves provides information about the nature of the speckle pattern generated by the other plane wave. For $C(q, k|q', k')$ zero, the two patterns provide no information relative to the other.

In the following the focus will be on understanding the speckle correlator defined in (4.191). This allows for an understanding of the change in speckle patterns as the incident and reflected waves of two different plane wave interact with the thin film. It also provides an understanding of the degree of correlations between the patterns generated by the two plane waves and how these correlations arise from the propagation of surface electromagnetic waves along the thin film.

Green's Function Approach

Using (4.184) and (4.189) in the correlation function in (4.191) and considering only the diffuse scattering in (4.189), the correlation function of the speckle patterns can be expressed in terms of the Green's functions for the propagation of surface plasmon-polaritons along the upper surface of the film in Fig. 4.17.

From (4.191) it then follows that [20, 21]

$$C(q, k|q', k') = \frac{16}{L^2}\left(\frac{\omega}{c}\right)^2 \alpha_0(q, \omega)\alpha_0(k, \omega)\alpha_0(q', \omega)\alpha_0(k'\omega) \times D(q, k|q', k') \tag{4.192}$$

where

$$D(q, k|q', k') = \left\langle |G_R(q|k)|^2 |G_R(q'|k')|^2 \right\rangle - \left\langle |G_R(q|k)|^2 \right\rangle \left\langle |G_R(q'|k')|^2 \right\rangle. \tag{4.193}$$

In (4.193) specular scattering terms involving $2\pi\delta(q-k)$ and $2\pi\delta(q'-k')$ have been omitted as these are not part of the diffuse scattering component from the rough surface. Only the diffuse scattering contains the interesting components of the correlation of the speckle pattern with surface wave properties of the thin film.

The surface averages in (4.192) and (4.193) involve many different averages of products of the surface profile functions. This provides a rich structure of correlated properties of the patterns of the scattered light from the two incident plane waves, arising in various level of the perturbation theory of the weak surface roughness characterizing the surface wave propagation in terms of their Green's functions. These correlations are now addressed for the contributions to the correlation functions at their various levels of approximation, obtaining an expression for the correlation function as a series in the surface profile functions.

It will be shown that the correlations between the patterns of the two incident beams exhibit a variety of memory and additional related effects arising from the presence of surface electromagnetic waves as they are associated with the scattering of each of the two incident wave by the thin film. These contributions are now considered in order of their significance in the perturbations expansion of $C(q,k|q',k')$ in terms of the weak surface roughness.

The terms in $C(q,k|q',k')$ which are of leading order powers of $\xi(x_1)$ are now treated. These terms are denoted as $C^{(1)}(q,k|q',k')$ and $C^{(10)}(q,k|q',k')$, and the leading contribution to both of these correlations are of order $\xi^4(x_1)$ and $\xi^6(x_1)$. The contribution $C^{(1)}(q,k|q',k')$ is determined to be proportional to $2\pi\delta(q-k-q'+k')$ while that of $C^{(10)}(q,k|q',k')$ is determined to be proportional to $2\pi\delta(q-k+q'-k')$. Consequently, each appears in a different region of the scattering phase space, and they both represent distinctly different types of scattering processes.

In terms of these two leading order terms it follows that the correlation function becomes

$$C(q,k|q',k') = C^{(1)}(q,k|q',k') + C^{(10)}(q,k|q',k') \tag{4.194}$$

where, as shown in the following, the terms $C^{(1)}(q,k|q',k')$ and $C^{(10)}(q,k|q',k')$ are easily separated out form (4.192) and (4.193). The separation is made by using a decoupling approximation on the surface averages of the production of four Green's functions, i.e., terms of the form

$$\left\langle |G_R(q|k)|^2 |G_R(q'|k')|^2 \right\rangle. \tag{4.195}$$

The decoupling approximation applied to (4.195) is similar to the decoupling procedure used in the Hartree-Fock treatment of the electron-electron interactions in a many-body treatment of electrons in metals and semiconductors. It is also similar to the decoupling of magnetic interactions in the many-body theory of magnons. In all of these treatments the four particle correlations functions are replaces by products of two particle correlations functions.

Applying these ideas to (4.192) and (4.193) the result in (4.194) follows from making the decoupling approximations [20, 21]

$$C^{(1)}(q,k|q',k') = F(q,k|q',k')\left\langle G_R^*(q|k)G_R(q'|k')\right\rangle\left\langle G_R^*(q'|k')G_R(q|k)\right\rangle \tag{4.196a}$$

and

$$C^{(10)}(q,k|q',k') = F(q,k|q',k')\left\langle G_R^*(q|k)G_R^*(q'|k')\right\rangle\langle G_R(q'|k')G_R(q|k)\rangle \tag{4.196b}$$

where

$$F(q,k|q',k') = \frac{16}{L^2}\left(\frac{\omega}{c}\right)^6 \cos\theta_s \cos\theta_i \cos\theta_s' \cos\theta_i'. \tag{4.196c}$$

The resulting decoupled expressions for $C(q,k|q',k')$ only involve averages of two particle Green's functions of the form $\langle G^*G\rangle$ and $\langle GG\rangle$. These remaining two

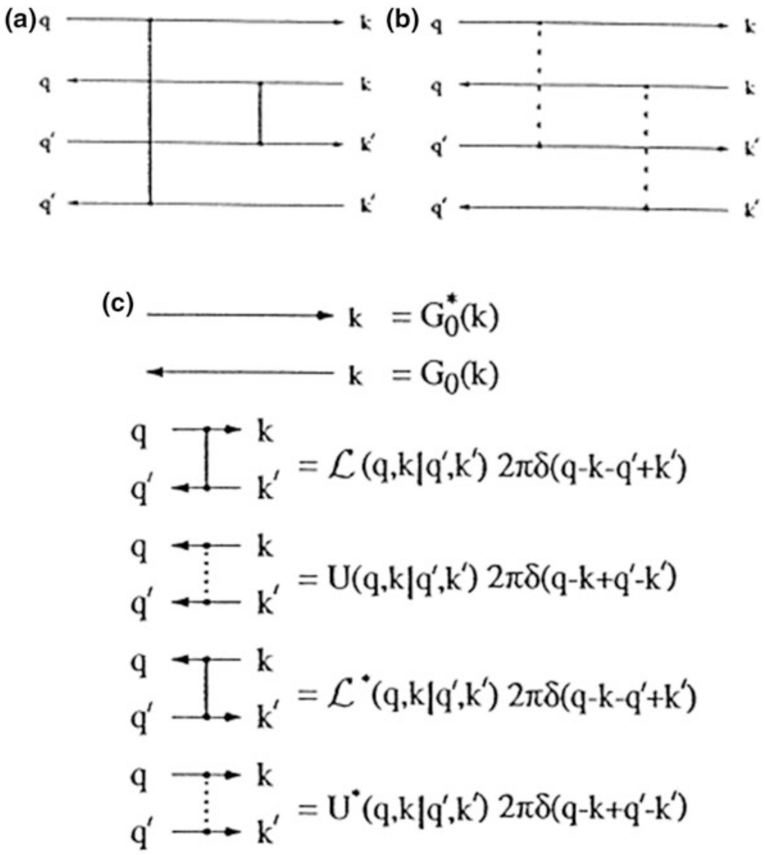

(a)
q
q
q'
q'
k
k
k'
k'
(b)
q
q
q'
q'
k
k
k'
k'
(c)
k = G₀*(k)
k = G₀(k)
q k q' k' = ℒ(q,k|q',k') 2πδ(q-k-q'+k')
q k q' k' = U(q,k|q',k') 2πδ(q-k+q'-k')
q k q' k' = ℒ*(q,k|q',k') 2πδ(q-k-q'+k')
q k q' k' = U*(q,k|q',k') 2πδ(q-k+q'-k')

Fig. 4.18 **a** Diagrammatic representation of the memory effect term, $C^{(1)}(q,k|q',k')$, given in (4.196a). **b** Diagrammatic representation of the term in (4.196b) denoted $C^{(10)}(q,k|q',k')$. **c** Definitions of the diagrammatic features representing the Green's functions, the complex conjugate Green's functions, and the irreducible vertex functions occurring in the diagrammatic representation of the terms entering into the correlation function $C(q,k|q',k')$ [20, 21]. Reproduced with permission from [20]. Copyright 1998 Elsevier

particle Green's functions are subsequently evaluated using the same methods employed to treat the diffuse scattering from the thin film in an earlier discussion.

In terms of a diagrammatic representation, the processes entering into (4.195) and (4.196) are represented as in Fig. 4.18. In particular, the processes in (4.196a) for the $C^{(1)}(q,k|q',k')$ terms are shown in Fig. 4.18a, and the processes in (4.196b) for the $C^{(10)}(q,k|q',k')$ terms are shown in Fig. 4.18b. The two different and distinct contributions are seen to be separated processes in the diagrammatic representation of the terms entering the correlation functions of light scattering. It should be noted that the diagrams listed are those used in the Hartree-Fock approximation of electron systems and in magnon decoupling.

In Fig. 4.18 the solid horizontal lines, as per the figure captions, represent Green's function propagators and their complex conjugates. The vertical lines represent the scattering interactions of the Green's functions with the surface roughness.

In the case of the solid vertical lines the scatterings represent a summation of an infinite series of single and multiple irreducible scattering effects. Some of these multiple scattering terms contain phase coherent processes arising from weak localization effects. These weak localization processes manifest themselves in sharp peaks in the correlation function as a function of wave vector and are related to the nature of the wave functions of the localized surface electromagnetic waves. As seen from Fig. 4.18 the weak localization effects are only present in the $C^{(1)}(q,k|q',k')$ contribution to $C(q,k|q',k')$.

The dashed vertical lines represent correlated single scattering processes where the correlation comes from the correlation of the surface profile functions of the rough surface. These processes are the sole contribution to $C^{(10)}(q,k|q',k')$ as it contributes to $C(q,k|q',k')$. Consequently, there are no sharp peaks related to surface wave localization in the $C^{(10)}(q,k|q',k')$ term, and it is generally found to be a smoothly varying function of the various wave vectors [20, 21].

Weak Localization

To understand the origin and nature of the weak localization effects and the enhancements in the speckle correlation function related to them, a brief review of the two-particle Green's function for the propagation of surface plasmon-polaritons along the rough surface of the thin film is necessary. For more detailed discussions the reader is referred to the subsection on diffuse transmission and reflection from the rough thin film. Following the review a presentation of numerical results for (4.194) applied to a silver film is given.

The important localization effect terms in the correlation function in (4.194) and (4.196a) involve diffuse scattering processes contained within the two-particle Green's functions of the form

$$\left\langle G_R^*(q|k)G_R(q'|k')\right\rangle. \tag{4.197}$$

For weak scattering these Green's functions are obtained within the ladder and maximally cross diagram approximations. These were treated earlier where they were used for the determination of the diffuse scattering of light from the thin film [20, 21].

In the ladder and maximally crossed diagram approximation it was shown that

$$\begin{aligned}\left\langle G_R^*(q|k)G_R(q'|k')\right\rangle &= 2\pi\delta(q-k)2\pi\delta(q'-k')G_R^*(k)G_R(k') \\ &\quad + 2\pi\delta(q-k-q'+k')G_R^*(q)G_R(q')G_R^*(k)G_R^*(k')L(q,k|q',k')\end{aligned} \tag{4.198}$$

where [20, 21] (Representing the solid verticle line in Fig. 4.18.)

$$\begin{aligned}L(q,k|q',k') &= V_0(q,k|q',k') \\ &+ \sum_{0<n}\left(\frac{f_n(q,q',k,k')}{(k-k')^2+4\Delta_n^2}+\frac{g_n(q,q',k,k')}{(q+k')^2+4\Delta_n^2}\right) \\ &+ \sum_{0<n<m}\left(\frac{f_{n,m}(q,q',k,k')}{\left(k-k'-K_{n,m}\right)^2+\Delta_{n,m}^2}+\frac{g_{n,m}(q,q',k,k')}{\left(q+k'-K_{n,m}\right)^2+\Delta_{n,m}^2}\right) \\ &+ \sum_{0,n,m}\left(\frac{f_{n,m}(q,q',k,k')}{\left(k-k'+K_{n,m}\right)^2+\Delta_{n,m}^2}+\frac{g_{n,m}(q,q',k,k')}{\left(q+k'+K_{n,m}\right)^2+\Delta_{n,m}^2}\right)\end{aligned} \tag{4.199}$$

In (4.199) the scattering potential term

$$V_0(q,k|q',k') = \sqrt{\pi}a\sigma^2 \exp\left[-\frac{|q-k|^2}{4a^2}\right] \times v_R^{(1)}(q|k)v_R^{(1)}(q'|k'), \tag{4.200}$$

where $v_R(q|k)$, which is defined in (4.188), has been encounter earlier in the discussions for diffuse scattering and transmissions from the thin film. The surface electromagnetic waves enter into (4.199) through the wave vectors of the forward and backward propagating surface plasmon-polaritons, $K_n(\omega)$, at the frequency of the incident light, ω, and the decay of these excitations in wave vector space is characterized by Δ_n. In terms of these dispersion characteristics $K_{n,m} = K_n - K_m$ and $\Delta_{n,m} = \Delta_n - \Delta_m$ enter into (4.199).

The scattering potential in (4.200) is a smoothly varying function of q, q', k, k' as are the various factors of f_n, g_n, $f_{n,m}$, and $g_{n,m}$, and the reader is referred to the literature for a detailed discussion of these terms. Here only an understanding of the rapidly changing features as a function of wave vectors will be a focus. These are the features that are of most importance in seeing the effects of surface electromagnetic waves on the development of correlations in the speckle pattern.

A strong wave vector dependence is introduced by the Lorentzian factor entering the three sums on the left hand side of (4.199). These factors introduce the interesting weak localization effects arising in the correlation functions from the properties of the surface electromagnetic waves that move along the thin film. The Lorentzian peaks that are introduced into (4.199), just as in the earlier study of the diffuse reflection and transmission of light from the thin film, have widths that are set by the rate of decay, $\Delta_{n,m}$, of the surfaces waves as they propagate along the rough film. As was found in those studies these width can be quite narrow, leading to rapid enhancement over narrow bands of wave vector space.

In (4.199) the first two sums on the right are terms for the ladder approximation and the last two terms on the right are the maximally crossed contributions. The ladder terms account for the usual diffusive transport properties found in disordered media, and the maximally crossed terms account for phase coherent processes related to weak Anderson localization effects. As shall be seen in the later discussions of numerical results, both types of contributions account for Lorentzian peaks enhancements over different regions of wave vector space.

Before a presentation of a numerical example, the contributions to $C(q,k|q',k')$ from the $C^{(10)}(q,k|q',k')$ term is outlined. This introduces effects that are less indicative of the properties of surface electromagnetic waves along the thin film than those from the $C^{(1)}(q,k|q',k')$ term. Both terms, however, contribute to the same order in the surface roughness.

The $C^{(10)}(q,k|q',k')$ term in the correlation function in (4.194) and (4.196b) involves diffuse scattering processes contained within two-particle Green's functions of the form [20, 21]

$$\langle G_R(q|k)G_R(q'|k')\rangle. \tag{4.201}$$

An important feature of the two-particle Green's function introduced in (4.201) is the absence of a complex conjugate pair of Green's functions forming the product. This has great significance in the treatment of the scattering entering into the Bethe-Salpeter equation for (4.201). In particular, for weak roughness, the maximally crossed diagrams no longer give significant contributions. Consequently, to obtain a good weak scattering result, one can deal mainly with the lowest terms of the ladder approximation for the surface scattering.

In the weak roughness limit, it is then shown that [20, 21]

$$\begin{aligned}\langle G_R(q|k)G_R(q'|k')\rangle &= 2\pi\delta(q-k)2\pi\delta(q-k')G_R(k)G_R(k') \\ &+ 2\pi\delta(q-k+q'-k')G_R(q)G_R(q')G_R(k)G_R(k')U(q,k|q',k')\end{aligned} \tag{4.202}$$

where (representing the dashed verticle lines in Fig. 4.18.)

$$U(q,k|q'|k') = \sqrt{\pi}a\sigma^2 \exp\left[-\frac{|q-k|^2}{4a^2}\right] \times v_R^{(1)}(q|k)v_R^{(1)}(q'|k') \tag{4.203}$$

and $v_R^{(1)}(q|k)$ is from (4.188).

Combining the above results in (4.196)–(4.203) for the $C^{(1)}(q,k|q',k')$ and $C^{(10)}(q,k|q',k')$ contributions to $C(q,k|q',k')$ as indicated in (4.194), it follows that

$$C^{(1)}(q,k|q',k') = F(q,k|q',k')D^{(1)}(q,k|q',k'), \tag{4.204a}$$

$$C^{(10)}(q,k|q',k') = F(q,k|q',k')D^{(10)}(q,k|q',k'), \tag{4.204b}$$

where

$$D^{(1)}(q,k|q',k') \cong L|G_0(q)G_0(q')G_0(k)G_0(k')|^2 \\ \times \{L(q,k|q',k')L(q',k'|q,k)\}2\pi\delta(q-k-q'+k') \tag{4.205a}$$

and

$$D^{(10)}(q,k|q',k') \cong L|G_0(q)G_0(q')G_0(k)G_0(k')|^2 \\ \times \{U(q,k|q',k')U^*(q',k'|q,k)\}2\pi\delta(q-k+q'-k'). \tag{4.205b}$$

These represent the dominant terms in the weak surface roughness which are now evaluated for a numerical example.

A simplified notation can be introduced to aid in the presentation of numerical results for the contributions to the correlation functions. This is done in (4.204) and (4.205) by extracting the delta-function factors involving the wave vectors and displaying them explicitly in the contributions to the correlation function. Making this separation it follows that [20, 21]

$$C^{(1)}(q,k|q',k') = 2\pi\delta(q-k-q'+k')C_0^{(1)}(q,k|q',k') \tag{4.206}$$

and

$$C^{(10)}(q,k|q',k') = 2\pi\delta(q-k+q'-k')C_0^{(10)}(q,k|q',k'). \tag{4.207}$$

The two functions are readily seen to be composed of an envelope function which is restricted in phase space by a delta-function condition on its wave vector components entering the scattering geometry. In the following, discussions will be made regarding the properties of the envelopment modulating the delta-functions.

Illustrative Example

The $C(q,k|q',k')$ correlation function in the approximation of (4.194) and (4.204)–(4.207) has been evaluated for a specific realization of the thin film model of Fig. 4.17. In this evaluation, light of wavelength $\lambda = 4579$ Å is incident on a silver film, and the dielectric constant of silver at this wavelength is found to be given by $\varepsilon(\omega) = -7.5 + i0.24$. For the numerical results the film has been taken with a mean thickness of 350 Å and the Gaussian random surface roughness is characterized by the parameters $a = 1000$ Å and $\sigma = 50$ Å. This is similar to the thin film parameters used in the earlier discussions of the diffuse reflection and transmission, and the small mean thickness allows for a diffuse transmission which exhibits its own speckle correlation function in addition to the speckle correlation function of the diffusely reflected light [20, 21].

In Fig. 4.19 a plot is presented of the $LC_0^{(1)}(q,k|q',q'+k-q)$ and $LC_0^{(10)}(q,k|q',q-k+q')$ terms for the light reflected from the thin film. Here L is the length of the rough surface and conditions have been set on the four wave

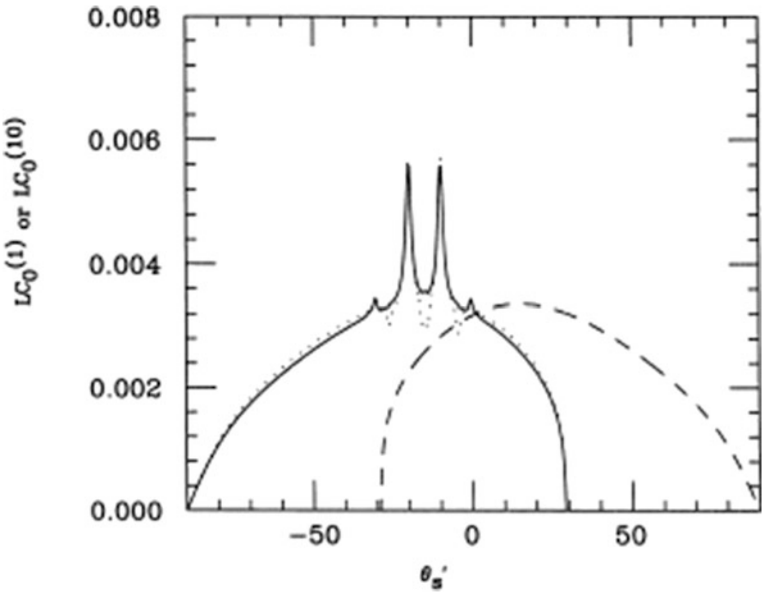


Fig. 4.19 Plots of $LC_0^{(1)}(q,k|q',q'+k-q)$ (solid line) and $LC_0^{(10)}(q,k|q',q'+k-q)$ (dashed line) as a function of θ'_s for fixed $\theta_s=-10°$ and $\theta_i=20°$. The angle θ'_i is set for $LC_0^{(1)}(q,k|q',q'+k-q)$ by the condition that $q-k-q'+k'=0$ and for $LC_0^{(10)}(q,k|q',q-k+q')$ by the condition that $q-k+q'-k'=0$ [20, 21]. Reproduced with permission from [20]. Copyright 1989 Elsevier

vectors so that the delta-functions are always non-zero, i.e., $2\pi\delta(q-k-q'+k')=L$ and $2\pi\delta(q-k+q'-k')=L$, respectively. This is used to reduce the wave vectors from the four (q,k,q',k') independent variables to three independent variables q, k, and q'. Consequently, the plots are only over the set of q, k, and q' for which the envelops of $C^{(1)}$ and $C^{(10)}$ are nonzero [20, 21].

In particular, from Fig. 4.17 the wave vector components parallel to the mean interfaces of the thin film are written in the forms [20, 21]

$$q=\frac{\omega}{c}\sin(\theta_s), \tag{4.208a}$$

$$k=\frac{\omega}{c}\sin(\theta_i), \tag{4.208b}$$

$$q'=\frac{\omega}{c}\sin(\theta'_s), \tag{4.208c}$$

and

$$k' = \frac{\omega}{c}\sin(\theta_i'). \tag{4.208d}$$

These expressions relate the (q, k, q', k') wave vector components to their associated angles $(\theta_s, \theta_i, \theta_s', \theta_i')$ used in the presentation of the plotted data in Fig. 4.19.

For the plots in Fig. 4.19 [20, 21] the angles $\theta_s = -10°$ and $\theta_i = 20°$ are fixed and the correlation functions are presented as a function of θ_s'. The solid line results are for $LC_0^{(1)}(q, k|q', q' + k - q)$ and the dashed lined results are for $LC_0^{(10)}(q, k|q', q - k + q')$. The plots for both contributions to the correlation function have a general smoothly varying component, but the $LC_0^{(1)}(q, k|q', q' + k - q)$ contribution has an addition structure consisting of some sharp peaks as a function of θ_s'.

Two large peaks are found in $LC_0^{(1)}(q, k|q', q' + k - q)$ at $\theta_s' = -10°$ and $\theta_s' = -20°$. These arise from the phase coherent Lorentzian processes found in (4.199) and are referred to, respectively, as the memory and time-reversed memory effects. The memory effect is observed to occur for $\theta_s' = \theta_s = -10°$ and the time-reversed memory effect for $\theta_s' = -\theta_i = -20°$. Both of these features originate in the multiple scattering effects of the surface electromagnetic waves along the interface of the thin film, and their widths are related to the lifetime for surface wave propagation along the thin film. The remaining smaller satellite peaks located on $LC_0^{(1)}(q, k|q', q' + k - q)$ come from higher order processes along the rough surfaces, and for a complete discussion of these the reader is referred to the literature.

A similar study can be made for the speckle correlation function of the light diffusely transmitted through the thin film. These results are now briefly outlined.

From Fig. 4.17 the wave vector components parallel to the mean interfaces of the thin film for diffuse transmission are written in the forms

$$q_t = \frac{\omega}{c}\sin(\theta_t), \tag{4.209a}$$

$$q_t' = \frac{\omega}{c}\sin(\theta_t') \tag{4.209b}$$

where (4.208b) and (4.208d) again represent the incident plane wave components parallel to the mean surfaces of the thin film. With these wave vectors for the diffusely transmitted waves, the correlation function of the transmitted diffuse scattering is defined for the transmitted fields in a similar fashion to that for the reflected speckle in (4.194). The diffuse scattering again can be written in terms of the averages of Green's functions for the motion of surface electromagnetic waves along the rough surface of the thin film. Following from this reformulation, the leading order terms of the correlations of the diffuse transmission can then be evaluated just as in the case of the reflected scattered fields.

In particular, to leading order the speckle correlation function of the diffuse transmitted waves has the form [20, 21]

$$C^{T}(q,k|q',k') = C^{T(1)}(q,k|q',k') + C^{T(10)}(q,k|q',k') \tag{4.210}$$

where

$$C^{T(1)}(q,k|q',k') = 2\pi\delta(q-k-q'+k')C_0^{T(1)}(q,k|q',k') \tag{4.211a}$$

is a term containing the phase coherent multiple scattering contributions to the speckle correlation function and

$$C^{T(10)}(q,k|q',k') = 2\pi\delta(q-k+q'-k')C_0^{T(10)}(q,k|q',k') \tag{4.211b}$$

has no interesting phase coherent processes. This is the analogue of the correlation function in (4.194) for the diffusely reflected light. The diagrammatic processes entering into $C^{T(1)}(q,k|q',k')$ are similar to those contributing to $C^{(1)}(q,k|q',k')$, while the processes entering into $C^{T(10)}(q,k|q',k')$ are similar to those contributing to $C^{(10)}(q,k|q',k')$. Both terms have the distinctive form of delta-function conditions on the scattering wave vectors multiplying an envelope function. The focus in the following will be on the envelope functions.

In Fig. 4.20 results [20, 21] are presented for $LC_0^{T(1)}(q,k|q',k')$ (solid line) and $LC_0^{T(10)}(q,k|q',k')$ (dashed line) plotted as a function of θ_t' for fixed values of $\theta_t = -10°$ and $\theta_i = 20°$. The thin film geometry is the same silver film as that for the plots of $LC_0^{(1)}(q,k|q',k')$ and $LC_0^{(10)}(q,k|q',k')$ in Fig. 4.19. As in Fig. 4.19 the light is of wavelength $\lambda = 4579$ Å.

As with the correlation function results in Fig. 4.19, the result in Fig. 4.20 for the $LC_0^{T(1)}(q,k|q',k')$ contribution to the correlation functions exhibits a series of sharp peaks while the $LC_0^{T(10)}(q,k|q',k')$ contribution is a smooth function of θ_t'. The two large peaks in $LC_0^{T(1)}(q,k|q',k')$ at $\theta_t' = -10°$ and $\theta_t' = -20°$ are, respectively, the memory and time-reverse memory effects. Both of these peaks arise from the phase coherent propagation of surface electromagnetic waves along the rough surfaces of the thin film and represent multiple scattering effects from the interaction of light with the thin film. The memory effect is observed to occur for $\theta_t' = \theta_t = -10°$ and the time-reversed memory effect for $\theta_t' = -\theta_i = -20°$.

Higher order satellite peaks which are smaller in amplitude than the memory and time-reversed memory peaks are also observed in $LC_0^{T(1)}(q,k|q',k')$ plotted in Fig. 4.20. These come from processes involving the dispersion of two different surface electromagnetic modes along the interface, and the reader is referred to the literature for a discussion of these process. As with the $LC_0^{(10)}(q,k|q',k')$ contribution to the correlation function of the reflected light, the $LC_0^{T(10)}(q,k|q',k')$ term in the diffuse transmitted light has none of the interesting phase coherent properties

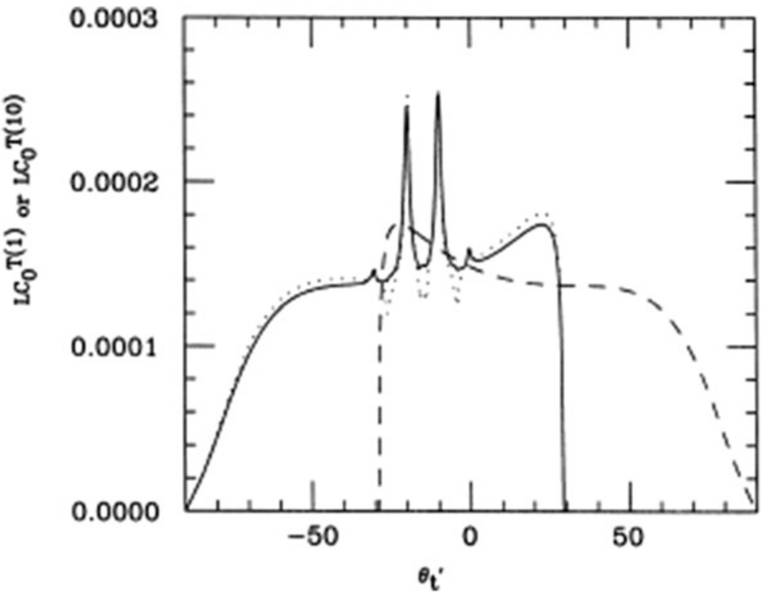


Fig. 4.20 Plots of $LC_0^{T(1)}(q,k|q',q'+k-q)$ (solid line) and $LC_0^{T(10)}(q,k|q',q'+k-q)$ (dashed line) as a function of θ_t' for fixed $\theta_t = -10°$ and $\theta_i = 20°$. The angle θ_i' is set for $LC_0^{T(1)}(q,k|q',q'+k-q)$ by the condition that $q-k-q'+k'=0$ and for $LC_0^{T(10)}(q,k|q',q-k+q')$ by the condition that $q-k+q'-k'=0$. The conditions on the geometry of the thin film and the wavelength characteristics of the incident light are the same as in Fig. 4.19. The film is a silver film with the same dielectric properties as in Fig. 4.19 [20, 21]. Reproduced with permission from [20]. Copyright 1998 Elseviers

found in the $LC_0^{T(1)}(q,k|q',k')$ terms. This is due to the weak roughness and the absence of phase coherence in the multiple scattering process involved in $C_0^{T(10)}(q,k|q',k')$.

Due to the delta-function factor, $2\pi\delta(q-k-q'+k')$ and $2\pi\delta(q-k+q'-k')$, in (4.194)–(4.196) and in (4.210)–(4.211), the leading order processes that have just been discussed have been defined over a very restricted region of phase space. Consequently, the $C^{(1)}$ and $C^{(10)}$ terms of the speckle correlation function are often referred to as short range contributions. In addition to these terms there are various long range and infinite range contributions to the multiple scattering correlations of the speckle correlator. These contributions enter as additive terms to the speckle correlation functions, and the nature of these additional terms entering the speckle correlation function will now be discussed [20, 21].

The Long Range Contribution to the Speckle Correlation Function

In the following a brief presentation of the nature of the longer range contributions to the speckle correlation function is made. To facilitate the presentation only the

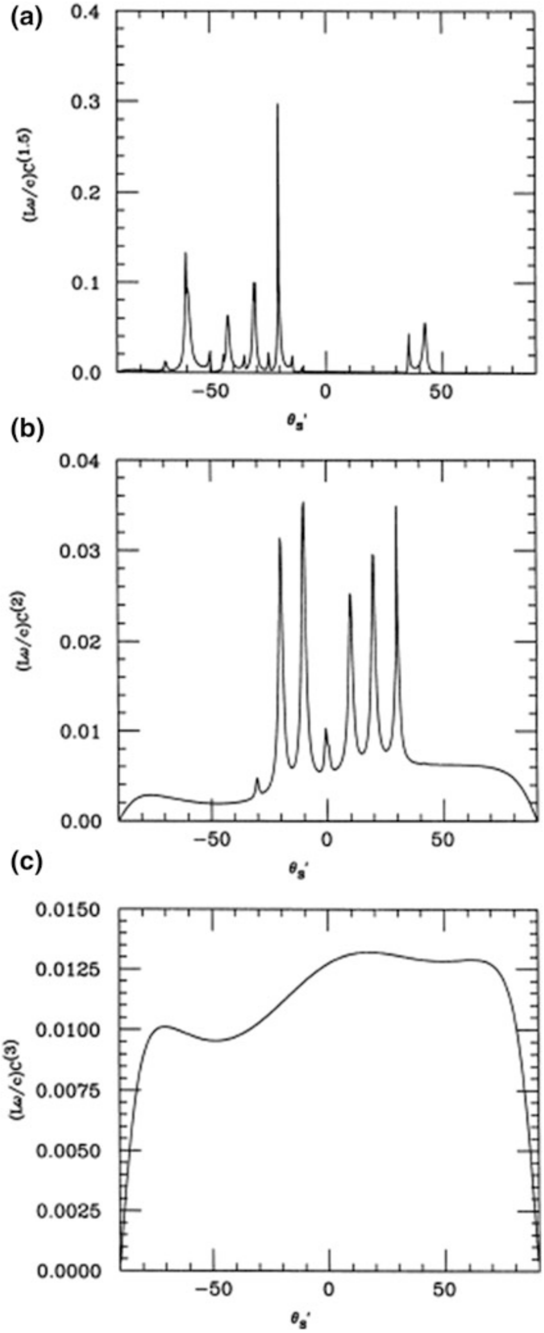


Fig. 4.21 Plots of: **a** $C^{(1.5)}(q,k|q',k')$, **b** $C^{(2)}(q,k|q',k')$, and **c** $C^{(3)}(q,k|q',k')$ as a function of θ_s' for fixed $\theta_s = -10°$, $\theta_i = 20°$, and $\theta_i' = 30°$. The film parameters and wavelength of light are as in Figs. 4.19 and 4.20 [20, 21]. Reproduced with permission from [20]. Copyright 1998 Elsevier

correlations in the reflected light are treated, and the reader is referred to the literature for a treatment of these correlations in the transmitted light.

Other contributions in addition to the $C^{(1)}$ and $C^{(10)}$ terms that enter additively into the speckle correlation function of the diffusely reflected light are the long range terms denoted $C^{(1.5)}$ and $C^{(2)}$ and an infinite range term denoted $C^{(3)}$. Each of these new components enters at a characteristic order of the perturbation sequence in the surface profile function, and unlike the $C^{(1)}$ and $C^{(10)}$ terms these terms give non-zero contributions for general values of (q, k, q', k') wave vectors [20, 21].

In particular, the $C^{(1.5)}$ and $C^{(2)}$ components enter at orders $\xi^6(x_1)$ and $\xi^8(x_1)$, respectively, while $C^{(3)}$ first enters at order $\xi^{12}(x_1)$. Both the $C^{(1.5)}$ and $C^{(2)}$ components are found to exhibit sharp peaks in their wave vector dependence arising from the excitation of surface electromagnetic waves. This, however, is not the case with the $C^{(3)}$ components which are generally smooth, nearly constant, functions of the wave vector variables. This accounts for their consideration as infinite ranged.

All of the processes forming the $C^{(1.5)}$, $C^{(2)}$, and $C^{(3)}$ components can be easily represented diagrammatically. However, since the number of diagrams entering into the accounting for these terms in quite large, the reader is referred to the literature for a detailed listing of them [20, 21].

In Fig. 4.21 some numerical results are presented for the $C^{(1.5)}$, $C^{(2)}$, and $C^{(3)}$ components of the speckle correlation functions, considered for the same system studied in Figs. 4.19 and 4.20 and for the same incident wavelength of light. For these plots the angles $\theta_s = -10°$, $\theta_i = 20°$, and $\theta_i' = 30°$ are fixed. (Notice that in Figs. 4.19 and 4.20 only two angles where fixed as the third was then set by the delta-function conditions on the wave vectors. The greater range of wave vectors over which the $C^{(1.5)}$, $C^{(2)}$, and $C^{(3)}$ components are non-zero requires the third angle be set in order to make a simple plotting of these functions.)

In Fig. 4.21a a plot is presented of $C^{(1.5)}(q, k|q', k')$ versus θ_s'. A number of sharp peaks are observed which find their origins in the excitation of surface electromagnetic waves along the thin film. The widths of the peaks are related to the lifetimes of the surface waves along the rough interfaces of the thin film. In Fig. 4.21b a similar plot of $C^{(2)}(q, k|q', k')$ versus θ_s' is presented. Again the sharp peaks are related to the excitations of surface electromagnetic waves and the widths of the peaks arise from the lifetime of the surface waves along the thin film [20, 21].

Unlike the plots in Fig. 4.21a, b, the plot in Fig. 4.21c of $C^{(3)}(q, k|q', k')$ versus θ_s' is smoothly varying and near a constant. The effects of surface electromagnetic waves are not directly evident from this contribution to the speckle correlation function [20, 21].

The preceding results are interesting as they demonstrate that there are detailed features of the statistics of the speckle of light scattering from rough surfaces and thin film. The speckle from a surface contains important components directly related to the surface electromagnetic waves excited on the surface. Consequently, the nuances of the speckle statistics are affected by the details of the electrodynamics of the surfaces reflecting and/or transmitting the diffusely scattered light.

4.3 Some Application of Plasmon-Polaritons

Plasmon-polaritons have a long history in the study of optical phenomena and form the basis of a number of important technological applications [29–63]. In this final section some of the recent applications of plasmon-polaritons will be described.

First the long studied technique of surface enhanced Raman Spectroscopy will be briefly reviewed [29–40]. This is one of the older applications of surface electromagnetic waves in technology and has been developed into an important laboratory tool. It involves using the large surface fields generated when surface electromagnetic waves are excited on interfaces to enhance the spectroscopic measurement for molecules deposited on the interface. This enhanced spectroscopy arises in part from the increased interaction of the fields at the surface with the molecules deposited on them.

More recently, important advances have also been made in the use of plasmon-polaritons for subwavelength light-guiding, in the design of plasmonic circuitry applications, and in the enhanced transmission of light by films and metamaterials [41–63]. These developments are a new class of applications that are in the process of formulation. The ideas involved in these technologies will also briefly be reviewed here.

4.3.1 Surface Enhanced Raman Spectroscopy

An important application of surface electromagnetic waves is in surface enhanced Raman spectroscopy (SERS) [29–40]. This is a spectroscopic technique applied to molecules that are adsorbed on surfaces supporting surface electromagnetic waves. It is used to study the inelastic scattering arising from the interaction of incident light with the molecules on the surface and from their binding with the surface. The surfaces in these studies are generally rough or formed with a layer of particulates, and a necessary feature is that resonances of the plasmon-polariton exist near the frequency at which the spectroscopic measurements are to be performed.

In addition, the surfaces must exhibit a mechanism by which the incident light used to perform the spectroscopy is coupled to the surface electromagnetic waves. This provides for the excitations of surface wave fields by the incident light. Common materials used in the design of surfaces for SERS are silver and gold, but the effect is found generally to a greater or lesser extent on any type of material providing for surface electromagnetic waves. For many applications, however, gold and silver are a preference as these exhibit good plasmon resonances in the range

400–1000 nm of the visible and near infrared. Aluminum has been used for studies in the ultra-violet [29, 34, 35, 38, 39].

The SERS has a great potential as a technique that extends the range in which Raman spectroscopy can be applied to the consideration of weakly dilute samples and trace analysis. In this regards it can be of importance in the design of sensors or detectors. Areas that have employed SERS are in biochemistry, food safety, threat detection, forensics, medical diagnostics, bacteria detection, and the study of cellular materials [29–33, 36]. It can be sensitive to the presence of materials on the order of single molecules [29–40].

In the development of the spectroscopy of the adsorbed molecules, the function of the surface is to allow the excitation of surface plasmon-polaritons at the frequency of the incident light so as to concentrate light at the frequency of the incident wave on the surface adsorbed molecule. The concentrated field of the surface waves then, through their interaction with the molecules, increases the molecular signal generated by the adsorbed molecules over the signal generated in the absence of the surface waves.

The surface plasmon-polaritons excited by the incident fields develop a concentrated surface plasmon-polariton field intensity at the interface between the two media which is much larger than that of the incident wave alone. This concentrated field intensity was demonstrated in the earlier discussions of the properties of surface plasmon-polaritons excited along the interface. There it was shown that surface electromagnetic waves exhibit a peak intensity at the position of the interface between two media and that the fields decay exponentially with the separation distance from the interface [29].

The signals generated in the light scattered from a molecule have intensities that depend on the intensity of the light of the frequency of the incident spectroscopic wave at the position of the molecule. This includes both the incident field and the plasmon-polariton that it excites on the surface. The field amplitude of the light radiated by the molecule is then proportional to the field amplitude at the site of the molecule generating the molecular response [29, 34, 38, 39].

In addition, the light generated during the Raman processes can also excite its own plasmon-polariton at the frequencies of the inelastically scattered wave. This leads to a further enhancement of the inelastic fields observed. Taking all of the amplification factors into account, the intensity of the signal generated is the square of the amplitude of the total inelastically radiated fields. This intensity is in turn proportional to the square of the intensity of the total spectroscopic fields of the incident and plasmon-polariton waves at the position of the molecule as well the amplification factors from the generation of the inelastic fields and their excited plasmon-polaritons [29, 34, 38, 39].

As a result of its dependence on the square of the field amplitudes and their amplification factors, large enhancements in the intensity of the Raman signal can be achieved with the presence of the plasmon-polariton resonance. In some instances Raman spectroscopy can even be performed on a single adsorbed molecule, and, for general surface coverages, enhancements of the Raman scattering peaks that are of order $10^4 \sim 10^6$ have been observed over those obtained using standard techniques in the absence of surface electromagnetic waves. Claims of enhancements to 10^{12} have been made [29, 34, 38, 39].

Due to molecular distortions and the geometric properties of the molecule itself as it is added onto the surface, symmetry considerations involving the molecule-surface adsorption enter into the spectroscopy of the modes observed. Some Raman modes found in the spectroscopy of molecules in the absence of the surface may be absent when the surface is present while other modes of the molecule are enhanced. In addition, the irregularity of the surface roughness can enter into the surface spectroscopy. Due to the non-uniformity of the surface, it may develop so-called hotspots in which the Raman modes display more enhancement than in other portions of the surface [29–40].

Originally it was thought that the enhanced area of a rough surface was the factor responsible for the enhancement of the Raman spectroscopy lines [29]. More molecules could be presented in a given region of the mean scattering plane than for a flat surface, but this has been shown not to be the prime factor in the enhancement. It is now generally agreed that the enhanced fields arising from surface plasmon-polaritons are a great contributing factor to SERS. In addition, certain chemical effects related to the surface adsorption may be in play.

To understand the nature of SERS and it basic mechanisms, in the following first a simple treatment of the elastic scattering of light will be given. This is followed by an elementary general discussion of the inelastic scattering processes of light known as Raman scattering and their relationships to the elastic processes also occurring at the surface. The treatment is very much simplified and employs a classical model involving classical electrodynamics. Following the discussions of the basic Raman mechanisms the effects of surface enhancement will be discussed along with an example of the effect. For a detailed quantum mechanical treatment the reader is referred to [28, 29, 34, 40].

Simple Model of Elastic Scattering

The elastic scattering of light from a molecule can be roughly understood by considering a highly simplified model. The basics of the model were put forward by considerations of the molecular polarizability made by Drude in the early twentieth century. Later these considerations were developed into a treatment of elastic scattering from molecules by Rayleigh and Thomson.

In the following version of Drude's model the molecule is considered as composed as an electron bound to an harmonic oscillator site. The model can be easily extended to treat multiple electrons at different frequencies of oscillation about the binding site, but the basic features of the response of the system are contained in the one electron system and are easily generalized to consider higher numbers of electrons.

In a simplistic Drude approach the equation of motion, for a single harmonically bound classical electron interacting with an external electromagnetic field, is given by

$$m\left[\ddot{x}+\gamma\ddot{x}+\omega_0^2 x\right]=-eE(t). \tag{4.212}$$

Here x is the displacement of the electron from equilibrium, ω_0 is the frequency of the harmonic motion of the electron upon its displacement from equilibrium, m is the effective mass of the electron, and the term involving γ represents energy dissipated to the environment by the atom through various statistical mechanical processes. In addition, the electric field $E(t)$ driving the electron motion is taken to be parallel to the electron displacement and represents an interaction from the total fields at the frequency of the incident wave. For the considerations of the elastic scattering given here, only radiation at frequencies of interest to Raman spectroscopy are treated.

For the frequencies of interest in typical Raman spectroscopy studies the electric field interacting with the molecule may be taken to be independent of the spatial coordinates. This follows as the wavelength of the incident radiation is much greater than the size of the molecules with which it interacts. At these same frequencies of light the magnetic components of the Lorentz force on the electron are ignored in (4.212). The forces of magnetic origin in general are small for applications of Raman scattering.

The electric field of an incident wave of frequency ω is then represented by $E(t)=E_0\mathrm{e}^{-i\omega t}$ and generates a response from (4.212) of the form $x(t)=x_0\mathrm{e}^{-i\omega t}$. Upon substitution of these two expressions for the field and electron displacement from equilibrium into (4.212) it follows that they are related by

$$m\left[-\omega^2-i\gamma\omega+\omega_0^2\right]x=-eE(t). \tag{4.213}$$

Considering the case of the single electron, its polarization is then described by

$$p=-ex=\frac{e^2}{m}\frac{1}{\omega_0^2-\omega^2-i\omega\gamma}E. \tag{4.214}$$

This is the Drude polarization result in the simple single electron limit.

The expression in (4.214) for the single electron polarization has the standard form of a Lorentzian resonance about the frequency, ω_0, of the electron oscillation, and the necessary damping term γ is seen to limit the singularity of the Lorentzian at the resonance frequency. The expressions in (4.213) and (4.214) represent the basic processes involved in the elastic scattering of light from the molecule, though for quantitative accuracy a quantum mechanical treatment is required along with extending the considerations to the case of molecules involving many electrons.

The classical model, however, can be easily extended to consider many electrons. Adding electrons to the treatment just results in a response that is a sum over Lorentzian resonances representing the oscillation frequencies of each of the separate electrons.

The generalization to molecules with many electrons is briefly mentioned next, followed by finishing the considerations of the elastic scattering from the single electron model. In the discussions of the total scattering from the molecule, first the elastic component of the scattering will be discussed. This is followed by discussions of the Raman scattering which is inelastic in nature and the modification of the (4.213) and (4.214) needed to understand Raman scattering. It will be seen that the Raman scattering frequency peaks are closely related to the elastic scattering peaks being treated here.

When generalized to treat N molecules per volume each of which consists of Z electrons, the electric polarization vector is

$$P = \frac{Ne^2}{m} \sum_{j=1}^{Z} f_j \frac{1}{\omega_j^2 - \omega^2 - i\omega\gamma_j} E. \tag{4.215}$$

Here ω_j is the frequency of the harmonic oscillator of the jth electron, γ_j is the rate of decay of the jth electron, and f_j is the degeneracy of the state of the jth electron. As mentioned earlier, the polarization is found to be composed of a sum of resonances occurring at the frequencies of the harmonic displacements of the electrons in the molecule. From (4.215) the dielectric constant representing the electron response of the system of molecules is then given by

$$\varepsilon(\omega) = 1 + 4\pi \frac{Ne^2}{m} \sum_{j=1}^{Z} f_j \frac{1}{\omega_j^2 - \omega^2 - i\omega\gamma_j}. \tag{4.216}$$

The properties of the many electron systems in (4.215) and (4.216) essentially display the basic features of the single electron system with the added treatment of multiple resonance frequencies existent in the model. The proceeding discussions will focus on the single electron system, presenting discussions of the elastic and then the Raman scattering of light from the single electron system.

Radiation from the Molecular Electron: Elastic Scattering

During the interaction of the electron with the incident fields, as described in (4.212), the electron undergoes an acceleration described by

$$\ddot{x} = \frac{e}{m} \frac{\omega^2}{\omega_0^2 - \omega^2 - i\omega\gamma} E(t). \tag{4.217}$$

In general, this motion is non-relativistic so that the power radiated by the accelerating charge is related by the Larmor formula

$$\frac{dU}{dt} = \frac{2e^2}{3c^3} |\ddot{x}|^2 . \tag{4.218}$$

Combining (4.217) and (4.218) then represents the elastically scattered radiation from the electron bound in the molecule.

In particular, in (4.218) $\frac{dU}{dt}$ is the rate of energy loss of an accelerating electron through radiation damping and c is the speed of light. From (4.217) and (4.218) it follows that the rate of energy radiated by the electron as it is driven by the incident field is

$$\frac{dU_r}{dt} = \frac{2e^2}{3c^3}\left(\frac{e}{m}\right)^2 \frac{\omega^4}{\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2} E_0^2 \,. \tag{4.219}$$

In this approximation, the total scattering cross section of the molecule interacting with the incident field is obtained by comparing the radiated power to the power flux of the incident plane wave as it interacts with the molecule. The average flux of the incident fields is then computed as the Poynting vector. For the incident fields, this has the form

$$S = \frac{c}{4\pi} E_0^2 \,. \tag{4.220}$$

Consequently, in terms of the scattered and incident fluxes the total elastic scattering cross section of the radiation incident on the single molecule is obtained as the ratio of (4.219) and (4.220) so that

$$\sigma_{Total} = \frac{\frac{dU_r}{dt}}{S} . \tag{4.221}$$

Following a little algebra it is found that

$$\sigma_{Total} = \frac{8\pi e^4}{3m^2c^4} \frac{\omega^4}{\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2} . \tag{4.222}$$

This represents the total of the elastic diffusely scattered radiation from the molecule at general frequencies of the incident light.

Two limits of the elastic scattering that are important are Rayleigh scattering and Thomson scattering. In the limit of Rayleigh scattering the frequency of the incident radiation and the rate of atomic damping obey $\omega,\gamma \ll \omega_0$. This is the case of elastic scattering from a strongly bound electron so that from (4.222) the Rayleigh limit becomes

$$\sigma_{Rayleigh} = \frac{8\pi e^4}{3m^2c^4} \frac{\omega^4}{\omega_0^4} . \tag{4.223}$$

The cross section exhibits the famous ω^4 scattering relation that among a variety of scattering effects is responsible for the blue coloration of the sky.

The other famous limit of (4.222) is that of Thomson scattering. This is the opposite limit to that of the Rayleigh limit. In this case $\omega_0, \gamma \to 0$ so that the electron is considered to be free, unbound, and experiencing no dissipative losses. From (4.222) it follows that the total cross section approaches a constant given by

$$\sigma_{T\ \mathrm{hom}\,son} = \frac{8\pi e^4}{3m^2c^4}. \tag{4.224}$$

This is the non-relativistic limit of scattering from electrons and its relativistic counterpart is Compton scattering.

The elastic scattering that is of most interest in Raman scattering is from Rayleigh scattering. This follows as the frequencies of the incident light in most Raman experiments are generally chosen to be much less that ω_0. Consequently, in general, the inelastic scattering known as Raman scattering is found at side band frequencies to the frequency of elastic scattering of the scattered light generated in the molecular sample by an incident plane wave of light. The Raman side bands are much weaker than the elastic scattering bands as they arise from perturbations of the equilibrium structure of the materials arising from the excitation of elementary interactions in the system.

Raman (Inelastic) Scattering

The Raman interactions leading to the excitation or absorption of elementary excitations are weak scattering processes. Processes involving the creation or destruction of single excitations are of small probabilities, and processes involving the creation and destruction of more than one elementary excitations are accordingly even less probable. While processes involving single excitations lead to frequency peaks in the inelastic scattering cross section, multiple excitation processes tend to be broad distributions in frequency [29–39].

A large variety of elementary excitations can be involved in generating the Raman effect. In this regard, processes involved in the inelastic scattering observed in Raman scattering can arise from the creation or absorption of more than one type of elementary excitations in the scattering medium. In the case that phonons are involved in the inelastic processes the scattering is often referred to as Brillouin scattering, but the excitations created may also be from a variety of electronic, plasmon, and polariton modes that may be present in the large variety of materials that are studied for their inelastic processes.

Due to the weak cross sections of Raman inelastic processes they can be separated into two different types. First order Raman processes involve a single elementary excitation, while second and higher order processes involve multiple sets of elementary excitations.

First Order Processes

First order Raman scattering involves the creation or destruction of a single elementary excitation in the scattering medium. The selection rules for these processes are that the energy or frequencies of the light and elementary excitations are conserved as well as their momenta. Specifically,

$$\omega' = \omega \pm \Omega \tag{4.225a}$$

where ω', ω, and Ω are, respectively, the frequencies of the Raman scattered light, the incident light, and the elementary excitation. In the case of the momenta conservations it follows that

$$\vec{k'} = \vec{k} \pm \vec{K} \tag{4.225b}$$

where $\vec{k'}, \vec{k}$, and $\vec{K}$ are, respectively, the wave vectors of the Raman scattered light, the incident light, and the elementary excitation. Here the upper signs in (4.225) are for the absorption of an elementary excitation and the lower sign is for the creation of an elementary excitation.

The conservation relationships in (4.225) and the basic idea of Raman inelastic processes itself can be easily obtained from considerations of the polarization response of a molecule to an applied electric field and the interaction of the molecular polarization with elementary excitations in the system. In the following a brief sketch is given of such an approach based on simple considerations. These discussions provide a rough theory that underlines some of the basic principles of the Raman effect. A completely correct treatment, however, requires quantum mechanics and the details of the crystal structure of the system being studied. Following these discussions, the role of surface plasmons in surface enhanced Raman spectroscopy will be discussed and some examples provided as an illustration.

In terms of the molecular polarizability,α, the dipole moment of a molecule in response to the applied electromagnetic field is

$$\vec{p} = \alpha \vec{E}_{applied}. \tag{4.226}$$

This provides the response of the molecule in terms of the molecular polarizability which in turn is affected by the environment of the molecule within the medium it is located. The molecular environment contains various types of elementary excitations which then influence the molecular polarizability and show up in the Raman spectroscopy.

If an elementary excitation of the scattering medium interacts with the molecules forming the Raman media, its effect on the polarizability can be described by introducing a generalized coordinate u for the amplitude of the interaction of the excitation with the molecule. The molecular polarizability is then a function of the amplitude, u. In this interaction, it is reasonable to assume that the polarizability can be expanded as a Taylor series in u. It then follows that in terms of the generalized coordinate

$$\alpha = \alpha_0 + \alpha_1 u + \alpha_2 u^2 + \cdots \ . \tag{4.227}$$

The forms in (4.226) and (4.227) are over simplifications as for most molecules the molecular polarizability is a tensor and the generalized coordinate is a vector so the (4.227) is a series involving a number of tensor terms. The basic idea of the Raman response, however, is contained in the treatment in (4.226) and (4.227) and will be followed throughout the discussions presented here.

If the generalized coordinate u of the elementary excitation interacting with the molecule has a time dependence of the form

$$u(t) = u_0 \cos \Omega t, \tag{4.228}$$

where Ω is the frequency of the elementary excitation, then for an applied field

$$E(t) = E_0 \mathrm{e}^{-i\omega t}. \tag{4.229}$$

it follows that the molecular dipole has a complicated time dependence composed of a number of harmonically varying terms in time. This is given by

$$\vec{p} = \left(\alpha_0 + \alpha_1 u_0 \cos \Omega t + \alpha_2 u_0^2 \cos \Omega t + \cdots\right) E_0 \mathrm{e}^{-i\omega t}. \tag{4.230}$$

The dipole in (4.230) is seen to have frequency components at frequencies ω, $\omega \pm \Omega$, and $\omega \pm 2\Omega$. These correspond to elastic processes, processes involving the creation and destruction of elementary excitations, and processes involving the creation and destruction of two elementary excitations, respectively.

First consider the terms involving a single elementary excitation. These are responsible for many of the important sharp frequency peak found in Raman scattering spectra. The various first order processes acting to form the spectra can be classified into a number of categories involved different types of inelastic transitions. Each of these enter into the scattering in its own way.

Inelastic processes which involve the creation of a single elementary excitation are known as Stokes processes and those involving the destruction of a single elementary excitation are known as anti-Stokes processes. All of these scatterings involve the elementary excitations that are quantized and obey Bose-Einstein statistics. Generally, at thermal equilibrium they are found to obey the Planck distribution. These properties are important in the following discussions.

For such a system involving first order scattering processes, if the system is initially in thermal equilibrium the intensities of the Stokes and anti-Stokes lines in Raman scattering satisfy the relation

$$\frac{I(\omega + \Omega)}{I(\omega - \Omega)} = \exp\left(-\frac{\hbar\omega}{k_B T}\right) \tag{4.231}$$

where ω is the frequency of the incident wave and $I(\omega')$ is the intensity of the Raman spectroscopy line at frequency ω'. It is important to note, however, that in the case of SERS the relation in (4.231) may be modified by the processes involved in the resonant excitation of surface plasmon-polariton waves. In general, however, it is seen that at low temperature the anti-Stokes lines are frozen out and the Raman spectroscopy is dominated by the Stokes peaks.

Higher order Raman Processes

Some inelastic processes involve the creation and destruction of two elementary excitations or the simultaneous creation and destruction of elementary excitations. These processes enter the scattering through the third term on the right in (4.230). The selection rules for these processes are that the energy or frequencies of the light and elementary excitations are conserved as well as their momenta, including the crystal momentum when applicable.

Specifically,

$$\omega' + \omega \pm \Omega \pm \Omega' \tag{4.232a}$$

where ω', ω, Ω, and Ω' are, respectively, the frequencies of the Raman scattered light, the incident light, and the two elementary excitations. The momentum conservation is then given by

$$\vec{k'} = \vec{k} \pm \vec{K} \pm \vec{K'} + \vec{G} \tag{4.232b}$$

where $\vec{k'}, \vec{k}, \vec{K}, \vec{K'}$ are, respectively, the wave vectors of the Raman scattered light, the incident light, and the two elementary excitation. In (4.232b) $\vec{G}$ is a reciprocal lattice vector that may enter into the as the component of crystal momentum in the case of a crystalline media or substrate.

In practice, the second order Raman affect does not manifest itself in sharp Raman frequency peaks. This is due to the multiplicity of ways that two elementary excitations can enter into satisfying the conservation relations in (4.232).

The earlier discussions were focused upon the basic ideas of Raman Spectroscopy as an inelastic scattering processed based on polarizabilities of the system. In the following it will be discussed how surface plasmons can be used to enhance the Raman effect. This provides an important technique for many important applications of this spectroscopy.

Mechanisms of Surface Enhanced Raman Spectroscopy

To conclude the discussion some remarks will now be made on the function of the surface in SERS. In particular the focus will be on the role of the surface in enhancing the fields to which the molecules are subjected. Discussions of the chemical effects of molecular adsorption on the surfaces will not be addresses. These considerations may be important in some instances and the interested reader is referred to the literature for such discussions [29–40].

The surface enhancement of the Raman effect arises from the resonant interaction of the molecules with surface plasmon-polariton resonances of the incoming and outgoing radiations as they interact and are generated by the molecule, respectively. Common types of surfaces that are used in SERS are rough surfaces or surfaces formed from metal particulates. For a good SERS effect rough surfaces typically are required to have subwavelength features that range from 1 to 100 nm. This is also the case with collections of particulates [29–40].

A Simple System for Surface Enhanced Raman Spectroscopy

A simple example of a SERS generating surface is a spherical metal particle. The molecules that are the objects of the SERS study would be located at or near the surfaces of the spherical particle. Since the particle is a subwavelength feature, its response to the applied incident field is modeled as that of a spherical dielectric interacting with a uniform electric field [29–40]. The polarization generated in a dielectric sphere by a uniform field is known to be given by

$$\vec{P} = \frac{3}{4\pi}\left(\frac{\varepsilon(\omega) - 1}{\varepsilon(\omega) + 2}\right)\vec{E}. \tag{4.233}$$

Here $\varepsilon(\omega)$ is the frequency dependent dielectric function of the material forming the sphere and $\vec{E}$ is the applied electric field from the incident wave.

It is readily seen from (4.233) that an enhancement of the polarization occurs in the regions of frequency for which

$$\varepsilon(\omega) \approx -2. \tag{4.234}$$

As discussed in an earlier chapter this is the condition for the excitation of a plasmon-polariton resonance in the sphere, and is the origin of the red coloration found in some glasses made to contain small gold particulates. In the case of the SERS affect, an incident field at this frequency will generate a large polarization response in the particle which will in turn create a large field at the surface adsorbed molecules. This polarization response is a result of the excitation of the plasmon-polariton mode in the sphere. In the case that the inelastically generated fields of the molecule are also at frequencies near the resonance, an enhancement of the fields radiated at the inelastic frequency will take place. Again, this effect on the inelastic wave is due to the plasmon-polariton excitation [29–40].

Figure 4.22a illustrates the polarization response of a spherical gold particle in a uniform time-dependent electric field. On the scale of the spherical particle, in general, the incident light for the Raman spectroscopy has wavelength much greater than the diameter of the sphere. This allows for the neglect of the spatial dependence of the wave and is responsible for the dipole response of the particle to the field. The single sphere is found to generate an intense local field that would allow the spectroscopic study of molecules adsorbed on its surface [29–40].

For many spectroscopic applications a collection of particles is used. These may be arranged in the deposition of a colloidal film or through some other deposition technique on a surface. In this case particles aligned parallel to the field of the incident field may assist in the generation of large enhancement fields. The aligned particles form an aligned set of dipoles such that the electric fields between the dipoles undergo enhancements in the regions separating the dipoles. This is illustrated in Fig. 4.22b.

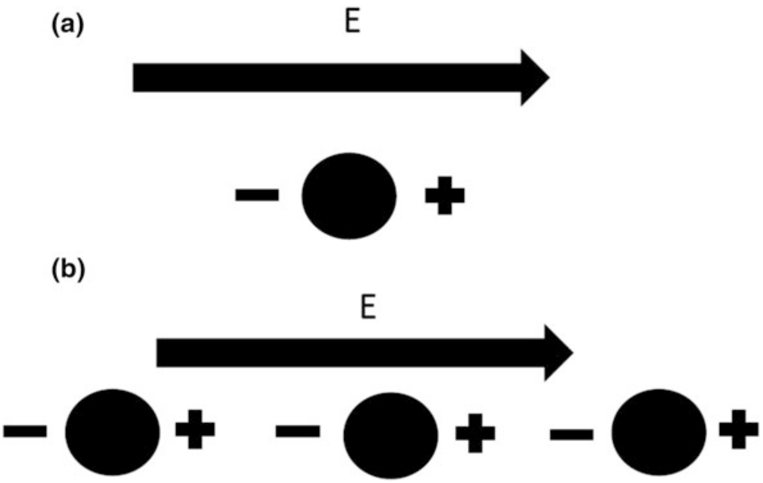


Fig. 4.22 Schematic plots of an applied electric field interaction with polarized spheres. In **a** a single sphere is polarized by the external field and in **b** multiple aligned spheres are polarized by the external field. The plus and minus signs indicate the polarized charge regions of the spheres

In the aligned system, the polarization generated depends on the field alignment. In asymmetric systems electromagnetic interactions between particles can result in collective plasmon-polaritons propagating through the arrays of particles. In this limit, the array would approach the limit of a rough surface which supports propagating plasmon-polaritons similar to the rough surfaces of the earlier discussions in this chapter. The plasmon-polariton resonance of surface electromagnetic waves excited on a randomly rough surface are another important arrangement used in applications of SERS.

To conclude, some experimental results of Raman spectroscopy performed on rough Ag, Au, and Cu surfaces is presented [40]. These results illustrate the dependence of Raman spectroscopy on surface roughness and on the composition of the surface upon which the molecules are adsorbed.

In Fig. 4.23 experimental results are presented for poly(3-hexylthiophene) and plyaniline-eneraldine base on various roughness of a Au supporting surface. In Fig. 4.23a the various rough surface profiles are shown for the periodic width a and the periodic peaks to valley height h. The figure is drawn to represent the ration a/h, and the reader is referred to the paper [40]. The metal surfaces are made using a deposition technique that allows for the surface profile to be modulated.

In Fig. 4.23b, c the SERS spectra are presented for (b) poly(3-hexylthiophene) and (c) polyaniline-emeraldine base structure. The numbered curves in each plot correlate with the number of the surface profiles in Fig. 4.23a. It is seen that as the surface roughness is increase the intensity of the spectra generated on the surface is found to increase significantly.

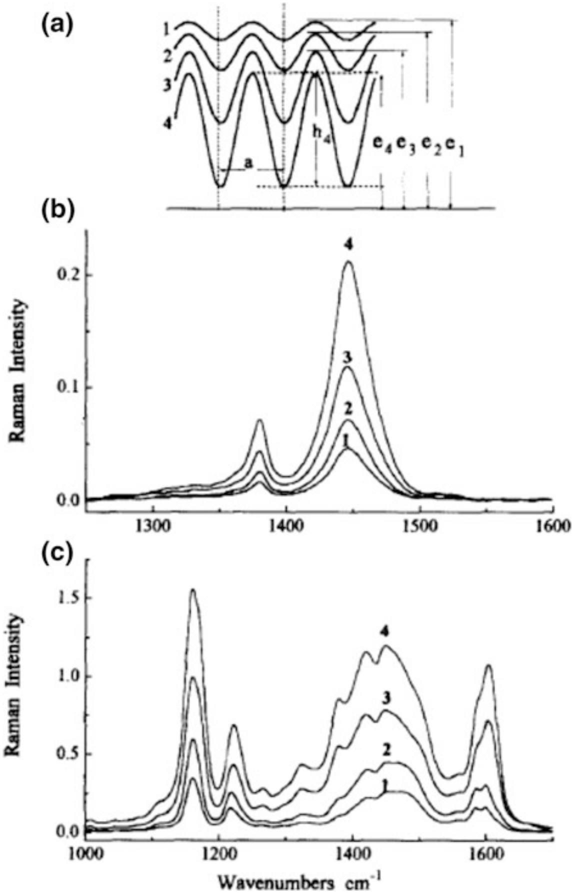


Fig. 4.23 **a** Idealized change in the metal roughness by means of different preparation techniques; roughness dependence of SERS intensity of **b** poly(3-hexylthiophene) and **c** polyaniline-emeraldine base structure. Spectra (1–4) from (**b**) and (**c**) correspond to the (1–4) roughness profiles in (**a**) [40]. Reproduced with permission from [40]. Copyright 1998 Elseviers

In Fig. 4.24 SERS spectro of emeraldine base films on Ag, Au, and Cu surfaces [40]. This offers a comparison of the spectra obtained from the excitation so surface plasmons on the three different metal surfaces. For a more detailed discussion of the chemistry and the detailed differences in the spectra generated on the different metal surfaces the reader is referred to the original paper [40].

4.3.2 *Subwavelength Properties in Light-Guiding, Spasers, and Plasmonic Circuitry*

In the following some important applications of the subwavelength properties of plasmon-polaritons to technology and device designs are discussed. These include: subwavelength properties in light-guiding [41–46], spasers [47–56], and plasmonic circuitry [41–46].

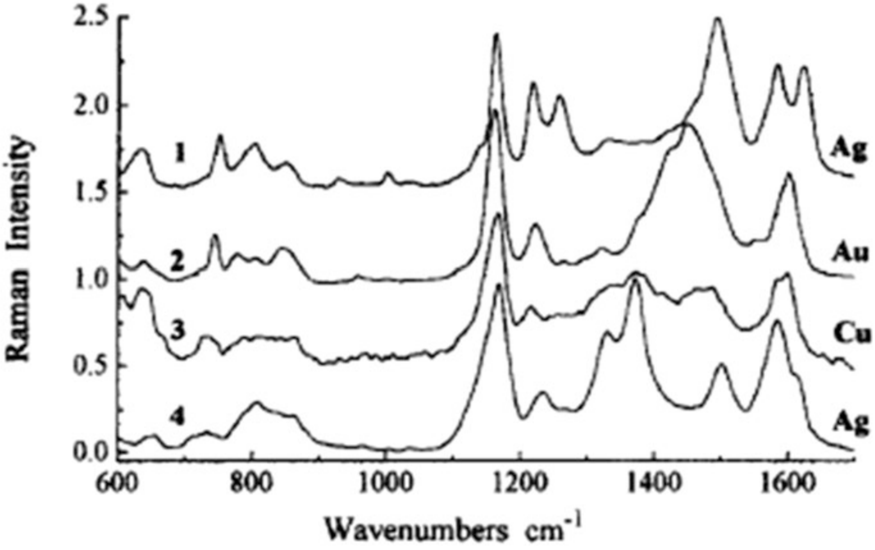


Fig. 4.24 SERS spectra of emeraldine base films of 30 nm thickness deposited on (1) Ag, (2) Au, and (3) Cu; (4) corresponds to the SERS spectrum of a 85 nm emeraldine salt film on Ag [40]. Reproduced with permission from [40]. Copyright 1998 Elsevier

The design of various types of waveguides for the steering of plasmonic excitations along channels laid out on the interface between media supporting plasmons has been a recent consideration [41–46]. In this regard, a large range of different types of waveguide channels created on planar surfaces have been proposed and tested to determine their effectiveness and efficiency characteristics in the transmission of plasmonic signals. Problems in regard to the propagation losses for both straight and bent waveguide channels have been found to be important components of the considerations needed in the design of these transmission devices. From the various waveguides that have been studied the next step has been the construction of plasmonic circuits created by the assembly and interconnection of arrays of conjoined waveguides.

The circuit complexity is again limited by the efficiencies of the single waveguide designs and of the joins that are made between waveguides [41–46]. These circuits and their waveguide components all require nanoscience dimension for their effective operation and are part of the current focus on nanoscience studies for the miniaturization of devices available for important technological applications.

In order to read into plasmonic circuits various inputs from the outside world and to restrict the flow of information within the plasmonic circuits, specially designed waveguide segments have been composed to act as heat and optical sensors and as polarization filters within the plasmonic circuitry [41–56]. These rely on the sensitivity of the materials used in waveguide design to external agents such as applied electric fields, the presence of chemical and biological agents, and the heat and optical interactions coming from the outside world to affect the way in which they manipulate plasmon-polaritons.

In addition, it is found that certain physical design patterns can be introduced into the waveguide channel components which make the waveguides sensitive to the transmission of different polarizations of guided modes introduced into them [41–56]. These allow for the modulation of polarized signal with plasmonic circuits.

Another important component of plasmonic circuits involves the design of plasmonic lasers and spasers which are used to introduce plasmonic signals into the circuits or the design of other efficient method for coupling signals from the outside world into the plasmonic circuitry [41–52]. These are needed to excite plasmon-polariton modes into circuits which will then use them to perform device operations.

A final aspect of the subwavelength interaction of light in plasmonic systems involves the enhanced transmission of light though plasmon supporting plates. These enhancers may be formed from a layer or plane of a single plane material or from composite metamaterials. The planar structures are used to perform important amplifications on optical signal which are incident on the full planar plasmon-polariton surfaces. This last example has less to do with optical circuits and more to do will general optical effects of materials. Consequently, it will be discussed separately in the following treatment in this chapter.

In the following, discussions are first given of plasmonic waveguides and their assembly into plasmonic circuits. The basic properties of waveguides and the various geometries of waveguide design are given. The design of plasmonic lasers for the excitation of plasmon-polariton waveguide modes are discussed and of various couplers for the coupling of external light into the waveguide in order to excite plasmon-polariton guided modes within waveguides. Waveguide losses and efficiency will be considered along with restrictions on the design of full plasmonic circuits. Routers and waveguide couplers will be explained and their use in plasmonics discussed. The features of waveguides needed in the design of polarizers, sensors of heat and chemical and biological agents, and waveguide modulation by external applied fields are explained.

A. **Plasmonic Waveguides and Circuits**

Recently there has been a great amount of effort focused on the development of plasmonic circuit technology. This is meant to promote plasmonics as another approach in optoelectronic methods for the replacement of electronic components by improvements based on optics or for improving the interfacing of optical and electronic systems. As with all such approaches to optoelectronics, plasmonics has its own particular set of advantages and disadvantages in attaining these objectives.

The advantages of plasmonics involves the ability to facilitate the miniaturization of circuits and the potential to increase the speed, over that found in electronics systems, at which signals and energy can travel in the systems intended to perform device functions. However, a major disadvantage of plasmonics is the large decay rate of signals in plasmonic systems and the need to develop efficient means of coupling signal into and out of plasmonic circuits [41–52]. The rapidly developing field of plasmonic circuitry is currently directed to the resolution of these problems and to the design and implementation of plasmonic devices in various fields of engineering, physics, and biology.

A fundament basis of this is the development of plasmonic waveguides for the transportation of plasmonic signals and energy through space [41–46]. In addition, various devices for the modification and manipulation of plasmonic signals are needed for the design of circuits which, similar to electronics circuits, produce meaningful outputs from a set of input signals. Efficient means of inputting and outputting optical signals from the plasmonic circuits as well as the development of plasmonic lasers for the generation of optical signals within plasmonic circuits are also required [53–56].

In the following the initial focus will be on discussions of the problems related to the function of plasmonic waveguides and the various type of waveguide designs considered for plasmonics. What are the advantages and problems with some of these waveguide designs? This is followed by the treatment of circuit devices and lasers, followed by circuit considerations and applications.

Plasmonic Waveguides

Plasmonic waveguides are in general one-dimensional features placed on or near the interface between two media [41–46]. The interface between the two media, along with the waveguide feature, support plasmon-polariton modes that travel along the waveguide channel. In particular, the channel of the waveguide feature is designed to steer the propagation of a plasmonic wave along its one-dimensional length.

The channel of the waveguide may be formed as a composite ridge taking the form of a strip on the surface, a surface groove, a surface edge, a nano-wire close to or on the interface between two media, or even as one-dimensional arrays of nano-particles on the interface of a surface [41–46]. In some instances, the guiding channel may include two parallel stripes. As a general rule, however, all of these channels are set to tightly confine propagating guided modes at scales that are beyond the diffraction limit of light. Each of the channels mentioned has been studied and characterized for efficiency in a variety of device applications and will be discussed later.

For the design of compact circuits, it is helpful if the guided plasmon-polariton modes are tightly bound about the waveguide channel. This facilitates the formulation of compact, localized circuits in space. The two aspects of the guided mode propagation length and the tightness with which the modes are localized about the guiding channel are fundamental design features taken into account in the formulation of waveguides for plasmon-polariton propagation. Unfortunately, they are often found to be conflicting aspects that need to be resolved to build optimal performing wave guides and circuits [41–46].

There are a large variety of designs and design considerations that have gone into the composition of the numerous one-dimensional waveguide channel designs that have been studied [41–52]. These considerations are needed to maximize the propagation lengths of the plasmon-polariton guided modes, increase their confinement characteristics about and within the channel media, and aid in the propagation of guided modes efficiently through channel bends meant to change the direction of guided mode propagation along the surface. In addition, the efficient interface between the waveguide and optical sources, inputted, and outputted signals which may be electrical, chemical, thermal, or optical in nature is necessary for many important circuit operations.

A common problem in the design of plasmonic waveguides is that of signal energy losses. These losses arise from both Joule heating and radiative scattering losses from the waveguide due to impurities or in the dielectric mismatches at the interfaces between waveguides or the devices with which they interact. As a result plasmon-polaritons have a range of propagation along the waveguide channel which, typically, is of the order of millimeters. Some of the approaches put forth to limit the losses are to incorporate low-loss materials into channel designs [41–52]. Applications in this regards that have been helpful have involved the use of metal oxides and nitrides as well as designs based on graphene interfaces.

Another source of radiative losses occurs in bended waveguides. These have channels which are not straight but at some points along the channel length are bent to propagate the guide modes in a different direction than that of the original channel before the bend. The radiative losses at the bend increase with the sharpness of the angle of the bend. Bragg mirrors have in some cases been employed to reduce these loss effects.

Related to the energy loss problems is the need to overcome the generation of heat within the waveguides. The heat generated in plasmonic circuits, in some instances, can approach the heat generated within electrical circuits [41–46]. In both plasmonic and electronics systems the generated heat can produce unwanted modifications in the properties which a crucial to the correct functioning of the circuits On a more positive note, heating effects may also be of interest in the design of certain types of sensors. Here the modification of the circuit operating characteristics may be a vital point in their design. This will be discussed later.

For an optimally functioning waveguide one would wish to develop a guide that both tightly confines the plasmon-polariton modes localized within and about the wave guide channel and also exhibits a long propagation length along the channel for its guided modes [41–46]. It is generally found, however, that these two features oppose one another. In particular, highly confined modes tend to exhibit higher energy losses. Consequently, the propagation length of guided modes along the channel is inversely related to the radius of the region about the channel in which the plasmon-polariton fields are confined. In a successful waveguide design a balance must be made between these two characteristics.

In Fig. 4.25 a schematic is presented showing a number of waveguide geometries [45]. In each figure the waveguide channel is taken to be perpendicular to the page so that the cross section of the localized electromagnetic fields for the guided

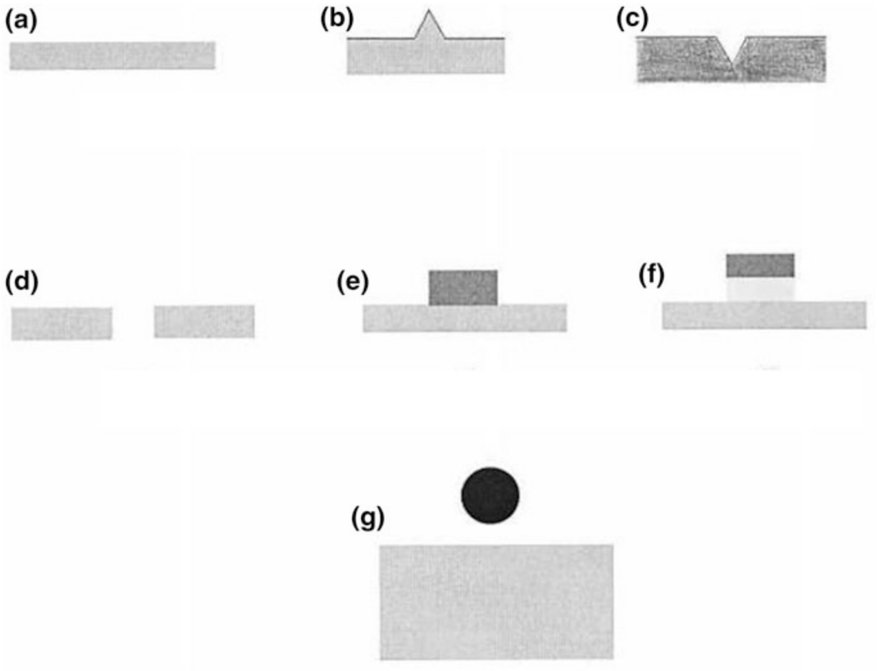


Fig. 4.25 A schematic of a number of different waveguide geometries that have been studied. In all of the figures the waveguide channel is perpendicular to the page. Waveguides shown include: **a** long range surface plasmon-polariton guided mode, **b** the metal wedge waveguide, **c** the V-groove channel waveguide, **d** the plasmonic slot waveguide (also illustrating the double-strip waveguide), **e** the dielectric loaded surface plasmon waveguide, **f** the hybrid plasmonic waveguide, and **g** the nano-wire waveguide[[41–46]

modes are bound to the features centered in each of the plots. The fields of the guided modes are concentrated in a region of finite radius within the plane of the page. Perpendicular to the page the guided modes are extended through the waveguide channel.

The geometry of the guide in Fig. 4.25a is for a long range surface plasmon-polariton wave guide. This is composed of a thin metal film which is surrounded by a dielectric. Due to the two parallel horizontal surfaces of the film and the reflection symmetry of the system about the horizontal center plane of the film, the guided plasmon-polaritons modes separate into guided modes that are symmetric and anti-symmetric in the direction normal to the metal film [41–52].

The symmetric mode is the mode of interest in applications. As the film is decreased in thickness the range of propagation of the symmetric mode increases so that the range of propagation can be tuned in this manner. However, with decreasing thickness the radius of the region of confinement in the plane of the page increases. Consequently, a balance is needed for the operation in this design between confinement and propagation length Due to the need to balance these opposing effects, the guide geometry in Fig. 4.25a is not always a good solution to the problem of waveguide design.

The channel designs in Fig. 4.25b, c are, respectively, metal wedge and V-groove channel waveguides [41–46]. In these two designs the lower medium is metal (e.g., silver) and the upper media is dielectric (e.g., SiO_2). The dielectric mismatch of the two media supports plasmon-polariton, with the wedge and grooves on the interfaces creating a set of localizing guided plasmon-polariton modes bound about them.

The wedge waveguide confines the energy of the guided modes moving into the page in Fig. 4.25b to locate about the channel length in the region surrounding the wedge. The localization about the wedge is determined by the angle at the apex of the wedge so that decreasing the apex angle increased the confinement about the wedge.

For the groove waveguide the guided modes that exist localized within and about the channel are referred to as channel plasmon-polaritons [41–52]. The location of the guided mode energy within the channel depends on the wavelength of the mode for its propagation along the channel. It is, in general, found that the depth of the channel groove should not be considerably less that the penetration depth of the fundamental guide mode the channel supports. This follows as part of the consequences of the positioning of the channel modes and their propagation distances along the channel being dependent on the apex angel of the groove.

Figure 4.25d illustrates the plasmonic slot waveguide composed as a metal plane with a slot channel into which a dielectric is introduced. In addition, the media surrounding the metal plane is also taken to be dielectric. The left and right features in the figure illustrate the slot in the metal plane within which guided plasmon-polariton modes propagate. The slot width can be much less than the wavelength of the guided mode confined within it, and the dielectric medium in the slot can be different than the dielectrics that surround the metal plane. In general, the system is designed so as to support fundamental guided modes which are highly confined within the slot [41–52].

A variant of the slot waveguide is the double-strip waveguide. In this design two thin metallic strips are deposited on a dielectric substrate. The two strips support guided modes propagating along their strip lengths. The surface charges generated by the guided modes as they move long the channel are found to be 180° out of phases.

In Fig. 4.25e the geometry is for a dielectric loaded surface plasmon-polariton waveguide. This waveguide is composed of a high refractive indexed (e.g., Si) which is deposited as a strip on a metal surface (e.g., Ag). In the figure the metal is at the bottom of the figure and the strip is the dark feature on the metal. The metal surface and strip are then surrounded in the upper region of the figure by a low index medium (e.g., SiO_2). The guided mode is limited to the surface region by metal low refractive media interface, and the width of the region about the strip in which the fields are confined is set by the size of the high refractive media strip within the low index cladding medium. The guided mode propagations perpendicular to the page. In this design the functioning of the system is somewhat similar to that found in cladded dielectric waveguide used in fiber optics [41–52].

Figure 4.25f illustrates the hybrid plasmonic waveguide. It is somewhat similar to the dielectric loaded surface plasmon-polariton waveguide but represents a general improvement of the design. Here a metal (e.g., Ag) occupies the lower region of the figure and the upper region is a low dielectric index material (e.g., SiO_2). The dark region is a high index (e.g., Si) strip of material and below the strip and the metal surface is another strip of a low refractive index medium. The guided modes of the high refractive index medium interacts with the surface plasmon-polariton modes of the metal surface to form hybrid guided modes that travel perpendicular to the page along the strip channels.

In general hybrid waveguides have been found to provide a combination of the propagation characteristics of cladded fiber optics waveguides and the confining characteristics of plasmon-polariton waveguides effectively to provide longer range between confinement than found in some other waveguide designs.

A feature of these waveguides is that they can confine guided mode of TE polarization within the high index medium and TM polarization within the low index medium. This then offers added capacity for treating signals of the two polarization differently within the plasmonic circuitry. Both polarization can be treated differently within the circuit through the applications of polarization filters. Such applications will be considered later.

Related to the hybrid plasmonic waveguides are nano-wire plasmonic guides. These are illustrated in Fig. 4.25g. The metallic nano-wire is place on or near the surface of the planar metal dielectric interface. Fast and slow modes appear, respectively, near the top and bottom of the nano-wire. An isolated wire will display a helical guided mode with a period that is related to the radius of the wire.

B. Plasmonic Devices for Circuit Applications

In the following some discussions are presented of plasmonic devices that may be connected together with plasmonic waveguides so as to form circuits. These are the mechanisms that are used to modify or initiate the information carried throughout plasmonic circuits to eventually be outputted to the world outside of the circuitry. In addition, the inputting and outputting devices themselves are important features in the development of plasmonic circuits and will also be considered.

Types of devices needed for circuit designs include [41–52]: Plasmonic lasers which act as light sources or signal amplifiers; surface plasmon-polariton logic gates for processing inputted logical signals into logical gate outputs in computer applications; sensor devices that modulate signals initialed or flowing through them as they interacted with externally applied fields, chemical or biological agents, or externally generated heat; surface plasmon-polariton multiplexers and routers for the transfer of signals into different channels of branched waveguide systems; plasmonic switches; and plasmonic interfaces with electronic circuits. Recently devices of all these types have been the focus of a great amount of research effort by a wide range of plasmonic research groups. Some of these basic components of circuit design are now briefly discussed, but no attempt is made to give a comprehensive review of the field. For such a review the reader is referred to the literature.

Couplers

To input energy into plasmonic waveguides, light from the outside world must either be directly coupled into a waveguide or there must be some type of plasmonic laser designed to introduce a signal into a waveguide. This requires a means of mediating the interaction between light from the bulk or from a plasmonic laser with the plasmonic guided modes of the waveguide. In both instances problems involving the dielectric mismatch and interface geometry between the waveguide and the bulk or lasing media are a difficult problem [1–6, 41–46].

As discussed in earlier sections of this chapter, at a flat metal-dielectric interface the dispersion relations of bulk light and surface plasmon-polaritons are distinct from one another. From those discussions it was found that it was the translational symmetry of the surface that allows the bulk and surface modes to exist independent of one another. In order to introduce bulk light into the plasmon-polariton system then requires some type of disruption of the translational symmetry at or near the surface of the metal dielectric surfaces or at the waveguide channel. In general, the disruption of the translational symmetry of the surface or waveguide is usually accomplished either by using prism coupling or the by the introduction of surface features which destroy the translation symmetry of the plasmonic surface or waveguide.

Prism coupling involves placing one of the flat surfaces of a prism on (Kretschmann geometry) or near (Otto geometry) the plasmon-polariton supporting surface [1–6] (see the schematics in Fig. 4.26 for an illustration of these two geometries). In each of these configurations light is sent into the prism so that it is internally reflected by the prism surface near the plasmon-polariton supporting surface. This internal reflection in the case of the Kretschmann geometry creates an evanescent wave in the metal of the supporting surface.

In the case of the Otto geometry the evanescent wave is created both in the region between the prism and the metal of the supporting plasmon-polariton surface and also within the supporting metal itself [1–6]. In both cases the evanescent wave

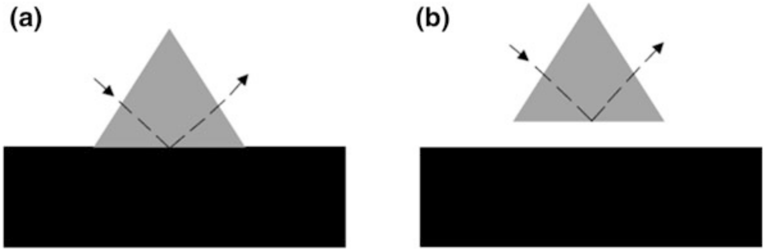


Fig. 4.26 Illustration of the prismatic coupling of bulk light to excite surface plasmon-polaritons at a dielectric-metal interface. The configurations shown are: **a** the Kretschmann geometry and **b** the Otto geometry

is used to interact with the supporting surface so as to create a plasmon-polariton wave along the supporting surface or in a waveguide constructed on the surface. Prismatic coupling, however, can be highly ineffective for the introduction of light into a nano-scale systems and other nano-scale means have recently been sought [41–46].

A signal can also be introduced into a plasmonic waveguide through the application of light generated by a plasmonic laser [53–56]. These lasers function as nanoscale devices and are often based on nano-wire designs. In the nano-wire design a segment of semiconductor nano-wire is used as a cavity resonator and also as a source of gain. A high gain material used to develop nano-wire lasers is based on semiconductors (e.g., CdS) and the nano-wire itself is generally separated, by a spacer of material with a low index of refraction (e.g., MgF_2), from the metal (e.g., Au) surface which is part of the surface plasmon support. A number of devices have been fashioned based on this design, but problems remain involving the efficient coupling of the laser output into plasmonic waveguides. These follow from the dielectric mismatch at the laser-waveguide interface and the mismatch of the geometric properties between the laser and the waveguide into which it couples. Methods to improve the efficiency of the coupling of the plasmonic laser into plasmonic waveguide is currently and focus of attention.

Recently, in the nano-wire design some success in the formulation of tunable lasers has been achieved [41–56]. These types of systems are based on the use of semi-conducting lasing media of the form $In_xGa_{1-x}N$ for the nano-wire cavity. In the $In_xGa_{1-x}N$ based lasers the outputted light can be tuned through the optical spectrum form blue to red.

Spaser

One of the important ideas that has been developed related to the design of nano-lasers is the spaser. The spaser uses stimulated emission to generate intense, coherent, surface plasmonic excitations. The functioning of the spaser is closely related to that of the laser, but whereas the laser creates an intense coherent beam of photons the spaser creates an intense coherent beam of surface plasmon-polaritons [49–56].

The principles of operation of the laser and the spaser are similar. This is due to the similarity in the physical properties of the photon and the surface plasmon-polariton excitations considered in each of the two systems. In particular, both the laser and the spaser deal with the amplification of electrically neutral spin one bosons, and, the plasmonic excitations, similar to photons, are weakly interacting with one another so that the modes of the system are highly linear.

In their operations both the laser and the spaser involve a type of resonant cavity and an externally excited gain medium. The external exciting fields in each case can be at different frequencies from the output modes that are generated.

The original idea for the spaser is illustrated in Fig. 4.27a. A type of resonant cavity is formed by a spherical nano-particle consisting of a dielectric core with an outer layer of silver. (Note that the composition of the spherical nano-particle could be revered, with the inner core formed of silver and the outer shell of dielectric.)

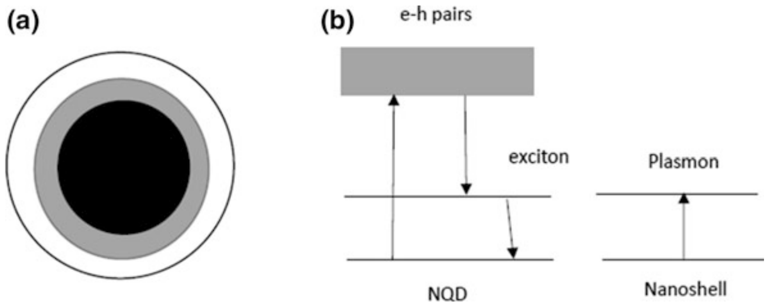


Fig. 4.27 The basic design of the spaser. A nano-particle based spaser is shown in (**a**). The black is a region of dielectric. The grey is a metal shell and the white outer shell is a region of nano-quantum-dots. The energy level diagram for spaser operation is shown in (**b**). The left of the figure **b** represents the nano-quantum—dot transitions in which an electron-hole gas is pumped from outside the system. The electrons and holes in the gas combine to form excitons. The excitons preferentially decay, giving up their energy to plasmons formed at the metal dielectric interface

The nanoparticle acts as a sort of resonant cavity for enhancing the surface plasmon-polariton fields that are generated in the surface plasmon-polariton modes on the nano-particle. The coherent enhancement of these cavity modes arises as a result of pumping provided by an externally applied bulk electromagnetic field [53–56].

To accomplish the pumping, in the spaser design, the nano-particle is surrounded by quantum dots. These are used to pump the system in order to generate the intense surface plasmon-polaiton modes attached to the nano-particle cavity. For this pumping an external electromagnetic field is sent into the system to excite the quantum dots, generating in them a gas of electron-hole pairs. The electrons and holes of the gas then combine to form excitons.

In the absence of the nano-particles the excitons on the quantum dotes would normally decay by emitting a bulk photon to propagate out of the system. If, however, the frequency of the surface plasmon-polaiton modes of the nano-particle is chosen to be the same as that of the bulk photon arising from the exciton decay a quite different process takes place. This is a crucial point of the spaser operation.

In the presence of the nano-particles the excitation energy of the exciton can be preferentially transferred to the surface plasmon-polariton modes of the nano-particle. This occurs due to the difference in the transition rates for the generation of bulk modes and the generation of cavity modes. It leads to an amplification of the surface plasmon-polariton modes on the nano-particle. The end result of the spaser operation, then, is the generation by stimulated emission of the surface plasmon-polaiton modes of the spaser (see Fig. 4.27b for a schematic summary of all of the processes outlined above for the spaser operation).

The spaser mechanism is quite useful in particular for the amplification of the so called ‘dark’ surface plasmon-polariton modes. These are surface plasmon-polariton modes that weakly couple to bulk electromagnetic modes. Due to their weak interactions with bulk electromagnetic modes, the dark modes are difficult to excite by other non-spaser based methods of coupling into them. In addition, the dark modes are very stable to decay into the bulk modes and this limits the radiative loss mechanism [49–56].

The resulting amplification of the surface plasmon-polariton modes allows for the generation of intense electric fields that are then available for Raman spectroscopy, flourescence imagining, and other chemical and biological investigations. In addition, due to their plasmonic nature they are subwavelength, focused, fields. This is an asset in the localized applications of the amplified fields as well as for plasmonic circuit applications.

It is seen that whereas the laser amplifies outputted photonic modes, the spaser amplifies outputted surface plasmon-polariton modes. The outputted modes of the spaser, however, can be converted to photonic laser modes by the introduction of a symmetry breaking to the nano-particles. Symmetry breaking provides for the conversion of the amplified surface plasmon-polariton modes into radiated bulk optical modes by inducing transitions from the surface plasmon-polaritons to the bulk electromagnetic modes. The spaser then acts as a nano-laser.

Other Nano-laser/spaser Configurations

A number of different types of nano-lasers have been engineered based on these and related ideas from laser technology. One type of laser design is based on the so called Metal-Insulator-Metal waveguide cavity [49–55]. As the name suggests this is basically a tri-layered system composed of an insulator layered between two metal plates, and it produces laser fields that are confined one-dimensionally. The metal-insulator interfaces support surface plasmon-polaritons which travel along the metal-insulator interfaces.

Consequently, there is a confinement of the surface wave modes in the direction normal to the metal insulator interfaces. As the plasmon-polariton modes propagate along the interfaces they are, in addition, confined by a partial reflective coating that is applied at the end of the layering and is positioned to be perpendicular to the direction of the surface mode propagation. The resulting system then forms a Fabry-Perot cavity.

In this arrangement, in addition to its function in forming a surface plasmon-polariton supporting interface between the metal plates, the insulating medium also acts as a gain medium which is pumped by external electromagnetic modes applied to the system. The surface plasmon-polaritons eventually arise from the pumped fields created within the cavity.

This is a rough description of the basic operational mechanism of the Metal-Insulator-Metal based laser. It is, however, a bit of an over simplification as a number of considerations and extra feature are involved in guiding and generating and pumping the surface plasmon-polariton lased within the cavity. For the details of this the reader is referred to the literature.

Related to the Metal-Insulator-Metal laser is the whispering gallery cavity based laser [53–55]. Here the surface plasmon-polariton exhibit a type of one-dimensional propagation as they are steered around a semiconductor disk which is coated by metal.

An example of a nano-laser involving two-dimensional confinement of the lasing modes was given earlier [49–56]. As mentioned at that time, another important laser structure is the hybrid Metal-Insulator Semiconductor laser composed as a nano-wire segment and a metal planar surface between which is a magnesium fluoride spacer. Like the Metal-Insulator-Metal laser this involves a gain medium, but in this case the confinement is in two dimensions with the fields confined within a two-dimensional region localized about the nano-wire of the system.

An example of a nano-laser having a full three-dimensional confinement of the radiation is the nano-particle laser, again related to the nano-particle system treated earlier [49–53]. Some recent fabrications of these types of systems have been made based on gold cores surrounded by sodium silicate shells which are in turn encased in a dye doped silica shell. The dye doped silica provides the gain medium which is pumped for the laser action. The modes are confined within the sphere of the nano-particle.

Other Photonic Nano-devices

Another important development in plasmonic devices are surface plasmon polariton logic gates [41–52]. These are based on the formulation of devices based on nano-wires, branchings of nano-wires, and the coupling of ring resonator waveguides. In all of such gates, logical inputs to the gate are represented by configurations of optical signals sent to the gate device. In turn a configuration of outputted optical signals is generated from the interaction of the gate with the inputted light and represents an outputted logical response to the logical input.

Logical gates are needed in the design of digital circuits for computers and have become particular important to nano-science for their use in quantum computing applications [41]. A large number of quantum systems have been investigated as possible sources of designs of quantum computers, and plasmonic is just one of many systems currently under investigation for such applications.

Recently some work has been done the development of circuitry that would act as AND, OR, and NOT gates [41]. In one formulation the polarizations of the guided plasmon-polariton modes were used to construct AND and OR logic gates [41–52]. In this approach, two polarized guided waves were inputted into a nano-wire waveguide with the two different logical inputs (i.e., 0 or 1) represented by the polarization of the guided waves. The result of the operation of the AND or OR logical gates was then indicated by the threshold intensity of the combined waves exiting the nano-wire. An illustration of such a combining of polarized signals within a nano-wire waveguide is show in Fig. 4.28a.

In another approach to the design of logic gates a branched system of interconnecting waveguides is used [41–52]. A simple illustration of this is given in Fig. 4.28b. In the Y-type structure shown in the figure two input waveguides meet a third output waveguide at a vertex common to all three. In the figure signals are inputted into the two waveguides labelled I1 and I2. These two input signals

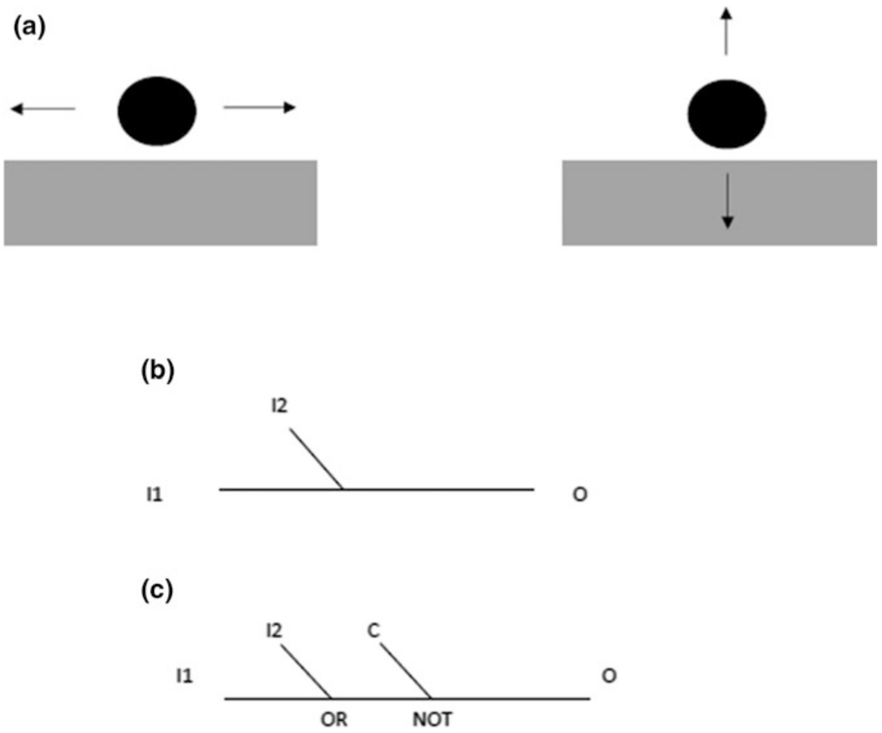


Fig. 4.28 Schematic plots of: **a** nano-wire plasmonic waveguides where the arrows indicate the two polarization of guided surface plasmonic waves along the nano-wire (Indicated in black.), **b** the three plasmonic waveguides forming an AND or OR logic gate where I1 and I2 are the input channels and the outputted phase added wave exits the logic gate at O, and **c** the NOR gate with inputs I1 and I2, control C, and the outputted signal at O

combine in the third wave guide O where they are outputted as a logical signal. By changing the relative phase of the two input signals, the intensity of the output signal is modulated. In particular, the output signal can be change from a signal maximum to a signal zero, providing an on off switching response. Continuing along these lines, in Fig. 4.28c a circuit has been designed by adding a fourth waveguide representing a control laser input. The resulting circuit can then be made to operate as a logical NOR gate.

A third approach involves a similar idea but it is based on the coupling between ring resonator waveguides which can be operated in a manner so as to act as a NOT gate. For a further treatment of these the reader is referred to the literature [41–52].

Other developments have involved the design of modulators which can change the flow of plasmonic guided modes that flow through them as they interact with outside stimuli [41–52]. An example is an electro-optical modulator. This is based on a type of hybrid waveguide structure shown in Fig. 4.25f. The dark layer is chosen to be an n-type silicone and the lower metal surface is silver. Between the silicone and silver a layer of dielectric with a strong Kerr nonlinearity is introduced forming a capacitor. The system is designed so that the guided mode field is contained within the Kerr dielectric. Modulation of the dielectric properties of the Kerr media can then be used to modulate the guided mode flow through modulating device.

In terms of the design of plasmonic circuits hybrid nanoplasmonic waveguides have displayed certain advantages over some of the other waveguide designs [41–52]. They provide for a tight confinement of the guided mode to the channel and also allow for ultra-sharp bending of the waveguide channel. Some circuit designs based on hybrid channels are the submicron-donut resonator which offers a resonant interaction with the modes of a straight waveguide channel. The donut resonator is a circular channeled feature (e.g., of radius 800 nm) which couples with the modes of a straight waveguide through its proximity to the waveguide channel (see Fig. 4.29a). Another design is a power splitter which redirects the energy flow in a single waveguide into two waveguides. An illustration of such a splitter in the Y-splitter shown in Fig. 4.29b. A number of different type of coupler have been developed based on similar types of designs to the Y-splitter including some directional couplers.

Couplers formed between two different waveguides which are distinct from one another have also been developed [41–52]. In these devices two different waveguides can be in close proximity to one another over a common length (see Fig. 4.29c). Along the region of close proximity the modes of the two waveguide can couple to one another through a weak interaction. Depending on the length of the region of proximity, a guided mode launched within one of the waveguides can be transferred by the weak coupling to travel in the other waveguide of the coupled pair. The transfer effect is also found to be dependent on the polarization of the modes launched into the system. This facilitates their use in the design of polarization beam splitters which allow plasmonic circuits to handle the propagation of TE and TM guided modes differently.

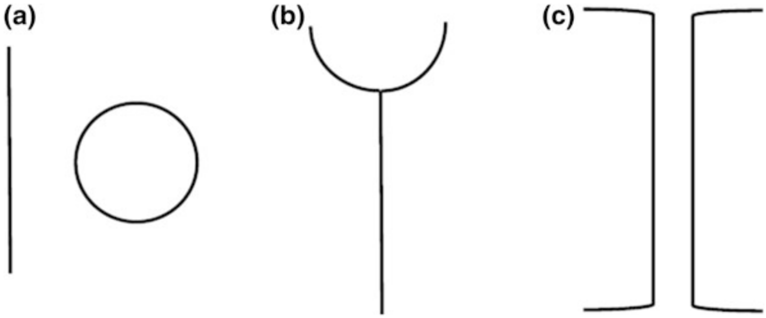


Fig. 4.29 Illustration of the waveguide channel designs looking down on the plane of the metal-dielectric interface which is the plane of the page. In these schematics: **a** represents the donut resonator interacting off-channel with a log straight waveguide, **b** a Y-splitter which sends energy in one channel to propagate within two other waveguide channels, and **c** a waveguide coupler formed by placing two waveguides with channel lengths in close proximity

Related to the coupler are guided mode polarizers [41–52]. Polarizing structures which preferentially allow one waveguide polarization to continue down a waveguide while reflecting the other polarization can be by made by introducing a region of grating into the waveguide. In this way polarizers which only pass either TE or TM modes have been fashioned. These structures allow for the design of plasmonic circuits that treat TE and TM components of inputted signals differently.

Another important application in plasmonic circuitry arises due to the sensitivity of the surface plasmon resonance effect to small changes within the parameters characterizing the plasmonic waveguides [41–52]. These sensitivities from the basis of the applications of plasmonic systems as heat, optical, chemical, and biological sensors. In addition, in the case of heat applications the sensitivity can be used to tune the properties of the circuitry introduced into device design.

As a final important circuitry feature that will be mentioned here, surface plasmon-polaritons are excitations that combine both an electromagnetic and an electronics component and this dual nature of the exciations facilitates the interaction of plasmonic signals with electronic systems [41–52]. In particular, the propagation of plasmonic modes along nano-wire wave guides can be used to induce interactions with electronic components. As an example, in some recent work electronic transistors have been designed in which plasmonic nano-wire waveguides form part of the base-collector-gate design of the transistor [41–52]. Plasmonic signals traveling along the nano-wire then effect the signal in the transistor and act as a signal connection between the plasmonic circuit and the electronic circuit of which they are a part.

4.3.3 Plasmonic Subwavelength Enhanced Transmission of Light

Another important feature of the subwavelength nature of the surface plasmon-polariton excitations is found in the enhanced transmission of radiation transmitted through a perforated metal film which supports surface plasmon-polaritons [57–63]. This is illustrated in Fig. 4.30 where light is normal incident from above a thin metal film that has a periodic patterning of holes penetrating the slab. Under certain conditions on the patterning, the fraction of the incident light transmitted through the slab into the region below the thin film is larger than expected from standard considerations of classical optics.

The phenomenon is termed extraordinary optical transmission and occurs for a periodic patterning of subwavelength apertures penetrating the metal film. A primary contributing mechanism in the transmission enhancement is believed to be the excitation of surface plasmon-polaritons on the metal film. These assist in the propagation of energy through the penetrating features of the metal film.

Before considering the extraordinary optical transmission from the array of holes, it is useful to review the results of diffraction from a thin film containing a

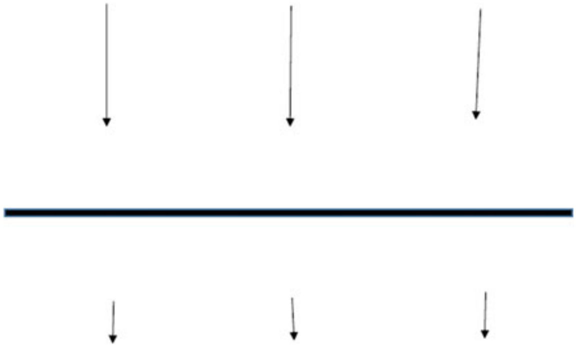


Fig. 4.30 Enhanced transmission of light through a thin metal film that has a periodic patterning of holes penetrating the slab. Light is normal incident on the film from above and partially transmitted below the thin film. The transmission is assisted by surface plasmon-polaritons excited on the thin metal film

penetrating single hole [63]. The study of the single holed systems provides insights into the nature of the physical processes entering into the extraordinary optical transmission phenomena and gives an idea of the significance of surface electromagnetic waves in the later discussions. Consequently, as a useful preliminary, the very basic system of a single hole in a film which does not support surface plasmon-polariton excitations will be discussed [63].

Following these remarks the enhancements arising in a system of periodically patterned holes on a surface plasmon-polariton supporting film will be presented and discussed in comparison with the results of the single holed film [57–62]. This allows for the direct observation of the importance of surface plasmon-polaritons in the phenomena while offering an indication of the extent of the transmission enhancement observed in extraordinary optical transmission. Specifically, the results will indicate how large the transmission effects are when compared to those found in the problem of a single hole.

Bethe considered the transmission problem in Fig. 4.30 for a thin film with a single aperture, treating the case in which the film is a perfect conductor [63]. In particular, due to the perfect conductivity, the film in this case does not support surface plasmon-polariton excitations so that these excitations do not enter into the problem. The film then only acts as a boundary condition to the waves propagating in the region outside the film, providing a limit keeping them from entering into the volume of the thin film.

The diffraction from a circular aperture is shown in Bethe's theory to be exactly solvable. As a result which is of importance to the discussions of enhanced transmission, a single subwavelength hole of radius, r, is shown to have a transmission efficiency that scales as $(r/\lambda)^4$ where λ is the wavelength of the incident radiation. Consequently, for optical radiation incident on a hole of radius of order of 150 nm, a transmission efficiency is obtain which is of order 10^{-3}. As seen in the later discussions, this is less by at least three orders of magnitude compared to experimentally determined transmission efficiencies of the extraordinary transmission results from a periodic patterns of such subwavelength holes.

In addition, the radiation transmitted through the hole is diffusely transmitted with an intensity distribution that is essentially uniform in the transmission angle below the thin film. In the case of periodic patterning the periodicity of the transmitted sources of radiation (i.e., the diffraction of the system) tends to project the light normally outward from the thin film [57–62]. This phase coherence, however, does not account for all of the enhancement observed in the light transmitted though the film. It is only part of the mechanism accounting for the three orders of magnitude difference mentioned earlier.

The case of a periodic patterning of subwavelength holes presents new feature which tend to enhance the transmission efficiency of the thin films. In general for radiation that is incident normal to the metal film, the ratio of the intensity of radiation transmitted through the area of the hole or, in the case of the periodic pattern, the holes in the thin film divided by the intensity of radiation incident on the area of the hole or holes measures the efficiency of the transmission through the thin film.

For films which do not support surface plasmon-polaritons the efficiency of transmission is found to be larger for a periodic pattering of holes than for a single hole. However, the efficiency of transmission is generally found to be much greater for periodic patterns in thin film which support surface plasmon-polariton than on systems of the same pattering which do not support surface plasmon-polaritons. This shows that the surface plasmon-polaritons are a fundamentally important feature in understanding the nature of extraordinary optical transmission.

Experiment

An early experiment that demonstrated the extraordinary optical transmission was performed by Ebbesen et al. [61]. In their experiment they considered a thin Ag film deposited on a quartz substrate. In a first study, the transmission of normal incident light through an Ag film with a thickness of $t = 200$ nm was measured. The Ag film had a pattern of subwavelength holes of diameter $d = 150$ nm arranged on a square lattice with a lattice constant $a_0 = 900$ nm. In Fig. 4.31 the results from their paper are presented [61].

In Fig. 4.31 the narrow peak at $\lambda = 326$ nm is from the excitation of a bulk plasmon in the silver. It is found to decrease as the thickness of the film is increased. A minimum in the transmission intensity versus wavelength is found at a_0 and for $\lambda > a_0\sqrt{\varepsilon}$ where ε is the dielectric constant of the quartz substrate there is no diffraction through the thin film either for a single hole or for the periodic array. Between $a_0 < \lambda < a_0\sqrt{\varepsilon}$ there are two extraordinary transmission peak maxima. These arise for the presence of surface plasmon-polaritons on the thin Ag film.

The largest transmission peak in Fig. 4.31 occurs at $\lambda = 1370$ nm. For this peak the transmission efficiency is found to be greater than two so that there is twice as much light transmitted through the thin film than is incident on the area of the periodic apertures of the thin film. This difference in the transmission efficiency from that of a film which does not support surface plasmon-polaritons, as shall now

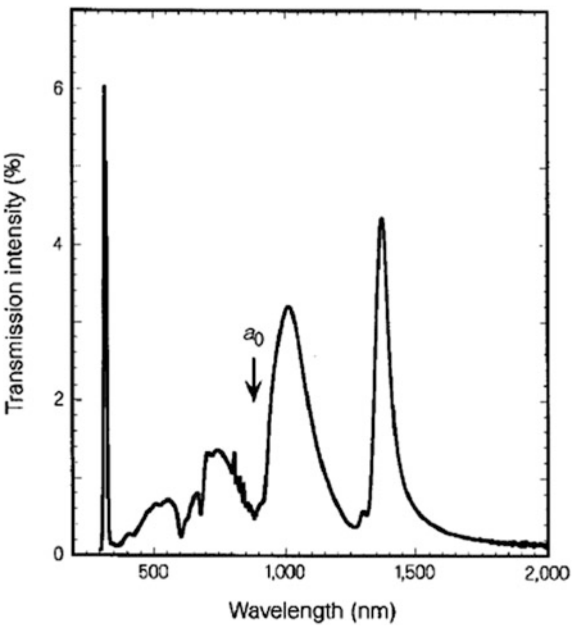


Fig. 4.31 Transmission intensity through a Ag thin film versus the wavelength of light. The thickness of the film is $t = 200$ nm, the holes of the film are cylindrical of diameter $d = 150$ nm, and the holes are arranged in a square lattice with the lattice constant $a_0 = 900$ nm. The result is for the zeroth order transmission [61]. Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Nature 391, 667), copyright (1998)

be seen from the experimental studies and later shown in theory, comes from the presence of the surface plasmon-polariton excitations as part of the transmission mechanism.

The importance of surface plasmon-polariton as the mechanism of the transmission enhancement in the results in Fig. 4.31 is evidenced experimentally by a further series of studies in which the transmission effect is determined as a function of the film parameters as well as the materials supporting the excitations. Some of the experimental studies of these factors are briefly reviewed in the following. This is followed by the presentation of a simple theory which indicates the origin of the transmission effects as being due to the surface plasmon-polaritons propagating along the thin film.

In Fig. 4.32 experimental results for the Transmission Intensity versus λ/a_0 are presented for a series of films formed of different metals deposited on quartz substrates. In these plots the film thicknesses, hole diameters, and lattice constants of the array of holes are different for each material that is investigated. It is seen, however, that in the presentation in the scaled variable λ/a_0, the curves for the different films have a similar system of transmission peaks and minima. In addition, the enhancement peaks are observed to scale precisely with the period of the square lattice array.

This scaling is to be expected in the case that the transmission effect is mediated by the surface plasmon-polaritons of the thin films and is a strong evidence that this, indeed, is the case. In particular, the surface plasmon-polaritons are the only waves propagating in the two-dimensional array of the surface which have a dispersion

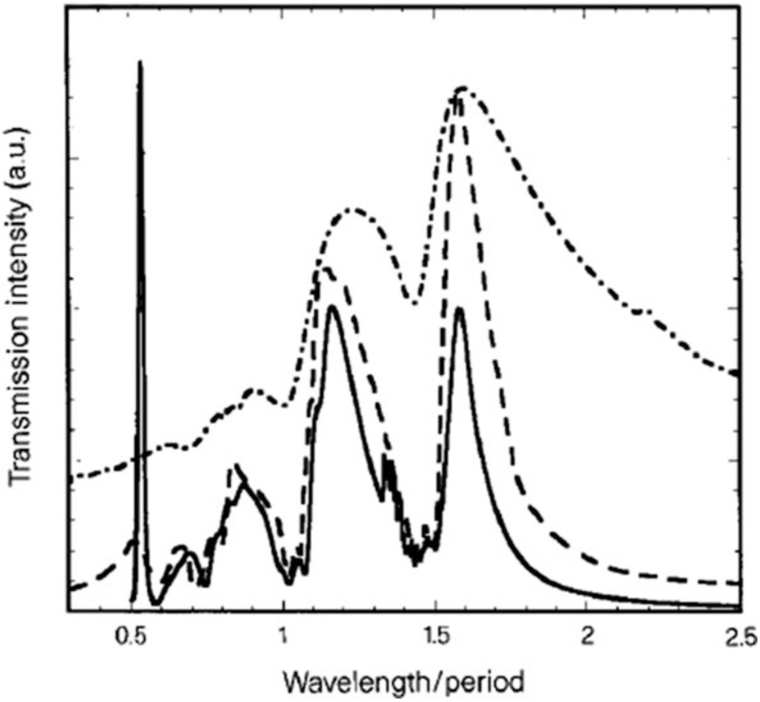


Fig. 4.32 Measured transmission intensity versus λ/a_0 for various square lattice arrays. Results are for: (solid) Ag with $a_0 = 600$, $d = 150$, $t = 200$ nm; (dashed) Au with $a_0 = 1000$ nm, $d = 350$ nm, $t = 300$ nm; (dashed-doted) Cu with $a_0 = 1000$ nm, $d = 500$ nm, $t = 100$ nm [61]. Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Nature 391, 667), copyright (1998)

relation depending on the ratio of their wavelength to the period of the square lattice array of holes in the thin film. Consequently, results dependent on surface plasmon-polaritons should exhibit this type of scaling behavior.

A feature of the transmission maxima in Fig. 4.32 which does exhibit a significant change from curve to curve in the plot is the difference in peak width of the various curves. The peak width of each of the curves is found to depend strongly on the ratio t/d. In Fig. 4.32, the width of the plotted curves is greatest for the case $t/d = 0.2$ and decreases to the sharpest peaks at $t/d = 1$. This is in accord with the general observation that the larger the diameter of the holes in the film compared to the film thickness the less sharp as a function of wavelength are the transmission peaks. It again indicates a strong dependence of the transmission effect on the wavelength of the surface electromagnetic waves.

An important feature in the transmission plots arising from the bulk Ag plasmons in the thin film is a transmission peak at the far left hand side of the Fig. 4.32 [61]. The intensity of this peak is found to decrease and eventually disappear with

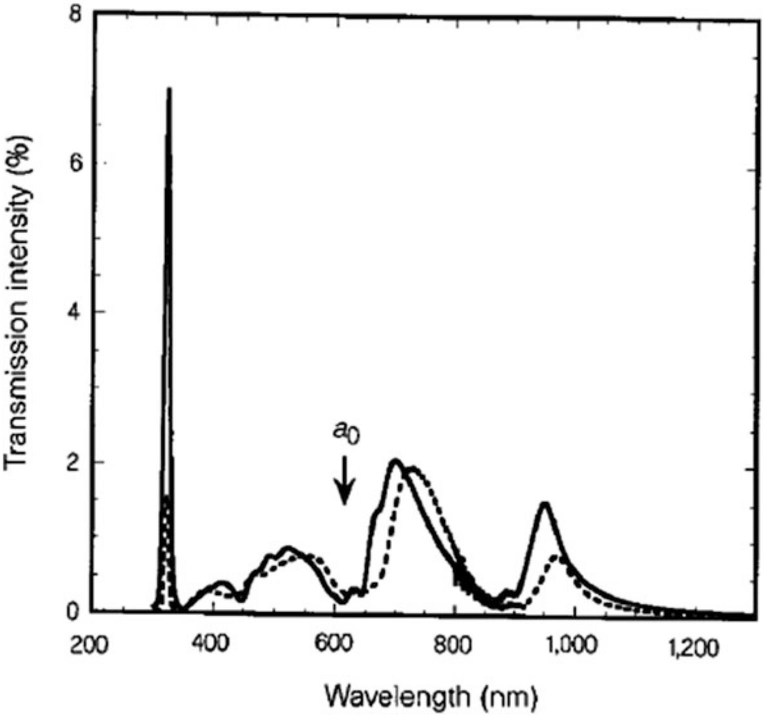


Fig. 4.33 Transmission versus wavelength for two Ag arrays. Both arrays have $a_0 = 600$, $d = 150$ nm. One array has $t = 200$ nm (solid) and the other has $t = 500$ nm (dashed). The dashed spectrum has been multiplied by 1.75 for the comparison [61]. Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Nature 391, 667), copyright (1998)

increasing slab thickness of the samples. The decrease in this feature is much more pronounced than that associated with the transmission features arising from the surface plasmon-polaritons and is evidence of the difference in the mechanisms it involves.

As a final experimental study [61], Fig. 4.33 presents some results for the transmission versus wavelength for two Ag arrays with identical square lattices of identical holes but for different film thicknesses. Both arrays have $a_0 = 600$, $d = 150$ nm. One array, however, has $t = 200$ nm (Solid) and the other has $t = 500$ nm (Dashed). For the comparison of the transmission results of the two arrays, the dashed spectrum has been multiplied by 1.75.

Again, the left most peaks from bulk Ag plasmons are seen to decrease significantly as the film thickness is increased. The other extraordinary optical transmission maxima at higher wavelengths decrease less quickly with increasing film thickness, exhibiting an approximately linearly decrease with the increase in film thickness. These last maxima arise from the surface plasmon-polaiton mechanism which includes propagation mediated by the subwavelength holes in the film. This mediation accounts for the lower sensitivity of these transmission maxima to the change in film thickness [61].

Simple Theory

The above experimental results for the extraordinary optical transmission maxima can be understood by a simple analytical theory based on a coupled mode approach [60]. This approach gives a rough treatment of the basic surface plasmon-polariton mechanism and provides an estimate of the gross features of the effect. It does not give a complete treatment such as would be handled, for example, in a computer simulation study. Nevertheless, the analytical treatment discussed provides a deeper understanding of the relationship of the phenomena to the surface plasmon-polaritons of the thin film than that provided by computer simulation studies.

In the following the coupled mode theory for extraordinary optical transmission at normal incidence through a thin metal film is presented [60]. In the presentation, following a description of the transmission problem to be treated, some general remarks about the nature of the coupled mode approach are made. It is shown that the coupled mode approach is a general method which is widely used to develop an understanding of the properties of many-body problems considered in physics. These types of problems include those found in phonon, electron, photon, and fluid systems. This is then followed by the application of the formalism of the coupled mode method to explain the surface plasmon-polariton treatment of the extraordinary optical transmission maxima. The explanation is made for the specific transmission problem earlier formulated for study. The section concludes with some references to computer simulation studies which have generated more precise treatments of extraordinary optical transmission problems.

The system that is considered for study is a metal film of thickness, t, that is patterned by a square lattice array of subwavelength holes [60] (see Fig. 4.34 for a schematic diagram). The surfaces of the thin film are both parallel to the x-y plane, and the axes of the subwavelength holes are in the direction of the z axis. The lattice constant of the square lattice array is a, and the square holes in the film have edges of length l that are aligned along the x and y directions. The metal film is surrounded by vacuum and the holes penetrating the film contain vacuum.

For this scattering system, light of wavelength $\lambda > l$ is incident on the film, traveling in the negative z direction. The light interacts with the film, having both a reflected and transmitted component. The focus on the treatment given later is on determining the transmission of electromagnetic fields through the film for transmitted light propagating in the negative z direction.

The electromagnetic fields in the above outline transmission problem can be thought of as being separated into the incident and transmitted waves propagating along the z direction, surface plasmon-polariton modes propagating parallel to the metal-vacuum interfaces, and the surface electromagnetic modes moving along the z directions of the hole surfaces [60]. Each of these components interact with one another through coupled transitions which move the electromagnetic fields through the thin film in a sequence of exchanges between the modes of the system.

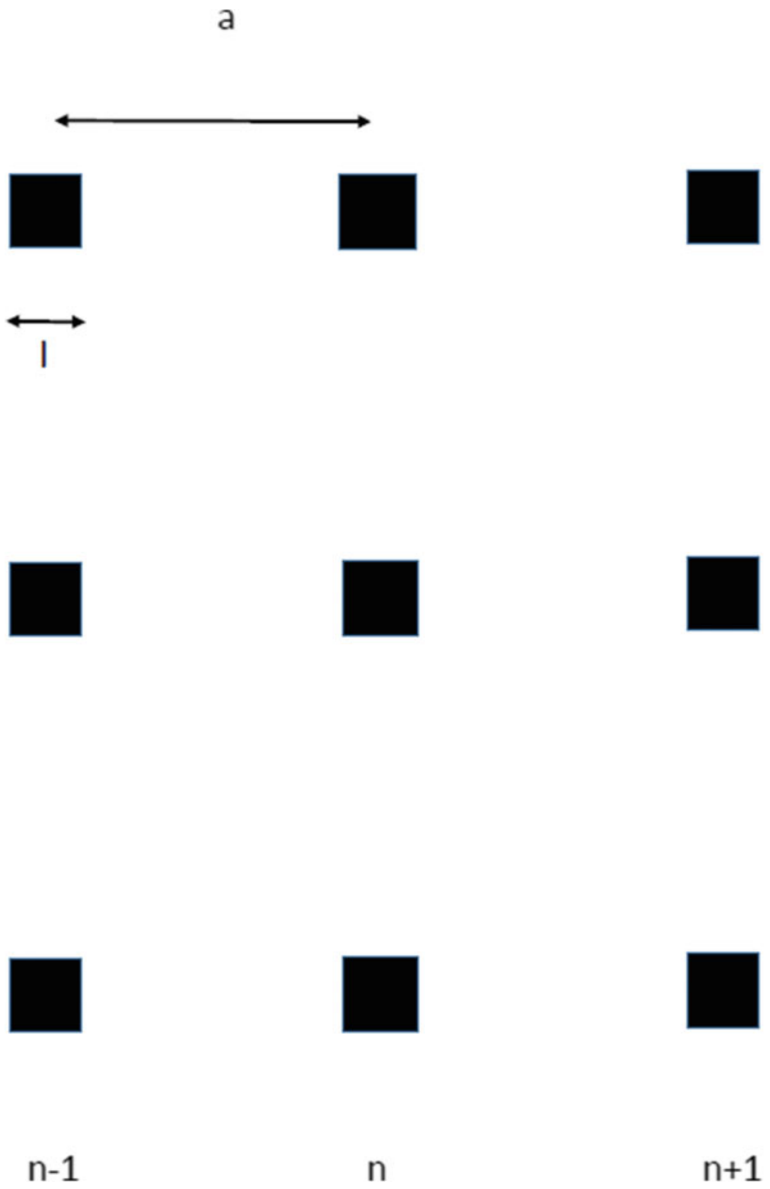


Fig. 4.34 Schematic of an infinite square lattice pattern of holes in a thin film. The pattern is in the *x*-*y* plane where the x axis is horizontal and the *y* axis is vertical. The lattice constant is a, and the edge length of the square holes is l. The vertical rows of holes are labeled by integers. The positive *z*-axis of the thin film geometry is out of the page

In this sequence of transfers the propagating incident fields transition into surface plasmon-polaritons which are in turn transferred into the fields propagating within the holes. The fields in the holes are finally transferred into the transmitted fields exiting below the thin film. In this manner the transmission through the film is described as a successive series of transitions between the various earlier mentioned modes or degrees of freedom [60].

The theory that describes these transitions between the modes or degrees of freedom of the thin film can be generated as a coupled mode theory [60, 64]. This type of theory is one that is common to many-body treatments that occur throughout all of physics and is appropriate for the dynamics of the electromagnetic scattering from the thin metal film. The coupled mode formulation describes the motion of excitations in a dynamical system as evolving through a series of transitions between the various degrees of freedom composing the system dynamics. The transitions are described by couplings between the degrees of freedom that determine the transition rate of one mode of the system into the other modes of the system. In general the couplings are linear, arising from the linear interactions present in a linear dynamical problem.

Coupled Mode Theories

A good example of a coupled mode theory is provided by the theory of phonons within a crystal lattice. This a basic system which, in the harmonic approximation, is commonly treated in texts on condensed matter physics. The simplest case of it is provided by an infinite chain of harmonically coupled atoms with a Hamiltonian given by [7]

$$H = \sum_n \left[\frac{m}{2} \dot{u}_n^2 + \frac{k_0}{2} (u_n - u_{n+1})^2 \right]. \tag{4.235}$$

Here m is the atomic mass, k_0 is the spring constant of the atomic interactions, and u_n is the displacement of the nth atom from its equilibrium position at na where a is the lattice constant of the chain.

The equations of motion obtained from the Hamiltonian are

$$m\omega^2 u_n = k_0 (2u_n - u_{n+1} - u_{n-1}) \tag{4.236}$$

where a time dependence of the form $e^{-i\omega t}$ is assumed. Rewriting (4.236) in the form

$$u_n = \frac{k_0}{m\omega^2 - 2k_0} (u_{n+1} - u_{n-1}) \tag{4.237}$$

it is seen that the displacement u_n is induced in the system as a transition from the displacement degrees of freedom u_{n+1} and u_{n-1}. The vibrational modes of the system are found to propagate as a series of tunneling transitions between the degrees of freedom of the dynamical system of the chain of atoms. The coupling coefficients that fix these transition amplitudes are given by

$$\frac{k_0}{m\omega^2 - 2k_0} \tag{4.238}$$

The next most important system treated in many-body physics is the thigh binding model of electrons [7, 65]. This is a basic model for understanding the properties of valence and conduction electrons in semi-conductor systems. It treats the electronic motion in these systems in terms of hopping jumps of the individual electrons between the atoms forming the crystal. These hopping transitions provide the basis for understanding the electronic current exhibited in materials with electron dynamics described by the thigh binding model.

The Hamiltonian of a one-dimensional tight binding model of electrons has the form [65]

$$H = \sum_n \left[t_0 a_n^+ a_n + t_1\left(a_n^+ a_{n+1} + a_n^+ a_{n-1}\right)\right]. \tag{4.239}$$

where t_0 and t_1 are electron hopping coefficients, and a_n^+ and a_n are the Fermi creation and destruction operators, respectively. The equations of motion obtained from (4.239) are

$$\varepsilon a_n^+ = t_0 a_n^+ + t_1\left(a_{n+1}^+ + a_{n-1}^+\right) \tag{4.240}$$

where ε is the energy of the mode which is assumed to have a harmonic time dependence. These are a set of difference equations which now involve quantum mechanical Fermion operators.

Rewriting (4.240) in the form

$$a_n^+ = \frac{t_1}{\varepsilon - t_0}\left(a_{n+1}^+ + a_{n-1}^+\right) \tag{4.241}$$

the electron at the nth site is found to arrive there by tunneling from the $(n+1)$th and $(n-1)$th sites. Again the transition amplitudes for these tunneling or hopping processes are given by the coupling coefficients

$$\frac{t_1}{\varepsilon - t_0}. \tag{4.242}$$

The transport in both the Boson and Fermion systems considered here are seen to be very similar, involving the tunneling of excitations along the sites of the one-dimensional chains. The fundamental difference is in the different quantum statistics of each system, but the dynamical interactions between the sites of the chain are handled the same in the theory of the two systems. Consequently, many of these basic ideas for treating a many-body system in terms of coupling and transitions between its various degrees of freedom can easily be extended to the study of other types of dynamical systems. Such considerations are now extended to treat the extraordinary optical transmission through the thin film problem associated with the system in Fig. 4.34.

Coupled Mode Treatment of Enhanced Transmission

For the thin film geometry in Fig. 4.34, the transmission through the film is studied for incident plane wave radiation. The radiation is taken to be traveling in the negative z-direction, incident normal to the surface of the thin film, and polarized with its magnetic field along the y-direction. This choice of polarization favors a coupling of the incident wave with surface plasmon-polaritons on the thin film that are propagating in the x-direction. Consequently, it will be assumed in the treatment presented here that the surface plasmon-polaritons excited in the system only involve surface modes moving in the positive and/or negative x-directions [60].

The degrees of freedom entering into the description of the dynamics of the transmission problem include: the initial incident and final transmitted plane waves traveling along the z-axis, the two surface plasmon-polaritons modes propagating along the x-direction in the film surfaces, and the electromagnetic modes propagating in each of the subwavelength holes penetrating the thin film. Each of these degrees of freedom is described by a normalized wave function which is defined within a specific spatial domain of the system. Consequently, the total normalized wave function of the system is written as a linear combination of the degrees of freedom separately weighted by an appropriate amplitude. By determining the amplitudes multiplying the component wave functions of the various degrees of freedom, a complete wave function description of the propagation of light though the system is obtained.

The basic assumption of the coupled mode theory presented here is that the propagation of the incident electromagnetic radiation through the thin film is mediated by a specific sequence of transitions. First there is a transition of the incident fields to the surface plasmon-polaritons which then transition into the modes of the individual subwavelength holes. Finally the fields in the holes exit the film by transitioning into the outgoing transmitted electromagnetic fields.

This sequence of processes is only a subset of all of the possible processes. For example, there are processes in which the fields of the incident plane wave transition directly into outgoing transmitted waves, processes in which the fields of the surface plasmon-polaritons transition directly into the transmitted fields, etc. The assumption, however, is that these are less dominant processes than those outlined above [60]. This can be eventually verified by comparing with results from a full computer simulation treatment.

In order to set up the system of coupled mode equations, consider the structure of vertical columns of holes in Fig. 4.34 that are labeled by the integers $n-1$, n, and $n+1$. First consider the surface plasmon-polariton modes propagating in the positive and negative x-directions with wave numbers k_{SP}. The wave functions of the surface plasmon-polaritons on the incident surface that are propagating to the right will enter the total wave function of the scattering problem with amplitudes

A_{n-1}, A_n, and A_{n+1}. These are the amplitudes for the right moving surface plasmon-polariton wave function in each of the columns labeled $n-1, n$, and $n+1$.

Similarly, the wave functions of the surface plasmon-polaritons on the incident surface that are propagating to the left will enter the total wave function of the scattering problem with amplitudes B_{n-1}, B_n and B_{n+1}. These are the amplitudes for the left moving surface plasmon-polariton wave function in each of the columns labeled $n-1, n$, and $n+1$.

In addition to the surface waves, the holes in the vertical columns $n-1, n$ and $n+1$ have amplitudes for the hole wave functions of the waves propagating down the holes. These amplitudes, respectively, are given by c_{n-1}, c_n, and c_{n+1}. The hole excitations deliver the electromagnetic fields from the incident to the transmission surface.

This completes the set of coefficients $\{A_m, B_m, c_m\}$ needed for a complete description of the total wave function of the scattering problem starting from the incident wave. It now remains to determine how these coefficients couple to one another in a linear relationship.

The set of $\{A_m, B_m\}$ coefficients are related to one another and the incident field by a set of two linear equations. These have the form [60]

$$A_n = \beta + \mathrm{e}^{ik_{SP}a}\tau A_{n-1} + \mathrm{e}^{ik_{SP}a}\rho B_{n+1} \tag{4.243a}$$

for right moving surface plasmon-polariton waves, and

$$B_n = \beta + \mathrm{e}^{ik_{SP}}\tau B_{n+1} + \mathrm{e}^{ik_{SP}a}\rho A_{n-1} \tag{4.243b}$$

for left moving surface plasmon-polariton waves.

In (4.243) β is the transmission amplitude for the normal incident plane wave radiation to transition into the left and right moving surface plasmon-polaritons. It governs the transition of incoming electromagnetic waves into surface plasmon-polaritons as well as the transition of surface plasmon-polaritons into outgoing electromagnetic waves. The τ term is the transmission amplitude for the tunneling of a surface plasmon-polariton wave to propagate along the surface as it encounters a column of holes. It provides for the continued propagation of the excitation in the direction it was going before it was incident on the column of holes. The ρ term is the transmission amplitude for the reflection of a surface plasmon-polariton wave as it encounters a column of holes. It provides for the reflection and propagation of the excitation in the direction opposite to that in which it was originally going before it was incident on the column of holes.

In addition, the various surface plasmon-polariton phase factors entering into the coefficients in (4.243) account for the phase changes between the $\{A_m, B_m\}$ as they are associated with different columns in Fig. 4.34. These phase factors account for the fact that the surface plasmon-polariton wave functions change phase as they pass from one column of holes to the neighboring column of holes.

A third equation needed to complete the description of the total wave function of the system is given by the linear form [60]

$$c_n = t + \mathrm{e}^{ik_{SP}a}\alpha A_{n-1} + \mathrm{e}^{ik_{SP}a}\alpha B_{n+1}. \tag{4.243c}$$

Here t is the transmission amplitude from the bulk plane wave into the modes of a single column of holes, and the α represents the modulus of the scattering transition coefficient of the surface plasmon-polaritons into the modes of the subwavelength holes penetrating the thin metal film. The phase factors again account for the phase differences of the surface plasmon-polariton wave functions between the columns of holes in the system.

Due to the subwavelength nature of the holes in the film, it is expected that $|\alpha|, |\rho| \ll |\beta|, |\tau|$. In particular, the arrays of holes in the thin film enter the problem as perturbations to the thin film system [60].

A solution of the set of equations in (4.243) can be found by assuming the forms [60]

$$A_n = A_0, \tag{4.244a}$$

$$B_n = B_0, \tag{4.244b}$$

and

$$c_n = c_0. \tag{4.244c}$$

This assumption for the form of the solution provides for a uniform transition of the incident mode over the entire surface.

The uniform mode in (4.244) represents the lowest order model solution in the infinite system and will be used as the basic transmission solution for the thin film model. In particular, experimentally, the transmission observed from the thin film is found to be uniform over the entire area of the thin film.

Substituting (4.244a) and (4.244b) into (4.243a) and (4.243b) gives solutions for $\{A_0, B_0\}$ that are of the form [60]

$$A_0 = B_0 = \frac{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right) + \rho\mathrm{e}^{ik_{SP}a}}{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right)^2 - \rho^2\mathrm{e}^{i2k_{SP}a}}\beta \tag{4.245}$$

It then follows from (4.243c) that

$$c_0 = t + 2\mathrm{e}^{ik_{SP}a}\alpha\frac{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right) + \rho\mathrm{e}^{ik_{SP}a}}{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right)^2 - \rho^2\mathrm{e}^{i2k_{SP}a}}\beta. \tag{4.246}$$

From (4.245) it is seen that the transmission amplitude for the transition of the incident plane waves into waves traveling down the system of holes is given by

$$t_a = c_0 = t + 2\mathrm{e}^{ik_{SP}a}\alpha \frac{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right) + \rho \mathrm{e}^{ik_{SP}a}}{\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right)^2 - \rho^2 \mathrm{e}^{i2k_{SP}a}}\beta. \tag{4.247}$$

Equation (4.247) then represent processes occurring at the incident surface of the thin film which is the first surface of the considerations given here.

The waves excited within the holes then travel away from the first surface down the holes until they encounter the second surface of the thin film. This is the surface from which the transmitted wave exits the thin film. At the second surface part of the waves in the holes are transmitted out of the thin film becoming radiation transmitted through the film and part are reflected back into the channel. The reflected waves travel back up the channel while the transmitted radiation contributes to the total transmission through the thin film.

Both the transmitted and reflected wave for radiation propagating in the holes are important in determining the total transmitted fields through the thin film. The nature of the reflected fields at the second surface are now addressed followed by the determination of the total transmission from the thin film.

The nature of the reflected wave in the hole modes must be determined before a complete solution of the transmission from the thin film is obtained. The modes in the holes of the thin film are reflected back and forth down their channels as they are in part transmitted and in part reflected at each end of the hole channels. For these modes the thin film acts as a Fabry-Perot oscillator, and, as with discussions of Fabry-Perot resonators, the fields that are passed by the resonator in each cycle of oscillation must be accounted for in determining the total output from the resonator. A discussion of the physics of the exit surface for the final transmitted wave is now addressed. This allows for an estimate of the reflection amplitude of the modes in the holes which then is used along with the transmission amplitude to determine the resonator output [60].

Similar considerations to those made in (4.243), treating the fields at the incident surface of the thin film, can be made at the second surface of the thin film from which the electromagnetic fields exit the thin film as transmitted waves. (In the following discussions the thin film surface receiving the original incident wave is referred to as the first surface, and the thin film surface from which the transmitted wave exits the thin film as the transmitted wave of the film is referred to as the second surface.) In the formulation at the second surface a set of relations is developed between the incident and reflected modes in the column of holes and the surface waves on the second surface. As stated earlier, the focus in the development of these relations is on obtaining an approximation for the reflection coefficient of the hole modes from the second surface.

In particular, an approximate relation between the surface electromagnetic modes at the second surface and the modes in a column of holes is given by the set of equations of the form [60]

$$A'_n = \alpha + e^{ik_{SP}a}\tau A'_{n-1} + e^{ik_{SP}a}\rho B'_{n+1} \tag{4.248a}$$

$$B'_n = \alpha + e^{ik_{SP}a}\tau B'_{n+1} + e^{ik_{SP}a}\rho A'_{n-1} \tag{4.248b}$$

$$c'_n = r + e^{ik_{SP}a}\alpha A'_{n-1} + e^{ik_{SP}a}\alpha B'_{n+1}. \tag{4.248c}$$

Here the primed coefficients $\{A'_m, B'_m\}$ refer to the surface plasmon-polaritons on the second surface of the thin film, and c'_n refers to the mode in the holes of the nth column that propagate from the second surface to the first surface. (Notice that in (4.243) and (4.248) the modes represented by c_n and c'_n move in opposite directions in the column of holes.)

In (4.248c) the coefficient r is the reflection coefficient for hole modes as they are reflected at the second surface of the thin film. In addition, the value of α, in the context of (4.248c), is taken as the amplitude for the transition of the incident hole mode on the second surface into the surface plasmon-polariton modes on the second surface. By reciprocity it is the same as the α in (4.243).

The phase factors in (4.248) are again chosen to account for the phase differences in the coefficients between the different columns. As with the discussion of (4.243) the holes are a perturbation on the system so that for the particular thin film being considered it is assumed that the transition amplitudes satisfy $|\alpha|, |\rho| \ll |\beta|, |\tau|$.

Proceeding as earlier in the case of (4.243), it follows from (4.248) that [60]

$$c'_0 = r + 2e^{ik_{SP}a}\alpha^2 \frac{\left(1 - e^{ik_{SP}a}\tau\right) + \rho e^{ik_{SP}a}}{\left(1 - e^{ik_{SP}a}\tau\right)^2 - \rho^2 e^{i2k_{SP}a}}. \tag{4.249}$$

As earlier in (4.247) it then follows that

$$r_a = c'_0 = r + 2e^{ik_{SP}a}\alpha^2 \frac{\left(1 - e^{ik_{SP}a}\tau\right) + \rho e^{ik_{SP}a}}{\left(1 - e^{ik_{SP}a}\tau\right)^2 - \rho^2 e^{i2k_{SP}a}} \tag{4.250}$$

provides the amplitude of reflection for the modes moving within the system of holes. It represents the reflection at the thin film surfaces for the uniform mode within the holes, just as (4.247) represents the transmission from these modes.

Equations (4.245)–(4.247) and (4.249) and (4.250) then provide the basis for the solution of the dynamics of a uniform mode of the thin film transmission problem. It represents a mode of the system with a uniform transmission and reflection of the incident radiation over the entire planar area of the infinite slab. Equations (4.247) and (4.250) do this for a single encounter of the hole mode with a thin film surface, providing both the transmission and reflection amplitude for this one encounter process. To correctly treat the total transmission of the thin film, however, (4.247) and (4.250) must be used to obtain the Fabry-Perot solution of the multiple scattering processes involved in the resonating motion of the hole modes in the slab.

The solutions in (4.247) and (4.250) will now be used as a basis for determining the transmission coefficient of the incident plane wave through the thin film, taking into account multiple scattering processes of the modes in the holes. These results follow upon examining the Fabry-Perot resonance of the modes oscillating in the system of holes. In these discussions, first the general nature of the Fabry-Perot resonance solution will be explained followed by its application to the extraordinary transmission problem.

Multiple Scattering Processes

Now that the transmission coefficient, t_a, and reflection coefficient, r_a, of the incident light on the thin film with the pattern of penetrating holes has been determine at both of the film surfaces, the problem of transmission through the thin film is greatly simplified. The new problem is that of the transmission through a homogeneous thin film that has the same transmission and reflection coefficients at its surfaces as the film with the pattern of holes and which is described by an effective refractive index n_e representing the propagation of the fields in the hole modes. This is the problem of transmission of light through a Fabry-Perot resonator. The solution of the transmission problem for a general Fabry-Perot resonator is treated next.

To understand the transmission through the Fabry-Perot resonator, consider the resonator problem based on the schematic figure shown in Fig. 4.35. The light in the system is normal incident on the film in the region above the film. The transmission and reflection coefficient of the upper surface of the film are t_1, r_1, respectively, and the transmission and reflection coefficient of the lower surface of

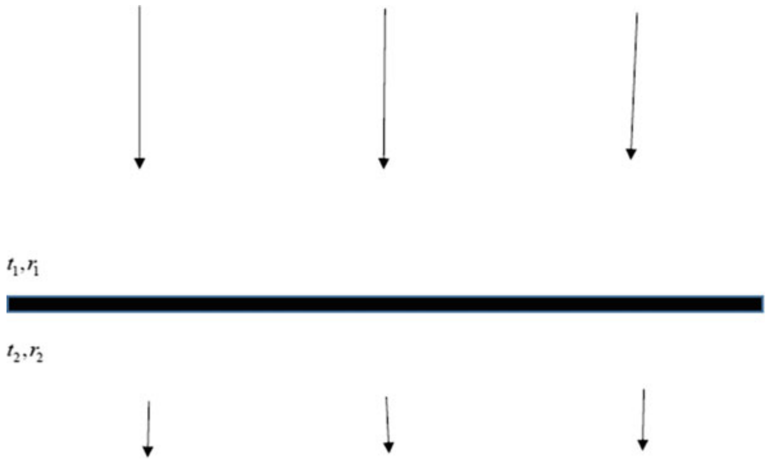


Fig. 4.35 Schematic of the Fabry-Perot resonator slab of effective refractive index n_e. The transmission and reflection coefficients t_1, r_1, respectively, are for the upper surface. The transmission and reflection coefficients t_2, r_2, respectively, are for the lower surface. The thickness of the thin film is denoted by t

the film are t_2, r_2, respectively. The film is taken to have an effective refractive index n_e and the film thickness is t.

For the transmission through the film the incident light is partially transmitted and partially reflected at the upper surface. The light transmitted through the upper surface then travels to the lower surface where it is partially transmitted and partially reflected. The partially reflected component at the lower surface in turn propagates back to the upper surface where it is partially transmitted and partially reflected. The partially reflected component of light at the upper surface then travels back to the lower surface where it undergoes partial transmission and partial reflection. In this way, the total transmission of light through the slab is built up as the sum of the various transmissions through the lower surface.

The sequence of transmissions and reflections outlined above are represented mathematically by an infinite series for the total transmission of light through the film, t_T, obtained as a sum of multiple scattering events at the film surfaces. This series is given by

$$\begin{aligned} t_T &= t_1 \mathrm{e}^{ikt}\left\{t_2 + r_2\mathrm{e}^{ikt} r_1 \mathrm{e}^{ikt} t_2 + \left(r_2\mathrm{e}^{ikt} r_1 \mathrm{e}^{ikt}\right)^2 t_2 + \cdots\right\} \\ &= t_1 t_2 \mathrm{e}^{ikt} \frac{1}{1 - r_1 r_2 \mathrm{e}^{i2kt}}, \end{aligned} \tag{4.251}$$

where k is the wave number of light in the thin film.

For the case of interest to the problem of the thin film with subwavelength penetration holes, it follows from (4.247) and (4.250) and the principle of reciprocity that $t_1 = t_2 = t_a$ and $r_1 = r_2 = r_a$. Specifically, the transmission and refection coefficients for the electromagnetic modes in the holes of the thin film are the same at both the first and second surfaces of the thin film. The total transmission for the thin film with subwavelength penetrating holes is then from (4.251) given by [60]

$$t_T = t_a^2 \mathrm{e}^{ikt} \frac{1}{1 - r_a^2 \mathrm{e}^{i2kt}}. \tag{4.252}$$

The result for the extraordinary optical transmission is obtained from (4.247), (4.250), and (4.252) with the input of the parameters $\alpha, \beta, \tau, \rho$ and the parameters of the surface plasmon-polariton wave number, etc. For these considerations, the coefficients $\alpha, \beta, \tau, \rho$ can be determined for a single column of penetrating holes studied as a function of the wavelength. When the results for the coupling coefficients determined in this way are used in (4.247), (4.250), and (4.252) a reasonable agreement is found with the results obtained from more general computer simulation studies. In addition, an important feature of the analytical result is that the extraordinary optical transmission effect can also be seen to arise from some of the analytical properties of (4.247), (4.250) and (4.252).

Extraordinary Transmission Effect

From the analytic studies presented in (4.252) the extraordinary transmission features are found to arise from the denominators in the second terms on the right hand side of (4.247) and (4.250). These denominators can be used to understand the transmission enhancement on the basis of its perturbation nature.

Each of the denominators in (4.247) and (4.250) are of the form $\left(1 - \mathrm{e}^{ik_{SP}a}\tau\right)^2 - \rho^2 \mathrm{e}^{i2k_{SP}a}$, containing an important factor of

$$1 - \mathrm{e}^{ik_{SP}a}(\tau + \rho). \tag{4.253}$$

This factor in the denominators accounts for the presence of the extraordinary optical transmission effect. In particular, for the case in which

$$1 - \mathrm{e}^{ik_{SP}a}(\tau + \rho) \approx 1 - \mathrm{e}^{ik_{SP}a}\tau \approx 0 \tag{4.254}$$

the denominators in (4.247) and (4.250) become small with the net result that the transmission in (4.252) becomes large, exhibiting an extraordinary optical transmission maxima. (Notice in (4.254) the perturbation limit of the holes, for which $|\rho| \ll |\tau|$, has been used.)

For the case of weak scattering by the column of subwavelength holes it is expected that $|\tau| \approx 1$ so that, in the case of a long propagation length for surface plasmon-polaritons along the surface of the film,

$$\mathrm{e}^{ik_{SP}a}\tau \approx \mathrm{e}^{i\left(\mathrm{Re}\left(k_{sp}a\right) + \phi_\tau\right)} \tag{4.255}$$

where $\tau = |\tau|\mathrm{e}^{i\phi_\tau}$. The extraordinary optical transmission maxima then occur under the condition

$$\mathrm{Re}(k_{SP}a) + \phi_\tau \approx 2\pi m \tag{4.256}$$

where m is an integer.

It is seen from (4.256) that the enhancement effect is intimately connected by the propagation properties of the surface plasmon-polariton modes on both of the two surfaces of the thin film. In addition, for the model presented, the effect requires that the holes be a perturbation on the propagation of the surface plasmon-polariton. The modes in the holes are also essential for the enhancement in providing a path for the electromagnetic waves through the film.

References

1. S.A. Maier, *Plasmonics: Fundamentals and Applications* (Springer, Berlin, 2007)
2. S. Enock, N. Bonod (ed.), *Plasmonics: From Basics to More Advanced Topics* (Springer, Berlin, 2012)

3. A.A. Maradudin, J. Roy, W. Barnes, *Modern Plasmonics* (Elsevier, Amsterdam, 2014)
4. V.M. Agranovich, D.L. Mills (ed.), *Surface Polaritons-Electromagnetic Waves at Surfaces and Interfaces* (Noth-Holland Publishing Company, Amsterdam, 1982)
5. R.F. Wallis, G.I. Stegeman (ed.), *Electromagnetic Surface Excitations* (Springer, Berlin, 1986)
6. M.G. Cottam, D.R. Tilley, *Introduction to Surface and Superlattice Excitations* (Institute of Physics Publishing, Bristol, 2005). G. Bergmann, Weak localization in thin films. Phys. Rep. **107**, 1–58 (1984)
7. C. Kittel, *Introduction to Solid State Physics*, 7th edn. (Wiley, New York, 1996)
8. A.A. Maradudin, W.M. Visscher, Electrostatic and electromagnetic surface shape resonances. Zeit Physik Condens. Matter **60**, 215–230 (1985)
9. A.R. McGurn, Enhanced retroreflectance effects in the reflection of light from randomly rough surfaces. Surf. Sci. Rep. **10**, 357 (1990)
10. A.R. McGurn, A.A. Maradudin, An analogue of enhanced backscattering in the transmission of light through a thin film with a randomly rough surface. Opt. Commun. **72**, 279 (1989)
11. G.C. Brown, V. Celli, M. Haller, A. Marvin, Surf. Sci. **136**, 391 (1984)
12. P.B. Johnson, R.W. Christy, Phys. Rev. **B6**, 4370 (1972). J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1998)
13. R.F. Wallis, Introduction to electromagnetic surface waves, in *Electromagnetic Surface Excitations,* ed. by R.F. Wallis, G.I. Stegeman (Springer, Berlin, 1986), pp. 2–7
14. A.A. Maradudin, Electromagnetic surface excitations on rough surfaces, in *Electromagnetic Surface Excitations*, ed. by R.F. Wallis, G.I. Stegeman (Springer, Berlin, 1986), pp. 57–131. B. Laks, D.L. Mills, A.A. Maradudin, Surface polaritons on large-amplitude gratings. Phys. Rev. **B23**, 4965–4978 (1981). A.A. Maradudin, W.M. Visscher, Electrostatic and electromagnetic surface shape resonaces. Z. Phys. B-Condens. Matter **60**, 215–230 (1985)
15. K.L. Kliewer, R. Fuch, Collective electronic motion in a metallic slab. Phys. Rev. **153**, 498–512 (1967)
16. A.A. Maradudin, T. Michel, A.R. McGurn, E.R. Mendez, Enhanced backscattering of light from a randomly rough grating. Ann. Phys. **203**, 255–307 (1990)
17. G. Zu-Han, R.S. Dummer, A.A. Maradudin, A.R. McGurn, Opposition effect in the scattering of light from a random rough metal surface, in *SPIE Proceedings,* vol. 1165 (1990). https://doi.org/10.1117/12.962835
18. J.W. Goodman, *Speckle Phenomena in Optics* (W. H. Freeman, San Francisco, 2007)
19. J.W. Goodman, Some fundamental properties of speckle. J. Opt. Soc. Am. **66**(11), 1145–1150 (1976)
20. A.R. McGurn, A.A. Maradudin, Speckle correlations in the light reflected and transmitted by metal films with rough surfaces: surface wave effects. Opt. Commun. **155**, 79 (1998)
21. A.R. McGurn, A.A. Maradudin, Speckle correlations in the light scattered by a dielectric film with a rough surface: guided wave effects. Phys. Rev. **B58**, 5022 (1998)
22. V. Malyshkin, A.R. McGurn, A.R. Maradudin, Features in the speckle correlation of light scattered from volume-disordered dielectric media, in *Proceedings of SPIE,* vol. 3426 (The International Society for Optical Engineering, 1998), p. 96
23. A.R. McGurn, A.A. Maradudin, Computer simulation studies of the speckle correlations of light scattered from a random array of dielectric spheres. Proc. SPIE **3426**, 134 (1998)
24. V. Malyshkin, A.R. McGurn, A.A. Maradudin, Features in the speckle correlations of light scattered from volume-disordered dielectric media. Phys. Rev. B **59**, 6167 (1999)
25. V. Malyshkin, A.R. McGurn, T.A. Leskova, A.A. Maradudin, M. Nieto-Vesperinas, Speckle correlations in the light scattered from weakly rough random metal surfaces. Waves Random Media **7**, 479 (1997)
26. A.R. McGurn, A.A. Maradudin, Speckle correlations in the light scattered and transmitted by dielectric and metal films with rough surfaces. Proc. SPIE **3141**, 255 (1997)
27. M. Francon, *Laser Speckle and Applications in Optics* (Academic Press, Amsterdam, 1979)
28. H.J. Rabal, R.A. Braga, *Dynamic Laser Speckle and Applications* (CRC Press, Boca Raton, 2008)

29. E. Le Ru, P. Etchegoin, *Principles of Surface Enhanced Raman Spectroscopy and Related Plasmonic Effects* (Elsevier Science, Burlington, 2008)
30. S. Schlucker (ed.), *Surface Enhanced Raman Spectroscopy: Analytical, Biophysical and Life Science Applications* (Wiley, Hoboken, 2013)
31. Y. Ozaki, K. Kneipp, R. Aroca, *Frontiers of Surface-Enhanced Raman Scattering Single Nanoparticles and Single Cells* (Willey, Hoboken, 2014)
32. M. Baia, S. Astilean, T. Iliescu, *Raman and SERS Investigations of Pharmaceuticals* (Springer, Berlin, 1974)
33. E. Alarcon, M. Griffith, K.I. Udekwu, *Silver Nanoparticle Applications in the Fabrication and Design of Medical and Biosensing Devices* (Springer, Berlin, 2015)
34. K. Kneipp, M. Moskovits, H. Kneipp, *Surface-Enhanced Raman Scattering Physics and Applications* (Springer, Heidelberg, 2006)
35. M. Fleischmann, P.J. Hendra, P. McQullan, Raman spectra of pyridine adsorbed at a silver electrode. Chem. Phys. Lett. **26**, 163–166 (1974)
36. S. McAughtrie, K. Faulds, D. Graham, Surface enhanced Raman spectroscopy (SERS): potential applications for disease detection and treatment. J. Photochem. Photobiol. C **21**, 40–53 (2014)
37. G. Frens, Controlled nucleation for the regulation of the particle size in monodisperse gold suspensions. Nat. Phys. Sci. **241**, 20–22 (1973)
38. U.K. Sur, J. Chowdhury, Surface-enhanced Raman scattering: overview of a versatile technique used in electrochemistry and nanoscience. Curr. Sci. **105**, 923–939 (2013)
39. E. Smith, G. Dent, *Modern Raman Spectroscopy: A Practical Approach* (Wiley, Hoboken, 2005)
40. M. Baibarac, M. Cochet, M. Lapkowski, L. Mihut, S. Lefrant, I. Baltog, SERS spectra of plyaniline thin film s deposited on rough Ag, Au, and Cu Polymer films thickness and roughness parameter dependence of SERS spectra. Synth. Mater. **96**, 63–70 (1998)
41. Y. Fang, M. Sun, Nanoplasmonic waveguides: towards applications in integrated nanophotonic circuits. Light Sci. Appl. **4**, e294 (2015). https://doi.org/10.1038/isa.2015.67. H. Wei, H. Xu, Nanowire-based plasmonic waveguides and devices for integrated nanophotonic circuits. Nanophotonics **1**, 155–169 (2012)
42. W.L. Barnes, A. Dereux, T.W. Ebbesen, Surface plasmon subwavelength optics. Nature **424**, 824–830 (2003)
43. W.L. Barnes, Surface plasmon-polariton length scales: a route to sub-wavelength optics. J. Opt. A Pure Appl. Opt. **8**, S87–S93 (2006)
44. D. Dai, H. Wu, W. Zhang, Utilization of field enhancement in plasmonic waveguide for subwavelength light-guide, polarization handling, heating, and optical sensing. Materials **8**, 6772–6791 (2015)
45. X. Sun, Hybrid plasmonic waveguides and devices theory, modeling and experimental demonstration. Masters Thesis, Department of Electrical Engineering and Computer Engineering, University of Toronto (2013). T.J. Davis, D.E. Gomez, A. Roberts, Plasmonic circuits for manipulating optical information. Nanophotonics. https://doi.org/10.1515/nanoph-2016-0131
46. R. Yang, Z. Lu, Subwavelength plasmonic waveguides and plasmonic materials. Int. J. Opt. **2012**, Article ID 258013, 12pp. (2012)
47. A. Yang, T.W. Odom, Breakthroughes in photonics 2014: advances in plasmonic nanolasers. IEEE Photonics J. **7**, 7000606 (2015). Y. Yin, T. Qiu, J. Li, P.K. Chu, Plasmonic nano-lasers. Nano Energy **1**, 25–41 (2012)
48. J.A. Anker, W.P. Hall, O. Lyandres, N.C. Shah, J. Zhao, R.P. Van Duyne, Optical, photonic and optoelectronic materials: sensors and biosensors, nanoscale materials. Nat. Mater. **7**, 442–453 (2008)
49. M.I. Stockman, The spaser as a nanoscale quantum generator and ultrafast amplifier. J. Opt. **12**, 024004 (2010). M.I. Stockman, Spasers explained. Nat. Photonics **2**, 327–329 (2008)
50. D.J. Bergman, M.I. Stockman, SPASER as ultrafast nanoscale phenomenon and device, in *Ultrafast Phenomena*, vol. XIV, ed. by T. Kobayashi, R. Okada, T. Kobyashi, K.A. Nelson,

S. De Sivestri (Springer, Berlin, 2014). D.J. Bergman, M.I. Stockman, Surface plasmon amplification by stimulated emission of radiation: quantum generation of coherent surface plasmons in nanosystems. Phys. Rev. Lett. **90**, 027402 (2003)

51. L. Zhang, J. Zhou, H. Zhang, T. Jiang, C. Lou, Ultra-strong surface plasmon amplification characteristic of a spaser based on gold-silver core shell nanorods. Opt. Commun. **338**, 313–321(2015). R.-M. Ma, R.F. Qulton, V.J. Sorger, X. Zhang, Plasmonic lasers: coherent light source at molecular scales. Laser Photonics Rev. **7**, 1–21 (2013)
52. S. Gwo, C.-K. Shih, Semiconductor plasmonic nanolasers: current status and perspectives. Rep. Prog. Phys. **79**, 086501 (2016). W.L. Barnes, Surface plasmon-polariton length scales: a route to sub-wavelength optics. J. Opt. A Pure Appl. Opt. **8**, S87–S93 (2006)
53. Y. Yin, T. Qui, J. Li, P.K. Chu, Plasmonic nano-lasers. Nano Energy **1**, 25–41 (2012)
54. R.-M. Ma, R.F. Oulyon, V.J. Sorger, X. Zhang, Plasmon lasers: coherent light source at molecular scales. Laser Photonics Rev. **7**, 1 (2013)
55. D.J. Bergman, M.I. Stockman, Surface plasmon amplification by stimulated emission of radiation: quantum generation of coherent surface plasmons in nanosystems. Phys. Rev. Lett. **90**, 027402 (2003)
56. M.I. Stockman, Spasers explained. Nat. Photonics **2**, 327–329 (2008)
57. F.I. Baida, M. Boutria, R. Oussaid, R. Van Labeke, Enhanced transmission metamaterials as anisotropic plates. Phys. Rev. B **84**, 035107 (2011)
58. J. Wang, W. Zhou, E.-P. Li, Enhanceing the light transmission of plasmonic metamaterials through polygonal aperture arrays. Opt. Express **17**, 20349–20354 (2009)
59. M. Beruete, M. Sorolla, I. Campillo, J.S. Dolado, L. Martin-Moreno, J. Bravo-Abad, F.I. Garcia-Vidal, Enhanced millimeter, wave transmission through quasioptical subwavelength perforated plates. IEEE Trans. Antennas Propag. **53**, 1897–1903 (2005)
60. H. Lui, P. Lalanne, Microscopic theory of the extraordinary optical transmission. Nature **452**, 728–731 (2008)
61. T.W. Ebbensens, H.J. Lezec, H.F. Ghaemi, T. Thio, P.A. Wolff, Extraordinary optical transmission through sub-wavelength hole arrays. Nature **391**, 667–669 (1998)
62. E. Altewischer, M.P. van Exter, J.P. Woerdman, Plasmon assisted transmission of entangled photons. Nature **418**, 304–306 (2002)
63. H.A. Bethe, Theory of diffraction by small holes. Phys. Rev. **66**, 163–182(1944)
64. H.-C. Huang, *Coupled Mode Theory: As Applied to Microwave and Optical Tranmssion* (Taylor & Francis, New York, 1984)
65. P.M. Yu, M. Cardonna, *Fundamentals of Semiconductors* (Springer, Berlin, 2010)

Chapter 5
Metamaterials

5.1 Basic Properties of Metamaterials

Metamaterials are engineered materials that are designed to manipulate light in ways that are not possible with materials taken directly from nature [1–6]. In their designs metamaterial have a synthetic structure introduced into them at subwavelength scales. As a result, for the wavelengths they manipulate, the metamaterials appear to be homogeneous media. A homogeneous medium, however, with previously unseen optical properties. In the development of this new class of optical materials, metamaterials have been studied in the frequency regions between the terahertz and optical regions. The essential limitation on the wavelengths of their applications being the ability to design and implement an appropriate subwavelength pattern of synthetic features which provides the source of the metamaterial response.

If chosen well their subwavelength structure allows metamaterials to exhibit optical responses and design characteristics otherwise not found in the study of traditional optics. The increase in the available types of optical responses shows up in a greater range of refractive properties and a variety of unusual energy transport properties exhibited by the new class of optical metamaterials. When added to the existent properties of naturally occurring optical materials, metamaterials increase the range of design characteristics that can be achieved in optical technology. This provides for an important variety of devices made possible solely through the novel optical characteristics of metamaterials.

The synthetic structure of metamaterials is designed by including specific artificial features and patterns of artificial features at subwavelength dimensions within an otherwise homogeneous background medium [1–11]. The artificial features and patterns are introduced to modify the material so that it exhibits a specific effective permittivity and permeability when interacting with radiation of wavelengths much greater than the artificial inclusions engineered into the material. Just as crystalline materials appear homogeneous to electromagnetic waves at optical wavelengths,

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_5

meta-materials are fashioned to appear homogeneous to the specific set of wavelengths they are designed to manipulate [1–12]. The discrete structure of both the natural occurring medium and the synthesized metamaterial is not directly noticed by the radiation with which they interact.

The newly fabricated metamaterials have found many possible and proposed applications in device technology [1–11]. Such tested and proposed uses of metamaterials include design applications in: certain schemes of making cloaking devices, antenna designs, high resolution lenses, systems displaying enhanced transmission effects, sensors, second harmonic generation schemes, applications to medical optics, and in the simulation of optical effects found in general relatively [13–22].

Many of these device applications depend on the greater range of refractive manipulation of light made possible by metamaterials in the applications of ray optics. Metamaterials, however, have a number of problems associated with the inherent nature of the designs upon which their subwavelength structure is based [1–22]. These include problems associated with the range of frequencies over which their design and implementation functions successfully and problems with the loss of optical energy as light moves through the metamaterials. In particular, energy loss is often found to occur as part of the characteristics of metamaterial design fundamentals. Consequently, there is currently a great endeavor focused on developing better, more efficient, metamaterials as well as in formulating metamaterial applications in the design and implementation of optical device technology.

Initially in the study of metamaterials the interest in the artificial subwavelength inclusions was based on their magnetic resonance properties [1–6]. Specifically, artificial inclusions can be designed so that they exhibit magnetic resonance responses to electromagnetic waves that are not found in the magnetic resonance properties of atoms and molecules occurring in natural crystals. The inclusions form artificial resonators which can be tuned to exhibit resonant properties with external electromagnetic fields over a much wider range of frequencies than those available in atomic and molecular interactions.

Systems containing these subwavelength features exhibit resonant frequencies and associated regions of negative permeability at frequencies in which atomic and molecular of crystalline materials do not [1–6]. This is an essential point, important in the design of metamaterials exhibiting the property of a negative refractive index at frequency regions of interest for device applications. It is an essential property in the new optics of these materials as, in naturally occurring materials, no material has been found that has a negative refractive index [1–6].

It was shown in the early twentieth century that a requirement for a homogeneous medium to exhibit a negative index of refraction is that at the frequency of the radiation both the permittivity and permeability of the media must simultaneously be negative. Frequency regions in naturally occurring media of simultaneous negative permittivities and permeabilities do not exist, and this has been a fundamental restriction on traditional optics.

In metamaterials, the situation is different. Negative index regions have now become available in designs of meta-materials. Consequently, the possibility of negative refractive index metamaterials extends what was once considered a fundamental limitation on optical design.

Metamaterials are a different class of artificial system from the periodic dielectric structures known as photonic crystals [1–22]. While the basic properties of photonic crystals arise from the diffraction of light in the periodic system, the object of metamaterials is in the design of a material which exhibits refractive (not diffractive) effects. The confining properties of photonic crystals, used to trap light in resonator cavities or waveguides is not a primary focus in metamaterial technology. Metamaterial design is focused on the gradual changes in the motion of light typically associated with ray optics. However, metamaterials offer an extended range of refractive properties to develop ray optics over the refractive properties available from naturally occurring materials. While the metamaterial can be classified as a new type of material, photonic crystals function more in the role of an optical device performing a diffractive function.

Recently the ideas of metamaterial technology have been extended to the study of another class of artificial materials. These are the so-called hyperbolic materials [23–25]. Hyperbolic materials are engineered to exhibit a specific type of electromagnetic dispersion relation for light propagating within them. They are composed of subwavelength features and again appear as homogeneous media for frequencies of light with which they are designed to interact.

In hyperbolic materials a focus is on the construction of a material with a particular form of dielectric tensor which sets the electromagnetic dispersion [23–25]. Specifically, the dielectric tensor is designed so that it leads to a hyperbolic form of the dispersion relation for light propagating within the material. As shall be shown later, materials with hyperbolic dispersion relations display a variety of interesting and useful effects on light moving within them. Whereas the earlier discussed metamaterials involved introducing structures that provide a magnetic resonance response, the structure of hyperbolic materials, as shall be seen, are based on a surface plasmon-polariton mechanism.

The subwavelength structures introduced into media in order to create hyperbolic materials are chosen to support surface plasmon-polaritons on their surfaces. The intense subwavelength fields of the surface plasmon-polariton excitations, distributed throughout the bulk of the hyperbolic material, enter as an important factor in the material design considerations. They are key in determining the interesting responses of the materials to applied electromagnetic fields.

The resulting hyperbolic materials, in some cases, are found to exhibit negative index of refraction [23–25]. In addition, they have also been shown to display enhanced transmission effects as well as properties useful in applications for the design of sensors and other optical device applications. These applications arise from a series of variations in the hyperbolic material designs.

In addition to the idea of creating metamaterials with extended optical properties, photonic crystals have been shown under certain conditions to exhibit properties that mimic some of the interesting effects found in metamaterials. These effects are usually limited and are localized about certain regions of the photonic crystal band structure in wave vector space. Some aspects of photonic crystals exhibiting unusual optical effects similar to those found in metamaterials will be briefly review at the end of this chapter.

In the following, discussions will first be given of metamaterials based on subwavelength features designed to give a magnetic resonant response to externally applied fields. The resonant response in these types of metamaterials is usually provided by a subwavelength feature known as a split ring resonator. There are many types of split ring resonator designs depending on the frequencies at which they are meant to operate as well as other design requirements peculiar to the material being formulated. In its basic form, it is just a structure with a resonant interaction when driven by an external frequency dependent field. The operation of the simplest form of a split ring resonator will be explained in terms of a simple modes of an LRC circuit which is driven by an externally applied magnetic field.

Following some basic discussion regarding the use of the split ring resonators in metamaterial designs, the essential properties exhibited by metamaterials are discussed. The most important of these is the property of negative refraction, and the behavior of light propagating in negative index media is reviewed. Examples of the behavior of negative index media presented include: the refraction of light at the interface between positive and negative indexed media, the properties of a perfect lens, the properties of radiation traveling within negative index media, and discussions of potential device applications.

The chapter is concluded with a presentation on the properties of hyperbolic-materials [23–25]. This includes discussions on the nature of the dielectric tensor in these systems and the properties of the associated dispersion relation of light in hyperbolic metamaterials. Applications of these materials are then explained in terms of the dispersive properties of light.

In addition, the use of photonic crystals to mimic some of the effects found in metamaterials is discussed. These photonic crystal properties are an analogy of various band structure effects found in the magneto-resistance of conduction electrons [12].

The first topic which is next addressed deals with the properties of a basic split ring resonator unit as it enters into meta-materials as a component and how split ring resonators can be arrayed in three dimensions to form bulk materials with engineered diamagnetic responses.

5.1.1 Properties of Split Ring Resonators and Split Ring Resonator Arrays

The split ring resonator is the basic artificial inclusion used in the formation of metamaterials [1–11]. There are many variations on its design details that are made for engineering material considerations, and some of these will be discussed later in

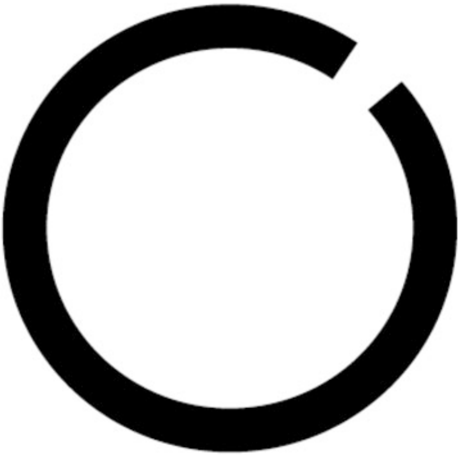


Fig. 5.1 Basic design of an idealized SRR unit. A metallic ring with a gap filled with dielectric media forms the basis of an LC circuit

this chapter. In its basic form, however, the split ring resonator is a conducting ring with a gap cut into it. (For a schematic of the simplest form of split ring resonator see Fig. 5.1.) The basic design of the split ring resonator causes it to function as an elementary forced resonator circuit of a kind that is commonly studied in introductory physics and electrical engineering courses.

The ring of the split ring resonator acts as an inductance, just as a ring of wire has a self-inductance when interacting with an external magnetic field or when a current is passed through the ring. The gap in the ring acts as a capacitor in which equal and opposite charges are developed on the surfaces that are separated by the ring gap. The split ring resonator then combines the inductance of the ring in series with the capacitance of the gap in the ring to form a simple LC resonant circuit [1–11].

The split ring resonator has no net charge on it, but, by interacting with a time-dependent external magnetic field applied perpendicular to the plane of the ring, it can develop a time-dependent current within the ring. As a result of the inductive interaction with the driving field, the current induced in the ring has the same frequency as that of the external field. The applied frequency driving the split ring current, however, is generally different from the natural resonant frequency set by the inductance and capacitance of the ring. To a large extent the difference between the resonant and applied frequencies determines the response of the ring to the applied field. This response characteristic is an essential feature of the nature of the split ring resonator as a harmonic oscillator forcibly driven by the externally applied field [1–6].

The natural frequency of the split ring resonator is determined from the inductance of the ring and the capacitance of the gap by a standard formula from the elementary physics of an LC circuit. This resonant frequency is then given by [1]

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{5.1}$$

where L and C are the self-inductance and capacitance of the split ring, respectively. The frequency in (5.1) is the resonant frequency of the interaction of the split ring resonator with an external driving electromagnetic wave, setting many important properties of the interaction of the ring with the external field.

The relationship of the resonant frequency in (5.1) to the frequency of the applied magnetic field will determine whether or not the magnetic moment induced in the split ring resonator is paramagnetic of diamagnetic in nature. In particular, the magnetic moment of the split ring resonator is determined by the electric current induced in the ring by the externally applied magnetic field. The phase of the induced current relative to that of the applied field is set by the resonant conditions of the split ring, and this is an important factor determining the dipole moment of the ring.

In a system formed as an array of split ring resonators, the nature of the paramagnetic or diamagnetic response of the array depends on the responses of the individual split ring resonators of the array as they interact with the external field and also with one another. This is the case with metamaterials formed as arrays of split rings. For these type of arrays, the collective response of all of the rings forming the arrays determines the important diamagnetic or paramagnetic response of the system, and this collective response depends on the self-inductance of the rings as well as the mutual inductance between rings.

In addition to the self-inductance, the neighboring rings in the meta-material have a mutual inductance coupling between one another. These mutual inductances couple the split rings together and gives them a collective dynamics of propagating electromagnetic waves between the rings. The mutual inductive couplings allow the magnetic fields generated from the currents flowing within one split ring to induce fields in its neighboring split rings. This process results in a traveling wave of such field transfers throughout the array of rings.

For an array of split rings formed on a lattice, the system (which is electrically neutral on the whole) exhibits what are termed magneto inductive waves [1–11]. These waves are electromagnetic waves which travel as plane waves in the array and exhibit a dispersion relation which depends on the mutual inductive couplings between the split rings. The study of magneto inductive waves has a long history prior to the study of metamaterials. Metamaterials, however, in part exhibit properties and characteristics which are functions of the physics of the magneto inductive waves.

The properties of magneto inductive waves will be discussed later. For now the focus is on a single split ring resonator and how it gives a paramagnetic or diamagnetic response to an applied frequency dependent electromagnetic field [1].

Consider the single split ring resonator in Fig. 5.1 driven by an external electromagnetic wave of frequency, ω, and with it magnetic field perpendicular to the plane of the page. The natural resonant frequency ω_0 of the split ring resonator is determined from (5.1) in terms of the self-inductance and capacitance of the split ring. A resonant interaction of the electromagnetic wave with the ring is observed as ω passes through ω_0 [1–6].

The essential physics of the split ring resonator interaction with the external field can be understood on the basis of the simple resonator circuits in Fig. 5.2. In Fig. 5.2a a simple resonator circuit composed as an inductor and a capacitor is shown. This models the basic split ring resonator where the inductance in Fig. 5.2a is the inductance of the ring and the capacitor is that of the gap in the ring. The resonance frequency for the free standing ring oscillator is given by (5.1).

A time dependent magnetic field introduced perpendicular to the plane of the split ring then, through an application of Faraday's Law, generates an electromotive force in the ring. In the Fig. 5.1 the applied magnetic field is perpendicular to the page with the positive sense of the field out of the page, and the positive current in the loop is in the anti-clockwise direction [1].

The induced emf in the split ring resonator from the magnetic field can be introduced into the circuit of Fig. 5.2a as an externally applied source of emf. The circuit representing the split ring resonator driven by the external magnetic field is that shown in Fig. 5.2b. Here the source of applied emf is from the emf generated in the ring by the changing magnetic field.

The induced emf from the time dependent field enters the LRC electromagnetic oscillator circuit shown in Fig. 5.2 as a driving force of the harmonic motion. The ring of the split ring feature acts both as the self-inductance of the circuit and the origin of the driving emf through the induced Faraday's Law interaction with the external field [1].

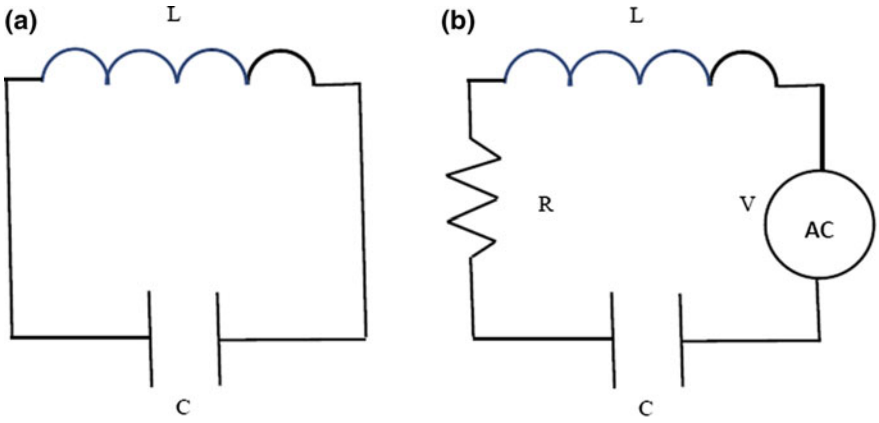


Fig. 5.2 Schematics of: **a** an LC ring with a time varying magnetic field perpendicular to the plane of the ring, **b** an LRC forced harmonic oscillator circuit

For the system represented in Fig. 5.2b, the equation of the driven oscillator is given by the standard form [1]

$$L\frac{d^2Q}{dt^2}+R\frac{dQ}{dt}+\frac{Q}{C}=V. \tag{5.2}$$

In this equation Q is the charge on the capacitor, V is the forcing emf from Faraday's Law, and the equation has been modified to include resistive losses.

Resistive losses are important in restraining the response of the split rings from becoming singular at the resonant frequency and also as a source of inefficiency in the split ring design. They come from two sources: Joule losses to the materials forming the ring and its capacitive gap and radiative losses from the ring. Care must be taking in the design of the system so that these two types of losses are kept low. In terms of radiative losses, the split ring itself forms an antenna, and an object is to make it as inefficient an antenna as possible [1–6].

Loss mechanisms are a major problem with metamaterials based on arrays of split ring resonators. Losses become particularly important in the system at and near resonance, and these are the regions of most interest for the design of metamaterials exhibiting the property of negative index of refraction. Another problem of resonator based metamaterials is the narrow band of frequencies over which the systems exhibit resonant behaviors. As shall be seen later, this limits much of the applications of these types of artificial materials.

To understand the response of the split ring resonator to an applied electromagnetic wave polarized with a magnetic field perpendicular to the page in Fig. 5.1, the physics of the driven oscillator in (5.2) is now treated. Through a straightforward application of Faraday's Law, a magnetic field component of the form $B(t)=B_0\cos\omega t$ interacting with the split ring is found to develop a time-dependent emf within the ring. In this way the driving emf is then given by [1]

$$V=V_0\sin\omega t, \tag{5.3}$$

where $V_0=\omega AB_0$ for a ring of area A, and the maximum flux through the split ring is given by AB_0. Equation (5.3) expresses the forcing emf of the model system in Fig. 5.2b which represents an equivalent circuit for the split ring interacting with the external field. It expresses this directly in terms of the applied magnetic field. By using (5.3) in (5.2) the dynamics of the two related systems are determined.

Using (5.3) as the driving term in the theory of the LRC forced oscillator, the response of the current in the ring is given directly from the impedance of the series inductor-resistor-capacitor circuit. This is found worked out in many elementary texts in physics and engineering. Specifically, it is shown that the relationship of the amplitude of the current I_0 to the amplitude of the voltage V_0 through the impendence Z is written as $V_0=ZI_0$. In addition, in this relationship the impedance Z of the forced LRC oscillator circuit has the form [1]

$$Z(\omega) = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{5.4}$$

in terms of the inductance, capacitance, resistance and the frequency of the driving potential.

A second important formula for relating the induced current in the circuit to the induced emf is that giving the phase difference between the current and driving potential. This phase difference, ϕ, between the current and the driving emf is obtained from the standard formula

$$\phi = \tan^{-1}\left(\frac{\omega L - \frac{1}{\omega C}}{R}\right). \tag{5.5}$$

In terms of the driving potential in (5.3) the current in the ring (and in its representative driven circuit in Fig. 5.2b) is then given by [1]

$$I(t) = I_0 \sin(\omega t - \phi) \tag{5.6}$$

where the amplitudes of the current and driving potentials are related through the standard form

$$I_0 = \frac{V_0}{Z(\omega)}. \tag{5.7}$$

From (5.3) through (5.7) the system is seen to exhibit a resonance when $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$. In particular, from (5.4) the impedance has a minimum value of $Z = R$ under these conditions so that for a fixed input voltage amplitude the output current amplitude $I_0 = \frac{V_0}{Z(\omega)}$ attains a maximum value. In addition, from (5.5), under the resonant condition $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$ the current and voltage are both in phase, i.e., $\phi = 0$. These two aspects of resonance are of fundamental importance for the design of metamaterials. One is useful to designs meant to display negative index effects, and one is a design problem which needs to be handled effectively [1].

As shall be shown later, the resonance maximum in the current is a good thing for the design of negative refraction media based on arrays of split ring resonators. On the other hand, the lack of phase difference between the applied voltage and current response at resonance gives rise to a design problem which is a fundamental difficulty in split ring metamaterials. It requires that the split ring exhibits its greatest losses at the frequency the ring exhibits the most interesting response properties [1].

As a consequence of the zero phase difference at the resonant condition, the time average resistive power losses in the circuit at resonance are a maximum given by [1]

$$P_{avg} = \left(\frac{I_0}{\sqrt{2}}\right)^2 R = \left(\frac{V_0}{\sqrt{2}}\right)^2 \frac{R}{Z^2}. \tag{5.8}$$

These loses arise directly from the presence of resistance in the system and are absent in systems without resistance. In the power loss formula the resistance R represents both the natural resistance of the materials forming the split ring and also the radiation resistance of the ring. The radiation resistance component of the total resistance arises from the behavior of the ring as an antenna when the ring carries a time-dependent current. It adds to the resistance coming from the dissipative properties of the materials forming the ring. Many of the proposed designs of split ring features are focused on dealing with the problem of removing or lessening both of these two types of resistance losses from the split ring array.

At resonance the relative phase difference between the driving potential and current is zero. In particular, as the frequency of the driving potential is changed through the resonance condition at $\omega = \omega_0 = \frac{1}{\sqrt{LC}}$ the relative phase changes sign, being different in sign on either side of its zero at the resonance frequency. As shall now be seen, the zero of the phase is the important point for the operation of the split ring resonator in affecting the magnetic response of the split ring to the externally applied field. The type of magnetic response exhibited by the split ring is strongly dependent on the sign of the relative phase of the driving potential and the induced current [1].

As the frequency of the applied field is passes through resonance, the magnetic response of the ring to the applied field passes from a diamagnetic to a paramagnetic type of response. The change from diamagnetic to paramagnetic response is directly linked to the change in sign of the relative phase, ϕ, at the resonance transition. The transition in the magnetic properties of the system is essential in the development of negative refractive index materials. It will now be discussed in this context.

To understand the magnetic response of the split ring to the applied magnetic field, it is necessary to study the magnetic moment of the ring generated by the current induced in it by the external field. This is done by calculating the time averaged magnetic moment studied as a function of the applied field. In particular, for the ring structure in Fig. 5.1 the magnetic moment of the loop, μ, is given by the standard formula [1]

$$\mu(t) = I(t)A, \tag{5.9}$$

in terms of the area of the ring A and the induced current $I(t)$ in (5.6) generated by the applied potential $V(t)$ in (5.3). Equation (5.9) gives the magnetic moment of the ring as a function of time. The object now will be to determine the time average magnetic moment of the split ring as it interacts with the applied time-dependent magnetic field.

The best approach to understanding the time averaged magnetic moment of the split ring is to study the potential energy of the split ring as it interacts with the external applied magnetic field. The magnetic potential energy of the induced magnetic moment of the ring interacting with the applied magnetic field is again obtained using the standard relationship [1]

$$U(t) = -\mu(t)B(t) \tag{5.10}$$

Substituting (5.6) for the induced current and the form for the external field given above (5.3) into (5.10), it follows that

$$U(t) = -AI_0B_0 \cos(\omega t) \sin(\omega t - \phi). \tag{5.11}$$

This represents the instantaneous potential energy of the magnetic moment induced by the applied time-dependent field.

For engineering applications the instantaneous potential energy is not as interesting as the time averaged potential energy which represents the properties of the ring over time scales of interest for device applications. The time scales of interest in device applications are much longer that those of the rapid fluctuations in the variation of the magnetic moment with the applied magnetic field.

Performing a time average of the expression in (5.11) over a period of the time variation of the applied magnetic field, it is found that the average potential energy of the split ring in the applied field is

$$\bar{U} = \frac{AI_0B_0}{2} \sin(\phi) \tag{5.12}$$

From (5.12) it is seen that the average potential energy is proportional to the factor of $\sin(\phi)$. Since the sine is an odd function, it follows that the sign of the potential energy is directly related to that of the phase difference ϕ.

A consequence of the zero phase difference between the induced current and emf at resonance is that at resonance $\bar{U} = 0$. This indicates that the effective magnetic moment of the split ring is zero at the resonant frequency. However, on either side of the resonance frequency the time averaged moment will be either diamagnetic (negative) or paramagnetic (positive).

The regions of interest for meta-material applications are the regions of non-zero magnetic moment. These are at frequencies slightly above or below the resonant frequency. To understand the properties of the average moment in these regions it is useful to expand the frequency of the applied electromagnetic wave about the resonant frequency of the split ring resonator.

Specifically, the frequency of the applied wave is written in the form [1]

$$\omega = \omega_0 + \Delta\omega. \tag{5.13}$$

Applying (5.13) for ω in (5.5), to leading order in $\Delta\omega$ the phase difference between the applied potential and the induced current takes the form [1]

$$\phi \approx 2\frac{L}{R}\Delta\omega. \tag{5.14}$$

In addition, from (5.4) the impedance relating the amplitude of the applied potential to the amplitude of the induced current becomes

$$Z \approx R\sqrt{1+4\left(\frac{L}{R}\right)^2(\Delta\omega)^2}, \tag{5.15}$$

again to leading order in terms of $\Delta\omega$.

The important issues to be addressed in the regions of frequency about the resonant frequency are the behavior of the power loss of the split ring and the determination of the average magnetic moment of the split ring. These two quantities are now studied as function of $\Delta\omega$, applying the expansions in (5.14) and (5.15).

The average power dissipation in (5.8) can be determined in terms of $\Delta\omega$ in the regions about the resonance condition. Applying (5.15) in (5.8) gives the average power to leading order in $\Delta\omega$ in the form [1]

$$P_{avg} = \left(\frac{V_0}{\sqrt{2}}\right)^2\frac{R}{Z^2} \approx \left(\frac{V_0}{\sqrt{2}}\right)^2\frac{1}{R\left(1+4\left(\frac{L}{R}\right)^2(\Delta\omega)^2\right)} \tag{5.16}$$

The maximum power loss is found at the resonant frequency at which $Z = R$. As the frequency changes from resonance, $\Delta\omega \neq 0$ and the average power in (5.16) becomes less than at the resonance condition. Consequently, the properties of the system at resonance are most conducive for the split rings to exhibit dissipative losses.

The average potential energy and the related magnetic moment of the split rings can be studied as functions of $\Delta\omega$ using (5.5), (5.12), (5.14), and (5.15). From these formulae it follows that the time averaged potential energy for the interaction of the split ring with the magnetic field is

$$\bar{U} \approx AI_0B_0\frac{L}{R}\Delta\omega. \tag{5.17}$$

Here the averaged potential energy is seen to be proportional to $\Delta\omega$, with the sign of the average potential energy depending on the sign of $\Delta\omega$.

From (5.17) it follows that for positive $\Delta\omega$ the split ring is diamagnetic and for negative $\Delta\omega$ the split ring is paramagnetic. The effective magnetic moment of the split ring is directly related to the average potential energy of the ring as it interacts with the applied field.

Consequently, the effective magnetic moment is obtained as [1]

$$\mu_{eff} = -\frac{\bar{U}}{B_0} \tag{5.18}$$

Using $V_0 = \omega A B_0$ and (5.7), (5.14), (5.15), and (5.17) in (5.18) it follows that the effective moment

$$\mu_{eff} = -\frac{L}{R^2}\frac{\omega A^2 B_0}{\sqrt{1+4\left(\frac{L}{R}\right)^2(\Delta\omega)^2}}\Delta\omega. \tag{5.19}$$

It consequently follows from (5.19) that the sign of the effective magnetic moment is opposite to that of $\Delta\omega$.

From (5.16) and (5.19) it is seen that near resonance the basic properties of the energy loss and effective magnetic moment of the split ring are simply related to the difference between the frequency of the electromagnetic wave and the resonant frequency of the split ring resonator [7]. These two properties are presented in Fig. 5.3 plotted for the region near the resonant frequency. The power loss in the LRC circuit from (5.16) is plotted in Fig. 5.3a in the region in the close neighborhood of the resonant frequency. The maximum of the power loss at the $\Delta\omega = 0$ resonance condition is clearly observed with a gradual decrease in the scaled frequency variable $2L\Delta\omega/R$. The full width at half maximum of the curve is determined from the condition $L\Delta\omega/R = 1$.

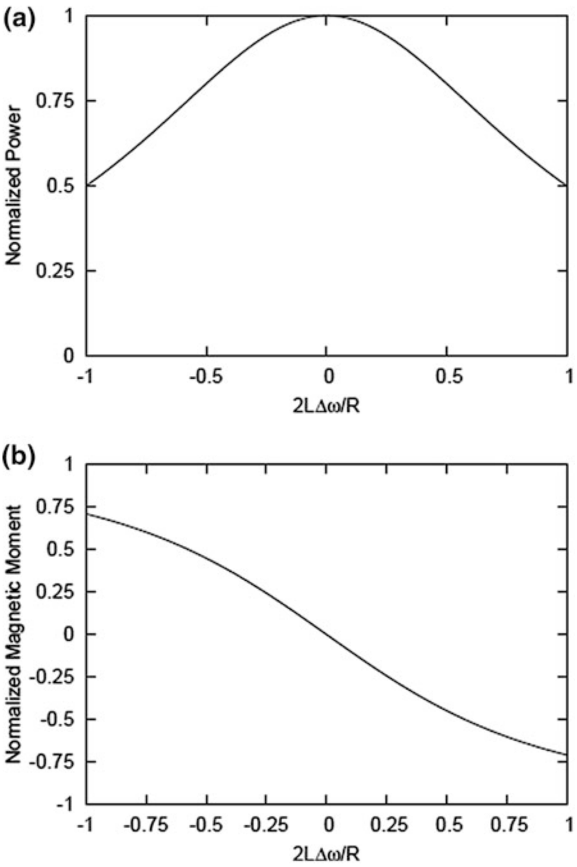
The effective magnetic moment from (5.19) is plotted in Fig. 5.3b where as a function of frequency it is seen to pass through a sign change as the frequency of the applied field goes through the resonance frequency. In this process, the response of the split ring resonator passes through regions of enhanced paramagnetic (positive effective magnetic moment) and diamagnetic responses (negative effective magnetic moment) as the frequency of the electromagnetic wave passes through the resonance of the split ring resonator.

The region of enhanced diamagnetic response has been of great recent interest as in this region it is possible to use split ring resonators to facilitate the design of materials with negative permeability. For the case that the diamagnetic response is made great enough, the split ring resonator will exhibit a negative permeability [1–10]. This is requisite to the design of systems with negative refractive index properties.

A focus in the design of negative refractive index metamaterials is to form an array of split ring resonators which as a whole gives a collective negative effective permeability response to applied electromagnetic waves. The theory presented earlier was focused on the response of a single split ring resonator. It is a pedagogical example. A real array of split ring resonator must take into account the interactions of each split ring with the other split rings of the array [1–10].

In addition, the above theory of the split ring resonator relies on a perturbation expansion in terms of $\Delta\omega$. This was done for the sake of presenting a simple

Fig. 5.3 Power loss of the field to the SRR and the effective magnetic moment of the SRR: **a** the normalized power, $2RP_{avg}/V_0^2$, and **b** the normalized magnetic moment, $2R\mu_{eff}/\omega A^2B_0$, both plotted versus $2L\Delta\omega/R$ where $\Delta\omega = 0$ at resonance



analytic discussion that indicates the basic behavior. In a more realistic treatment a greater range of frequencies in terms of $\Delta\omega$ must be studied. The widening of the region of frequencies is needed in order to enter a region of frequencies where higher order terms of $\Delta\omega$ enter. It is, generally, in this region that the split ring gives a large enough response of the system so that it exhibits a full negative permeability instead of just a diamagnetic response. The separation, however, of the frequency response at the resonant frequency into diamagnetic and paramagnetic regions is not effected by the inclusion of the higher order terms in $\Delta\omega$.

In its self, the simple model of the split ring given in Fig. 5.1 is an over simplification of the type of resonators that are used in the designs of real metamaterials. Various types of modifications are necessary for the effective application of the ideas represented by the simple split ring discussed earlier, and some of these modified forms and the reasons for their modified forms will be discussed later.

Design Considerations

The results in Fig. 5.3 illustrate not only the basic development of the diamagnetic response of the split ring but also a problem with the implementation of split ring resonator meta-materials for their negative refractive index possibilities. The maximum of the losses in the materials occurs in close proximity to the region of enhanced diamagnetism, extending the region of losses into the region of negative refractive index. One of the design problems of meta-materials is to find ways of lessening the losses of the materials while obtaining a good negative index of refraction. The resonance nature of the negative index presents another design problem. In particular, the resonances associated with the negative index response are typically associated with instabilities of the system at a single frequency. This tends to limit the negative index effect to frequency regions close to the isolated resonant frequencies. Consequently, the applications which are discussed later in split ring resonator based metamaterials have predominantly been studied for narrow frequency bands [1–10]. In applying these magnetic design considerations, it must also be remembered that for a material to display a negative index of refraction it must simultaneously have a negative permittivity and permeability [4, 5], and this presents a focus on the associated properties of the permittivity of the total system.

The discussions above have shown that a single split ring resonator can be tuned to exhibit a negative permeability for an electromagnetic wave with a magnetic field polarized perpendicular to the plane of the split ring resonator. However, new problems arise with the incorporation of these properties into the design of a bulk three dimensional metamaterial. Specifically, it is necessary to make an effective three dimensional array of split ring resonators forming a material displaying a three dimensional homogeneous isotropic diamagnetic response to electromagnetic waves propagating in the material.

An example of such a three dimensional array is obtained by placing split ring resonators periodically arranged in the *x*-*y*, *y*-*z*, and *x*-*y* planes of a bulk media. The sets of split ring resonators in these three different planes then forms a three dimensional crystal created by the repetition of the three split ring resonators basis. In addition, the resonance conditions of the split ring resonators of the array must be tuned to give an enhanced diamagnetic response for long wavelength electromagnetic waves for which the metamaterial appears homogeneous. For these wavelengths the dielectric resonance must be sufficiently great that the material exhibits a net negative permeability.

In addition, to the engineered diamagnetic response the background medium component of the metamaterials in which the split ring resonators are arrayed must exhibit a negative permittivity response [1–10]. This may in itself require designs based on inclusions with a background supporting medium. The negative permittivity response, however, is not as problematic as the negative permeability response, and the reader is referred to the literature for the details of these [1–10].

The two conditions of negative permittivity and permeability when successfully achieved combine to create a bulk material with a negative permeability. The theoretical basis for the combination of negative permittivity and negative permeability to form a negative index medium shall be demonstrated in the next sections of this chapter.

As a final point regarding the practical applications of split ring resonators in the construction of metamaterials for specific device applications, remarks are made on some of the structures recently used in engineering realizations of split ring resonator based meta-materials. The first structure proposed as a split ring resonator unit was based on the split ring feature presented in Fig. 5.1 but with just a little more complicated structure.

In particular, Pendry et al. [26] initially proposed a split ring resonator in which two of the C shaped structures in Fig. 5.1 are composed into a single unit. In this structure one large C contains a second small C. Both C's lie in the same plane, and the gaps of the two C's are arranged to be 180° in opposition to one another [3, 26]. These were, subsequently, shown [3, 27, 28] to exhibit essentially the same type of properties as the single ring resonator shown in Fig. 5.1.

In the two C split ring resonator proposed by Pendry et al. [26] the magnetic resonance feature of the unit depends on more parameters of the system than in the single C resonator in Fig. 5.1. Specifically, with the added geometric variables of the resonator structure, variations can be made in the inner and outer radii of the rings, the gaps in the rings, and the gap between the inner and outer C's. This allows for a more flexible system which can be adjusted to meet specifications need for the design of metamaterials for engineering applications.

As an example, the double C structure can be used to reduce electric dipole effects associated with the capacitive gaps of the C's. In the double ring structure, the opposite directed rings tend to cancel the dipole effects from the single ringed structure. This reduces the electric field interactions between neighboring rings from being electric dipole interactions to being of the order of electric quadrapole interactions. In addition, the capacitive effects in the system now include those arising from the capacitance between the inner and outer C rings as well as between the ring gaps [2, 3]. This allows for a greater variety of resonant frequencies to be available from the resonator design.

The Pendry et al., proposal has been applied in a number of experimental treatments. Soon after the double C structure was proposed an experimental application to the study of magneto-inductive waves in a one-dimensional chain based on the double C resonators [19–23, 29]. In addition, the double C resonators were employed [30] as well in the first designs of negative indexed media and cloaking devices [31].

Since Pendry's suggestion of the double C resonator structure, a number of other resonator configurations have been put forth. Some of the modification made in the split ring resonator structure are specifically tailored for the resonator to handle different frequency ranges of applications than those treated in the studies of Pendry et al. These include split ring resonators involving essentially different geometric structures than that of the Pendry double C.

Examples of such different geometry types encountered in resonator unit designs involve features based on: U-shape geometries [32], omega type geometric structures [33], and S-shape resonator geometries [24]. Even the single C-shape structure which in its basic form is shown in Fig. 5.1 has found some device applications [34]. These geometric variations have been developed to handle design problems encountered at various frequency bands and in specifications required in the designs of devices for which reparameterization of the original split ring resonator geometry is not effective.

The various representation of these structures and their resonant properties have been characterized in terms of equivalent circuit models [3, 27, 28, 35]. These models express the nuances of the basic resonance interaction of the resonator with an applied external electromagnetic wave, yielding response properties in terms of their specific structural details. The frequency ranges over which the different types of split ring resonator structures operate to provide a negative permeability response, however, in all cases are found to be limited to a narrow region about a single resonance frequency. This continues to be a basic problem in the design of negative indexed materials on the basis of magnetic resonators.

In the above treatments the basic ideas of the operation of magnetic resonators in the design of negative index metamaterials. For more detains, the reader is referred to the original literature for this field [1–36]. The discussions now turn to the properties of negative refractive index materials themselves, including some remarks about device application of these materials.

5.1.2 Negative Refractive Index Metamaterials

In this subsection the propagation characteristics of an electromagnetic wave traveling in a medium described in terms of a negative permittivity and a negative permeability are discussed [1–5]. The medium is taken to be uniformly homogeneous and isotropic, and the electromagnetic wave is considered to be of the form of a plane wave propagating one-dimensionally through the medium. The resulting dynamics of the system is shown to be characterized as that for a negative refractive index material, with the solution of the simple one-dimensional motion illustrating many of the basic features of electromagnetic waves in a negative indexed medium.

For this system some of the interesting properties of the electromagnetic wave dynamics are determined. These include the unusual relationship between the three vectors of the electric field, the magnetic field, and the wave vector of the propagating electromagnetic wave solutions. Another interesting property involves the relationship between the Poynting vector for the energy flux of the electromagnetic wave through the medium and the wave vector of the electromagnetic wave. These are determined and their unusual properties linked to a display of negative index of refraction [1–5].

Later, these permittivity and permeability properties that characterize the propagation of electromagnetic waves through negative refractive index media will be seen to be of great importance in describing the refraction of waves at the planar interface between two different types of media. In particular, new previously unobserved refractive effects are found in the light traveling between positive and negative indexed media. These effects are at the basis of some of the important proposed device applications that have been set forth in the new metamaterial technology [1–5, 13–23].

To understand the nature of negative refractive materials first consider the characteristics of a plane wave propagating in one dimension within a uniform medium of general permittivity and general permeability. This will be done mathematically in a way so that the sign of the permittivity and permeability never enter into the considerations during the derivation of the dynamics of the wave propagation. At the end of these considerations, however, the permittivity and permeability can be chosen negative. The results at the end of the process then reveal the characteristics of materials with negative index of refraction. These considerations begin by reviewing the mathematics for an electromagnetic plane wave [1, 5, 15] moving along the x-axis with $\vec{E}$ and $\vec{B}$ polarized, respectively, along the y- and z-axes.

For this system it follows from Faraday's law that

$$\frac{\partial E_y}{\partial x} = -\frac{\partial B_z}{\partial t} \tag{5.20}$$

and from Ampere's law that

$$-\frac{\partial B_z}{\partial x} = \mu\varepsilon\frac{\partial E_y}{\partial t} + \mu\sigma E_y \tag{5.21}$$

where ε is the permittivity, μ is the permeability, and σ is the conductivity. In (5.20) and (5.21) there are no restrictions on the ε, μ, and σ other than that they are real and $\sigma \geq 0$. The conductivity in most of the following discussion is considered to be small or zero so that it involves either a very small perturbation or no perturbation on the system.

Under the stated conditions, the plane wave solutions of (5.20) and (5.21) take the general form

$$E_y = E_0 \mathrm{e}^{i(kx-\omega t)} \tag{5.22a}$$

and

$$B_z = B_0 \mathrm{e}^{i(kx-\omega t)}. \tag{5.22b}$$

Upon substituting (5.22) into (5.20) and (5.21) a matrix equation is generated for the field amplitudes. It has the standard matrix form

$$\begin{vmatrix} k & -\omega \\ -(\mu\varepsilon\omega + i\mu\sigma) & k \end{vmatrix} \begin{vmatrix} E_0 \\ B_0 \end{vmatrix} = 0. \tag{5.23a}$$

The solution to the set of linear homogeneous algebraic equations in (5.23a) yields two modal solutions for the amplitude of the electric and magnetic inductance.

The dispersion relation of the plane wave solutions is obtained from the zeros of the determinant of the matrix in (5.23a) so that [1, 5, 15]

$$k^2 - (\mu\varepsilon\omega^2 + i\mu\sigma\omega) = 0. \tag{5.23b}$$

Equation (5.23b) can be treated as an equation for the wave vector k determined at fixed frequency $\omega > 0$. In this way, expanding to first order in σ results in the following set of complex wave vectors

$$k = \pm\sqrt{\mu\varepsilon}\omega \pm \frac{i\mu\sigma}{2\sqrt{\mu\varepsilon}}. \tag{5.24}$$

From (5.24) it is seen that the upper signs come from taking the square root to be a positive number and the lower signs come from taking the square root to be a negative number. The positive sign gives a phase velocity in the positive x-direction while the negative sign gives a phase velocity in the negative x-direction. It should be noted for the following discussions that for the cases in which both the permittivity and permeability are positive or negative, no further considerations of the complex form of the wave vectors are needed.

For the case of a mixture of positive and negative signs in the permittivity and permeability in (5.24), however, the situation is more complex. The added complication in these cases is due to the $\sqrt{\mu\varepsilon}$ factors in (5.24). Nevertheless, such systems involving mixed sign are not of interest for the discussions here and, consequently, will be ignored in the following considerations.

Substituting (5.24) into (5.22) for systems in which the permittivity and permeability are of the same sign gives [1, 5, 15]

$$E_y = E_0 \mathrm{e}^{i(\pm\sqrt{\mu\varepsilon}\omega x - \omega t)} \mathrm{e}^{\mp\frac{\mu\sigma}{2\sqrt{\mu\varepsilon}}x} \tag{5.25a}$$

and

$$B_z = B_0 \mathrm{e}^{i(\pm\sqrt{\mu\varepsilon}\omega x - \omega t)} \mathrm{e}^{\mp\frac{\mu\sigma}{2\sqrt{\mu\varepsilon}}x}. \tag{5.25b}$$

In the case of the upper signs in the exponentials in (5.25), the corresponding field amplitudes determined from (5.23a) are related by

$$\begin{vmatrix} E_0 \\ B_0 \end{vmatrix} = \begin{vmatrix} \frac{1}{\sqrt{\mu\varepsilon}} \\ 1 \end{vmatrix} A \tag{5.26a}$$

where A is a normalizing amplitude of the wave. For solutions with the lower sign in the exponentials in (5.25), the corresponding field amplitudes determined from (5.23a) are related by

$$\left|\begin{matrix} E_0 \\ B_0 \end{matrix}\right| = \left|\begin{matrix} -\frac{1}{\sqrt{\mu\varepsilon}} \\ 1 \end{matrix}\right| A. \tag{5.26b}$$

Here A, again, is the normalized amplitude of the magnetic induction.

From (5.25a) and (5.25b) it is found that in the $\sigma \to 0$ limit the Poynting vector for the plane wave solutions propagating along the x-axis is given by [1, 5, 15]

$$\vec{S} = \frac{1}{2}\frac{1}{\mu}\vec{E} \times \vec{B}^* = \pm\frac{1}{2}\frac{1}{\mu}\frac{1}{\sqrt{\mu\varepsilon}}|A|^2\hat{\imath}. \tag{5.27}$$

The form on the far right hand side of (5.27) is a one-dimensional Poynting vector representing an energy flow along the x-axis.

Furthermore, from the far right expression in (5.27) the upper (lower) sign is for waves propagating to the right (left). This is an important point that will enter into the consideration of negative indexed materials. In particular, it is seen that the net direction of the energy flow along the axis is ultimately determined by the sign of the permeability of the medium. The sign of the permeability can act in (5.27) to reverse the energy flow from being in the direction of the plane wave propagation to being in the opposite direction.

To illustrate this point, consider the following: From (5.27), it follows that in the case of a positive indexed material (i.e., for $\varepsilon, \mu > 0$) the flow of energy in the system is parallel to the wave vector of the plane wave. Specifically, from (5.25) the wave vector for the positive index material is given by $\pm\sqrt{\mu\varepsilon}\omega\hat{\imath}$ which for positive permeability is parallel to the vector in (5.27).

However, from (5.27) in the case of a negative indexed material (i.e., for $\varepsilon, \mu < 0$) the flow of energy in the system is anti-parallel to the wave vector. Under the same considerations as those made for the positive index medium, again for the negative index material the wave vector is given from (5.25) by $\pm\sqrt{\mu\varepsilon}\omega\hat{\imath}$ which for negative permeability is opposite to the vector in (5.27). The signs of the one-dimensional wave vector and Poynting vector, consequently, agree in a positive indexed medium and disagree in a negative index medium.

When a small non-zero conductivity is included in the electromagnetic solutions of (5.25) they exhibit spatially decaying fields. The decay in the fields is a direct result of the dissipated losses in the system introduced by the electric conductivity. These losses transfer the field energy to other degrees of freedom in the system. As a result of this mechanism both the electric field and the magnetic induction solutions given in (5.25) exhibit a decay governed by factors of the form $e^{\mp\frac{\mu\sigma}{2\sqrt{\mu\varepsilon}}x}$, and in the case of zero conductivity the fields are seen to revert to purely propagating plane wave forms.

For the case of a positive indexed material with $\mu\sigma > 0$, the amplitude of the wave is seen from (5.25) and (5.27) to decay in the direction of the energy flow. In particular, it follows from (5.27) that, in the limit of zero conductivity, the energy flow is parallel to the direction of the wave vector, and this is the direction of the amplitude decay. The small conductivity expressions then describe the flow of energy in the direction of the wave vector of the plane wave solution, and as the wave moves through space both its energy and amplitude decrease. These are common behaviors with which physicists and engineers are familiar.

For the case of a negative indexed material with $\mu\sigma < 0$, a quite different, unexpected set of behaviors are found. In particular, as with positive index materials, it follows from (5.25) and (5.27) that the amplitude of the waves again decay in the direction of the energy flow. However, it is seen from (5.27) that in the limit of zero conductivity, the wave vector of a wave propagating in the negative indexed medium is anti-parallel to the energy flow of the wave. This follows from (5.27) because the permeability of the negative index material is negative.

Both the positive and negative indexed media solutions just discussed make physical sense. In both solutions the field dynamics describes energy flows in which the energy of the waves decay as the flow of energy spatially advances through the respective positive or negative index dissipative media. The essential difference in the two types of media, however, is that in a positive index medium the wave vector is parallel to the energy flow whereas in a negative index medium the wave vector is anti-parallel to the energy flow. It shall be seen later, in the discussions of refraction between media, that the property of the parallel or anti-parallel nature of the wave vector and Poynting vector introduces fundamental differences in the refractive behaviors of positive and negative index media.

The anti-parallel nature of the wave vector and Poyning vector in negative refractive index media may at first seem unusual or counterintuitive. That this is not the case can be understood by considering the motion of a broad energy pulse within a negative refractive index medium. This provides an illumination of the detailed motion of energy in a negative index medium. In the following, this motion will be discussed and a comparison made with pulse propagation in a positive index medium.

To establish a comparison, first consider the treatment of a plane wave pulse of radiation propagating in a positive indexed medium, i.e., for $\varepsilon, \mu > 0$. The medium is taken to be non-dissipative, and the pulse is broadly localized in space. Due to the positive index of the medium, the motion of the energy components forming the pulse are then in the direction of their wave vectors. In addition, as a simplification it will be assumed that the pulse envelop is very broad so that only a limited set of wave vectors are summed into the representation of the pulse.

In particular, the electric and magnetic energy density of the pulse propagating in the positive index medium are given, respectively, by [1, 5, 15]

$$U_E = \frac{1}{2}\varepsilon |E(x,t)|^2, \tag{5.28a}$$

$$U_B = \frac{1}{2}\frac{1}{\mu}|B(x,t)|^2. \tag{5.28b}$$

The magnitude of the Poynting vector giving the flow of energy in the direction of the wave vector in the absence of dispersion or for weak dispersion, is related to the energies in (5.28) by

$$S = 2|U_E|v = 2|U_B|v. \tag{5.29}$$

In (5.29) v is the speed of light in the positive index medium and the equality of the energy density of the electric and magnetic fields has been used.

Since $\varepsilon, \mu > 0$ the Poynting vector in (5.28) through (5.29) represents a net flow of positive energy carried by the pulse. This is a standard result found from classical electrodynamics for the propagation of a very broad pulse through space. In particular, in (5.29) the pulse is considered in the limit that it is essentially of uniform energy density extended throughout space.

Now the same treatment is extended to consider the flow of energy of a spatially localized field pulse propagating in a negative indexed medium. In the negative index medium the results in (5.28) and (5.29) also hold for a propagating pulse. However, now $\varepsilon, \mu < 0$ so that the energy densities in (5.28) are negative. As a consequence, the pulse, written in the field amplitudes $|E(x,t)|$ and $|B(x,t)|$, shows up as a spatially localized decrease in energy. This is opposite the case in the positive index $\varepsilon, \mu > 0$ case in which the pulse represented a spatially localized increase in energy.

Though the pulse in the negative index medium moves in a direction parallel to the wave vector, it is a pulse of energy decrease not of energy increase. Consequently, as it moves the pulse decreases the energy in the region to which it travels. This is opposite to the pulse in a positive index medium that increases the energy in the region to which it travels.

When comparing the Poynting vectors of pulses moving in the positive and negative index media, it is useful to note that the flow of an amount of negative energy past a point in space can be viewed as the flow of an equivalent amount of positive energy in the opposite direction. Applying this, the amplitude of the Poynting vector in (5.29) for a negative index medium, then, represents a pulse of energy decrease traveling parallel to the wave vector. It is equivalent to the flow of an energy pulse in the direction anti-parallel to the wave vector. These properties illustrate the essential difference between the Poynting vectors in positive and negative index media.

The earlier discussions of the Poynting vectors in positive and negative index media have a certain analogy with the properties of the current density of electrons and holes in semiconductors. In particular, in the treatment of the current density in semiconductors, positive charged holes move parallel to the electric field to create the same electric current as negative charged electrons moving anti-parallel to the electric field. Consequently, currents arising from positive and/or negative charges cannot, in general, be distinguished. Such a distinction between electron and hole currents can only be made through the application to the system of a symmetry breaking magnetic field.

In the case of the optics problem the motion of pulses in positive and negative index media behave like electrons and holes in their respective n and p type materials. The motion of net electrical charge through the semiconductors is similar to the motion of energy in the optical systems.

In our later discussions, the analogy of the positive and negative index problem in optics with the problem of electrons and holes in semiconductors will be extended to considerations of the flow of light energy past an interface between a positive and negative index medium. In particular, this type of optical interface problem is reminiscent of the problem of the flow of electrical current consisting of electrons and holes passing through a p-n semiconductor junction. At the optical interface a pulse of light energy is an energy particle in a positive indexed optical medium, and in a negative indexed optical medium a pulse of decrease of light energy is an energy hole. The energy moves through positive (negative) indexed media by the motion of energy particles (energy holes), just as electrons carry the current in n-type semiconductors and holes carry the current in p-type semiconductors [12]. Similar to electrons and holes meeting at a p-n junction, the energy particles and holes, upon meeting, can combine to destroy one another.

In the next section, the problem of the interface between a positive and negative index medium are discussed. This problem exhibits many of the interesting properties that arise for technological applications of these types of optical media. In addition, the analogy with the p-n junction in semiconductor physics will be further discussed.

5.1.3 Refraction Between Positive and Negative Index Media

The next level of problem from that of the propagation of a plane wave in a negative index medium is the refraction of a wave at the interface between two different media, one of which having a negative index. This involves the study of a boundary value problem which, at the interface between the two media, applies the standard electrodynamic boundary conditions on the four electromagnetic fields. In their application, the boundary condition relations employed are taken directly from classical electrodynamics without amendment. They are quite general and appropriate for interfaces between either positive or negative index media.

This section focuses on the example of the refraction of a plane wave at the planar interface between a positive and a negative indexed material [1, 5, 15]. In a first study the refraction of light originating in a positive indexed medium is treated as it is refracted into a negative index medium. This is followed by a treatment of the refraction of light originating in the negative indexed medium as it is refracted into a positive index medium.

For these two cases a detailed study of the relationships of the wave vectors and energy flows at the interface is given, and a comparison of these relations with those for a wave traveling between two positive index media is made. The wave vectors and energy flows at the interface between a positive and a negative index medium are found to have an unusual relationship compared to those at the interface between a positive and another positive index medium or between a negative and another negative index medium. The unusual properties at the interface between a positive and a negative index medium are shown to arise from the flow of energy. Specifically, the energy flow and wave vector are oppositely directed in a negative index medium.

The discussions here will only consider the electromagnetic modes of the refraction problem. The important topic of the related surface electromagnetic wave solutions on an interface between positive and negative media will not be treated in this section. The study of surface waves is a very important aspect of the properties of electromagnetic waves at an interface between media and will be treated elsewhere in discussions related to device applications. It is also an important topic in the study of plasmonics.

In particular, following the treatment of refraction in this section, the refraction results are applied as a part of the discussion of the design of a perfect lens [1–5, 13]. The perfect lens involves the refractive properties discussed below as well as the properties of surface electromagnetic waves traveling along the planar interface between a positive and a negative index medium [13]. The additional considerations of surface electromagnetic waves are needed so that all of the optical waves originating from the optical object in the lens system can be assembled by the perfect lens into the image generated by the lens. This involves propagating as well as evanescent waves arising in the object-lens-image system. The discussions of surface electromagnetic waves will, however, be included in the course of the presentation of materials on the perfect lens.

For the most part, in the discussions of refractive effects, surface electromagnetic waves, and the perfect lens the focus will be on lossless media. Some discussions at the end of the chapter will be made on attempts to overcome the losses in metamaterials implementing magnetic resonance effects in their designs.

Light Originating in the Positive Index Media

In this subsection, a treatment is given of the refraction of light at the planar interface between a positive and negative index medium. For the considerations, a plane wave of light is incident on the interface from the side of the positive index medium, with part of the incident light being refracted at the interface while part is reflected at the interface. To keep the consideration brief a focus will be on the case of light polarized with the magnetic field perpendicular to the plane of incidence. Once this case is understood the other case of the electric field polarized perpendicular to the plane of incidence can be easily worked out.

In the geometry of the interface problem (given in the schematic diagram in Fig. 5.4) the positive indexed medium called "medium 1" is in the region $y > 0$, and the negative indexed called "medium 2" is in the region $y < 0$. Medium 1 is described by permeability and permittivity parameters $\mu_1, \varepsilon_1 > 0$ of a positive index medium, and medium 2 is described by the corresponding parameters $\mu_2, \varepsilon_2 < 0$ of a

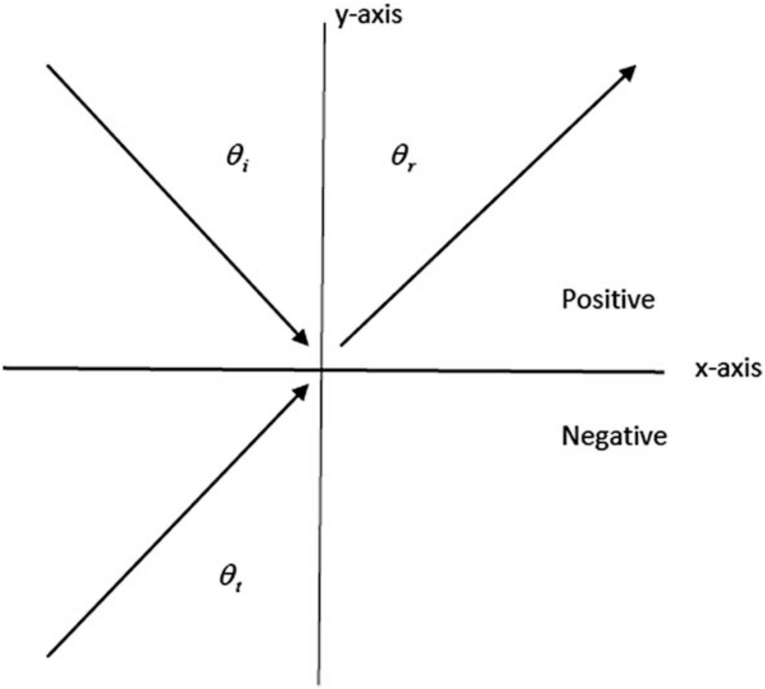


Fig. 5.4 Refraction at planar interface for incident light in a positive medium and refracted light in a negative medium. The arrows indicate the wave vectors and as represented in the figure $\theta_i, \theta_r, \theta_t > 0$

negative index medium. Medium 1 is taken to be the medium containing the incident wave.

The expressions for the incident and reflected waves in medium 1 are obtained as general forms of the solution of the Maxwell equation in the uniform homogeneous medium characterized by ε_1, μ_1. For the expression of these solutions, the various positive angles describing the incident, reflected as well as the refractive plane waves about the interface are indicated in the schematic diagram given in Fig. 5.4. The boundary value problem at the interface is then completely specified within the context of the variables defined in Fig. 5.4.

In terms of these angles the wave incident on the interface from medium 1 is of the form [1, 5, 13, 15]

$$\vec{E}_I = E_I(\cos\theta_i\hat{i} + \sin\theta_i\hat{j})e^{i[k(\sin\theta_i x - \cos\theta_i y) - \omega t]}, \tag{5.30a}$$

$$\vec{B}_I = E_I\frac{k}{\omega}\hat{k}e^{i[k(\sin\theta_i x - \cos\theta_i y) - \omega t]}, \tag{5.30b}$$

where the positive angle θ_i is shown in Fig. 5.4. The wave is assumed to be polarized with the magnetic field perpendicular to the plane of incidence.

Similarly, the wave reflected at the interface back into medium 1 is given by

$$\vec{E}_R = E_R(-\cos\theta_r\hat{i} + \sin\theta_r\hat{j})e^{i[k(\sin\theta_r x + \cos\theta_r y)-\omega t]}, \tag{5.31a}$$

$$\vec{B}_R = E_R\frac{k}{\omega}\hat{k}e^{i[k(\sin\theta_r x + \cos\theta_r y)-\omega t]}, \tag{5.31b}$$

where the positive angle θ_r is shown in Fig. 5.4 and the magnetic field is again perpendicular to the plane of incidence.

In both the expression for the incident and reflected waves the wave number for propagation in the positive index medium is given in terms of the frequency, permeability and permittivity by the expression

$$k = \sqrt{\mu_1\varepsilon_1}\omega. \tag{5.32}$$

For the solutions in medium 1, described by (5.30) through (5.32), the relationship between the incident and reflected wave amplitudes (i.e., E_I, E_R, respectively) are obtained as part of a boundary value problem considered later at the interface between media 1 and 2.

From the form of the solutions in (5.30) and (5.31), the wave vectors of the incident and reflected waves in the positive medium are given by [1, 5, 15]

$$\vec{k}_i = k(\sin\theta_i, -\cos\theta_i, 0), \tag{5.33a}$$

$$\vec{k}_r = k(\sin\theta_r, \cos\theta_r, 0). \tag{5.33b}$$

The wave vector in (5.33a) represents a wave with a phase velocity directed towards the interface whereas the wave vector in (5.33b) represents a wave with a phase velocity directed away from the interface. From (5.30) through (5.33) the respective Poynting vectors of the incident and reflect waves is obtained from the standard expression for the Poynting vector in classical electrodynamics.

In this way it follows that [1, 5, 15]

$$\vec{S}_I = \frac{1}{2}\frac{1}{\mu_1}\frac{1}{\omega}|E_I|^2\vec{k}_i \tag{5.34a}$$

is Poynting vector of the incident wave and

$$\vec{S}_R = \frac{1}{2}\frac{1}{\mu_1}\frac{1}{\omega}|E_R|^2\vec{k}_r \tag{5.34b}$$

is the Poynting vector of the reflected wave. Equations (5.33) and (5.34) show the standard result of the interface problem, indicating an energy flow, respectively, into and out of the interface between the two media. Note that since $\mu_1 > 0$ in the positive index medium, the Poynting vectors of both the incident and reflected waves are parallel to their respective wave vectors.

The next part of the solution to consider is the refracted wave in medium 2. This part of the problem is in the region located below medium 1 and the interface. In this region medium 2 has a negative index (i.e., $\varepsilon_2, \mu_2 < 0$) so that additional care must be taken in its treatment. This must be done while choosing a solution for the refracted wave in medium 2 that represents an energy flow away from the interface. In particular, the energy refracted at the interface into medium 2 must travel away from the interface so that energy does not collect at the interface between the two media. At the same time the wave vector of the solution in medium 2 must correctly match the interface boundary conditions with the two waves in medium 1.

Under these conditions, the correct form of the solution for the transmitted wave is

$$\vec{E}_T = E_T(-\cos\theta_t\hat{i} + \sin\theta_t\hat{j})e^{i[q(\sin\theta_t x + \cos\theta_t y) - \omega t]}, \tag{5.35a}$$

$$\vec{B}_T = E_T\frac{q}{\omega}\hat{k}e^{i[q(\sin\theta_t x + \cos\theta_t y) - \omega t]}, \tag{5.35b}$$

where the positive angles defined by Fig. 5.4 are used for the case of a negative medium 2. Consequently, the positive angle θ_t is located in the third quadrant, being measured from the y-axis.

In both (5.35) expressions for the refracted wave the wave number for propagation in the negative index medium is given in terms of the frequency, permeability, and permittivity by the expression

$$q = \sqrt{\mu_2\varepsilon_2}\omega. \tag{5.35c}$$

In the system in medium 2, described by (5.35), the relationship between the incident and transmitted wave amplitudes (i.e., E_I, E_T, respectively) are obtained as part of a boundary value problem considered later at the interface between media 1 and 2.

From the general form of the transmitted wave solution in (5.35), it is seen that the wave vector in medium 2 is given by [1, 5, 15]

$$\vec{q}_t = q(\sin\theta_t, \cos\theta_t, 0) \tag{5.36}$$

where $q = \sqrt{\mu_2\varepsilon_2}\omega$. In this case the wave vector of the transmitted wave in (5.36) is found to exhibit the counterintuitive property that it points from the medium containing the refracted wave towards the medium containing the incident wave. This is not a difficulty as medium 2 is a negative refractive index medium.From earlier discussions, it is known that the flow of energy in a negative index medium is in a direction opposite that of the wave vector. Consequently, the direction of the wave vector in (5.36) is necessary to account of the flow of energy away from the interface.

From (5.35) and (5.36) and the standard form of the Poynting vector from classical electrodynamics it follows that the Ponyting vector of the transmitted wave solution take the form [1, 5, 15]

$$\vec{S}_T = \frac{1}{2}\frac{1}{\mu_2}\frac{1}{\omega}|E_T|^2\vec{q}_t. \tag{5.37}$$

It is seen from (5.36) that since $\mu_2 < 0$ in the negative index medium, the transmitted wave Poynting vector represents an energy flow opposite q_t. In particular, the energy in the transmitted wave travels away from the interface and into the third quadrant.

Another important feature of the form of the solution in (5.35) through (5.36) involves the component of the wave vector of the solution along the interface. The electromagnetic boundary conditions at the interface require that the component of the wave vectors of the incident, reflected, and refracted waves agree at the interface. This follows directly as a consequence of the translational symmetry of the interface. The form of the wave vector chosen in (5.36) is made so as to provide a basis for this agreement. The details of this condition and its fulfilment will be discussed again later.

The agreement of the interface wave vectors and the requirement of energy flow away from the interface are seen to set the general form of the solution. Before finishing the boundary condition problem to determine E_r, E_T in terms of E_I, the refraction problem in which medium 2 is a positive index medium will be set up. This facilitates a comparison of the two problems.

In order to provide an understand of the new aspects of negative index media in refraction problems, it is useful to make a comparison of the results obtained in (5.30) through (5.37) for a negative index medium 2 to the case in which medium 2 is a positive indexed medium. This provides for a detailed study of where, specifically, the differences in the two problems arise. In the following, the calculations in (5.30) through (5.37) are repeated, considering the case in which medium 2 is positive indexed (i.e., $\varepsilon_2, \mu_2 > 0$).

In the case of a positive index medium 2, the form of the solution in (5.35) for the transmitted wave must be changed. The transmitted wave in media 2 is now given by

$$\vec{E}_T = E_T(\cos\theta_t\hat{i} + \sin\theta_t\hat{j})e^{i[q(\sin\theta_t x - \cos\theta_t y) - \omega t]}, \tag{5.38a}$$

$$\vec{B}_T = E_T\frac{q}{\omega}\hat{k}e^{i[q(\sin\theta_t x - \cos\theta_t y) - \omega t]}, \tag{5.38b}$$

where, in the case of a positive medium 2, the positive angle θ_t must be in the fourth quadrant and measured from the y-axis.

From the form of the solution in (5.38), it is now seen that the wave vector of the transmitted wave written as [1, 5, 15]

$$\vec{q}_t = q(\sin\theta_t, -\cos\theta_t, 0) \tag{5.39}$$

where in terms of the positive permeability and permittivity and the frequency of the wave

$$q = \sqrt{\mu_2\varepsilon_2}\omega. \tag{5.40}$$

The wave vector of the refracted wave is now directed from medium 1 to medium 2, pointing away form the interface and into medium 2.

Computing the Poynting vector of the solution in (5.38) through (5.40) the flux of energy from through the interface is given by

$$\vec{S}_T = \frac{1}{2}\frac{1}{\mu_2}\frac{1}{\omega}|E_T|^2\vec{q}_t. \tag{5.41}$$

Since for the positive index medium $\mu_2 > 0$, the wave vector and Poynting vector are now parallel to one another. As a result the energy of the transmitted wave again flows away from the interface.

The energy flow generated from (5.38) is now into the fourth quadrant. This comes from the boundary conditions at the interface and the translational symmetry of the system along the interface. Specifically, the components of the wave vectors parallel to the interface between the two media must be the same for the incident, reflected, and refracted waves. The form for the fields in (5.38) are the only possible solutions satisfying the conditions on the energy flow and the wave vector.

To summarize the two problems discussed above: First consider the solutions in (5.35) through (5.37) and (5.31) through (5.34) for the refraction between the positive and negative index media. Start by comparing the solutions in (5.35) through (5.37) for the negative indexed medium 2 with the solutions for the reflected wave for the positive indexed medium 1 given in (5.31), (5.33b), and (5.34b). Both solutions have wave vectors with positive y-components. However, the two solutions both represent energy flows away from the $y = 0$ plane in opposite directions along the y-axis. This is due to the sign difference between $\mu_1 > 0$ and $\mu_2 < 0$ in the Poynting vectors obtained from their solutions.

From (5.35) and (5.36) for the transmitted wave for the $\mu_2 < 0$ medium the wave vector and Poynting vectors are anti-parallel. In this case the wave vector has a positive component along the y-axis, while from (5.36) the energy flow of the transmitted wave is in the negative y-direction. The energy flow of the transmitted wave is, consequently, directed away from the interface. The reflected wave, however, is in medium 1 so that its wave vector and Poynting vectors are parallel and the reflected wave has an energy flow along the positive y-axis.

For the case in which medium 2 is a positive index medium, the behavior of the wave vectors and energy fluxes change from the negative index case. For the solution of this case, treated in (5.38) through (5.41), all of the energy flows are in the direction of the wave vectors of the solutions. The reflected and refracted waves have oppositely directed y-components of their wave vectors and, consequently, their energy fluxes are oppositely directed.

As mentioned earlier the translational symmetry of the planar interface at $y = 0$ between media 1 and 2 requires all of the x-components of the wave vectors of the incident, reflected, and refracted waves to agree. [In particular, this follow from applying boundary conditions to the solutions in (5.30), (5.31), (5.35), and (5.38).] This requires that in the reflected solutions

$$\theta_i = \theta_r \tag{5.42a}$$

and in the refracted wave solutions

$$k \sin\theta_i = q \sin\theta_t. \tag{5.42b}$$

These results arise solely from the translational symmetry and are independent of whether or not media 1 and 2 are positive or negative indexed materials.

In applying (5.42b) it should be remembered that the angle θ_t has been defined differently for the positive and negative index problems. In the above discussions for positive index media postive θ_t has been measured anti-clockwise from the negative y-axis, while for negative index media positve θ_t has been measured clockwise from the negative y-axis. If in the negative index media positive θ_t is redefined to be measured anti-clockwise from the negative y-axis, then in Snell's law in (5.42b) $\theta_t \to -\theta_t$. As a consequence of this Snell's law for the redefined angle now reads

$$k \sin\theta_i = -q \sin\theta_t \tag{5.42c}$$

or

$$\sqrt{\mu_1\varepsilon_1}\sin\theta_i = -\sqrt{\mu_2\varepsilon_2}\sin\theta_t. \tag{5.42d}$$

Written in terms of the newly defined θ_t Snell's law in (5.42c) and (5.42d) appears to determine the refraction for an interface between a positive and negative index medium where the sign of the refractive index is made clear. In this formulation, the refractive index of medium 2 is negative.

The field solutions for the refraction problem at the interface of positive and negative index media are formulated in (5.30), (5.31), and (5.35), while the field solutions at the interface of two different positive index media are formulated in (5.30), (5.31), and (5.38). In order to determine the electric field amplitudes from these two sets of solutions, it is necessary to match the boundary conditions at the interface between the two media. The boundary conditions are the same at the interface between a positive and negative index medium and between a two positive index media. They involve the continuity of the component of the electric field tangent to the interface and the component of the magnetic field tangent to the interface. Following from this application the fields E_R, E_T are expressed in terms of E_I.

From the continuity of the electric fields tangent to the interface it follows that

$$E_I - E_R = \alpha E_T, \tag{5.43a}$$

where $\alpha = \frac{\cos\theta_t}{\cos\theta_i}$ for a positive indexed medium 2 and $\alpha = -\frac{\cos\theta_t}{\cos\theta_i}$ for a negative indexed medium 2. Similarly, from the continuity of the component of the magnetic field tangent to the interface it follows that

$$E_I + E_R = \beta E_T, \tag{5.43b}$$

where $\beta = \frac{\mu_1\sqrt{\mu_2\varepsilon_2}}{\mu_2\sqrt{\mu_1\varepsilon_1}}$. In (5.43b) β is the same for both a positive or a negative indexed medium 2.

Equations (5.43) are then general expressions, valid for an interface between a negative and a positive index medium or between two different positive index media. The difference between these two problems is that α and β are defined differently for the two cases. Solving (5.43) the three field amplitudes are given by

$$E_T = \frac{2}{\alpha+\beta}E_I, \tag{5.44a}$$

$$E_R = \frac{\beta-\alpha}{\alpha+\beta}E_I \tag{5.44b}$$

for the appropriate α and β defined below (5.43a) and (5.43b).

Equations (5.44) solves the problem of the refraction of light for light incident from a positive index medium onto an interface with a medium of negative index of refraction. For the study of the perfect lens it is useful to also consider the solution in the case that the light is incident from the negative index of refraction onto an interface with a positive index of refraction. These solutions along with a treatment of the properties of surface plasmon-polaritons at the interface of positive and negative index media are a basis for the treatment of the properties of a perfect lens. The next section handles refraction going from a negative index medium into a positive index medium.

Refraction of Light Originating in the Negative Indexed Media

In the case of light going from a negative index medium to a positive index medium, the considerations are somewhat similar to those in the refraction problem just treated. The planar interface is between a negative indexed medium 1 (i.e., $\mu_1, \varepsilon_1 < 0$) in the region $y > 0$, and a positive indexed medium 2 (i.e., $\mu_2, \varepsilon_2 > 0$) is in the region $y < 0$. This is illustrated by the schematic figure presented in Fig. 5.5.

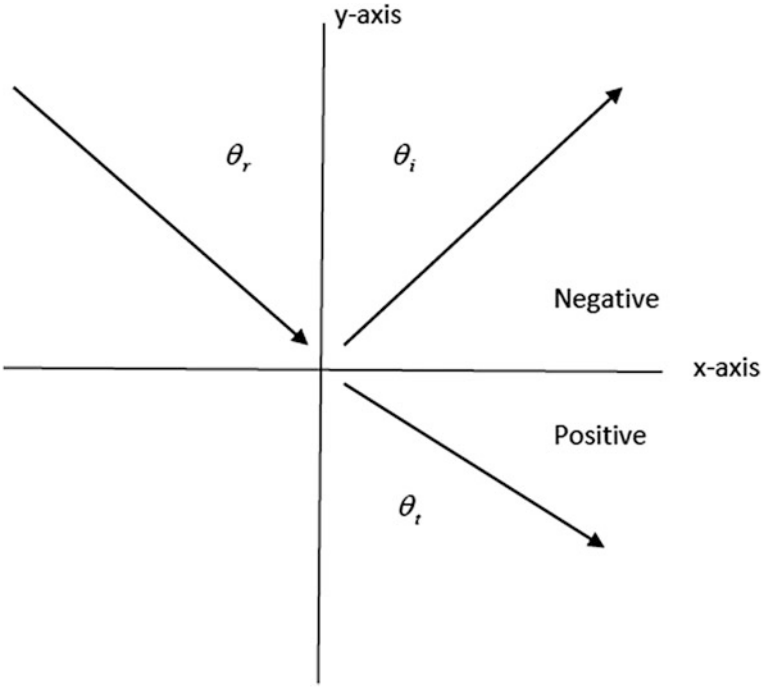


Fig. 5.5 Refraction at a planar interface for incident light in a negative index medium and refracted light in a positive index medium. The arrows indicate the wave vectors and as represented in the figure $\theta_i, \theta_r, \theta_t > 0$

Now, however, the incident wave is in the negative indexed medium located above the interface, and the transmitted wave is in the positive indexed medium [1, 5, 15] located below the interface.

For the positive angles defined in Fig. 5.5 the wave incident on the interface from medium 1 has the form

$$\vec{E}_I = E_I(\cos\theta_i\hat{i} - \sin\theta_i\hat{j})e^{i[k(\sin\theta_i x + \cos\theta_i y) - \omega t]}, \tag{5.45a}$$

$$\vec{B}_I = -E_I\frac{k}{\omega}\hat{k}e^{i[k(\sin\theta_i x + \cos\theta_i y) - \omega t]}. \tag{5.45b}$$

Here again only the polarization with a magnetic field perpendicular to the plane of incidence is studied. The other polarization is left to be worked out by the reader. The form of the reflected wave in medium 1 is then given by

$$\vec{E}_R = E_R(\cos\theta_r\hat{i} + \sin\theta_r\hat{j})e^{i[k(\sin\theta_r x - \cos\theta_r y) - \omega t]}, \tag{5.46a}$$

$$\vec{B}_R = E_R \frac{k}{\omega} \hat{k} \mathrm{e}^{i[k(\sin\theta_r x - \cos\theta_r y) - \omega t]}, \tag{5.46b}$$

where the polarization of the reflected wave is maintained during the interaction with the interface between the two media.

The wave vectors of the incident and reflected waves in (5.45) and (5.46) are, respectively,

$$\vec{k}_i = k(\sin\theta_i, \cos\theta_i, 0), \tag{5.47a}$$

$$\vec{k}_r = k(\sin\theta_r, -\cos\theta_r, 0), \tag{5.47b}$$

where the wave number in (5.47) is related to the frequency, permittivity, and permeability by the expression

$$k = \sqrt{\mu_1 \varepsilon_1}\omega. \tag{5.48}$$

From (5.47a) the wave vector of the incident fields within medium 1 points away from the interface whereas the wave vector of the reflected fields within medium 1 points towards the interface. The directions of the two wave vectors seem anomalous but are due to the negative index of refraction of the medium 1. This is clarified from a discussion of the energy flux in these two waves, obtained by considering their Poynting vectors.

The Poynting vector of the incident and reflected waves is computed using the standard expression from classical electrodynamics. Applying this expression for the Poynting vector to the fields in (5.45) and (5.46) the energy flux of the incident field is given by [1, 5, 15]

$$\vec{S}_I = \frac{1}{2}\frac{1}{\mu_1}\frac{1}{\omega}|E_I|^2 \vec{k}_i, \tag{5.49a}$$

and the energy flux of the reflected fields is

$$\vec{S}_R = \frac{1}{2}\frac{1}{\mu_1}\frac{1}{\omega}|E_R|^2 \vec{k}_r. \tag{5.49b}$$

Equations (5.47) and (5.49) show the energy flux into and out of the interface of the two media.

In the negative index medium 1, $\mu_1 < 0$. Consequently, the energy flux of the incident and reflected waves in medium 1 is opposite to their wave vectors. While the wave vector of the incident wave points away from the interface the energy flux of the incident wave is towards the surface, and while the wave vector of the reflected wave points towards the interface the energy of the reflected wave is away from the interface. This is a general property of propagation in a negative indexed medium.

In the problem considered in Fig. 5.5, medium 2 containing the transmitted wave is a positive index medium. For this case $\varepsilon_2, \mu_2 > 0$, and the transmitted wave in medium 2 is given by

$$\vec{E}_T = E_T(\cos\theta_t\hat{i} + \sin\theta_t\hat{j})e^{i[q(\sin\theta_t x - \cos\theta_t y) - \omega t]}, \tag{5.50a}$$

$$\vec{B}_T = E_T\frac{q}{\omega}\hat{k}e^{i[q(\sin\theta_t x - \cos\theta_t y) - \omega t]}. \tag{5.50b}$$

In (5.50) the wave vector of the transmitted wave is seen to be given by the form

$$\vec{q}_t = q(\sin\theta_t, -\cos\theta_t, 0), \tag{5.51}$$

where the wave number is written in terms of the frequency, permittivity, and permeability as

$$q = \sqrt{\mu_2\varepsilon_2}\omega. \tag{5.52}$$

For the case of the positive index medium the transmitted wave vector is observed to point away from the interface between the two media.

From the fields in (5.50) the Poynting vector of the transmitted wave is given by

$$\vec{S}_T = \frac{1}{2}\frac{1}{\mu_2}\frac{1}{\omega}|E_T|^2\vec{q}_t, \tag{5.53}$$

where now for the positive permeability μ_2 the Poynting vector is parallel to the wave vector of the transmitted wave. In the positive index media below the interface both the wave vector and the energy flux are directed away from the interface.

Again, a consideration of the translational symmetry parallel to the interface of the two media requires the equality of the x-components of the wave vectors of the incident, reflected, and transmitted waves. A consequence of this is that the incident, reflected, and transmitted waves obey the angular conditions

$$\theta_i = \theta_r \tag{5.54a}$$

and

$$k\sin\theta_i = q\sin\theta_t. \tag{5.54b}$$

The first relation in (5.54a) is the standard condition between the angle of incidence and the angle of reflection known as the law of reflection. The second relation is Snell's law which, as in the discussion of (5.42), can be rewritten to manifestly exhibit a negative index of refraction. This involves redefining the positive sense of the angles in Fig. 5.5 and is left to the reader.

Both the incident and reflected field amplitudes can be written in terms of the amplitude of the incident wave. This involves an application of the boundary conditions at the interface between the two media. The first condition is the continuity of the tangential component of the total electric field at the interface. Applying this condition to (5.45), (5.46), and (5.50) requires that

$$E_I + E_R = \alpha E_T. \tag{5.55a}$$

Here $\alpha = \frac{\cos\theta_t}{\cos\theta_i}$ for the case in which medium 1 has a negative index of refraction and medium 2 has a positive index of refraction.

The second boundary condition is the continuity of the tangential component of the total magnetic field at the interface. From (5.45), (5.46), and (5.50) this requires

$$-E_I + E_R = \beta E_T. \tag{5.55b}$$

Here $\beta = \frac{\mu_1\sqrt{\mu_2\varepsilon_2}}{\mu_2\sqrt{\mu_1\varepsilon_1}}$ between a negative index medium 1 and a positive index medium 2.

Solving (5.55) gives [1, 5, 15]

$$E_T = \frac{2}{\alpha - \beta} E_I, \tag{5.56a}$$

$$E_R = \frac{\alpha + \beta}{\alpha - \beta} E_I, \tag{5.56b}$$

relating the reflected and transmitted wave amplitudes to that of the incident wave. Using (5.56) the full results, discussed above, for the field amplitudes and energy flux can now be expressed in terms of the amplitude of the incident fields.

An Analogy with Semiconductors

The refraction of light at an interface between positive and negative indexed media has an analogy with electron and hole currents passing through an n-p or p-n semiconductor junction [1, 12]. In a semiconductor the electrons and holes are discrete particles of charge with motions that can both transfer a net amount of charge from one point to another in an electrical system. A current of electrons flows in an n-type semiconductor, while a current of holes flows in a p-type semiconductor. At an n-p or p-n junction the nature of the carries and their drift velocities changes as one passes through the interface of the junction.

Similarly, an electromagnetic pulse in positive or negative indexed media, respectively, either carries energy or an energy decrease. In this way the electromagnetic pulses, in their motion, accomplish a net energy transfer from point to point in an optical system. As with electrons and holes in the electrical system, pulses of energy or an energy decrease in the optical system move differently in the two types of dielectric media. In particular, pulses in positive index media move in the direction of the energy flow whereas pulses in negative indexed media move opposite the direction of energy flow. At the interface between a positive and negative indexed medium the energy flow in the light transferred through the system must change from pulses of energy to pulses of energy decrease in the respective supporting media. They do this in a manner which provides for the net flow of energy through the system.

Consider light traveling from a positive to a negative index medium. The light incident on the interface from the positive indexed medium is in the form of an energy pulse with its motion directed towards the interface. As the pulse travels towards the interface, it interacts with a pulse of energy decrease coming towards the surface from the negative indexed medium. This response of the system arises through the solution of the electromagnetic boundary value problem at the interface which links the fields in the positive and negative indexed media. In this way the pulse of energy decrease in the negative indexed medium exists as a response to the fields from the incident energy pulse at the interface.

The incident energy pulse and the pulse of energy decrease destroy one another as they meet at the interface. In the process of their destruction, however, they create a reflected energy pulse at the interface in the positive index medium. The reflected pulse created in this manner travels away from the interface and into the bulk of the positive indexed medium. The net result of the refraction process in the system is the creation of two separate pulses providing a net energy flow moving away from the interface in both the positive and negative indexed media.

In the case in which light traveling in a negative index medium is incident on the interface with a positive index medium, a pulse of energy decrease propagates in the negative medium away from the interface. This creates an incident energy flow towards the interface. From the boundary value problem in the two media, the fields of the pulse of energy decrease in the negative index medium causes an energy pulse to be created in the positive index medium. The created pulse of energy moves away from the interface and into the bulk of the positive indexed medium. This represents the transmitted energy flux in the positive index medium.

In addition, as part of the same boundary value problem, an additional pulse of energy decrease is created in the negative index medium. The additional pulse of energy decrease travels towards the interface through the negative indexed medium and represents the flow of reflected wave energy in the negative index medium. The reflected wave energy, from the pulse of energy decrease traveling in the negative index medium towards the interface, represents a reflected energy flow from the interface.

The energy flows, then, correctly describe the flux of incident, reflected, and transmitted energy even though the pulse propagation in the media appears to be unorthodox. For a net incident flow of energy to the interface, the net result in the final state of the system is energy moving away from the interface and into the bulk of both the positive and negative indexed media.

To conclude, in both of the refraction problems discussed earlier, the energy transport properties are determined by the differing natures of the energy pulses in the positive indexed media and the pulses of energy decrease in the negative indexed media. The essential consideration in treating energy flow in these two different media is that the energy flux is parallel to the wave vector in positive indexed media and anti-parallel to the wave vector in negative indexed media. This is due to the differing nature of the energy content contained by the pulses in the two types of media which is ultimately related to the different signs of the permittivity and permeability in the two media.

In an analogy with the physics of n-p and p-n semiconductor junctions, at the interface between positive and negative index media of the optical system, the energy pulses and pulses of energy decrease in part destroy one another at the interface between positive and negative index media. The destruction at the interface then gives rise to various reflected and transmitted flows of energy in the dielectric media [1, 12].

5.2 Perfect Lens

The refraction of light at the interface between different optical media is commonly used in the important application of lens design [1–5, 13]. For these optical functions a basic lens is composed of two interfaces that are designed so that light from the position of an object or source is steered through the lens to form an image at another position in space. The image created by the lens may be magnified as well as shifted in its angular orientation in space relative to that of the object. In addition, as the light flows through the system and passes through the position of the image it appears to come from the image rather than the object from which it originated. This shifts the position at which one perceives the location of the object.

Magnification, location, and orientation are the essential relationships of the image to the object that are managed by the lens. In this way lenses form the fundamental components in optical microscopes, telescopes, spectrometers, and many other such optical instruments meant to manipulate the appearances of optical images in space.

In classical optics, in which the design of lenses is based on materials with positive index of refraction, lenses are composed of curved surfaces. These surfaces are arranged so that the lens consists of a finite bounded volume of optical medium. As a result such lenses usually present a circular aperture to the incident light, and this turns out to be one of the many fundamental limitations on their designs and on the optical properties they exhibit. In particular, the image resolution of the lens is related to the finite aperture of the lens [1–5, 13].

There are many other problems to be overcome in the design of a lens based on media of positive index of refraction. These must be addressed in order for the lens to provide a good image, i.e., a nicely focused representation of the object presented with a high degree of accuracy. From the standpoint of fabrication the presence of impurities in the surface curvature and in the optical medium need to be addresses as well as energy losses and the dispersive properties in the optical material. The two last mentioned of these factors provide a fundamental restriction on the lens design while the first two are problems of engineering implementation or fabrication. All of these factors have been addressed in the many years that classical optical systems have been studied [1, 4, 13] so that optics has become one of the most accurate areas of study in the sciences.

In the optics of positive indexed materials there is a fundamental limitation on the refraction at the interface between two media. This was discussed earlier where it was shown that in such systems the refraction of a ray of light incident on a planar interface from the first or second quadrant can only be refracted, respectively, into the third or fourth quadrants [1, 4, 13]. Due to these restrictions, a focusing lens from positive index material must have at least one concaved or convex curved surface. The presence of curvature in the lens surface allows for an additional bending of a ray of light from that arising solely from the difference in the index of refraction between the media forming the interface. This limitation on the optics of positive index materials is seen in the simple case of a single planar surface.

While a dielectric mismatch between two positive index media at a planar interface is unable to focus an image, a dielectric mismatch between a positive and negative index medium is by itself able to focus an image. The reason for this is due to the much larger change in the path of the refracted light from that of the incident beam that is obtained at the interface between positive and negative index media. In the system of two positive index media a planar interface between the media cannot cause the refracted light to cross the optical axis of the system. This is a fundamental difference between positive-positive and negative-positive interfaces.

Unlike negative index systems, however, in positive index media a result of the curvature of the lens surfaces is that the two surfaces intersect one another in a circle of radius R. This intersection is the aperture of the lens. The finite radius of the lens aperture is known to provide a fundamental limitation on the resolution of the image formed by the lens. Only wavelengths less that the diameter of the lens, when emitted by a source, can pass through the lens and arrive at the focus of the lens in a way so as to contribute to the formation of a well resolved image of the source.

A general criterion for the focusing power of a lens in positive index media systems involves the wavelengths of the light from the source and the radius of the lens aperture. This occurs in classical optics in the form of the Rayleigh focusing criteria. In forming an image of an object, only wavelengths of light, λ, satisfying

$$\lambda < R \tag{5.57}$$

are resolved by the lens in the focused image of the object.

Lens of finite apertures are then essentially imperfect due to this limitation of their focusing ability arising from finite R of the lens aperture. To give a perfect image of the object, the image formed by the lens should contain all Fourier components of light rather than a restricted subset satisfying (5.57). As shall be shown later, a perfect image must include all of the propagating components for the source as well as all of the evanescent components generated by the source.

The new optics of negative indexed materials allows for the design of lenses which, in principle, can have infinite apertures. This can be seen from the earlier discussions of the refraction of light at the interface between a positive indexed medium to a negative indexed medium. At such an interface the refraction of a ray of light incident on a planar interface from the first or second quadrant of a positive indexed medium is refracted into the fourth or third quadrants of the second negative indexed medium, respectively. This means that the additional beam bending arising from a curved surface between the two media is not necessary and precludes the need for additional bending from a curved surface. A consequence of this, which shall be treated below, is that it is possible to design a lens that is composed as an infinite slab of negative refractive medium. The details of such a design will be discussed later.

Here it shall only be pointed out that a lens with an infinite slab geometry displays an infinite aperture with $R \to \infty$ in (5.57) of the Rayleigh focusing criterion. As a result, in principle the lens is able to focus into the image it forms all of the propagating wavelength components of the object or source received into the lens system. By a careful choice of the permittivity and permeability in the optical system, the slab lens can be adjusted so that the phases of the propagating waves and the decay rates of the evanescent components of light from the object are reassembled at the image in such a way as to give a complete characterization in the image of the object [1, 4, 13]. The image and the object are the same but only differ in their positions in space.

In an ideal theoretical sense a slab lens of negative refractive index medium can be engineered to give an image with a perfect resolution of the source or object features. The lens is perfect only in the ideal sense, just as in classical optics perfect surfaces and dispersionless media allow for an optimal lens under the limitations of positive indexed media optics. It is important to note that the perfect lens model has some practical complications in its experimental realization. These come from losses in the system due to the magnetic resonance origins of engineered negative indexed materials and the dielectric and current losses in the system components involved in the design of metamaterials.

The magnetic resonances required in the metamaterial design are required by the Kramer-Kronig relations to also be associated in increase energy losses in the systems. These energy losses are greatest at the resonant frequency around which the system operation is of most interest. Another problem with metamaterials based on magnetic resonances is that the regions of negative permeabilities they create only exist over a narrow band of frequencies. Consequently, due to the difficulties in material losses and the nature of magnetic resonances, the idea of a perfect lens has only been very narrowly realized experimentally [1–5, 13].

In the remainder of this section the above qualitative discussions will be firmed up, with a theoretical treatment given of the properties of a particular formulation of the perfect lens. The formulation considered here was originally proposed by Pendry [13].

In the considerations, first the slab system will be defined and its geometric optics properties discussed. This will be followed by a treatment of the details of how the phase and evanescent wave properties of the object are related to the image formed by the lens. Finally, it will be shown that the phase and evanescent nature of the object and image of the lens are identical. This is the basic property exhibited by a perfect lens, i.e., the perfect reconstruction of the object in the image.

5.2.1 Ideas of the Perfect Lens

The geometry of the perfect lens and its imaging system is shown in Fig. 5.6a. In the figure, the lens is designed as a slab formed of a negative refractive index medium. An object is placed at $x = 0$, and the slab forming the lens is located between $x = d_0$ and $x = d_0 + d$ so that the width of the lens slab is d. Both of the surfaces of the slab are taken to be parallel to the y-z plane [13]. As shall be shown later, an image of the object is formed by the lens at $x = 2d$.

As a simple source or object for the treatment of the optics of the lens, a point dipole is located at the origin of coordinates. The dipole is located to the left of the lens, and the image of the dipole formed by the system is on the right of the lens. The horizontal lines in the figure represent the optical axis and the geometric considerations of the components forming the imaging of the dipole source by the lens. These are presently to be discussed.

In the following treatment, the region outside the slab is vacuum. In order to make a perfect lens for this particular choice of surrounding medium, a specific requirement is set on the value of the negative refractive index of the slab medium. In particular, the negative index medium of the slab for a surrounding vacuum must be set to a refractive index of −1. The necessity of this choice will become apparent in the course of the following presentation [13].

With this particular set of refractive indices the image of the dipole will be shown to be formed at $x = 2d$. This is indicated in Fig. 5.6, and an explanation of the geometric optics of the lens and its focusing features is now presented [13].

The location of the dipole image can be understood from a consideration of Fig. 5.6b. In the figure two rays are shown leaving the dipole source (object) at $x = 0$. These propagate in the x-y plane and eventually encounter the left planar surface of the infinite slab lens. Upon encountering the lens an application of Snell's law for an interface between vacuum with an index of refraction of 1 and a medium of negative index of −1 requires that the magnitudes of the incident angle equals that of the transmission angle.

The resulting refraction is indicated in the figure. Later it shall be explained that there is no reflected wave generated by the interaction with the surface. As they travel through the lens the refracted rays next meet the optical axis of the system at $x = 2d_0$.

Both rays shown in the figure pass through $x = 2d_0$ and proceed towards the second surface of the lens. After traveling a distance $d - d_0$ they arrive at the

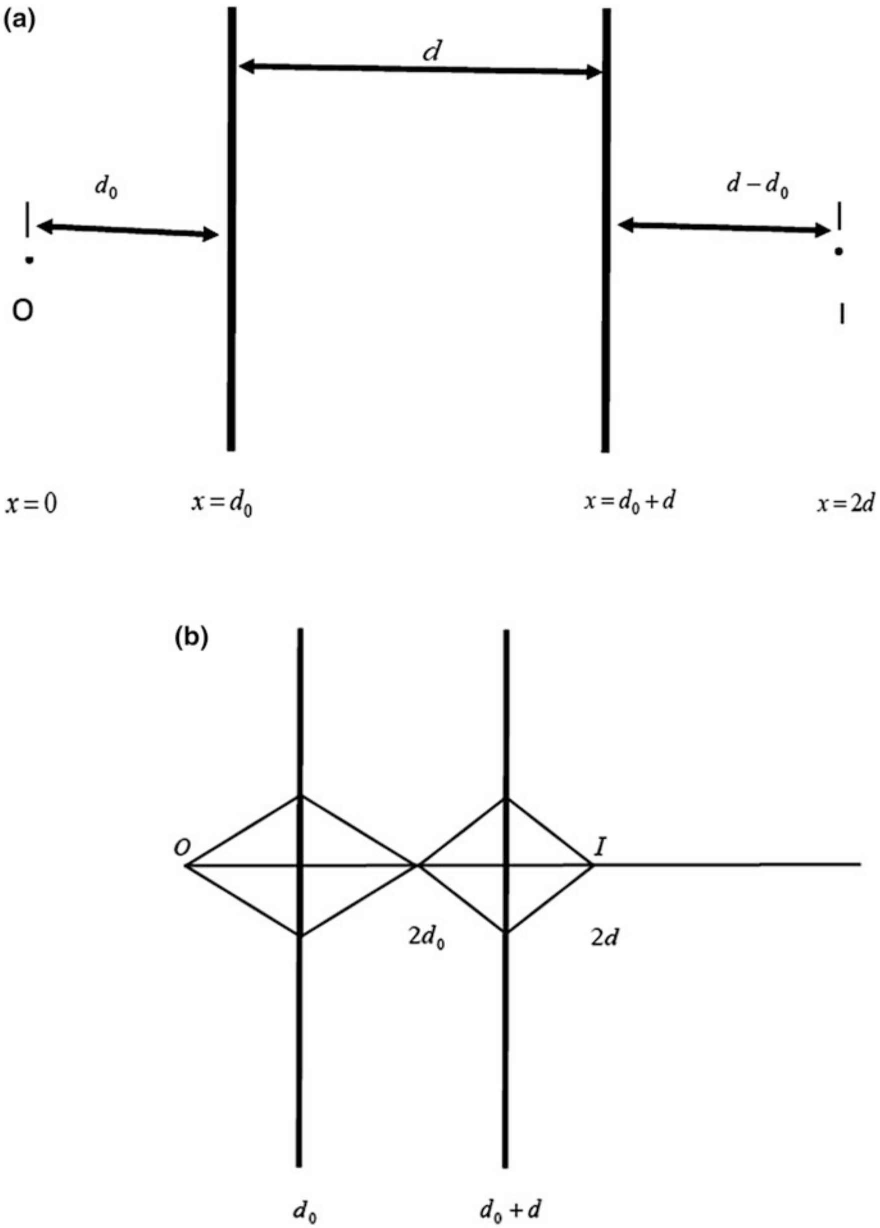


Fig. 5.6 Schematics for: **a** the object (at $x = 0$) and image (at $x = 2d$) of a perfect lens of thickness d located in the region between $x = d_0$ and $x = d_0 + d$, and **b** a ray optics diagram for the formation of an image I of and object O by the perfect lens diagramed in (**a**). The optical axis in (**b**) is the horizontal line

second surface of the lens. At their encounter with the second surface the rays are again refracted. Later it shall be explained that there is again no reflected wave generated by the interaction of these rays with the second surface [13].

At the second surface of the lens the two rays again are refracted in a manner so that the amplitudes of the angles of incidence equal those of the angles of refraction. This is indicated in the geometric opitcs figure, Fig. 5.6b. As a result of this refraction, the light from the point on the optical axis a distance $d - d_0$ to the left of the second surface of the lens shows up as an image point a distance $d - d_0$ to the right of the second surface. The final image is then formed at $x = 2d$.

The treatment just outlined in terms of Fig. 5.6b gives a basic geometric optics presentation of the motion of propagating light in the system. It shows the position of the image in its relation to the source or object and also the position of another image formed within the lens. However, to understand in detail the nature of the image formed by the lens it is necessary to study the changes in phase of the light as it moves through the imaging process of the lens [13]. These phase changes are not addressed in the geometric optics discussions just given.

Such phase considerations allow for a comparison of the light amplitude and phase at the image as it is related to that at the source. In the following the phase changes in the radiation from the source will be determined as it passes through the lens and forms an image. It will also be explained that no reflection occurs to waves propagating through the lens system.

Next consider the nature of the fields as the electromagnetic wave propagates through the perfect lens. This includes the full representation of the fields in terms of a Fourier series consisting of the full solutions of the Maxwell equations for the light moving through the system. A discussion is given of the changes in amplitude and phase of these various components.

For such considerations, the magnetic fields in the system can be represented by the general form [13]

$$\vec{H}(\vec{r}, t) = \sum_{k_y, k_z} \vec{H}(k_y, k_z) \mathrm{e}^{i\left[k_x x + k_y y + k_z z - \omega t\right]}. \tag{5.58}$$

This expression for the magnetic fields determines them as series composed of plane waves and evanescent waves. This is an important point as the presence of both propagating and evanescent waves in the series comes from the nature of the dispersion relation of the electromagnetic modes of the system. This feature is now discussed.

In vacuum, the wave vector components of the electromagnetic modes satisfy

$$k_x^2 + k_y^2 + k_z^2 - \frac{\omega^2}{c^2} = 0 \tag{5.59a}$$

where $c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$, and within the slab of refractive index n the wave vectors satisfy

$$k_x^2 + k_y^2 + k_z^2 - n^2 \frac{\omega^2}{c^2} = 0 \tag{5.59b}$$

where $\frac{c}{n} = \frac{1}{\sqrt{\mu\varepsilon}}$. From (5.58) and (5.59) it is found that the propagating waves at the dipole source are those for which k_x, k_y, k_z are all real. However, in the complete representation in (5.58) evanescent waves, which have at least one imaginary wave vector component, are also included in the sum. It is necessary to understand how both of these types of modes are handled by the system to completely determine the imaging properties of the lensing system.

Behavior of the Propagating Modes

To understand the nature of the image of the point dipole that is created by the lens, consider the light propagating in the *x*-*y* plane containing the dipole. The magnetic field of the waves propagating in this plane can be divided into two polarizations. In one polarization the magnetic field vector is perpendicular to the plane of incidence and in the other the magnetic field vector is parallel to the plane of incidence. Due to this separation, the refraction of each of these polarizations can be studied separately. For simplicity in the following considerations only one of these components will be discussed in detail.

In the treatment now presented, the details of the refraction of the fields which have magnetic fields polarized perpendicular to the *x*-*y* plane are given. A similar treatment for the other polarization is left to the reader, who can also obtain the details from the literature [13].

The general form of the magnetic field in the *x*-*y* plane arising from the dipole source located at $x = 0$ is given from (5.58) by [13]

$$H_z(x, y, z = 0, t) = \sum_{k_y, k_z} H_z(k_y, k_z) \mathrm{e}^{i\left[k_x x + k_y y - \omega t\right]}. \tag{5.60}$$

For properly chosen Fourier coefficients, this represents the magnetic field radiated by the dipole source of frequency ω everywhere in the *x*-*y* plane. In particular, at x = 0 the dipole fields of (5.60), which include those at the position of the dipole at $x = y = z = 0$, are then written as

$$H_z(x = 0, y, z = 0, t) = \sum_{k_y, k_z} H_z(k_y, k_z) \mathrm{e}^{i\left[k_y y - \omega t\right]}. \tag{5.61}$$

As the x coordinate in (5.61) is changed from zero, the properties of the fields exhibit a change related to their properties as the radiation fields of the dipole source. This comes from the phase factors $\mathrm{e}^{i\left[k_x x + k_y y - \omega t\right]}$ which multiply the Fourier coefficients in (5.60). By studying these phase factors, the changing nature of the field amplitudes can be determined as functions of time and position.

Of particular interesting in the following is to determine how the $\mathrm{e}^{i\left[k_x x + k_y y - \omega t\right]}$ phases change as light from the source passes through the lens and forms an image. In the following it will be demonstrated that the phase generated at the source, when passed through the perfect lens, remains unchanged at the image. This is a very important point in the reconstruction of the image of the source. This will be explained later.

An interesting calculation is to consider one of the Fourier components in (5.61) and to determine how it changes during the propagation of the wave through the lens in Fig. 5.6. In particular, for this determination start with the component in (5.61) given by [13]

$$\mathrm{e}^{i\left[k_y y - \omega t\right]}. \tag{5.62}$$

This is just one of the phase terms multiplying the Fourier coefficients in (5.61).

As mentioned earlier, in general the phase factors in (5.61) separate into two different types. These are classified as propagating and evanescent waves, and the classification of the two types depends on the values of

$$k_x^2 = -k_y^2 - k_z^2 + \frac{\omega^2}{c^2} \tag{5.63}$$

In particular, if the right side of (5.63) is positive the waves propagate through the system while if the right side of (5.63) is negative the waves are evanescent in nature. In the discussion, first a treatment will be given of propagating wave terms in (5.61). Afterward, this will be followed by a treatment of evanescent terms.

Considering the propagating terms: As the wave from the dipole source propagates towards the first surface of the lens, located at $x = d_0$, the factor in (5.62) becomes [13]

$$\mathrm{e}^{i\left[k_x d_0 + k_y y - \omega t\right]}. \tag{5.64}$$

The wave is seen to pick up a phase factor related to the distance it must travel to arrive at the $x = d_0$ surface of the lens.

Once the wave arrives at the surface it undergoes an additional phase shift as it passes from one side to the other of the interface. This phase shift is given by the transmission amplitude through the planar interface. It was determined earlier in this chapter in the discussion of refraction from a positive indexed to a negative indexed medium.

The transmission amplitude at the left surface of the lens can be evaluated using (5.44a). For the system in Fig. 5.6b it is found from this formulas that the relative transmission amplitude is -1 and the reflection amplitude is zero. However, an account must be made of the factor of -1 introduced in the field amplitude definition in (5.35a). Specifically, the field amplitude was defined with a factor of -1 introduced. An adjustment is needed so that the signs in the definitions of the field amplitudes in (5.35a) and (5.45a) agree as the first becomes the input for the second equation.

Consequently, upon applying the transmission amplitude and the correction to the definition of the field amplitude, the net transmitted wave retains the form given in (5.64) upon entering the negative index medium. The wave transmitted into the lens then propagates through the lens and becomes the incident field on the second surface of the lens.

Upon entering the medium of refractive index −1, the wave propagates through the negative index medium from the first to the second surface. In doing so it changes its phase by a factor of e^{-ik_xd}.

The reason for the negative sign in the argument of the exponent is that the wave moves in a negative index medium. Consequently, for the energy of the wave to move from left to right through the lens the wave vector of the wave must point from right to left. From earlier discussions, this is a known general property of waves in a negative index medium.

Upon arriving at the second surface the factor in (5.64) transform into [13]

$$\mathrm{e}^{-ik_xd}\mathrm{e}^{i\left[k_xd_0+k_yy-\omega t\right]}=\mathrm{e}^{i\left[k_x(d_0-d)+k_yy-\omega t\right]}. \tag{5.65}$$

After its arrival, it must undergo transmission from the negative index medium to the vacuum. Upon doing this it acquires a phase shift which is governed by the transmission amplitude determined in (5.56a) and (5.56b).

The transmission amplitude at the second surface from (5.56a) and (5.56b) is a factor of 1, and the reflection amplitude is zero. Consequently, the phase of the wave upon passing through the second surface of the lens and emerging into the vacuum is

$$\mathrm{e}^{i\left[k_x(d_0-d)+k_yy-\omega t\right]}. \tag{5.66}$$

After passing through the lens the wave propagates through vacuum to arrive at the image which is located $x=2d$.

The next consideration is to understand how the wave propagates from the second surface of the lens to the image located at a distance $d-d_0$ to the right of the second surface of the lens. The propagation is through vacuum and introduces a factor of $\mathrm{e}^{ik_x(d-d_0)}$. Now the energy flow and the wave vector are parallel to one another as the medium is positive index media.

Applying the $\mathrm{e}^{ik_x(d-d_0)}$ phase change to (5.66), the phase at the image becomes [13]

$$\mathrm{e}^{i\left[k_yy-\omega t\right]}. \tag{5.67}$$

This is now the phase of the propagating components multiplying the Fourier amplitudes in the representation of the magnetic field at the image.

However, (5.67) is also the phase factor for the propagating components at the original position of the dipole source. Consequently, the resulting contribution to the field from the propagating Fourier components at the image is

$$H_z(x = 2d, y, z = 0, t) = \sum_{k'_y, k'_z} H_z(k'_y, k'_z)\mathrm{e}^{i\left[k'_y y - \omega t\right]}. \tag{5.68}$$

where the sum over primed wave vectors is restricted to the propagating waves in the system. This is exactly the same expression as the propagating wave part of (5.61) at the dipole source. The propagating waves contributions therefore are the same in both the source and the image.

As a result, the propagating fields from the dipole are exactly identical with those from the image. Their origin has only been relocated in space.

The propagating fields, however, are not the only fields associated with the dipole. In particular, the dipole is also a source of evanescent fields which decay in space as they pass through the lens and are projected into the image. These fields also arise as the set of solutions obtained from the Maxwell equations and are passed through the lens to contribute to the fields at the source and at the image. The change in these components must be examined to understand the complete relationship of the source to image. It shall be shown in the following that the terms involving the evanescent waves are also identical in their contributions to the source and image fields.

Behavior of the Evanescent Fields

To begin the study of the interaction of the negative index lens in Fig. 5.6 with the evanescent waves from the source, it is necessary to understand the transmission and reflection of these waves at the interfaces. Consequently, in the following a study will be given of the reflection and transmission of evanescent waves that are incident on the interface between positive and negative index media. These results will then be used to determine the interaction of the evanescent waves with the lens system in Fig. 5.6.

In the following a study is presented of the evanescent wave solutions at the interface between a positive and negative index medium, treating the case in which the magnetic field is polarized perpendicular to the plane of incidence. A focus is on the determination of the transmission and reflection coefficients of these waves as they pass from a positive to a negative indexed medium and as they pass from a negative to a positive indexed medium.

First consider an evanescent wave incident from a positive index medium passing into a negative index medium. For these treatments the appropriate Maxwell equations are of the form [13]

$$\nabla \times \vec{B} = -i\mu'\varepsilon'\omega\vec{E} \tag{5.69a}$$

and

$$\nabla \times \vec{E} = i\omega\vec{B} \tag{5.69b}$$

Here the waves are considered to have frequency ω, and $(\mu', \varepsilon') = (\mu_0, \varepsilon_0)$ in vacuum, and in the negative index medium $(\mu', \varepsilon') = (\mu, \varepsilon)$ with $\mu\varepsilon = \frac{1}{c^2}$ where c is the speed of light in vacuum.

These forms of the Maxwell equations will be used to study the transmission of an evanescent wave through the surface between a positive and negative index medium. The interface in the treatment is represented schematically in Fig. 5.7. In this figure, the interface between the two different media is the x-z plane, and the evanescent fields are evanescent along the y-axis. The geometry is consistent with that used in the refraction treatments for propagating waves given in Sect. 5.1 and has a different coordinate arrangement than that in Fig. 5.6. Since only the transmission amplitudes are of interest here, the coordinate system used for the calculation is not important.

The solutions of (5.69) for the incident evanescent waves in vacuum are [13]

$$\vec{E}_I = i\frac{q}{\mu_0\varepsilon_0\omega}B_I e^{qy+ikx}\hat{i} + \frac{k}{\mu_0\varepsilon_0\omega}B_I e^{qy+ikx}\hat{j} \tag{5.70a}$$

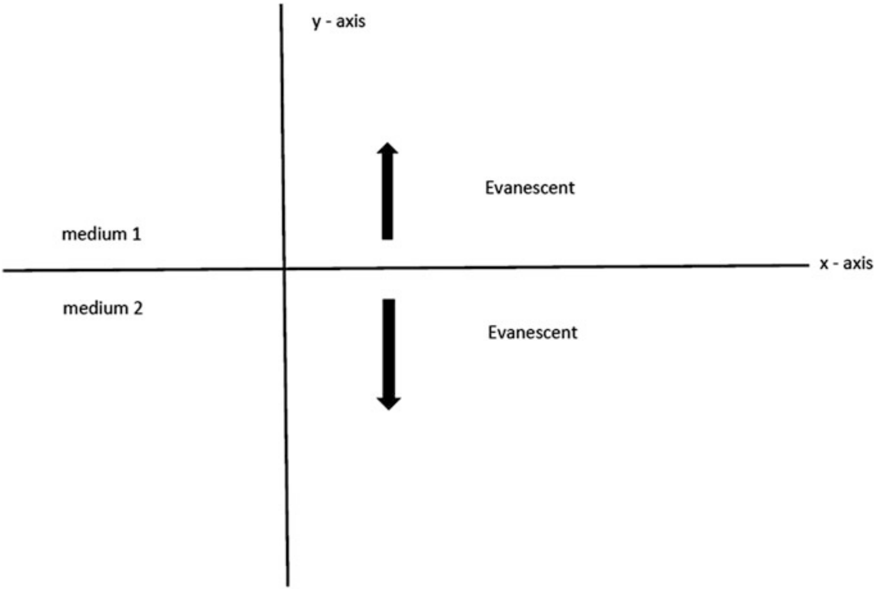


Fig. 5.7 Evanescent waves at the x-z interface. Waves either decay or increase along the y-axis as they leave the surface, depending upon the conditions of the treatment for the transmission amplitudes being considered in the text

and

$$\vec{B}_I = B_I e^{qy + ikx} \hat{k}. \tag{5.70b}$$

In both of these expressions

$$q = \sqrt{k^2 - \mu_0 \varepsilon_0 \omega^2} \tag{5.71}$$

where $k^2 > \mu_0 \varepsilon_0 \omega^2$. With these definitions, the solutions for the reflected evanescent waves in vacuum are

$$\vec{E}_r = -i \frac{q}{\mu_0 \varepsilon_0 \omega} B_r e^{-qy + ikx} \hat{\imath} + \frac{k}{\mu_0 \varepsilon_0 \omega} B_r e^{-qy + ikx} \hat{\jmath} \tag{5.72a}$$

and

$$\vec{B}_r = B_r e^{-qy + ikx} \hat{k} \tag{5.72b}$$

The incident fields in (5.70) are seen to decrease as they approach the surface while the reflected fields in (5.72) are seen to decrease they leave the surface.

The solutions for the transmitted evanescent waves in the negative index medium are [13]

$$\vec{E}_t = i \frac{q'}{\mu \varepsilon \omega} B_t e^{q'y + ikx} \hat{\imath} + \frac{k}{\mu \varepsilon \omega} B_t e^{q'y + ikx} \hat{\jmath} \tag{5.73a}$$

and

$$\vec{B}_t = B_t e^{q'y + ikx} \hat{k} \tag{5.73b}$$

In these expressions the factor governing the decay of the waves in the medium is given by

$$q' = \sqrt{k^2 - \mu \varepsilon \omega^2} \tag{5.74}$$

where $k^2 > \mu \varepsilon \omega^2$. The transmitted waves in (5.73) are again found to decay as one moves away from the interface and the source of the fields in the positive index medium.

Applying the electromagnetic boundary conditions to the fields in (5.70), (5.72) and (5.73) at the interface, the following occurs: The continuity of the component of the electric field at the $y = 0$ interface between the two media yields the condition [13]

$$B_I - B_r = \frac{q'}{q}\frac{\mu_0 \varepsilon_0}{\mu \varepsilon} B_t. \tag{5.75a}$$

The continuity of the component of the magnetic field at the $y = 0$ interface between the two media yields a second condition

$$B_I + B_r = \frac{\mu_0}{\mu} B_t \tag{5.75b}$$

In dealing with these equations a useful relationship between the permeability and permittivity of the vacuum and the -1 negative index medium is [13]

$$\frac{\mu \varepsilon}{\mu_0 \varepsilon_0} = 1 \tag{5.76a}$$

It follows from this that the permeabilities and permittivities in the two media are related by [13]

$$\frac{\mu}{\mu_0} = \frac{\varepsilon_0}{\varepsilon}. \tag{5.76b}$$

Both of these conditions are helpful in simplifying the transmission and reflection amplitudes obtained from (5.75a) and (5.75b).

From (5.75) and (5.76) it follows that the reflected and transmitted magnetic fields are given in terms of the incident magnetic field by

$$H_r = \frac{-q' + \frac{\varepsilon}{\varepsilon_0} q}{q' + \frac{\varepsilon}{\varepsilon_0} q} H_0 \tag{5.77a}$$

and

$$H_t = \frac{2 \frac{\varepsilon}{\varepsilon_0} q}{q' + \frac{\varepsilon}{\varepsilon_0} q} H_0 \tag{5.77b}$$

These expressions relate through (5.70), (5.72), and (5.73) the magnetic and electric fields of the reflected and transmitted fields those of the incident fields for the case of an incident field in the positive index medium.

Repeating the above for the case in which the incident wave is in the negative index medium, the reflected and transmitted magnetic fields in this case are expressed in terms of the incident magnetic field by [13]

$$H_r = \frac{-\frac{\varepsilon}{\varepsilon_0} q + q'}{\frac{\varepsilon}{\varepsilon_0} q + q'} H_0 \tag{5.78a}$$

and

$$H_t = \frac{2q'}{\frac{\varepsilon}{\varepsilon_0}q + q'}H_0. \tag{5.78b}$$

These expressions relate through (5.70), (5.72), and (5.73) the magnetic and electric fields of the reflected and transmitted fields those of the incident fields for the case of an incident field in the negative index medium.

Summarizing the above: At the interface for an evanescent wave to go from a positive index medium to a negative index medium, the relative reflection amplitude is [13]

$$r = \frac{-q' + \frac{\varepsilon}{\varepsilon_0}q}{q' + \frac{\varepsilon}{\varepsilon_0}q} \tag{5.79a}$$

and the relative transmission amplitude is

$$t = \frac{2\frac{\varepsilon}{\varepsilon_0}q}{q' + \frac{\varepsilon}{\varepsilon_0}q} \tag{5.79b}$$

At the interface for an evanescent wave to go from a negative index medium to a positive index medium, the relative reflection amplitude is

$$r' = \frac{-\frac{\varepsilon}{\varepsilon_0}q + q'}{\frac{\varepsilon}{\varepsilon_0}q + q'} \tag{5.80a}$$

and the relative transmission amplitude is

$$t' = \frac{2q'}{\frac{\varepsilon}{\varepsilon_0}q + q'}. \tag{5.80b}$$

Combining the results in (5.79) and (5.80) both the reflection and transmission through the lens in Fig. 5.6 can be determined for the evanescent fields. This allow for the determination of the nature of the evanescent fields at the image in terms of the evanescent fields at the dipole source.

The above transmission and reflection amplitudes are for a single interface. To account for the two interfaces involved with the slab transmissions a sequential process of single interface transmissions and reflections must be addressed. These are now discussed.

The transmission and reflection through the lens must take into account the multiple reflections and transmission as the electromagnetic fields bounces back and forth between the surfaces of the lens. In terms of the transmission amplitude in (5.79) and (5.80) the transmission through the lens can be written as an infinite series summing the various encounters of the evanescent wave with the lens surfaces. The total transmission amplitude through the surface is given by the series [13]

$$T = tt'\left\{\mathrm{e}^{-qd} + (r')^2\mathrm{e}^{-3qd} + (r')^4\mathrm{e}^{-5qd} + \cdots\right\} = tt'\frac{\mathrm{e}^{-qd}}{1-(r')^2\mathrm{e}^{-2qd}}. \tag{5.81}$$

In (5.81) the factor tt' represents the transmission of the wave through the first surface followed by transmission through the second surface. The terms in the brackets represent a sum of multiple reflections within the lens. The first term, e^{-qd}, represents the decay of the fields going from their entry at the first surface to their exit at the second surface. The second term of the series, $rr'\mathrm{e}^{-3qd}$, represents the fields in the lens going from their entry at the first surface, to a reflection at the second surface, followed by a reflection at the first surface, and then a transmission through the second surface. The exponential term accounts for the decay along the path represented in the second term. The successively high terms in the series represent successively longer paths in the negative index medium.

Applying a similar reasoning to the reflections from the lens, the total reflection amplitude is given by [13]

$$R = r + tt'r'\{\mathrm{e}^{-2qd} + (r')^2\mathrm{e}^{-4qd} + (r')^4\mathrm{e}^{-6qd} + \cdots\} = r + tt'\frac{r'\mathrm{e}^{-2qd}}{1-(r')^2\mathrm{e}^{-2qd}}. \tag{5.82}$$

This, again, sums all of the multiple reflection processes as the decaying wave amplitude is reflected back and forth within the lens to represent the total reflection from the lens.

As a final consideration, the permeability and permittivity parameters characterizing the negative index medium are chosen so that the negative index of refraction of the lens material is −1. This is done for the transmission through the slab by substituting (5.79) and (5.80) into (5.81) and taking the limit of $\mu \to -\mu_0$ and $\varepsilon \to -\varepsilon_0$. The transmission amplitude reduces to the limiting form

$$\begin{aligned} T &= \lim_{\substack{\mu\to-\mu_0\\ \varepsilon\to-\varepsilon_0}}\left\{tt'\frac{\mathrm{e}^{-qd}}{1-(r')^2\mathrm{e}^{-2qd}}\right\} \\ &= \lim_{\substack{\mu\to-\mu_0\\ \varepsilon\to-\varepsilon_0}}\left\{4\frac{\varepsilon}{\varepsilon_0}qq'\frac{\mathrm{e}^{-q'd}}{\left(q'+\frac{\varepsilon}{\varepsilon_0}q\right)^2-\left(q'-\frac{\varepsilon}{\varepsilon_0}q\right)^2\mathrm{e}^{-2q'd}}\right\} \\ &= \mathrm{e}^{qd}. \end{aligned} \tag{5.83}$$

This limiting process is also done for the reflection amplitude yielding [13]

$$
\begin{aligned}
R &= r + tt'\frac{r'\mathrm{e}^{-2qd}}{1-(r')^2\mathrm{e}^{-2qd}} = \lim_{\substack{\mu\to-\mu_0\\ \varepsilon\to-\varepsilon_0}}\left\{r + tt'\frac{r'\mathrm{e}^{-2qd}}{1-(r')^2\mathrm{e}^{-2qd}}\right\}\\
&= \lim_{\substack{\mu\to-\mu_0\\ \varepsilon\to-\varepsilon_0}}\left\{r + 4\frac{\varepsilon}{\varepsilon_0}qq'(-r)\frac{\mathrm{e}^{-2q'd}}{\left(q'+\frac{\varepsilon}{\varepsilon_0}q\right)^2-\left(q'-\frac{\varepsilon}{\varepsilon_0}q\right)^2\mathrm{e}^{-2qd}}\right\}\\
&= 0.
\end{aligned}
\tag{5.84}
$$

Consequently, no reflected wave is generated.

The transformation of the evanescent waves as they originate in the source, pass through the lens, and arrive at the image can now be determined. To treat the evanescent waves, consider one of the Fourier components in (5.61) in which the component of wave vector in the x-direction is imaginary. [Notice the different coordinate system used in (5.61).] In particular, in this case it is assumed that

$$k_x = iq. \tag{5.85}$$

where $q = \sqrt{k_y^2 - \mu_0\varepsilon_0\omega^2}$ is positive real and k_y is also real. The focus in the following with be on the transformations in such a term as the light goes through the lens system.

To determine how the evanescent component changes during the propagation of the wave through the lens in Fig. 5.6, start with the component in (5.61) given by

$$\mathrm{e}^{i\left[k_y y-\omega t\right]}. \tag{5.86}$$

Remember that now and in the following the coordinates in Fig. 5.6 are being used. These are not the same as those of the system in Fig. 5.7 used to determine the transmission amplitude of an evanescent wave as it encounters an interface between a positive and negative index media. The transmission amplitude determined from Fig. 5.7 is independent of the coordinates used in its determination so that this is not a problem here.

Considering the evanescent terms in Fig. 5.6: As the wave from the dipole source propagates towards the first surface of the lens, located at $x = d_0$, the factor in (5.62) becomes

$$\mathrm{e}^{-qd_0+i\left[k_y y-\omega t\right]}. \tag{5.87}$$

From (5.83) it was determined that as the evanescent term passes through the lens the wave amplitude is multiplied by a factor of

$$e^{qd}. \tag{5.88}$$

Emerging from the second surface of the lens the amplitude is of the form

$$e^{qd}e^{-qd_0+i\left[k_y y-\omega t\right]} = e^{q(d-d_0)i\left[k_y y-\omega t\right]}. \tag{5.89}$$

After passing the lens the wave must continue on a distance $d - d_0$ to arrive at the image. The resulting amplitude at the position of the image is [13]

$$e^{i\left[k_y y-\omega t\right]}. \tag{5.90}$$

This, however, is just the factor in the Fourier term that we started with at the source. Consequently, both the propagating and evanescent factors in the Fourier representation of the source signal are reproduced identically at the position of the image. This is the function of the perfect lens.

5.2.2 Other Applications of Positive–Negative Refractive Properties

Aside from the perfect lens, there are a number of other applications that have come from the development of systems exhibiting a negative refractive index. As in the case of the perfect lens such applications arise from the increased deflection of the incident ray as it encounters an interface between positive and negative refractive media. Through application of this increase range of optical properties, in principle it has become possible to guide light at will in its motion through space. Consequently, the increased guiding of light facilitated by the diffractive photonic crystal technology is now complemented in the guiding of light through the design of continuum limit refractive metamaterials [14–43].

An example of such a use of the guiding ability of metamaterials is found in the development of electromagnetic cloaking [1–5, 37, 38]. In electromagnetic cloaking an engineered medium is designed which displays a gradual spatial variation of the refractive index. As the index varies though space, taking on a range of positive and/or negative values, light can be steered around an object hidden in the medium. This is arranged so as to make the object essentially invisible to an observer receiving light from the cloak.

In a typical cloaking arrangement the cloaking device is designed as a hollow dielectric shell. The object to be hidden is put in the hollow of the shell, and the material of the shell is formulated to have a variation of the index of refraction which guides light around the hollow of the shell. In this way the index variation is made so that parallel incident rays on the outside of the dielectric shell are steered around the hollow containing the object to be hidden. As the rays encounter the shell surface opposite that of their incidence they are sent off into the region outside the shell in a direction parallel to that of their original direction before the encounter with the cloaking device [37, 38].

To accomplish the steering of light in this manner requires the use of both positive and negative indexed media [37, 38]. Again, as with the development of the perfect lens, the efficiency of the cloaking device is dependent on the properties of the materials used in its design and the degree that the system can be accurately assembled. Due to the restrictions from magnetic resonance effects, the losses near resonance, and the dielectric and joule losses of the materials, this has only been done in the case of a narrow band of frequencies of incident light and with some degree of signal degradation.

These ideas can be extended in many ways. As a recent very interesting example of employing the ideas of designing media with a spatial variation of positive and negative index, suggestions have also been made for mimicking certain optical effects in special and general relativity [19–22]. These ideas are based on the relativistic invariance of the Maxwell equations and their transformations between various different coordinate systems. Such transformations can result in Maxwell equations with renormalized index of refractions.

These types of examples may seem extreme. However, they suggest many other applications in the design of passive optical systems for device applications that, though more mundane, may be important to technology. In addition to the study of passive optical interactions, the index of refractions of metamaterials offer many new design features in active optical systems.

It is seen that the increase in the range of index of refraction of metamaterials offers many opportunities for passive optical systems. Now the focus will be placed on the properties of active optical systems. In particular, all of the earlier discussed topics of metamaterial technologies have focused on the interaction of existing radiation fields with devices designed from metamaterials. Another set of important application of metamaterials involves radiation problems dealing with the generation of radiation fields. Specifically, how does the presence of metamaterials affect the generation of radiation from charge and current sources. These types of origin problems not only cover the unusual phenomena exhibited by electromagnetic fields radiated into metamaterials by the typical features treated in classical electrodynamics (i.e., point dipoles,, accelerating charges, Doppler effect) but also the broad and important technological applications of antenna theory [1–5, 13–17]. In the following these topics will be outlined.

5.3 Radiation in a Negative Indexed Medium

The unusual properties of electromagnetic fields in negative refractive index media lead to some novel behaviors observed in the standard radiation problems of classical electrodynamics. An important determiner of the properties of these radiation fields arises from the fact that the wave vector and Poynting vector of the radiation in a negative indexed medium are anti-parallel. This relationship is responsible for altering some of the behaviors of the radiation fields generated within negative refractive index media by accelerating charges from those found in positive refractive index media. In the following, the radiation fields of an electric dipole antenna located in a negative indexed medium and of a point change moving within a negative index medium are treated [1–5, 13–17]. Some additional discussions will also be given of Cherenkov radiation and of the Doppler effect observed in negative refraction index media.

Radiation problems in classical electrodynamics are best treated in terms of the electromagnetic vector and scalar potentials, $(\vec{A}(\vec{r},t), V(\vec{r},t))$, rather than directly in terms of the electromagnetic fields which are related to these potentials by [1]

$$\vec{E}(\vec{r},t) = -\nabla V - \frac{\partial \vec{A}(\vec{r},t)}{\partial t} \tag{5.91a}$$

and

$$\vec{B}(\vec{r},t) = \nabla \times \vec{A}(\vec{r},t). \tag{5.91b}$$

Applying (5.91) in the Maxwell equations and working in the Lorentz gauge in which the potentials are further specified by the gauge relationship

$$\nabla \cdot \vec{A} = -\frac{1}{c_m^2}\frac{\partial V}{\partial t} \tag{5.92}$$

the radiation equations for electromagnetic waves generated from time varying charge and current sources can be written in terms of $(\vec{A}(\vec{r},t), V(\vec{r},t))$ [38]. Upon doing this one finds the following relationships

$$\nabla^2 V - \frac{1}{c_m^2}\frac{\partial^2 V}{\partial t^2} = -\frac{1}{\varepsilon}\rho, \tag{5.93a}$$

$$\nabla^2 \vec{A} - \frac{1}{c_m^2}\frac{\partial^2 \vec{A}}{\partial t^2} = -\mu\vec{J}, \tag{5.93b}$$

where $(\vec{A}(\vec{r},t), V(\vec{r},t))$ are the vector and scalar potentials and $(\vec{J}(\vec{r},t), \rho(\vec{r},t))$ are the current and charge densities of the source terms. In these equations, c_m is the speed of light in the medium.

The solutions of the inhomogeneous equations in (5.93) are

$$V_{\mp}(\vec{r},t) = \frac{1}{4\pi\varepsilon}\int \frac{\rho(\vec{r}',t_{\mp})}{|\vec{r}-\vec{r}'|}d^3r', \tag{5.94a}$$

$$\vec{A}_{\mp}(\vec{r},t) = \frac{\mu}{4\pi}\int \frac{\vec{J}(\vec{r},t_{\mp})}{|\vec{r}-\vec{r}'|}d^3r', \tag{5.94b}$$

where $t_{\mp} = t \mp \frac{|\vec{r}-\vec{r}'|}{c_m}$. In (5.94) the upper (lower) signs are known as the retarded (advanced) potentials of the fields generated by the source terms. These two different types of solutions will be very important in the treatment of positive and negative indexed media.

It should be noted that homogeneous solutions for which $(\vec{J}(\vec{r},t), \rho(\vec{r},t)) = (0,0,0,0)$ in (5.93) can be added to (5.94) to match particular boundary conditions which may be placed on the radiation problem. For the presentations in the following, however, homogeneous solutions will not be needed. In this regard, it shall be assumed that the charge and current sources are localized in infinite space so that the fields go to zero at infinite separation from the sources.

In positive indexed media the radiation fields in (5.94) from the retarded potentials describe the motion of radiation away from the sources generating them, and the advanced field solutions describe a time reversed state in which the radiation fields return from infinity to arrive at the charge and current sources. Later, it shall be shown that these relationships are reversed in a negative refraction index medium. In particular, for radiation in a negative refraction index medium the advanced potentials are those of interest in the description of radiation generated at charge and current sources and propagating to infinitity, and the retarded potentials describe the time reversed fields traveling from infinite to arrive at the charge and current sources. Consequently, in the following both retarded and advanced potentials will be considered.

Upon evaluating (5.94) for a given set of localized charge and currents, the electric field and the magnetic induction are then obtained from the vector and scalar potentials in (5.94) through and application of (5.91). This then accounts for a complete solution of the electric and magnetic radiation fields of a localized source in infinite space. As an interesting limiting case of these results, notice that in the limit as $c_m \to \infty$ in (5.94) the potentials, appropriately, reduce to the scalar and vector potentials of static electric and magnetic induction fields.

As a first example of the application of the above radiation formulation consider the fields generated by an electric point dipole source. For these considerations the dipole is taken to be located at the origin of coordinates, having a dipole moment given by the harmonic time dependent form [1]

$$\vec{p}(t) = p_0 \cos(\omega t)\hat{k}, \tag{5.95}$$

and with the focus of the treatment being on the $kr \gg 1$ far field limit. Here k is the wavenumber of the radiation fields, and in this limit the point of observation is a great distance from the source. In particular, it is much greater than a wavelength.

Treating (5.91) through (5.94) for the dipole source in (5.95) by considering the $kr \gg 1$ far field limit in which the point of observation is a great distance from the source, the electric field and magnetic induction are [1–5]

$$\vec{E} = -\frac{\mu p_0 \omega^2}{4\pi}\frac{\sin\theta}{r}\cos\omega\left(t \mp \frac{r}{c_m}\right)\hat{\theta}, \tag{5.96a}$$

$$\vec{B} = \mp\frac{\mu p_0 \omega^2}{4\pi c_m}\frac{\sin\theta}{r}\cos\omega\left(t \mp \frac{r}{c_m}\right)\hat{\phi}, \tag{5.96b}$$

where (r, θ, ϕ) are the standard polar coordinates centered at the dipole source which is located at the origin of coordinates. From (5.96) the Poynting vector of the electromagnetic radiation fields is given by

$$\left\langle \vec{S} \right\rangle = \pm\frac{\mu p_0^2 \omega^4}{32\pi^2 c_m}\frac{\sin^2\theta}{r^2}\hat{r}. \tag{5.97}$$

Using (5.97) the average power radiated from the dipole can be computed using standard methods. Upon doing this the average power is found to be given by

$$\langle P \rangle = \pm\frac{\mu}{12\pi}\frac{p_0^2\omega^4}{c_m}. \tag{5.98}$$

Notice from (5.97) and (5.98) that for a positive indexed medium the permeability in each of these expressions is positive. Consequently, in this limit the upper sign from the retarded solution gives an energy flow away from the dipole source. In the other limit of a negative indexed medium the permeability is negative. A consequence of this is that the lower sign from the advanced solution gives an energy flow away from the dipole source. These results indicate the fundamental difference of the radiation generated by a source in the positive and negative index media.

A related important type of radiation problem is the determination of the radiation from a point charge accelerating in a dielectric medium. This again exhibits important differences in the treatment of the moving source as it passes through a positive or negative refractive index medium. It shall be the focus of the next discussions [1–5].

As a starting point in the consideration of this system, the results in (5.94) for the radiation fields of a general time-dependent localized charge distribution is used. The case of (5.94) for a general localized charge and current density are then specialized to the problem of a single accelerating point charge. In this application the limit $r' \ll r$ is considered, and the charge distribution is considered to be located about the origin of coordinates. Under these restrictions and in the non-relativistic limit, (5.94) become [1–5]

$$V_{\mp}(\vec{r}, t) = \frac{1}{4\pi\varepsilon}\left[\frac{Q}{r} + \frac{\hat{r}\cdot\vec{p}(t_{\mp})}{r^2} + \frac{\hat{r}\cdot\dot{\vec{p}}(t_{\mp})}{c_m r}\right], \tag{5.99a}$$

$$\vec{A}_{\mp}(\vec{r},t) = \frac{\mu}{4\pi}\frac{\dot{\vec{p}}(t_{\mp})}{r}. \tag{5.99b}$$

where in these equations Q represents the net charge of the distribution and $\vec{p}(t)$ is the electric dipole moment of the charge distribution about the origin of coordinates. The electric and magnetic induction fields are then related to these potentials using (5.91) and from these the general properties of the radiation fields are obtained.

To apply (5.99) to an accelerating point charge it is only needed to determine the electric dipole moment of the charge relative to the origin of coordinates. Once this is done, the radiation fields of the charge distribution are expressed in terms of the dipole moment and found to be given by

$$\vec{E}(r,\theta,\phi,t_{\mp}) = \frac{\mu}{4\pi}\ddot{\vec{p}}(t_{\mp})\frac{\sin\theta}{r}\hat{\theta}, \tag{5.100a}$$

$$\vec{B}(r,\theta,\phi,t_{\mp}) = \pm\frac{\mu}{4\pi}\frac{\ddot{\vec{p}}(t_{\mp})}{c_m}\frac{\sin\theta}{r}\hat{\phi}. \tag{5.100b}$$

Here (r,θ,ϕ) are standard polar coordinates chosen so that the charge is accelerating along the z-axis.

From the fields in (5.100) it follows from the standard expressions of classical electrodynamics that the Poynting vector of the radiation fields from the accelerating charge is [1–5]

$$\vec{S} = \pm\frac{\mu[\ddot{p}(t)]^2}{16\pi^2 c_m}\frac{\sin^2\theta}{r^2}\hat{r}. \tag{5.101}$$

From this a standard consideration gives the net radiated power from the accelerating charge in the form

$$P = \pm\frac{\mu}{6\pi}\frac{q^2a^2(t)}{c_m}, \tag{5.102}$$

where $a(t)$ is the acceleration of the charge.

From the result in (5.101) and (5.102) the properties of the radiation in the positive and negative refractive media directly follow. In the case of a positive indexed medium the permeability and permitivity are both positive. Consequently, the upper sign from the retarded solution gives an energy flow away from the dipole source. In the case of a negative indexed medium the permeability and permittivity are negative. Now, in this case, the lower sign from the advanced solution gives an energy flow away from the dipole source. It is found that, in the positive indexed medium the Poynting vector and wave vector are parallel while they are anti-parallel in the negative indexed medium. This is a general result that is evident from the earlier discussions of energy flow in positive and negative refractive index media.

Both of the problems treated above are basic to the study of radiation and offer a direct comparison of the flow of radiation generated in positive and negative refractive media. Next a different set of problems will be addressed that arise in relativity. These are basic problems in electrodynamics which display surprising differences between the physics of positive and negative indexed media.

Specifically, the Doppler effect and the properties of Cherenkov radiation will be treated. The Doppler effect is an effect involving the frequency shift measured in the light from a moving source of radiation. This is a significant and important result in spectroscopy of atoms and molecules. Following this a return to more complex radiation problems will be made in the consideration of the details of the Cherenkov radiation. Cherenkov radiation is encountered in the motion of relativistic accelerating point charges. A comparison will be made of the Cherenkov radiation in positive and negative refractive indexed media [1–5].

5.3.1 Doppler Effect

Another interesting property of negative index media is the nature of the Doppler effect. This involves a source and observer located within and in relative motion inside a negative index medium [1–5, 44–46]. It turns out that the frequency shift of the source radiation at the observer is quite different depending on whether the source and observer are located in a positive or a negative index media. In the following discussions of both types of media will be treated and the observed Doppler shifts in the two different media compared.

To understand the Doppler effect in both positive and negative index media, consider the Lorentz transformation between the two reference frames in Fig. 5.8. In Fig. 5.8, a source of radiation of frequency ω is located at the origin of the unprimed frame and an observer is located at the origin of the moving primed frame. The primed frame is moving with a velocity v relative to the unprimed frame. The two frames are in a uniform homogeneous medium which can be either positive or negative index media.

The question posed regarding the Doppler effect is how are the frequency ω in the rest frame of the source and the frequency ω' of the source in the observer's rest frame related to one another. This can be determined through the nature of the plane wave form as it transforms between the primed and unprimed frame.

The relationship between the primed and unprimed frequencies are obtained by considering the transformation of the plane wave form $e^{i(\vec{k}\cdot\vec{r}-\omega t)}$ in the unprimed frame to the plane wave form $e^{i(\vec{k}'\cdot\vec{r}'-\omega' t')}$ in the primed frame. Both of these forms are scalars under the Lorentz transformation. Consequently, it follows that

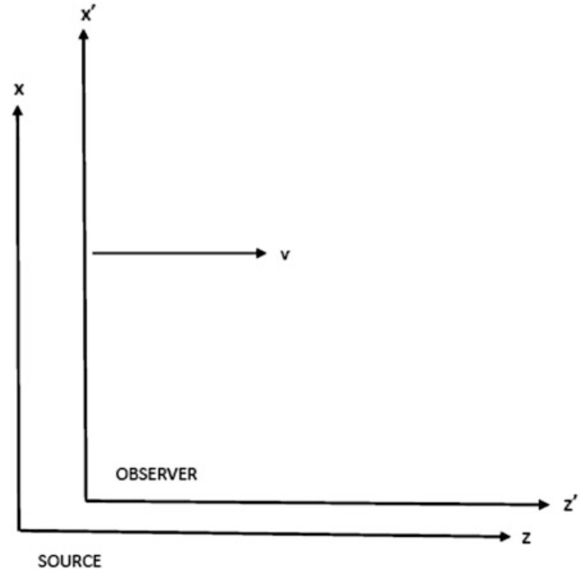


Fig. 5.8 Source and Observer frames of reference for the discussions of the Doppler effect

$$\vec{k} \cdot \vec{r} - \omega t = \vec{k}' \cdot \vec{r}' - \omega' t' \tag{5.103}$$

between the two frames.

To accomplish this transformation, both the 4-vector of position and time and the 4-vector of wave vector and frequency must be transformed by the appropriate Lorentz formulae. Under the Lorentz transformation for the coordinates and time, it follows that between the primed and unprimed frames the variables are related by [44–46]

$$t' = \gamma\left(t - \frac{v}{c^2}z\right), \tag{5.104a}$$

$$z' = \gamma(z - vt), \tag{5.104b}$$

$$x' = x, \tag{5.104c}$$

$$y' = y, \tag{5.104d}$$

where $\gamma = \left[1 - \frac{v^2}{c^2}\right]^{-\frac{1}{2}}$. Similarly, the wave vectors and frequencies in the two frames are related by the Lorentz transformation forms given by [44–46]

$$\omega' = \gamma(\omega - vk_z), \tag{5.105a}$$

$$k_z' = \gamma\left(k_z - \frac{v}{c^2}\omega\right), \tag{5.105b}$$

$$K'_x = K_x, \tag{5.105c}$$

$$k'_y = k_y. \tag{5.105d}$$

From an application of these two sets of relationships, the frequencies in the primed and unprimed frames follow directly.

As an initial point consider the Doppler effect in a positive index medium. From Fig. 5.8 consider an electromagnetic wave originating from the source of frequency ω and propagating to deliver energy to be received by an observer located at the origin in the primed frame. The observer is located to the right of the source and is moving with a velocity v away from the source.

For the source to transmit energy to the right of the source in the unprimed frame, the wave vector in the unprimed frame must be positive, i.e., it should be a vector pointing to the right. For this case under the Lorentz transformation in (5.105a), both $v > 0$ and $k_z > 0$.

From (5.105a) it then follows that

$$\omega' = \gamma(\omega - |vk_z|) < \omega. \tag{5.106a}$$

The frequency perceived by the moving observer in the rest frame of the observer is at a lower frequency than the frequency in the rest frame of the source. As a result, the radiation is Doppler shifted to the red as it is received by the moving observer.

Now consider the Doppler shift in the case that the medium in which the radiation propagates is a negative refractive index medium. Again consider the situation of the source and observer in Fig. 5.8. For the source to transmit energy to the right of the source in the unprimed frame, the wave vector in the unprimed frame must now be negative, i.e., it should be a vector pointing to the left.

This follows as in a negative index medium the energy flow is in the opposite direction to that of the wave vector. For this modified case under the Lorentz transformation in (5.105a), $v > 0$ but $k_z < 0$. From (5.105a) it then follows that [44–46]

$$\omega' = \gamma(\omega + |vk_z|) > \omega. \tag{5.106b}$$

The frequency perceived by the moving observer in the rest frame of the observer is now at a higher frequency than the frequency in the rest frame of the source, i.e., it is shifted to the blue as it is received by the moving observer.

In general, the Doppler shift in a negative index of refraction medium is opposite the sense of the shift in a positive index of refraction medium. Both of the cases treated in (5.106) are for observers receding from the source. The cases in which the observer approached the source is left to the considerations of the reader.

Now consider another relativistic effect. This is the generation of Cherenkov radiation by a moving radiating point charge.

5.3.2 Cherenkov Radiation

A final important radiation problem to treat is that of Cherenkov radiation from a charge moving in a dielectric medium. This involves the study of the configuration of the radiation fields emitted by the charged particle. For the case that the speed of the particle is greater than the speed of light within the medium an important new effect is found. In particular, a characteristic cone of radiation, known as Cherenkov radiation, is generated by the particle with the axis of the cone centered about the velocity vector of the particle. It turns out that there are qualitative as well as quantitative differences in the Cherenkov effect as it is observed in positive and negative refractive index media.

To understand the Cherenkov effect in a negative refractive index medium, first a review of the Cherenkov effect in a positive index of refraction medium is developed. Following this the Cherenkov effect within a negative refractive index medium will be discussed, applying similar arguments as those used in the discussions for the positive index medium. A focus will be given only to the aspects of the Cherenkov effect that exhibit a difference between the two different media.

In Fig. 5.9a a schematic diagram is presented for a radiating charged particle moving in a positive index of refraction medium having a refractive index $n > 0$. The particle is traveling horizontally to the right with a velocity $v > 0$ which is greater than the speed of light, $\frac{c}{n}$, within the medium. In order to radiate energy, the particle must also be accelerating, but in the following discussions the acceleration of the particle will not enter the discussions other than through the assumption that radiation fields are emitted.

Referencing the figure it is seen that in a time Δt the particle will move horizontally on its trajectory through a distance $v\Delta t$. During this same time the radiation emitted at the position of the particle at the beginning of the time interval will

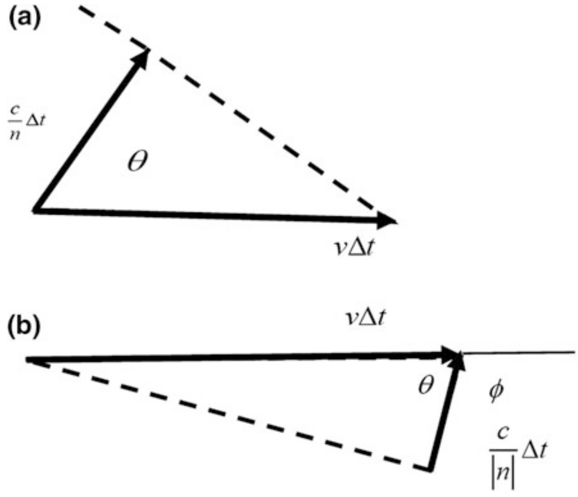


Fig. 5.9 Schematic diagram for the Doppler effect in: **a** a positive index of refraction medium and **b** a negative index of refraction medium. In both figures the particle is moving to the right

propagate a distance $\frac{c}{n}\Delta t$. The radiation is emitted as a spherical wave, but applying Huygen Principle it is found that the waves radiated along the trajectory of the particle add to form a wave front indicated in the figure as a dashed line.

Considering the triangle in Fig. 5.9a formed from the distance traveled by the particle and the distance from the point of emission of the radiation to the wave front, the angle θ of the wave vector cone of radiation emitted from the moving particle is obtained. This cone of wave vectors formed about the velocity vector of the particle is the Cherenkov radiation which, from the right triangle in Fig. 5.9a, makes an angle θ with the velocity vector and is given by [44–46]

$$\cos\theta = \frac{c}{nv} \tag{5.107}$$

Since in the positive index of refraction medium the Poynting vector is parallel to the wave vector of the propagating radiation, θ defines the angle with the velocity vector of the cone of radiated energy emitted as Cherenkov radiation.

The treatment for the Cherenkov radiation from a charge moving in a negative refractive index medium parallels that of the above discussions for a positive refractive index medium. However, unlike the positive index solution, where the radiation is obtained from a retarded solution of the Green's function, the negative index solution uses the advanced solution of the Green's function. Whereas the retarded solution represents spherical waves propagating away from the source, the advanced solution represents spherical waves converging on the source. This is not a problem as in positive index medium the energy flow is away from the source and in the direction of the wave vector, but in a negative index medium the energy flow is away from the source and in the direction opposite the wave vector.

In Fig. 5.9b a schematic diagram is presented for a radiating charged particle moving in a negative index of refraction medium with refractive index $n < 0$. The particle is again traveling horizontally to the right with a velocity $v > 0$ which is greater than the speed of light, $\frac{c}{|n|}$, within the medium.

Referencing the figure it is seen that in a time Δt the particle will move through a distance $v\Delta t$ horizontally towards the right on its trajectory. At any time during this journey the phase of the wave received at the position of the particle originates from the dashed line drawn on the figure. For example, the phase received at the particle after it passes through the distance $v\Delta t$ is located a distance $\frac{c}{|n|}\Delta t$ from the dashed line. This is shown in the figure. The dashed line is again the origin of the phase or wave vector carried by the wave as it is received at the source. The wave vector ultimately is directed towards the source.

In this process, the radiation is received at the location of the sources in the form a spherical wave. The location of the dashed line of phase origin in the figure is a consequence of applying Huygen Principle operating in the time reverse [44–46].

An important point in all of this is that the energy flow in the negative index of refraction medium is opposite to the direction of the wave vector. As the wave vector is received by the source, the energy is propagating away from the source.

Considering the triangle in Fig. 5.9 b formed from the distance traveled by the particle and the distance from the point of phase origin on the dashed line, the angle θ in the figure is given by (5.107). This however, is not the angle of interest.

The angle of interest in the Cherenkov problem is the angle between the particle motion and the cone of radiation emitted by the particle. This angle is denoted by $\phi = \pi - \theta$ in the figure. From the schematic in Fig. 5.9b it follows that [44–46]

$$\cos \phi = \frac{c}{nv}. \tag{5.108}$$

Since in the negative index of refraction medium the Poynting vector is anti-parallel to the wave vector of the propagating radiation, ϕ defines the angle with the velocity vector of the cone of radiated energy emitted as Cherenkov radiation.

5.4 Application of Metamaterials in Antenna Design

On a point of technological application, it is important to note that metamaterials have entered into a number of useful design proposals on antennas [1–5, 32, 33, 40–43]. Antenna engineering is a complex technology that is constantly being modified to meet new developmental requirements in the electrical engineering of devices. In this regards, the implementation of metamaterials has been made to aid in the design of smaller sized antennas with high gain operating over greater bandwidths.

The implementation of metamaterials is proposed based on the novel properties they exhibit as a class of materials. In regards to their new properties, however, not all of the applications of metamaterials in antenna design are based on negative index properties. As an example, in some technologies the development of metamaterials with zero permittivities and/or permeabilities have found application in the design of antennas which manifest interesting radiative properties.

In this sense, the formulation of metamaterials offers another opportunity to extend the range of properties found in naturally occurring materials. Metamaterials are not just designed to display properties absent in naturally occurring materials. They can extend the range of properties known to be available through conventional means [1–5, 32, 33, 40–43].

In the following some of these ideas of metamaterial designs are presented by offering examples from early studies of antennas incorporating metamaterials. For the discussions, a design based on negative index of refraction media and a design based on an application of a zero permeability metamaterial are focused upon. These discussions are given as illustrations of the potential available based on metamaterial technology, and a comprehensive treatment of antenna design is not intended.

An early indication of the potentials of metamaterials in antenna design was made by Ziolkowski and Kipple [40]. They treated the problem of a radiating dipole

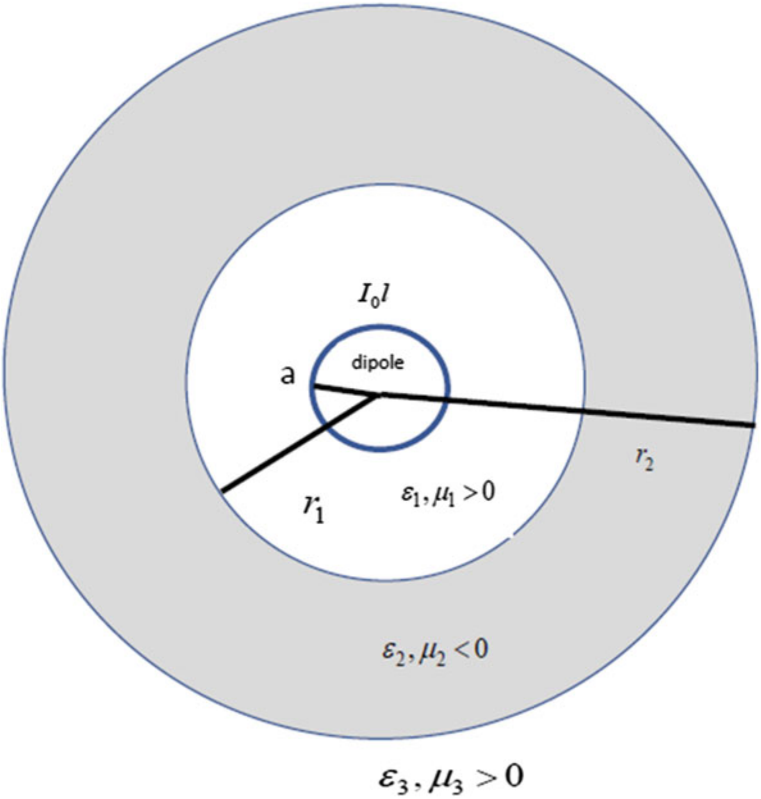

I_0l
a
dipole
r_1
r_2
$\varepsilon_1, \mu_1 > 0$
$\varepsilon_2, \mu_2 < 0$
$\varepsilon_3, \mu_3 > 0$

Fig. 5.10 An infinitesimal electric dipole of strength I_0l centered within a negative index of refraction metamaterial of inner radius r_1 and outer radius r_2 [40]

centered within a shell of negative refractive index medium using both analytical and computer simulation methods. (The schematic for the problem they studied is shown in Fig. 5.10.)

Through their considerations they demonstrated that negative refraction index media can offer important opportunities in antenna design. In the following, their study will be summarized as an example of an important result for radiation problems and antenna design. Specifically, it will be seen that negative index media, if properly employed, can greatly increase the radiation efficiency of antennas.

The problem treated by Ziolkowski and Kipple [40] considered the specific configuration of a dipole located within a metamaterial shell shown in Fig. 5.10. In the figure an electric dipole source of radius a, frequency ω, and dipole moment I_0l is located at the center of a dielectric shell of inner radius r_1 and outer radius r_2. The shell is composed of an homogeneous negative index of refraction medium, but, otherwise, the media inside and outside of the shell are positive refractive index media.

The object of the calculations was to study the radiation fields and the power radiated by the dipole within the shell. This can be used to determine the effects of the negative index of refraction medium on the radiation properties of the system. In their original paper a variety of other geometric configurations were consider, but here the focus is on the result for the problem in Fig. 5.10.

For the system in Fig. 5.10, a particular set of permittivity and permeability parameters of the form

$$(\varepsilon_1, \mu_1) = (-\varepsilon_2, -\mu_2) = (\varepsilon_3, \mu_3) = (\varepsilon_0, \mu_0) \tag{5.109}$$

where considered. In particular, the positive index of refraction media are vacuum and the negative index of refraction medium of the shell has a permittivity and permeability which are the negative of those of the free space values. It should be noted that this arrangement of the permittivities and permeabilities is essentially the same as that used in the treatment of the problem of the perfect lens. In the perfect lens problem studied earlier a negative index slab lens with permittivity and permeability parameters that are the negative of those of free space was surrounded by vacuum.

In the numerical studies of the system in Fig. 5.10, the frequency of the source was $f = 10\,\text{GHz} = c/\lambda_0$ and $I_0 l = 2\lambda_0/1000\,\text{A m}$ with $I_0 = 1\,\text{A}$. These values were selected as a reasonable set of parameters that would characterize a radiation problem of interest in nano-photonics. The metamaterial shell which forms part of the antenna structure had an inner radius $r_1 = 100\,\mu\text{m}$ while the outer radius was varied in the region $r_2 > r_1$. This allowed for a determination of the value of r_2 at which the most power is radiated from the dipole-shell antenna array, i.e., what value gives the best antenna for the radiation generated at the dipole [40].

For the system and parameters just given the total power radiated to infinity by the dipole within the shell was determined. In addition, as a comparison, the total power radiated to infinity by the same dipole placed alone in vacuum was also determined. From the ratio of these two powers the change in relative radiative efficiency of the antenna with and without the shell of metamaterial was determined.

In Fig. 5.11 the ratio of the radiated power for the shell system divided by the radiated power for the dipole placed alone in vacuum is presented as a function of r_2. The antenna with the metamaterial shell is found to be a much improved antenna over that of the dipole alone. The results of Fig. 5.10 are now discussed in some detailed considerations of the particulars [40].

The results plotted in Fig. 5.11 indicated that a tremendous enhancement of the power radiated from the dipole is achieved in the presence of the metamaterial. This is found over a wide range of values for the outer radius of the shell. As a result, a considerable leeway is available for the design of the antenna.

As a particular point, note that over the wide range of outer radii giving enhanced radiative power, a maximum in the power radiated is observed at

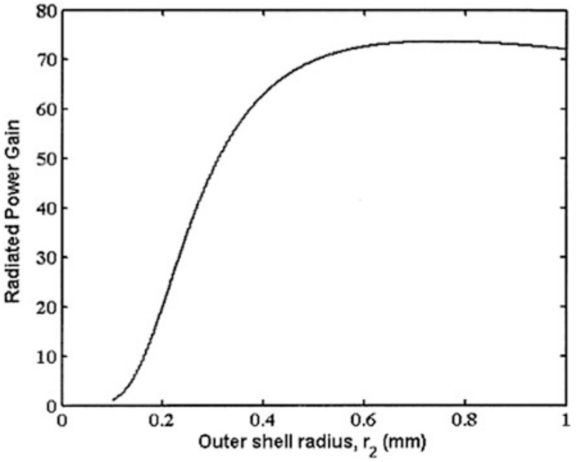


Fig. 5.11 Plot of the ratio of the power radiated by the dipole located within the shell of negative refractive index medium divided by the power radiated by the dipole located within free space versus the outer radius of the dipole shell. The plot is made for a fixed inner shell radius. The details of the parameters of the problem are given in the text [40]. Reproduced with permission from [40]. Copyright 2003 IEEE

$r_2 = 748.8\,\text{mm}$. At this outer radius an enhancement in the radiated power is observed of a little over 70 times the radiated power of the system in the absence of the metamaterial shell.

These results illustrate the potential amplification properties metamaterials with negative index of refraction offer in the design of antennas. It also provides a great potential for enhancing the fields radiated from other types of antennas which, in the absence of metamaterials in their designs, would otherwise be poor radiators. In this regard, in general, conventional antennas require a half wavelength dimension to be efficient radiators whereas metamaterials have allowed for a reduction of the antenna dimensions by 25–50 times.

The above example employed negative index of refraction metamaterials as a means to improve antenna performance. Some useful effects can also be developed using metamaterials that exhibit zero permittivities and/or permeabilities. These dielectric properties of the metamaterial can be used to affect the directionality of the radiation patterns generated by antennas as well as to affect their radiation efficiencies [40].

As an example of this type of system consider a slab of material composed of such a metamaterial. It shall be shown that by judiciously choosing the permittivities and permeabilities a directional antenna can be designed. The discussions presented in the following are based on a study given by Tang, Mei, and Cui [42]. In Fig. 5.12 the properties of the radiation fields in an antenna system, based on a design employing a metamaterial in which the permeability tensor has a specific form, is considered. The system is based in part on an antenna formed from an infinite wire carrying a harmonically time varying current. The metamaterial enhanced antenna is then formed by placing the wire centered within a slab of metamaterials [40].

To establish a point of comparison it is good to begin by considering the fields generated by the wire alone. In Fig. 5.12a the radiation fields from an infinitely long

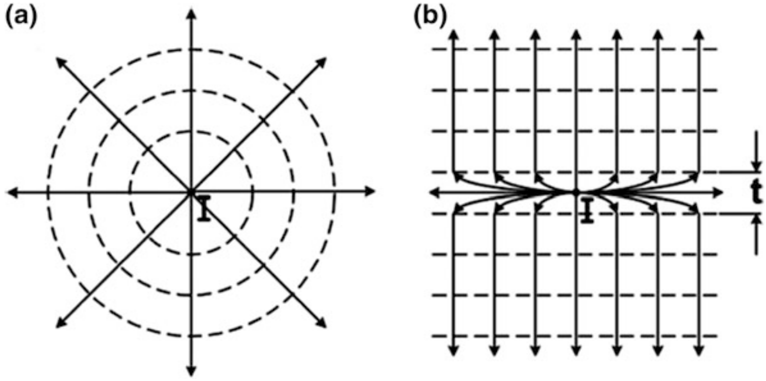


Fig. 5.12 The radiation fields from: **a** an infinite wire placed along the z-axis in free space and **b** the same infinite wire located in a metamaterials which has a zero y-component of its magnetic permeability tensor [42]. Reproduced with permission from [42]. Copyright 2015 Springer

wire carrying an harmonically varying current are illustrated. The wire generates spherical waves which are radiated from the wire and travel out from it as dashed cylindrical wave crests. The circles of increasing radii in the figure then represent the crests of the waves at a particular point in time.

Next a metamaterial slab with a zero permeability component medium is introduced into the system so that the radiating wire in contained within the slab. In particular, the y-component of the permeability tensor is set to zero in the metamaterial slab. For the metamaterial antenna the infinite wire is place at the center of the infinite metamaterials slab of thickness at t as shown in Fig. 5.12b.

As a consequence of the zero permeability component of the slab the radiation pattern is distorted into the new field pattern shown in Fig. 5.12b. Now the dashed radiation wave crests move off in planes parallel to the surfaces of the metamaterial slab. The radiation generated by the infinite wire in every direction has been redirected to move off from the antenna in a very restricted sense. In addition, the antenna also focuses the total power from the antenna, concentrating it along one axis in space.

This particular system provides an illustration of the transformation of the radiation fields available with the applications of metamaterials. The metamaterial does not necessarily need to display negative refractive index properties as the focus of its application.

Of course, there are many other important applications of metamaterial ideas to obtain antenna solutions [40–43]. Some of these include the development of strips of metamaterial transmission lines as well as smaller configurations of resonators formed in a variety of combinations and configurations [42]. The resonance properties of these can be designed to minimize their reflected power properties and set the frequency bands over which they exhibit these minima. This is an aid in the antenna designs. In addition, some ideas of antenna design involve the layering of positive and negative refractive index media [42]. This is done in such a way that the negative refractive index slabs compensate for phase changes in the positive layers. Such layers rid the system of phase changes which inhibit the radiation fields generated by these types of structures. Also applications based on two and three dimensional arrays of nano-circuits exist and are found to display important radiative features [42]. As well as radiative systems operating over a wide range of frequencies, some metamaterial systems have been developed to inhibit thermal radiation for surfaces and coating. This is also an important function of the ideas of antenna operations.

There are many such ideas that have been put forth for designs based on ideas from the foundations of metamaterials, and these cannot be further gone into here. For details of metamaterial antennas and their further properties the reader is referred to the literature [40–43].

5.5 Photonic Crystal Solutions to the Negative Refractive Index Problem and Hyperbolic Materials

In this section discussions are presented of systems based on photonic crystals which can exhibit some of the properties of negative index of refraction metamaterials [47–49]. The first systems considered are photonic crystals exhibiting electromagnetic band structures and which are operated in the diffractive limit. A focus will be on these systems considering the case in which the electromagnetic dispersion relation displays regions in wave vector space in which the electromagnetic modes of the photonic crystal have oppositely directed phase and group velocities. In these regions the phase velocity and wave vectors of the modes are opposite the direction of the energy flow in the system. As note in the earlier discussions of the basic properties of metamaterials, this feature of the energy flow in the electromagnetic modes is an essential property of negative index of refraction media. For the photonic crystal systems the negative index is a band structure effect and the wave vectors of the modes of interest are of order of the inter-atomic spacing.

A second type of system considered is a metamaterial based on photonic crystal like layerings or on arrays of cylinders or nano-wires. Due to their periodicity properties these systems look like photonic crystals, but they are designed to act as metamaterials for light which has much larger wavelengths than the fundamental lengths characterizing the periodicity of the media. For these wavelengths of light the materials appear to be homogeneous media with dielectric properties which are anisotropic in space. Such materials are known as hyperbolic metamaterials and depending on their specific structure they exhibit metallic (dielectric) behavior along one axis of space and dielectric (metallic) properties in the plane perpendicular to this axis. The essential feature that characterizes these materials is that light propagating in them has a hyperbolic dispersion relation. This gives them many interesting properties that are related to negative index media and to the formation of lenses displaying features of prefect lenses [47–49].

5.5.1 Photonic Crystals

A photonic crystal functioning in the diffractive limit of its interaction with electromagnetic radiation modes can exhibit properties of a negative refractive index material. This comes from the form of the band structure of the electromagnetic modes in the photonic crystal. In particular, in some cases it is possible to have regions of the photonic crystal wave vector space in which there are modes which have the modal phase velocity opposite to that of its group velocity.

In Fig. 5.13 an example of such a dispersion relation is presented. Shown in the figure is a hypothetical plot of the frequency of electromagnetic modes as a function of the wave vector. Two bands are presented as an illustration, and the upper band is the one used to understand the functioning of the photonic crystal as a negative refractive index medium. The dispersion relation of this band is designed so as to exhibit oppositely directed phase and group velocities [47–49].

In the region near the origin in the frequency versus wave vector plot, the upper band has a dispersion relation in which the frequency of the modes increase in frequency as the wave vector is decreased to zero. The dispersion relation in this region is then approximated by the form [47–49]

$$\omega(k_x, k_y, k_z) = \omega_0 - \frac{1}{2}\omega_1(k_x^2 + k_y^2 + k_z^2). \tag{5.110}$$

Computing the group velocity obtained from (5.110) at the wave vector (k_x, k_y, k_z) gives the result

$$\vec{v}_g = \nabla\omega(k_x, k_y, k_z) = -\omega_1(k_x, k_y, k_z) \tag{5.111}$$

It is seen that the direction of the group velocity is opposite that of the wave vector and, consequently, parallel to the phase velocity. As a result, the energy flux of the mode is opposite the mode vector of the mode. This is a fundamental signature feature of a negative refractive index medium.

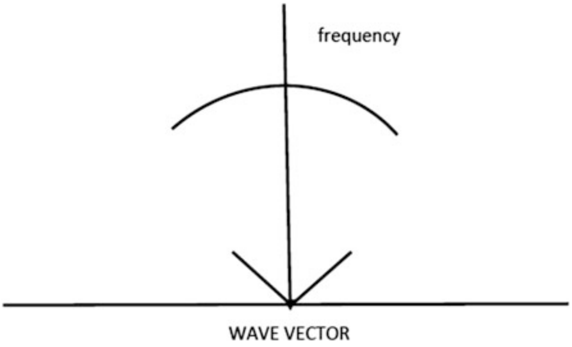


Fig. 5.13 Schematic of the dispersion relations in two bands of a photonic crystal in a region near and centered about the center of the Brillouin zone

A similar effect is found in the transport properties of electrons in metals and semi-conductions. In these systems the presences of regions of modal solutions in which the phase and group velocity are opposite one another has been studied early on. The origin of the effect, as with the photonic crystal effect, arises due to the strong Bragg scattering of the modes from the periodic lattice of the system. Consequently, it is a diffraction effect whereas in metamaterials the negative index property is a refractive rather than a diffractive effect.

5.5.2 Hyperbolic Materials

A second application of photonic crystals to negative refractive index technology is in the development of hyperbolic metamaterials. These are designed to function in the refractive limit of photonic crystal implementation and to exhibit similar properties to negative refractive index media. They are artificial structures formed as layered media or composite media based on nano-wire arrays, and their interesting properties derive from the form of the dispersion relation of electromagnetic waves propagating within them.

The basis of the interesting properties of hyperbolic metamaterials is their hyperbolic dispersion relations [23–25, 42, 43]. In a hyperbolic medium the frequency versus wave vector dispersion relation of electromagnetic radiation exhibits constant frequency surfaces which are hyperbolas of revolution in phase vector space. To understand the nature of the properties arising from these unusual dispersion relations consider the nature of the dispersion relation of light in a uniform positive medium as compared to those of a metamaterial medium with a hyperbolic dispersion relation.

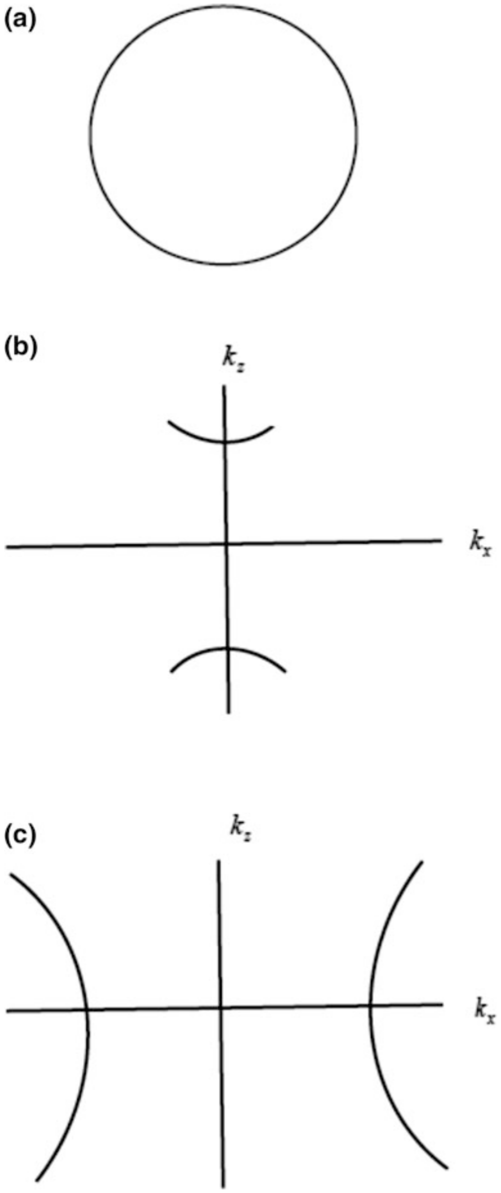
First consider light in a uniform medium with a positive index of refraction. The dispersion relation of light in such a medium is written as [23–25, 42, 43]

$$\frac{k_x^2 + k_y^2 + k_z^2}{\mu\varepsilon} = \omega^2. \tag{5.112}$$

As readily seen, the dispersion relation in (5.112) represents the propagation of modes which have constant frequency surfaces in the form of spheres.

For a useful later comparison with the dispersion relations of the hyperbolic media, in Fig. 5.14a a plot of the dispersion relation of the spheres of the positive index medium are shown in the x-z plane. This gives a cross section of the dispersion relation which appears as a circular cross section intersecting in the x-z plane for which $k_y = 0$ in (5.112). The circular nature of the cross section of the constant frequency surface fixes many of the qualitative behaviors of the wave propagation in the medium. These are now addressed.

Fig. 5.14 Schematics of constant frequency surfaces in: **a** positive index refractive medium, **b** Type I hyperbolic metamaterial, and **c** Type II hyperbolic metamaterial. The rotation symmetry of the surfaces is about the vertical axis



To understand the qualitative aspects of the propagation of waves of frequency ω, rewrite (5.112) considering the $k_y = 0$ surface. In this way (5.112) takes the form [23–25]

$$k_z^2 = \mu\varepsilon\omega^2 - k_x^2. \tag{5.113}$$

Next consider the nature of wave propagation in the z-direction as it depends on the value of k_x. Two types of behaviors along the z-direction are observed. For $\mu\varepsilon\omega^2 > k_x^2$ the wave propagates along the z-direction as a plane wave, but for $\mu\varepsilon\omega^2 < k_x^2$ the wave is evanescent along the z-axis. Consequently, only small k_x propagate whereas larger k_x decay in space.

An important point is to note the different behaviors of the k_x in the transfer of an image through the optical system. If a wave is emitted from a source on the z-axis and travels on the z-axis along the positive z-direction, plane wave states will travel indefinitely along the z-axis. Evanescent waves, however, will decay and be gradually lost to the radiation fields as they move along the z-direction.

The loss of the evanescent waves in the radiation emitted by the source represents a loss of information about the source, and the loss of information from the evanescent components increases with distance from the source along the z-direction. In terms of the information radiated into space, only the propagating plane wave component of the radiation from the source can faithfully transfer information about the source along the z-direction. In its final form, the resulting image generated from the energy radiated by the source and sent through the system will be incomplete when it is reassembled into the form of an image.

In conclusion, it follows that if a vector field pulse of the form $\vec{f}(x, y, z = 0)$ is created in the x-y plane, it will increasingly lose evanescent information in its propagation for increasing $z > 0$. It shall now be seen that this is not the case if the medium into which the source radiates is a hyperbolic medium. This is done by following the above considerations, treating the form of the hyperbolic dispersion relation. First some introduction will be given regarding the hyperbolic form of dispersion relation and how it can occur.

Consider the propagation of radiation in the case of a hyperbolic medium. Hyperbolic media have permittivity and permeability tensors that are of the general forms

$$\begin{vmatrix} \varepsilon_\perp & 0 & 0 \\ 0 & \varepsilon_\perp & 0 \\ 0 & 0 & \varepsilon_{||} \end{vmatrix} \tag{5.114a}$$

$$\begin{vmatrix} \mu_\perp & 0 & 0 \\ 0 & \mu_\perp & 0 \\ 0 & 0 & \mu_{||} \end{vmatrix} \tag{5.114b}$$

These matrices yield dispersion relations for two modes which differ in polarization and are known as the ordinary and extraordinary waves. The waves of interest to the following discussions are the extraordinary waves, and it is assumed that only these modes are excited for the properties discussed. The dispersion relation of the extraordinary modes is given by [23–25]

$$\frac{k_x^2 + k_y^2}{\mu\varepsilon_{||}} + \frac{k_z^2}{\mu\varepsilon_{\perp}} = \omega^2. \tag{5.115}$$

where it is assumed that $\mu_{||} = \mu_{\perp} = \mu$.

In the dispersion relation in (5.115) the cases of interest in the study of hyperbolic materials are for $\mu\varepsilon_{||} < 0$ and $\mu\varepsilon_{\perp} > 0$ in Type I hyperbolic materials and for $\mu\varepsilon_{||} > 0$ and $\mu\varepsilon_{\perp} < 0$ in Type II hyperbolic materials. For positive μ it is seen then that Type I materials are metallic in the $\varepsilon_{||}$ direction and dielectric in the plane perpendicular to this direction. In the case of Type II materials for positive μ these materials are dielectric in the $\varepsilon_{||}$ direction and metallic along the $\varepsilon_{\perp}$ plane perpendicular to this direction.

The essential feature of hyperbolic materials in the determination of their electromagnetic properties is their dispersion relation as obtained from (5.115). A schematic plot of the dispersion relation in the *x*-*z* plane for Type I and Type II materials is shown in Fig. 5.14. For the plots of both Type I and Type II materials the dispersion relations are found to intersect the *x*-*z* plane in hyperbolic surfaces of constant ω.

The important point in these plots is that the constant frequency surfaces, unlike those of the uniform positive refractive index medium, are unbounded, i.e., the constant frequency surfaces extend to regions in wave vector space in which the length of the wave vectors are arbitrarily large. This has significant consequences for the propagation of radiation from sources and the density of electromagnetic modes in hyperbolic materials. It is at the basis of two of the most important technological applications of hyperbolic materials.

Before further discussing the radiative and electromagnetic density of states properties of hyperbolic metamaterials, it is important to take a break here and to note some practical material science aspect to these media. Conventional crystalline (non-artificial) solids exist which display hyperbolic tensors of the form in (5.114). These, however, are limited in the frequency range over which they exhibit hyperbolic properties. The metamaterial format is needed to increase the range of frequencies over which the hyperbolic properties are available. In addition, the detailed nature of the composite geometry in the formulation of the composite material is fashioned to match the technology of it application and can take many different forms and formats.

Some examples of the materials used in layered and nanowire hyperbolic composite media and their applications to the electromagnetic spectra as media composed of two components are featured in the following. The ranges of applications that have been made for the indicated frequency range are: (1) In the ultraviolet Au/Al$_2$O$_3$ and Ag/Al$_2$O$_3$, (2) In the visible Au/TiO$_2$ and Ag/TiO$_2$, (3) In the ultraviolet Au/Al$_2$O$_3$ and Ag/Al$_2$O$_3$, (4) In the near-infrared Ti/N and Zr/N, and (5) In the mid-infrared and terahertz III–V semiconductors. Hyperbolic media have been formed from these composites in both layered and nanowire composites, and for the particulars of the metamaterial designs employed in these formulations the reader is referred to the literature.

To understand the qualitative electrodynamic properties of composite hyperbolic materials, revisit the considerations in (5.112) and (5.113) but now applied to the hyperbolic media. For these discussions consider the case of a Type I hyperbolic material. A similar argument can be easily extended to treat Type II materials.

In Type I media the general from of the dispersion relation of light now is written in the form

$$\frac{k_x^2 + k_y^2}{\mu\varepsilon_{||}} + \frac{k_z^2}{\mu\varepsilon_{\perp}} = \omega^2. \tag{5.116}$$

This dispersion relation is to be considered in the x-z plane in phase space. With this restriction (5.116) can be rewritten for $k_y = 0$ into the form

$$k_z^2 = \mu\varepsilon_{\perp}\left[\omega^2 - \frac{k_x^2}{\mu\varepsilon_{||}}\right] \tag{5.117}$$

for which $\mu\varepsilon_{||} < 0$.

Following the earlier treatment for the positive index media in (5.112), again consider the nature of wave propagation in the z-direction as it depends on the value of k_x. This will reveal the important qualitative nature of the propagation of radiation for a source located in a hyperbolic medium and the nature of the information it carries with it. It is seen for the hyperbolic material in (5.117) that, unlike the positive refractive index medium, the wave vector k_z is always real. In this medium, there are no evanescent waves.

As a consequence, if a pulse in the form of a vector field $\vec{f}(x, y, z = 0)$ is created in the x-y plane, it does not lose evanescent information from the source as it travels in the $z > 0$ direction. Unlike in the positive medium, there is no evanescent information to lose. All of the radiated modes from the source are plane wave, propagating, states. Consequently, all of the information from the source is retained in some form as it is propagated from the source.

A similar result is obtained from a consideration of the propagation in Type II media, and the arguments are essentially the same in both Type I and Type II media. Based on these results one can design a type of waveguide of hyperbolic metamaterial in the form of a finite slab. By placing a dipole source appropriately at one side of the slab, the evanescent modes from the source as they enter the slab of hyperbolic medium are made to propagate as plane waves in the hyperbolic slab. Upon encountering the opposite side of the slab from that of the source, these waves are extracted. The extracted waves retain the information originally generated at the dipole source.

Since the hyperbolic metamaterial delivers all components of the radiation field it is natural to try to apply it in the design of a perfect lens. An attempted to make a perfect lens in the form of a slab of hyperbolic material, however, has not been successful. This is because of the anisotropy of the dielectric medium. In particular the absence of evanescent waves is not effective for every propagation direction of the materials and, under general conditions, a properly focused image is not possible.

Attempts have been made at the design of slab lenses similar to the perfect lens of negative refraction medium discussed in earlier sections. Some aspects of negative refraction are present in a slab lens in which $\varepsilon_{||} < 0$ is perpendicular to the slab surface and $\varepsilon_{\perp} > 0$ is parallel to the surface. For this medium the rays from the source that make small incident angles on the slab surface exhibit negative index of refraction properties. In particular, the phase and group velocities are opposite one another. As the incident angle is increase, however, the system eventually loses this property and the incident rays are no longer focused by the slab. The perfect lens becomes imperfect in this sense.

Some advances in lens designs have been made through the combination of hyperbolic materials and curved geometries to form so-called hyperlenses. Applying these ideas an imaging lens can be made that acts on the near-field radiation generated by a source on one side of the lens and forms on the other side of the lens a subwavelength image of the source or object. To do this the layered hyperbolic metamaterial is bent into a half cylinder [23–25].

In this arrangement, the layering of the metamaterial is perpendicular to the radian vector going out from the axis of the cylinder. For this geometry, the near field object is transformed by the lens into a far field image with a greater resolution than that which is theoretically attainable from lenses of classical optics. In the last stages of the generation of the final far field image the radiation exiting the metamaterial is acted on by a conventional lens. On the whole, however, the resulting image is more resolved than that obtained using standard optical methods. This type of lens arrangement is often referred to as a hyperlens.

A final development of hyperbolic metamaterials that is of interest for technology is the effect they have on the electromagnetic density of states within the hyperbolic medium. Unlike the frequency density of states of radiation in free space, which is related to the bounded spherical surfaces of constant frequency, hyperbolic materials have constant frequency surfaces based on the hyperbolic constant frequency surfaces obtained from (5.116) and (5.117). This means that the constant frequency surfaces of hyperbolic materials are of infinite extent in wave vector space. An interesting consequence arises from this in the decay of excited atoms and molecules located within these two different types of materials [23–25].

The density of frequency states for modes of frequency ω are the number of modes between the ω and the $\omega + \Delta\omega$ surface of the phase space dispersion relation in the limit as $\Delta\omega \to 0$. This is very large for hyperbolic materials and represents an increase in their density of states from that of the density of states of frequency ω in free space.

The rate of decay of an excited atom or molecule is found to be proportional to the density of frequency states of the electromagnetic mode radiated into by the atom during its decay. The rate of decay will then depend on the density of frequency modes in the material within which the atom is located. For the enhanced density of states provided by the hyperbolic medium there is an increase in the rate of decay of atoms over that found in free space which has a lower density of states.

The properties of materials with enhanced frequency density of states carries over to the case of an atom or molecule in the proximity of the hyperbolic metamaterials. Recent studies have been made of the radiation rates of atoms as a function of their separation from a hyperbolic medium. An enhancement is observed arising from the proximity of a medium with an enhanced number of modes for the atom or molecule to radiate into.

References

1. A.R. McGurn, *Nonlinear Optics of Photonic Crystals and Meta-Materials* (Claypool & Morgan, San Rafael, 2015)
2. W. Cai, V. Shalaev, *Optical Metamaterials: Fundamental and Applications* (Springer, New York, 2010)
3. N. Engheta, R.W. Ziolkowski (eds.), *Metamaterials: Physics and Engineering Explorations* (IEEE Press, Wiley-Interscience, Wiley, Canada, 2006)
4. S.A. Ramakrishna, Physics of negative index materials. Rep. Prog. Phys. **68**, 449 (2005)
5. V.G. Veselago, The electrodynamics of substances with simultaneously negative values of ε and μ. Sov. Phys. Usp. **10**, 509 (1968)
6. P. Giri, K. Choudhary, A. Sen Gupta, A.K. Bandyopadhyay, A.R. McGurn, Klein-Gordon equation approach to nonlinear split-ring resonator based metamaterials: one-dimensional systems. Phys. Rev. B **84**, 155429 (2011)
7. M. Eleftheriou, N. Lazarides, G.P. Tsironis, Magnetoinductive breathers in metamaterials. Phys. Rev. E **77**, 036608 (2008)
8. G.V. Eleftheriades, EM transmission-line metamaterials. Mater. Today **12**, 30 (2009)
9. I. Kourakis, N. Lazarides, G.P. Tsironis, Self-focusing and envelope pulse generation in nonlinear magnetic metamaterials. Phys. Rev. E **75**, 067601 (2007)
10. M. Lapine, M. Gorkunov, K.H. Ringhofer, Nonlinearity of a metamaterial arising from diode inserts into resonant conductive elements. Phys. Rev. E **67**, 065601 (2003)
11. I.V. Shadrivov, S.K. Morrison, Y. Kivshar, Tunable split-ring resonators for nonlinear negative-index metamaterials. Opt. Express **14**, 9344 (2006)
12. M.P. Marder, *Condensed Matter Physics*, 2nd edn. (Wiley, Hoboken, 2010)
13. J.B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966 (2000)
14. V.M. Agranovich, Y.N. Gartstein, Spatial dispersion and negative refraction of light. Phys. Usp. **49**, 1029 (2006)
15. V.M. Agranovich, Hybrid organic-inorganic nanostructures and light-matter interaction, in *Problems of Condensed Matter Physics*, ed. by A.L. Ivanov, S.G. Tikhodeev (Clarendon Press, Oxford, 2006) (Chapter 2)
16. I.V. Shadrivov, A.A. Zharov, Y.S. Kivshar, Second harmonic generation in nonlinear left-handed metamaterials. J. Opt. Soc. Am. B **23**, 529 (2006)
17. Y. Dong, T. Itoh, Metamaterial-based antennas. Proc. IEEE **100**, 2271 (2012)

18. J. Wang, W. Zhou, E.-P. Li, Enhancing the light transmission of plasmonic metamaterials through polygonal aperture arrays. Opt. Express **17**, 20349 (2009)
19. A.V. Kildishev, V.M. Shalaev, Transformation optics and metamaterials. Phys. Usp. **54**, 53 (2011)
20. U. Leonhardt, T.G. Philibin, General relativity in electrical engineering. New J. Phys. **8**, 247 (2006)
21. U. Leonhardt, T.G. Philibin, Transformation optics and the geometry of light. Prog. Opt. **52**, 69 (2009)
22. J. Van Bladel, *Relativity and Engineering* (Springer, Berlin, 1984)
23. V.P. Drachev, V.A. Pololsky, A.V. Kildishev, Hyperbolic metamaterials: new physics behind a classical problem. Opt. Express **21**, 15048–15064 (2013)
24. A. Poddubny, I. Iorh, P. Belov, Y. Kivshar, Hyperbolic metamaterials. Nat. Photonics **7**, 959–967 (2013)
25. P. Shekhar, J. Atkinson, Z. Jacob, Hyperbolic metamaterials: fundamentals and applications. Nano Conver. (2014). https://doi.org/10.1186/s40580-014-0014-6
26. J.B. Pendry, A.J. Holden, D.J. Robbins, W.J. Stewart, Magnetism from conductors and enhanced nonlinear phenomena. IEEE Trans. Microw. Theory Tech. **47**, 2075 (1999)
27. S. Hrabar, J. Bartolic, Capacitively loaded loop as basic element of negative permeability metamaterial, in *Proceeding of European Microwave Conference*, vol. 2, (Milan, 2002), p. 357
28. S. Hrabar, J. Bartolic, Simplified analysis of split ring resoanator used in backward meta-materials, in *Proceedings of International Conference on Mathematical Methods in Electromagnetic Theory*, vol. 2 (Kiev, 2002), p. 500
29. E. Shamonina, V.A. Kalinin, K.H. Ringhofer, L. Solymar, Magnetoinductuve waves in one, two, and three dimensions. J. Appl. Phys. **92**, 6252 (2002)
30. I.V. Shadrivov, A.N. Reznik, Y.S. Kivshar, Magnetoinductive waves in arrays of split-ring resonators. Phys. B **394**, 180 (2007)
31. D.R. Smith, W.J. Padilla, D.C. Vier, S.C. Nemat-Nasser, S. Schultz, A composite medium with simultaneously negative permeability and permittivity. Phys. Rev. Lett. **84**, 4184 (2000)
32. C. Enkrich, M. Wegener, S. Linden, S. Burger, L. Zschierich, F. Schmidt, J.F. Zhou, Th Koschny, C.M. Soukoulis, Magnetic metamaterials at telecommunication and visible frequencies. Phys. Rev. Lett. **95**, 203901 (2005)
33. B.D. Braaten, R.P. Scheeler, M. Reich, R.M. Nelson, C. Bauer-Reich, J. Glower, G.J. Owen, Compact metamaterial-based UHF RFID antennas: deformed omega and split-ring resonator structures. ACES J. **25**, 530 (2010)
34. T.J. Yen, Y.-C. Lai, A plasmonic biosensor demonstarates high sensitivity and long-distance detection. SPIE News. (2011). https://doi.org/10.1117/2.1201107.003782
35. J.D. Baena, J. Bonache, F. Martin, R.M. Sillero, T. Lopetegi, M.A.G. Laso, I. Gil, M.F. Portillo, M. Sorolla, Equivalent-circuit models for split-ring resonator and complementary split-ring resonators coupled to planar transmission lines. IEEE Trans. Microw. Theory Tech. **53**, 1451 (2005)
36. H.S. Chen, L.X. Ran, J.T. Huangfu, X.M. Zhang, K.S. Chen, T.M. Grzegorczyk, J.A. Kong, Magnetic properties of S-shaped split-ring resonantors. Progr. Electromang. Res. PIER **51**, 231 (2005)
37. J.B. Pendry, D. Schurig, D.R. Smith, Controlling electromagnetic fields. Science **312**, 1780–1782 (2006)
38. D. Schurig, J.J. Mock, B.J. Justice, B.J. Cummer, J.B. Pendry, A.F. Starr, D.R. Smith, Metamaterial electromagnetic cloak at microwave frequencies. Science **314**, 977–980 (2006)
39. M. Anand, Applications of metamaterials in antenna engineering. Int. J. Tech. Res. Appl. **2**, 49–52 (2014)
40. R.W. Ziolkowski, A.D. Kipple, Application of double negative materials to increase the power radiated by electrically small antennas. IEEE Trans. Antennas Propag. **51**, 2626–2640 (2003)

41. M.A. Henari, Compact meta-material antenna with a-shaped topology for ultra wide band microwave communications. SOP Trans. Wirel. Commun. **1**, 32–39 (2014)
42. W.X. Tang, Z.L. Mei, T.-J. Cui, Sci. China: Phys. Mech. Astron. **58**(12), 127001 (2015). K. V. Ajetrao, A.P. Dhande, Review on metamaterial and its application as antenna. Adv. Eng. Technol. 95–100 (2015)
43. S. Yan, Metamaterial design and its applications for antennas. Ph.D. thesis in Electrical Engineering, KU Leuven, Science, Engineering & Technology (2015)
44. V. Ginis, J. Danckaert, I. Veretennicoff, P. Tassin, Transforming Cherenkov radiation in metamaterials. Proc. SPIE **9546**, 9546Q-1 (2015)
45. J.-K. So, J.-J. Won, M.A. Sattorov, S.-H. Bak, K.-H. Jang, G.-S. Park, D.S. Kin, F. J. Garcia-Vidal, Cerenkov radiation in metallic metamaterials. Appl. Phys. Lett. **97**, 151107 (2010)
46. H. Chen, M. Chen, Flipping photons backward: reversed Cherenkov radiation. Mater. Today **14**, 34–41 (2011)
47. C. Luo, S.G. Johnson, J.D. Joannopoulos, All-angle negative refraction without negative effective index. Phys. Rev. B **65**, 201104 (2002)
48. E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopoulou, C.M. Soukoulis, Electromagnetic waves: negative refraction by photonic crystals. Nature **423**, 604–605 (2003)
49. V. Mocella, Negative refraction in photonic crystals: thickness dependence and Pendellosung phenomenon. Opt. Express **13**, 1361–1367 (2005)

Chapter 6
Force

In this chapter forces that are often taken into consideration in the study of nanoscience are discussed. These are forces that can be used in the manipulation of individual nanoparticles or in the assembly of systems formed from an ordered arrangement of nanoscale features. Such interactions are very important for technological applications as well as in developing an understanding of how small particles interact with their environments. In this regard, there are many applications in self-assembly processes and in the manipulation of both biological and non-biological particles.

In the following, some discussions of the use of magnetic and electric fields in particle manipulation are given. Treatments of the properties of ferromagnetic, paramagnetic, and diamagnetic particles manipulated by the use of externally applied magnetic fields are presented [1–7]. Each of these three types of magnetic systems is found to display its own characteristic properties in an interaction with an applied field, and these interactions facilitate their technological uses in biological [1, 5, 7] and non-biological systems [1–7].

Similar discussions are also given of the interaction of ferroelectric and non-ferroelectric particles in an applied electric fields [8–10]. Again, these have various technological applications which distinguish between the dielectric properties of the particles involved.

These discussions will be followed by considerations of the trapping of individual ions in space. In order to trap ions in space it is necessary to use time varying electromagnetic fields, and the trapping of such ions is very important in many studies in quantum optics [11, 12]. Such quantum optics treatments include both technologically based discussions and in tests of the foundations of quantum mechanics.

An important device in the technology of particle trapping and manipulation is the optical tweezer [13–20]. This has many applications in the study of biological and non-biological particles where it exhibits important means of positioning and orienting nano-particles. A brief discussion of the basic theory for the operation of the optical tweezer will be given.

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_6

In a final presentation the Casimir effect will be treated [21–28]. This is observed as a weak short range force that exists between surfaces and which may be attractive or repulse in nature. The Casimir force arises solely from the properties of the electromagnetic density of states as they are modified by the geometry and dielectric properties of the surfaces. These surface interactions can be important in various mechanical considerations of nanoscience machines.

6.1 Magnetic Forces for the Manipulation of Nanoparticles

One means of manipulating nanoparticles is through the application on them of static external magnetic fields [1–7]. This can be a very important device in nanoscience for the physical transport of particles through space or for their alignment in space. In such applications the applied fields acting on nanoparticles with magnetic properties are found to develop forces and torques on the particles. These interactions cause the particles to align or to be propelled in space relative to the lines of applied magnetic induction. They are basic ideas which set the stage for many important applications in the engineering of nano-machines, the manipulation of biological cells such as bacteria, for the sorting and assembly of nanoparticles, and for applications in drug delivery systems [1–7].

Generally, interactions of this type are particularly important in the nanoscience mechanics where they enter into considerations of forces or torques on particles exhibiting paramagnetism, diamagnetism, or ferromagnetism. Forces and torques in these systems ultimately arise from the changes in the energy of interaction of the particle as the configuration of the system is changed. In particular, the force on the particle is related to the negative of the derivative of the particle energy with respect to some generalized space coordinated used to represent the orientation or motion of the particle in space. Ultimately, the theory of these nanoparticle interactions has an origin in the most basic fundamentals of the theory of magnetostatics [1–7].

In the following, after a brief review of some basic theoretical ideas of magnetostatics, the mechanical effects that are commonly employed in the manipulation of paramagnetic, diamagnetic, and ferromagnetic particles are discussed. The first system to be treated is that of ferromagnetic nanoparticles. These exhibit permanent magnetic moments that can be influenced by an applied magnetic field but are not dependent on the applied fields for their existence. After this discussions of paramagnetic and diamagnetic particles are given. In these types of particles the magnetic moments of the particles are induced by the applied fields so that in the absence of an applied field the particle has no magnetic moment [1–7].

In the discussion of the particle dynamics, two different field types are addressed. In uniform fields the particles can be oriented in space. For spatially non-uniform fields, however, the particles can be both oriented and propelled along trajectories in space.

After the theoretical treatment of the forces and torques on the three different types of particles, some of the recent applications in nanoscience systems are given as illustrations. These include: applications to mechanical nano-projectiles that can be steered through space along predetermined trajectories, the design of drug delivery system, the guiding of bacteria through space, and the separation of nanoparticles on the basis of their various different sizes, etc. [1–7].

6.1.1 Review of Magnetostatics

In the study of magnetism three basic fields are encountered. These are the magnetic field, $\vec{H}$, the magnetic induction, $\vec{B}$, and the magnetization, $\vec{M}$. Each of these three vectors arises from different currents flowing in the materials they characterize, and, ultimately, the energy and force on particles in an applied magnetic field are all expressed in terms of these vectors.

In the following, the origins and relations between the three fields are reviewed. This is followed by some general considerations of the energy and forces on nano-particles interacting with an external magnetic field.

Properties of the Three Magnetic Vectors

From Maxwell's equations, it is found that the magnetic field is generated by the currents of free charge moving in the system, e.g., conduction electrons internal or external to the media. The magnetization arises, however, from a different set of charges that are bound to atoms and molecules and are confined to move within them. These bound charges then generate atomic or molecular currents.

The bound charge currents, in their circulation about the atoms and molecules, create magnetic moments which are ultimately related the magnetization as the vector sum of the magnetic dipoles of the individual atoms. Finally, the magnetic induction is related to both the magnetic field and the magnetization by the relationship [1]

$$\vec{B} = \mu_0\left[\vec{H} + \vec{M}\right] \tag{6.1}$$

where μ_0 is the permeability of free space.

For paramagnetic and diamagnetic media the magnetization is linearly related to the magnetic field through the relationship [1]

$$\vec{M} = \chi\vec{H}, \tag{6.2}$$

where the constant of proportionality χ is the magnetic susceptibility. In the case that $\chi > 0$ the medium is paramagnetic, and in the case that $\chi < 0$ the medium is diamagnetic.

For both of these types of linear magnetic media $|\chi| \ll 1$ so that the effect of the magnetization on these systems is small. This is seen by applying (6.2) in (6.1). From this substitution, the magnetic induction in a linear media is found to be related to the magnetic field through the linear form [1]

$$\vec{B} = \mu_0[1 + \chi]\vec{H} \equiv \mu\vec{H}. \tag{6.3}$$

Since $|\chi| \ll 1$ it follows that in paramagnetic and diamagnetic media $\vec{B} \approx \mu_0\vec{H}$ to a good approximation. In some cases considered in the following, however, it shall be seen that the small magnetization does result in significant physical effects in nano-systems.

Important physical effects in nanoscience related to the magnetic susceptibility show up in mechanical effects of nanoparticles interacting with applied magnetic fields. These effects are intimately related to the paramagnetic and diamagnetic nature of the particles. In particular, it will soon be seen that the difference in the sign of the susceptibility in paramagnetic and diamagnetic systems results in very different mechanical behaviors for particles made from these two different media. This is found from the difference in the way the diamagnetic and paramagnetic magnetization contributes to the particle energy in an applied magnetic field. To begin these considerations, first consider the interaction of a magnetic particle with an applied magnetic field.

Energy of a Particle in an Applied Magnetic Field

In the presence of an applied magnetic field, the energy density within the media of a magnetic particle is of the form [1]

$$u = \frac{1}{2}\vec{H} \cdot \vec{B}, \tag{6.4}$$

where $\vec{B}$ is the magnetic induction within the particle. The energy density in (6.4) consists of two contributions. Specifically, it includes an energy associated with the creation of the applied field and an energy associated with the magnetization of the particle. The focus in the following will be on the energy associated with the magnetization of the particle.

From a consideration of the relationships in (6.1) and (6.3), the total energy density in (6.4) including the field energy and the energy of interaction with the magnetization of the particle then becomes [1]

$$u = \frac{1}{2}\left[\frac{1}{\mu_0}\vec{B} - \vec{M}\right] \cdot \vec{B}. \tag{6.5}$$

It is seen from (6.5) that the energy density associated with the particle magnetization can be separated out from the energy density associated only with the applied field. In particular, the energy associated with the magnetization disappears for zero magnetization.

Making this separation gives to leading order in the small magnetization of the particle [1], the energy density associated with the particle magnetization given by

$$u = -\frac{1}{2}\vec{M}\cdot\vec{B} = -\frac{1}{2}\chi\vec{H}\cdot\vec{B} = -\frac{1}{2\mu_0}\chi B^2. \tag{6.6}$$

In addition, since $\vec{M}$ is small $\vec{B}$ is essentially the same inside and outside the particle as, similarly, is the case with $\vec{H}$.

Forces on Particles

The force acting on the particle is related to the change of the particle energy, determined from (6.6), as it is displaced in the field. Assuming that u is constant over the particle the energy of the particle is given by [1]

$$U = Vu, \tag{6.7a}$$

where V is the volume of the particle. Consequently, from the particle energy it follows that the force on the particle is given by [1–7]

$$\vec{F} = -\nabla U = \frac{V}{\mu_0}\chi\left(\vec{B}\cdot\nabla\right)\vec{B}. \tag{6.7b}$$

The result in (6.7b) assumes that the particle is in a background medium which is neither paramagnetic nor diamagnetic so that the energy of the particle is related only to the susceptibility of the particle itself. This is not always the case and nano-particles are often found suspended within media which exhibit their own magnetic properties. When this is the case, additional considerations must be made.

If the background medium suspending a nano-particle has a paramagnetic or diamagnetic interaction with the applied field this must also be taken into account. In case of a linear background magnetic media, if $\chi_{particle}$ is the susceptibility of the particle and $\chi_{bacground}$ is the susceptibility of the background, the modified form of the force becomes [1–7]

$$\vec{F} = \frac{V}{\mu_0}\left[\chi_{particle} - \chi_{background}\right]\left(\vec{B}\cdot\nabla\right)\vec{B}. \tag{6.7c}$$

As seen from (6.7c) the presence of a linear magnetic background media can have a significant effect on the force due to interaction with an external field. In the case where that particle and background are of opposite types of linear media, the force on the particle can be enhanced over the force of the particle given by (6.7b). When the particle and background are of the same type of media, the force can be decreased or even reversed from the force on the particle given in (6.7b). These results can have interesting effects on the dynamics of the particle.

Case of Ferromagnetic Particles

For ferromagnetic materials the relationship between the magnetization and magnetic field is not quite as simple as that of linear media. In general, the relationship is not linear but is highly nonlinear so that

$$\vec{B} = \vec{B}(\vec{H}) \tag{6.8}$$

exhibits a hysteresis loop of the form shown in Fig. 6.1. Specifically, the hysteresis loop is a multiple valued function and the behaviors of the system under small changes in $\vec{H}$ depends on where the initial configuration of the system is on the curve and on the past history of the system.

For a permanently magnetized particle, the particle energy in a magnetic field is given by a formula which is well known in electrodynamics and statistical physics. The particle energy in an applied magnetic induction is expressed as [1–7]

$$U = -\vec{m} \cdot \vec{B}, \tag{6.9a}$$

where $\vec{m}$ is the dipole moment of the particle. The force on the particle is then given by [1]

$$\vec{F} = (\vec{m} \cdot \nabla)\vec{B} \tag{6.9b}$$

Notice that the energy in (6.9a) does not have the factor of $\frac{1}{2}$ found in (6.6). This difference is due to the dependence of the dipole moment of linear media particles on the magnetic field. Consequently, the magnetization in the linear media is

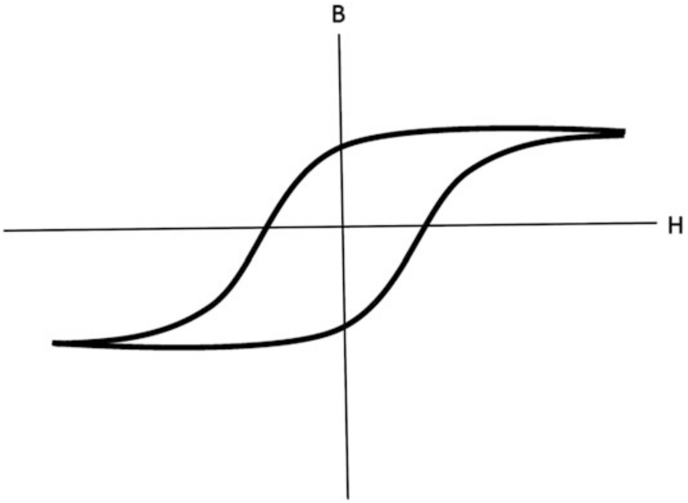


Fig. 6.1 The form of a typical hysteresis curve for a permanent ferromagnetic particle

directly proportional to the applied field whereas the magnetic moment of a ferromagnetic particle is almost independent of the applied field.

The differences in (6.7b) and (6.9b) in their applications are now treated. The difference of the magnetostatic physics of paramagnetic, diamagnetic, and ferromagnetic are related to the properties of these equations for both uniform and inhomogeneous applied magnetic fields. In the following, a focus is on nanoscience applications.

6.1.2 Forces on Ferromagnetic Particles

The force on ferromagnetic particles is given by (6.9b). It is seen from this formula that, in a non-uniform magnetic induction, the ferromagnetic particle experiences a force pushing it from a region of weak magnetic induction to one of strong magnetic induction [1–7]. The force is also found to be proportional to the strength of the magnetization of the particle and proportional to the strength of the applied magnetic induction.

Due to the nature of the hysteresis curve, the magnetization of a ferromagnetic particle is dependent on the applied field $\vec{B}$. In nanoscience applications, however, the applied field is often arranged so that the magnetic properties of the particle are on a portion of the curve where the magnetization is relatively independent of changes in the magnetic induction. This means that the particle magnetization is stable and the force arises from changes in the magnetic induction with changes in the spatial configuration of the particle-field system. In the following, the focus will be on systems under these conditions.

In a uniform applied magnetic field the particle does not experience a net force, but the permanent magnetic dipole of the particle experiences a torque. From the energy in (6.9a) the torque is given by [1–7]

$$\vec{\tau} = \vec{m} \times \vec{B} \tag{6.10}$$

and is seen to rotate the dipole to become parallel to $\vec{B}$.

The force and torque relations have been applied to a number of nanoscience applications. These involve interesting designs in nano-mechanical systems and machines that are first steps in exploring the possibilities of the technology. An outline of recent efforts is now given.

Nano-particle Projectiles

A particular interesting set of experiments have recently been performed on the design of micro particles that can be propelled through fluids as steered particles [1–3]. In one experiment a 1.5 μm long rod was fabricated and driven along a trajectory through a hydrogen peroxide solution. The rod was composed of a series of metallic segments, layered along the length of the rod. Specifically, a series of alternating segments of gold and magnetized nickel were ordered along the axis of the rod, and at one end of the rod was placed a cap composed as a layer of platinum [1–7].

The purpose of the platinum cap was for it to act on the hydrogen peroxide solution as a catalyst. In this capacity the platinum decomposed the hydrogen peroxide of the solution, forming oxygen bubbles at the capped end of the rod. The bubbles generated at the cap acted to propel the rod, making it into a type of nano-rocket [1, 2].

The nickel layers of the rod were permanently magnetized, with the permanent magnetic moment of each magnetic segment being parallel to all the others. The direction of the magnetic moments of the segments were set perpendicular to the axis of the rod. This created a total permanent magnetic moment normal to the direction of motion of the rod as it was propelled by the peroxide engine [1–4].

In the presence of an applied uniform magnetic field the total magnetic moment of the rod aligned along the applied field. This assure the rod would always move perpendicular to the field. Consequently, by changing the direction of the field the rod could be steered along a trajectory within the solution.

For the rod in this experiment 55 mT fields were used to guide the motion of the rods through the fluid. The applied fields were always uniform so that the torque force in (6.10) was the operable mechanism of magnetic control of the system [1–3].

In a second experiment, a similar nano-rocket design was treated. In this experiment the projectile was formulated as a rod of layers of silicone and cobalt with one end of the projectile capped with platinum. The rod formed in this way was 5 μm in length [1–3].

The propulsion was again provided by the catalysis of hydrogen peroxide by the platinum, and the steering was accomplished with a uniform applied magnetic field. For this experiment the applied magnetic induction was 5 mT [1–3].

Both of these experiments are based on designs utilizing uniform magnetic fields and the steering is accomplished by changing the direction of the applied field. Additional important mechanisms in nano-mechanics involve forces derived from spatially dependent fields and time-dependent fields.

Systems Involving Time-Dependent Fields

Some examples of the applications of time-dependent fields fall into two general classes: One of these uses the fields to directly generate motion of particles through magnetic forces and the other uses the fields to generate heat. While some applications are based solely on steering the projectile motion of particles through mechanical means, others are based on the generation of heat in the nano-particles or on combinations of the two approaches. These include important medical applications.

Ferromagnetic particles have found some applications in cancer research. In these techniques the interest is in the applications of time-dependent fields to generate heat in the ferromagnetic particles. The heat arises from the work done on the magnetic moment of the particle by the externally applied field [1–7]. The nano-particles provide a localized deleivery system for exacting medical procedures.

Similar ideas based on the generation of heat in a particle have been involved in some dynamic applications. Specifically, these are applications in which a particle coated on one side by a permanent magnetic layer can be propelled by the heat generated in the permanent magnet by an oscillating field. The directed heat from the particle is the mechanism creating the motion [1–7].

Another example of particle motion generated by a time-dependent rotating fields is based on ideas from the mobility of single cell organisms. Here the generation of heat is not a factor in the mobility, but the mechanical motion of a flagella leads to the propulsion. Some of microorganisms move through the use of flagella which function in a way as to thrust the organism through the fluid in which it lives. These ideas can be directly translated into nanoscience applications [1–7].

If the nano-rods discussed earlier are bent and not uniformly straight, it is possible to apply a rotating field to rotate the nano-rod in space. The rotational motion can be used to propel the system. Coupled with a peroxide motor this design can be used to propel the particle through a fluid in one direction using the peroxide motor or in an opposite sense using the rotational motor [1, 2].

A magnetic propulsion can also be achieved by attaching a permanently magnetized ferromagnetic particle to a helix tail. Using a time-dependent rotating field the dipole moment of the magnetized particle can be rotated in space [1]. The rotation of the particle translates into a rotation of the helix which then acts similar to a propeller on a ship. This provides a linear translation of the total structure [1].

6.1.3 Forces on Paramagnetic Particles

An interesting effect of the interaction of a uniform applied magnetic induction on paramagnetic particles is related to their dynamical orientation. The interaction of the particles with the field leads to an orienting response of the particles to the field. The particular orientation of the particle with respect to the applied field depends on the details of the geometry of the particles and the nature of the susceptibility tensor of the particular medium employed in the particle design [1–7].

For example, in some recent experiments elliptical particles of a paramagnetic material are cover on one side by platinum. In the dynamics of the particle, the platinum is used as a propellant. Putting the side coated particles in a hydrogen peroxide solution activates a peroxide engine through the catalyzed generation of oxygen. The oxygen expelled from the platinum surface then acts to propel the particle through space [1, 4].

From (6.7c) it is found that the application of an external magnetic field to the particles can orient them, with their induced dipole moments directed in space relative to the direction of the magnetic induction vector. The oriented particles will then move in a fashion directed by the applied field and the clever design of the particle geometry and susceptibility tensor [1–7].

In the case that the applied fields are inhomogeneous, a net magnetic force is exerted on the particles. This force is given by (6.7c). Such types of forces from inhomogeneous fields have been applied, for example, to deposit ions on gel plates and surfaces. They have also been used in the steering of paramagnetic particles along the magnetic boundaries of garnet films, and in the separation of particles of varying size distributions that are suspended in fluid media. The paramagnetic properties of bacteria and other microorganisms can also be used to manipulate and direct their motions through their fluid environments [1, 5, 7].

All of these effects are based on the fact that for $\chi > 0$ in (6.7b) or $\left[\chi_{particle} - \chi_{background}\right] > 0$ in (6.7c). From these equations, it is found that paramagnet particles are attracted by regions of high magnetic fields and repelled by regions of low magnetic fields. As well as these regions of attractions and repulsions, the particles are also oriented in space by the torque created by the relative orientations of the magnetic field and the dipole moment vectors [1].

In addition to static inhomogeneous fields time-dependent field effects have been used in the treatment of suspensions of nano-particles. In particular, a rotating magnetic field arrangement has be used to assemble a rotating chain composed of linked paramagnetic particles [1].

6.1.4 Forces on Diamagnetic Particles

As with paramagnetic particles a uniform applied magnetic induction on diamagnetic particles is an orienting mechanism. This has been used for the alignment of diamagnetic molecules and in some biological studies. In particular, an interesting study has been made on the organism Paramecium Caudatum [1, 5]. This is a self-propelled single celled organism which moves randomly through its fluid environment. It has been shown that the diamagnetic properties of the organism can be used to steer the direction in which is the organism travels within a fluid [1–7]. The orienting effect occurred from fields of 3 T.

Another example from biology is the deformation effects on liposome structure due to diamagnetic orientation of its constituent molecules. Some of these deformations are found to exhibit important changes with changing temperatures at fixed magnetic fields.

In inhomogeneous magnetic fields the force on nano-particles is given by (6.7b) or (6.7c). In these interaction for $\chi < 0$ in (6.7b) or $\left[\chi_{particle} - \chi_{background}\right] < 0$ in (6.7c) so that diamagnetic particles are attracted by regions of low magnetic fields and repelled by regions of high magnetic fields [1]. As well as these regions of attractions and repulsions, the particles are also oriented in space by the torque created by the relative orientations of the magnetic field and the dipole moment vectors.

The repulsion of diamagnetic particles from regions of high magnetic fields has been used as a levitating device. A diamagnetic particle can be suspended by a magnetic pole. In this case the magnetic force pushing the particle upward must be equal and opposite the gravitational force attracting the particle downward [1].

Such interactions have been used in demonstrations of the levitation superconducting particles and in the separation of nanoparticles of varying sizes which are suspended in a fluid. Outside the realm of nanoscience there are some additional interesting applications of diamagnetic levitation. One example is in transportation technology. In this area the levitation effects as applied to superconductors have been proposed as part of the technology in the design of super-trains. In a less technologically important second example, the diamagnetic force has been used to levitate frogs. Frogs are of a particular physiological type which cause them to exhibit a high degree of diamagnetism [1, 7].

Time-dependent magnetic fields have also been applied in the study of diamagnetism. These have found applications in, e.g., the orientation of nylon fibers [1–7].

In the next subsections, the focus will be turned to the application of electric fields to manipulate nano-particles and to trap atoms for applications in quantum optics. This will be followed by discussions of the application of intense laser beams in the manipulation of nano-particles. The laser trapping employs a number of phenomena in electrodynamics to realize a device known as the optical tweezer [13–20]. These technologies all involve the use of applied electromagnetic fields to control the motion of particles.

6.2 Electric Forces for the Manipulation of Nanoparticles

A similar manipulation of nanoparticles to that found from the application of magnetic fields can be obtained in the application of electric fields to nanoparticles. The reason for this is that the mathematics of the two systems display an isomorphism so that much of the theory in the early treatment of magnetic forces can be taken over to the study of electric systems [8–10].

For example, the electric field energy density, $u = \frac{1}{2}\varepsilon E^2$, is mathematically very similar to that of the magnetic field energy density, $u = \frac{1}{2}\frac{1}{\mu}B^2$. Likewise, for particles with electric dipole moments, $\vec{p}$, the interaction energy in the presence of an external electric field is $U = -p \cdot \vec{E}$ whereas the interaction energy of a magnetic dipole $\vec{m}$ in external magnetic induction is $U = -\vec{m} \cdot \vec{B}$. The torque experience by an electric dipole moments, $\vec{p}$, is $\vec{\tau} = \vec{p} \times \vec{E}$ while that on a magnetic dipole $\vec{m}$ is $\vec{\tau} = \vec{m} \times \vec{B}$.

In both electric and magnetic systems the force acting on a particle in the presence of these fields is obtained as a spatial gradient of the energy of interaction of the particles with the fields. Consequently, the effects of uniform, inhomogeneous, and time-dependent electric fields are similar to those found in the earlier discussions of magnetic interactions [8–10].

Some recent application of the electric force in nanoscience and microbiology include: The generation of forces on bacteria and viruses that can be used to manipulate and separate them from one another, the self-assembly of nanosctructures for device applications, applications in the design of biological and chemical sensors, modification of particle motion within nano-channels and capillaries, and in the applications of dielectrophoretic forces used to separate molecular species from one another [8–10].

6.3 Ion Traps Based on Electric Forces: Paul and Penning Traps

In the earlier discussions of this Chapter a focus has been on the manipulation of particles through the application of magnetic and electric fields [11, 12]. Another application of electric fields, however, is in the design of field configurations which can trap ions. Such a trapping configuration is used to suspend individual isolated ions or isolated interacting groups of ions in a background of vacuum. This allows them to be studied spectroscopically and forms the basis of the investigation of many of the fundamental properties of quantum mechanics. These studies offer tests of the properties that distinguishes the nature of quantum mechanical systems from those of classical mechanical systems and have been used to develop an understanding of many of the early paradoxes of quantum theory. In the following, some of the most basic elements of the theory of ion trapping are presented.

6.3.1 Earnshaw's Theorem

The first thing to point out is that it is impossible to trap an ion in three dimensions using a static configuration of electric fields. This restriction is known as Earnshaw's theorem and can be seen from an applications of the Laplace equation for the electrostatic potential ϕ and the relationship of the electrostatic potential to the electric potential energy V. Specifically, Laplace's equation states that [11]

$$\nabla^2\phi = 0 \tag{6.11a}$$

and the electric potential energy of a charge q in ϕ is given by

$$V = q\phi \tag{6.11b}$$

To trap an ion about the origin of coordinates in three-dimensional space, the electrostatic potential, ϕ, at the origin of coordinates would need to be at a minimum for $q > 0$ or a maximum for $q < 0$. In particular, consider a polynomial solution of (6.11a) of the form [11]

$$\phi(x, y, z) = \frac{\phi_0}{r_0^2}\left(\alpha x^2 + \beta y^2 + \gamma z^2\right), \tag{6.12}$$

where ϕ_0 and r_0 are constant with units of energy and distance, respectively. This gives a general expression for the behavior of the electrostatic potential near the origin in terms of the set of coefficients (α, β, γ). For a minimum of (6.12) the (α, β, γ) must all be positive while for a maximum (α, β, γ) must all be negative.

The set of coefficients (α, β, γ) are then determined to make (6.12) a solution of Laplace's equation. This places a set of restrictions on the set (α, β, γ) from the theory of electrostatics which limits the ability of (6.12) to represent a minimum or a maximum [11].

In particular, substituting (6.12) into (6.11a) it follows that [11]

$$\alpha + \beta + \gamma = 0. \tag{6.13}$$

In order for (6.13) to have a non-zero solution at least one of the set (α, β, γ) of coefficient must differ in sign from the others. This is inconsistent with (6.12) exhibiting a minimum or maximum at the origin of coordinates. Consequently, a solution for a trapping potential centered about the origin of coordinates does not exist.

6.3.2 Time-Dependent Potentials

While a static trapping potential does not exist as a solution of the Laplace equation, it is possible to modify the potential in (6.12) to make a trapping potential. Specifically, a modification of (6.12) involving the addition of a time-dependence can result in an electric potential that will trap ions, localizing them about the origin [11].

To understand how this works consider a specific solution of (6.13) given by [11]

$$(\alpha, \beta, \gamma) = (1, 1, -2). \tag{6.14}$$

Entered into (6.12) the electrostatic potential in terms of these coefficients becomes [11]

$$\begin{aligned}\phi(x, y, z) &= \frac{\phi_0}{r_0^2}\left(x^2 + y^2 - 2z^2\right) \\ &= \frac{\phi_0}{r_0^2}\left(\rho^2 - 2z^2\right),\end{aligned} \tag{6.15}$$

where the cylindrical coordinate $\rho^2 = x^2 + y^2$ has been introduced. From (6.11), (6.12) and the relation of the force to the electric potential energy, the force on the charge q is [11]

$$\vec{F}(x, y, z) = -2q\frac{\phi_0}{r_0^2}\left(x\hat{i} + y\hat{j} - 2z\hat{k}\right). \tag{6.16}$$

From (6.16) it is found that, in the case that $q\frac{\phi_0}{r_0^2} > 0$, the force is attractive and harmonic about the origin in the x-y plane but repulsive from the origin along the z-axis. In the case that $q\frac{\phi_0}{r_0^2} < 0$, however, the force is attractive and harmonic about the origin along the z-axis but repulsive from the origin in the x-y plane. Neither of these two cases allows the particle to be trapped three-dimensionally about the origin of coordinates.

While the two cases considered do not independently confine the particles, if they were intermittently applied over alternating periods of time, they might succeed in confine the particle. In this case they would act similar to how two ping-pong player confine the ping-pong ball to the game table. This turns out to be the case for a time mixture of the two cases discussed above [11, 12].

To generate such a time-dependent potential, consider the case in (6.15) and (6.16) that ϕ_0 is time-dependent and of the form

$$\phi_0(t) = U_0 + V_0 \cos \Omega t. \tag{6.17}$$

This represents such a time-dependent transition between the two cases involving harmonic attraction in the x-y plane followed by harmonic attraction along the z-axis. From Newton's laws and (6.16) it then follows that the particles motion obeys the dynamical equation [11]

$$m\frac{d^2}{dt^2}\left(x\hat{i} + y\hat{j} + z\hat{k}\right) = \vec{F}(x, y, z) = -2q\frac{\phi_0(t)}{r_0^2}\left(x\hat{i} + y\hat{j} - 2z\hat{k}\right). \tag{6.18}$$

The solutions of (6.18) are now studied with the intent of determining the conditions for finding tightly localized solutions about the origin of coordinates. This search is facilitated as (6.18) can be transformed into one of the standard equations of mathematical physics.

To see the trapping behavior in (6.18), (6.18) can be rewritten into the form of a standard Mathieu equation. The Mathieu equation is a classic equation of mathematical physics, and its solutions have been well studied and are under certain conditions known to yield trapped solutions. In particular, under a change of variables, (6.18) takes the form [11]

$$\frac{d^2u}{d\xi^2} + [a_u - 2q_u\cos(2\xi)]u = 0. \tag{6.19}$$

where $u = x, y, z$, $\xi = \Omega t/2$, and

$$\begin{aligned} a_x = a_y &= \frac{8qU_0}{mr_0^2\Omega^2} = -\frac{a_z}{2}, \\ q_x = q_y &= -\frac{4qV_0}{mr_0^2\Omega^2} = -\frac{q_z}{2}. \end{aligned} \tag{6.20}$$

To realize the trap configuration (U_0, V_0) are chosen so that (6.19) and (6.20) represent bounded solutions about the origin of coordinates in all three spatial directions. These can be worked out from the tabulated results and solutions of the Mathieu equation [11].

6.3.3 Paul and Penning Traps

The above results describe the binding of an ion in a Paul trap. The trapping mechanism comes solely from the application of electric fields. It is also possible to design traps that involve an arrangement of electric and magnetic fields.

Another type of trap of interest is the Penning trap which relies on the application of a magnetic field for the formation of a three-dimensional trapping effect. In the Penning trap an electric field confines the particles harmonically along one axis in space and a combination of electric field along with a uniform magnetic field parallel to the axis of harmonic motion confines the particles in the plane perpendicular to the harmonic axis.

Both of these types of traps can be formulated based on designs involving electrode plates in the form of hyperbolic shaped caps or cylindrical forms generated from hyperbolas rotated about and external axis. As discussed, the Penning trap also requires the presence of a uniform applied magnetic field. For the details of the Penning trap theory and the experimental realizations of the Penning and Paul traps the reader is referred to the literature.

As a final note: The above discussions have focused on traps that can be used to isolate single ions. In some applications it is of interest to isolate a linear array of interacting ions. A modification of the Paul trap known as a linear radio frequency trap can be used in studies of these types of systems.

For this type of traps the three dimensional potential in (6.15) is replaced by a two dimensional form $\phi(x,y) = \frac{U_0 - V_0\cos\Omega t}{r_0^2}(x^2 - y^2)$ and the linear array of ions is set out along the z-axis. End caps at DC potentials can be places along the z-axis to stabilize the ion array, fixing it stationary on the z-axis.

Arrays formed in this manner have been a focus in spectroscopy and in some quantum computer schemes. The reader is referred to the literature for further details [11, 12].

6.4 Optical Tweezer

Another important interaction is that between beams of focused laser light and dielectric particles [13–20]. The physics of this system forms the basis for the design of an optical device known as an optical tweezer. This is a device which employs the focused laser beam to trap a dielectric particle within a region located near the focus of the beam. The principle operating in the trapping effect is the binding of the particle due to an interaction energy that exists between the particle and the applied light in the laser beam. The binding in this interaction is a harmonic force centered at a point within the focus of the beam. Essentially, then, the force on the particle is a result of the gradient of the energy density of the focused laser beam.

This can be a useful effect, for example, in biology where cells suspended in a fluid can be fixed in space and held for examination. Similarly, any other type of dielectric nanoparticle can be trapped and held for study by an optical tweezer [13–20].

The optical tweezer effect arises in the interaction of particles with laser beams over a wide range of wavelengths of light. The physics of the trapping, however, requires different approaches for its study, depending on the ratio of the lengths scale characterizing the particle as compared to the wavelength of the light. As examples, two limits will be considered in the following for the consideration of spherical nanoparticles. For the case that the radius of the particle is much greater than the wavelength of the light, an approach based on geometric optics can be made. For the case that the radius of the particle is much smaller than the wavelength of the light, an approach based on a dipole approximation for the particle is made [13–20].

In the following, after some preliminary remarks on the momentum carried by electromagnetic fields, both of these limits will be treated. This will be followed by a discussion of some of the applications of the optical tweezer to the nanosciences.

6.4.1 Momentum Considerations

In both wavelength limits of the optical tweezer to be studied, the energy and force on particles arising from interaction with an electromagnetic field are important considerations. In particular, as light travels between to different media it not only carries energy with it, but it also carries momentum between the two media. This can show up as a mechanical force acting between the two media. To see this, it is necessary to consider the Poynting vector and its relationship to the momentum carried by light.

In classical electrodynamics the radiation pressure exerted on a surface is known to be simply related to the Poynting vector of the radiation, where in terms of the radiation fields $\vec{E}$ and $\vec{H}$ the Poynting vector is given by [13]

$$\vec{S} = \vec{E} \times \vec{H} \tag{6.21}$$

In particular, for the system in (6.21) the radiation pressure carried by the fields is given by

$$\vec{P} = \frac{1}{c_m} \vec{E} \times \vec{H} \tag{6.22}$$

where c_m is the speed of light in the medium in which it is traveling.

Both (6.21) and (6.22) are standard results in classical electrodynamics, but they can also be expressed in terms of quantum mechanics considerations. The quantum mechanical approach treats the motion of individual photons propagating in the system, and the momentum and energy processes in the system arise from a study of the dynamics of each of the photons as it moves between media in the system. Viewed in this regard, a description in terms of the motion of photons provides a deeper understand of the tweezer effect than that offered in classical electrodynamics.

Looking at the dynamics of particles and light from the standpoint of the quantum theory of light, the momentum carried by a plane wave (photon) of light with a wave vector $\vec{k}$ is [13]

$$\vec{p} = \hbar \vec{k}. \tag{6.23}$$

In this view, as the photon moves through an optical media it carries with it a momentum which it transfers from one part of the system to another during the course of its journey. As a simple example of this process, (6.23) then represents the momentum transferred to an object upon its absorption of the photon in question.

A second important example of photon transfer of momentum is the case of the reflection of light. In the case that the plane wave is incident on a perfect reflecting surface, it is not absorbed but its trajectory is reversed. Upon reflection of the light from the surface, the surface then acquires a net momentum

$$\vec{p}_s = 2\hbar \vec{k} \tag{6.24}$$

transferred to it.

The magnitude of the momentum transferred to the surface, denoted Δp, can then be related to the energy of the light by applying Planck's relation in quantum electrodynamics and the $\omega = c_m k$, dispersion relation of light in the medium. This gives

$$\Delta p = p_s = 2\frac{\hbar\omega}{c_m} = 2\frac{\Delta E}{c_m}. \tag{6.25}$$

where $\Delta E = \hbar\omega.$ is the energy of the plane wave of light in terms of its frequency.

The magnitude of the force transferred to a perfect reflecting surface is given by the impulse [13]

$$F = \frac{dp}{dt} = \lim_{\Delta t \to 0} \frac{\Delta p}{\Delta t} = 2\frac{1}{c_m}\frac{dE}{dt} = 2\frac{1}{c_m}P. \tag{6.26}$$

where P is the power carried in the plane wave. Dividing both sides of (6.26) by an area in the plane normal to $\vec{k}$ gives the relation for the momentum per area sec to the energy per area sec obtained in (6.21) and (6.22) for light reflected at normal incidence from a perfect reflecting surface.

An important point of the quantum treatment is that the momentum transfer as light passes through a surface can be obtained in terms of the photon wave vector passed through the interface between two media. In particular, consider a photon as it passes through an interface at normal incidence between two different dielectric media. If the wave vector of the incident photon is $\vec{k}$, the reflected photon is $-\vec{k}$, and the refracted photon is $\vec{k}$, then the momentum transferred between the incident and refraction media can be expressed in terms of these and the reflection and transmission amplitudes of the problem. This is the general problem in which the light is partially transmitted and partially reflected at the interface it is incident upon.

Specifically, for a normal incident wave with momentum

$$\Delta\vec{p}_I = \hbar\vec{k} \tag{6.27a}$$

incident on a planar interface, the momentum carried in the transmitted wave [13] is

$$\Delta\vec{p}_T = \hbar\vec{k}T \tag{6.27a}$$

where T is the transmission coefficient of the surface between the two media. By the same reasoning, the momentum of the reflected wave from the surface is

$$\Delta\vec{p}_R = -\hbar\vec{k}R \tag{6.27b}$$

where R is the reflection coefficient and $R + T = 1$.

The sum of the momentums of the transmitted and reflected waves and the mechanical momentum given to the media must equal the incident momentum delivered in the incident wave. In the following these momentum considerations will be used to study the momentum transferred by light to a slab and to a spherical particle.

6.4.2 Momentum Consideration of Light Incident on Slabs and Spherical Particles in the Geometric Optics Limit

In the case of light at normal incidence on an infinite slab of a lossless dielectric medium, the total momentum impulse transferred to the slab can be determined in terms of wave vector considerations. To understand this determination, consider the schematic drawing in Fig. 6.2a. The figure represents a dielectric, lossless, slab surrounded on both of its siders by vacuum, and the light is incident normal to the slab surfaces.

In the drawing, the incident and reflected rays of light on the slab as well as the transmitted light through the slab are shown. The total momentum delivered to the slab by these rays is obtained by considering the momentum change imparted to the slab by the light at each of its two surfaces.

The momentum carried to the slab by the normal incident wave is

$$\Delta\vec{p}_I = \hbar\vec{k}. \tag{6.28}$$

Upon interacting with the surface the incident wave transfers momentum to the wave reflected from the slab and the wave transmitted by the slab as well as providing a momentum to the slab itself [13].

The momentum given to the slab then follows from the conservation of momentum for the system of fields and the slab [13]. This is expressed as

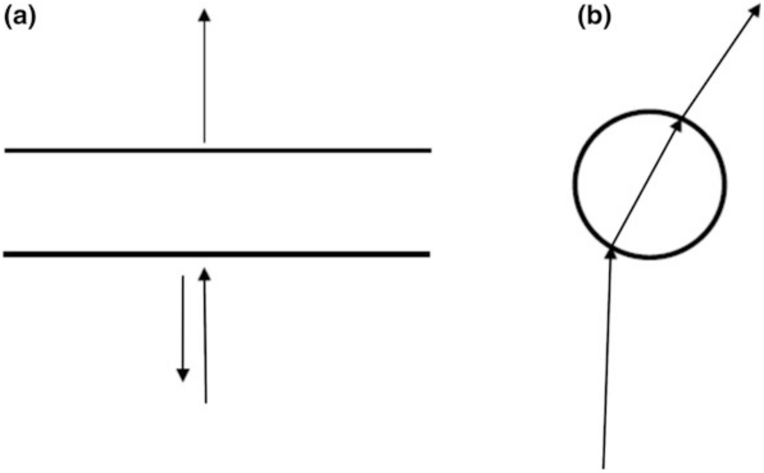


Fig. 6.2 Ray optics schematic for: **a** light at normal incidence on the slab as it is reflected and transmitted through the slab, **b** the lowest order transmission process for an incident ray transmitted by a spherical dielectric particle

$$\left[(T-R)\hbar\vec{k}+\vec{S}\right]-\hbar\vec{k}=0 \tag{6.29}$$

where $\vec{S}$ is the momentum of the slab. The terms in the brackets are the momentum of the system after the incident wave interacts with the surface and $\hbar\vec{k}$ is the momentum of the incident wave.

From (6.28) and (6.29) it follows that the momentum to the slab is obtained from

$$\vec{S}=\hbar\vec{k}+(R-T)\hbar\vec{k}. \tag{6.30}$$

Considering the result in (6.30) it is found that both of the limits of photon absorption and photon reflection in (6.23) and (6.24) are seen to be given by (6.30). In addition, (6.30) holds as well for all elastic processes involving the slab.

The slab problem gives an idea of the balancing of momenta between the various slab surfaces, but it is not of interest for technological applications. A problem of more technological interest is that of a spherical particle interacting with an incident light. This is a practical example of how light can be used to manipulate and position a nano-particle, illustrating many of the principles involved in the tweezer technology.

Another case of direct interest to optical tweezers is that of a uniform incident plane wave of light incident on a spherical dielectric particle. A schematic figure for the problem is given in Fig. 6.2b. For the considerations, the particle is assumed to have a radius much larger that the wavelength of light. As a consequence, then, of the plane wave nature of the radiation and the size of the wavelength compared to the particle radius, the treatment of the light in the system is essentially a problem in the geometric, ray optics, limit.

Figure 6.2b shows one of the ray optics trajectories of light in the system. It indicates an incident ray on the sphere which is refracted by the sphere, exiting the sphere as a transmitted wave at an angle to the optical axis. For this case, the transmitted momentum in the plane normal to the optical axis does not, as was found in (6.30) for the case of the flat surfaces at normal incidence, sum to zero.

Unlike the rays considered in the infinite slab problem, the ray shown in Fig. 6.2b now imparts a net momentum to the particle in the plane perpendicular to the optical axis. The reason for this is that the dielectric sphere has curved surfaces which bend the flow of light through space. In the sphere problem, the motion of light is a two-dimensional motion. This is unlike the case of the dielectric slab in which the light had only a one-dimensional motion.

For the dielectric slab the incident, reflected, and refracted waves all moved parallel to the same axes in space. Consequently, all of the momentum in the system flowed along the optical axis [13]. Now the sphere alters the net momentum of the ray of light shown in Fig. 6.2b, and by conservation of momentum the momentum of the sphere must in turn be changed. The momentum change of the sphere allows it to be moved in the plane perpendicular to the optical axis. This forms the basis of the optical tweezer application.

The ray in the drawing in Fig. 6.2b imparts a net momentum to the particle in the plane perpendicular to the optical axis. If the total radiation incident on the sphere is a plane wave of uniform intensity in the plane perpendicular to the optical axis, however, the total momentum from all the incident rays must be treated. In particular, rays related to one another by rotational symmetry about the optical axis will have components of the momenta in the plane perpendicular to that axis which add so as to cancel one another. Consequently, the sphere will experience no net force in the plane perpendicular to the optical axis.

If the incident beam of light on the spherical particle, however, is not of uniform intensity in the plane perpendicular to the optical axis, the cancelation of rays related by rotational symmetry about the optical axis will not be complete. This follows because while the sphere has rotational symmetry about the optical axis the illuminating beam itself lacks this symmetry. Consequently, the sphere experiences a net force in the plane perpendicular to the optical axis.

It should be noted in these discussions that an infinite set of higher order refractions and reflections within the sphere have been omitted. These higher order effects are seen, for example, in the multiple bows observed from the rainbow phenomenon and arise from the multiple transmits of light within the individual water droplets. Only the most dominant, leading order, scattering of light entered into the earlier discussions based on Fig. 6.2b. In the full treatment of the illuminated dielectric sphere all of these higher order effects are summed over to obtain the detailed result for the total force on the sphere from its interaction with the laser beam.

Upon making this detailed analysis it is generally found that for an application of a focused laser beam the dielectric particle experiences a harmonic force pulling the particle to the most intense point of the focused beam. The particle, then, becomes fixed near the focus experiencing a harmonic restoring force acting about the optical axis in the plane perpendicular to the axis [13].

An illustration of these features, for the geometric optics problem considered in Fig. 6.2b, is now made by presenting the results from some recent numerical simulation studies [13]. The simulations were performed on a system involving an optical tweezer in the form of a Gaussian laser beam which is applied to an oil drop.

As an example of the tweezer force in the ray optics limit, some numerical simulation results are presented in Fig. 6.3 [13] for a study of the force exerted on an oil drop by a Gaussian laser beam. The plot is given of the force in the plane perpendicular to the optical axis of the beam on an oil drop of radius $a = 1$ μm suspended in water. The light for the optical tweezer was of wavelength $\lambda = 832$ nm incident on the droplet as a Gaussian beam of a width $\sigma = 1.2$ mm and with a power of $P_l = 4.8$ mW.

In Fig. 6.3 a harmonic force is observed with typical force of order of tenths of *pN*. For the simulation, the plane of the force is near the focus of the laser beam, and Δx is a measure of the displacement of the drop from the center of the beam. The results show the degree of trapping that can be achieved by a laser beam.

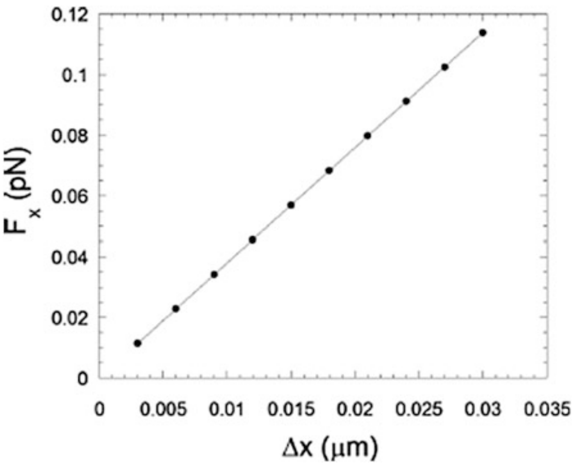


Fig. 6.3 Plot of the force versus the displacement from the optical axis of the laser beam in the plane perpendicular to the optical axis. The results are from a computer simulation study presented in [13]. Reproduced with permission from [13], with the permission of the American Association of Physics Teachers

6.4.3 Force on a Dielectric Sphere When the Wavelength of Light Is Large Compared to the Sphere Radius

The second limit that will be treated here is that in which the wavelength of the light is much greater than the typical length scale of the particle being manipulated. For these consideration it is assumed that the dielectric particle is formed of a homogeneous isotropic medium, and the wavelength of the light is long enough that the particle can be treated as interacting with a uniform electric field with a harmonic variation in time [13].

Considering a dielectric particle of radius a and dielectric constant K suspended in a fluid medium, it is a standard result from electrodynamics that the dipole moment, $\vec{p}$, induced by the applied electric field, $\vec{E}$, is related to the field by [13]

$$\vec{p} = \frac{K-1}{K+2} a^3 \vec{E} \tag{6.31}$$

where

$$K = \frac{\varepsilon}{\varepsilon_m} \tag{6.32}$$

with ε the permittivity of the dielectric sphere and ε_m the permittivity of the suspending medium.

Treating the particle as a point dipole, the potential energy of interaction with the polarizing field of the suspended dipole is [13]

$$U = -\vec{p} \cdot \vec{E}. \tag{6.33}$$

The force and torque on the dipole are then related to the spatial derivatives of the interaction energy with

$$\vec{F} = -\nabla U = \nabla\left(\vec{p} \cdot \vec{E}\right) \tag{6.34}$$

Using (6.31) the force can be written solely in terms of the applied electric field as

$$\vec{F} = \frac{K-1}{K+2} a^3 \nabla E^2. \tag{6.35}$$

From this is it is seen that the force is related to the energy in the applied electric field. For paramagnetic particles the force is directed towards the region of increasing field intensity, but for diamagnetic particles the force is directed away from the region of increasing field intensity.

Consequently, the systems of interest for the application of (6.35) are paramagnetic particles. These will be attracted to the intensity maximum at the focus of the laser beam. The force in the plane perpendicular to the optical axis (denoted by $\vec{F}_{||}$ as it is parallel to the plane perpendicular to the optical axis) is given by [13]

$$\vec{F}_{||} = \frac{K-1}{K+2} a^3 \frac{\partial E^2}{\partial x_{||}}, \tag{6.36}$$

and the stiffness of the harmonic force is

$$\kappa_{||} = -\left(\frac{\partial F_{||}}{\partial x_{||}}\right)_{equilibrium} = -\frac{K-1}{K+2} a^3 \frac{\partial^2 E^2}{\partial x_{||}^2}_{equilibrium}. \tag{6.37}$$

Here the equilibrium position is on the optical axis of the focused beam, and this is the axis on which the derivatives in (6.37) is evaluated.

The important result from the application of (6.35) and (6.37) are: Dielectric particles are pushed towards the focal point of the focused laser beam. This sets the particle interacted with to be located in a plane perpendicular to the optical axis of the beam and near its focal point. Within the plane the particle experiences a harmonic force of attraction towards the optical axis of the laser beam.

Some important uses of the optical tweezer technology include [13–20]: The trapping, transporting, and patterning of nanoparticles for particle manipulation and assembly. These can be important in the building of devices and ordered patterns. A variety of applications of interest in biological studies include the use of trapping effects on bacteria, viruses, and even DNA. Ideas of tweezer technology have also been found to be of use in studies of molecular motors and the properties of DNA and membranes [13–20].

6.5 Casimir Effect and Casimir Forces

Another important effect that enters into problems of nano-science is the interaction between surfaces, known as the Casimir effect [21–28]. This involves the generation of a quantum mechanical force between surfaces that is operative at nanoscales. The Casimir force is a particular component of inter surface forces arising from the Heisenberg uncertainty principle. In particular, it comes purely from the presence of vacuum fluctuations found in the quantum electrodynamic fields of the physical system. These fluctuations are present in the system both at zero and non-zero temperatures [21–28].

As a consequence of its origins in quantum fluctuations, the Casimir force is not present in classical electrodynamics nor does it have an analogy in the typical interactions between surfaces as treated in classical theory. For example, the force from the Casimir effect does not come from a classical charge distribution or polarization on the surfaces, but it arises from the nature of the quantized electromagnetic modes of the system. It is always present within the physical system, even at zero temperatures, as a very short range interaction operative at the length scales of nano-science applications.

The source of the force is from the influence of surface boundary conditions on the quantized modes of the electromagnetic fields in the space containing the surfaces [21–28]. Ultimately the restrictions on the fields at the surfaces modify the zero point energy of the electromagnetic modes. It changes the modes so that their zero point energy differs from the zero point energy of the quantized fields in infinite free space. In practice, the force on a surface is then related to the change in the zero point energy of the system under slight perturbations acting on the surfaces being considered.

The relation of the Casimir force to the changes in the zero point energy of the system has recently become of great interest in the study of quantum field theory and nano-science. In the earliest work on quantum theory, it was once thought that the zero point energy in quantum field theory had no physical manifestations. This has proved not to be the case as evidenced by a number of other different physical manifestations of the presence of the vacuum in relativity and high energy physics. In the development of the theory of the vacuum, however, the Casimir effect was the first indication of the importance of the vacuum fluctuations [21–28].

Another problem in the development of quantum theory involving the zero point energy was that of the nature of the vacuum itself. In particular, the zero point energy of the universe was found to be infinite. This did not present an overwhelming difficulty to quantum field theory because the zero point energy did not enter into the measurement of physically important properties of the system. As the infinity did not enter into the calculation of the interesting physics of the system, it was initially just ignored [21–28].

However, it can be shown that changes in the boundary conditions between two systems can result in a finite well defined change is the zero point energy between the two systems. This is the case even though the total zero point energy of each of the two systems taken separately is infinite.

Before proceeding to an explanation of the of the Casimir theory [21], it should be noted that some analogies of the Casimir force are found in classical physics. In classical physics, there are examples in which classical fluctuations can mediate a force between surfaces. The fluctuations in classical systems, however, are not caused by a zero point energy arising from the Heisenberg uncertainty principle.

An example from fluid mechanics is the force between two surfaces separated by water. In the presence of wave motion in the water a force is found between the surfaces. Specifically, this is an effect that can draw two ships together that are parked close to one another in water. The fluctuations in the water are the source of the attraction, based on a similar argument as that used in the study of the Casimir force. Similar effects arise in other classical media at non-zero temperatures. These all have an origin in the temperature fluctuations present in the system.

In the following, a simple theory of the Casimir force between two perfectly conducting plates is discussed [21]. This is followed by discussions of applications to nanoscience.

6.5.1 Theory of Casimir Effect

An example of the Casimir effect and the associated Casimir force is provided by the solution of a basic problem in quantum electrodynamics [21–28]. For these considerations, the easiest Casimir problem to treat is that of the force between two parallel perfect conducting plates that are surrounded by vacuum. The development of the theory involves the study of the electromagnetic modes of the system and their vacuum energy as a function of the plate separation. The zero temperature treatment of the parallel plate problem just outlined is the focus of the theoretical presentation in this section.

The problem of the perfect conducting plates illustrates many of the essential points of consideration in the development of the Casimir effect. Complications to its applications to technology involve the treatment of real metals, dielectrics, systems of general surface geometry, and the introduction of temperature effects. Consequently, following this discussion of the perfect conducting plates some generalizations that have been made to handle variations to other case of technological interest will be indicated [21–28].

Two Parallel Perfect Conducting Plates

Consider two parallel perfect conducting plates that are separated by a distance a and surrounded by vacuum. A schematic representation of the system is given in

Fig. 6.4a. The system is studied at zero temperature so that the only energy outside of the plates is the zero point energy of the electromagnetic fields within the regions of vacuum. The energy in the perfect conducting plates is not of interest to the problem as the electromagnetic fields are excluded from entering the plates. Aside from the boundary conditions the fields experience no interactions with the interior media contained within the plates.

In order to determine the zero point energy of the fields, the electromagnetic modes in the regions of vacuum must be computed. These modes are free space propagating modes subject to the boundary conditions that the electric field vanishes within the perfect conducting planes.

As a simplification of the treatment of the parallel plate system, it is best to proceed by initially considering the fields between the two perfect conducting plates which are separated by the distance a. Once these fields and their zero point energy are determined, the solution of the total system of the parallel plates within the three regions of vacuum is obtained as a composition from the initial considerations.

The two infinite perfect conducting plates in Fig. 6.4a are located at $x = 0$ and $x = a$ where they set a requirement on the electromagnetic fields that the electric field components vanish on these planes. As a result of this, the modal free space solutions for the electric fields between the plates are of the form [21]

$$\vec{E}(x, y, z) = \vec{E}_{m,n,p} \sin\left(\frac{m\pi}{a}x\right) \exp\left\{ i\left[\frac{2n\pi}{L_y}y + \frac{2p\pi}{L_z}z\right]\right\} \tag{6.38}$$

where the solutions are defined over the region $0 < x < a$, $-\frac{L_y}{2} < y < \frac{L_y}{2}$, and $-\frac{L_z}{2} < z < \frac{L_z}{2}$.

In (6.38) periodic boundary conditions have been applied over the region of the y-z plane in limit $L_y, L_z \to \infty$. Consequently, for the totality of the boundary conditions, m ranges over the positive integers, accounting for the perfect conduction of the plates. In addition, n and p run over all of the integers, accounting for the periodic boundary conditions in the y-z plane.

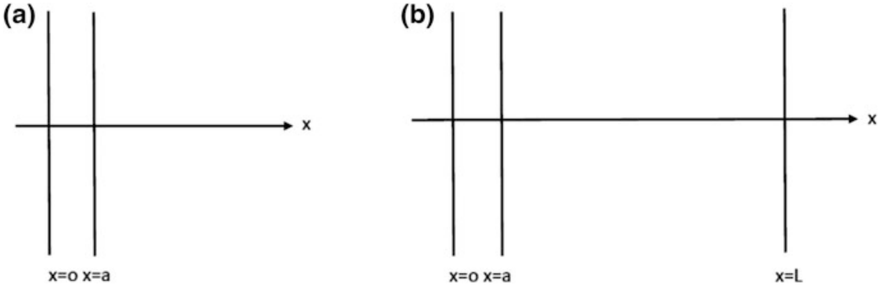


Fig. 6.4 Schematic drawing of: **a** two parallel perfect conducting parallel plates and **b** the addition of a third perfect conducting plate to the configuration of two plates in Fig. 6.4a

With this notation the wave vector components of the modal solutions in the *y*-*z* plane are then represented as

$$\left(k_y, k_z\right) = \left(\frac{2n\pi}{L_y}, \frac{2p\pi}{L_z}\right). \tag{6.39}$$

Adopting this notation, (6.38) then becomes

$$\vec{E}(x, y, z) = \vec{E}_{m,n,p} \sin\left(\frac{m\pi}{a}x\right) \exp\left\{i\left[k_y y + k_z z\right]\right\}. \tag{6.40}$$

Consequently, in vacuum the dispersion relation of the modes between the plates is given by [21]

$$\omega_{m,n,p} = c\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_y^2 + k_z^2}. \tag{6.41}$$

This form employs the standard frequency-wave vector relation in free space. Essentially, it is a statement that the phase velocity of each mode is equal to the speed of light in vacuum.

In the quantum mechanical treatment of light, the dispersion relation in (6.41) is very important in determining the energy contained in each electromagnetic mode. Specifically, from a fundamental result in quantum electrodynamics it is known that the energy of a mode of frequency ω is related to its frequency by [21]

$$E = \left(n_{ph} + \frac{1}{2}\right)\hbar\omega \tag{6.42}$$

where n_{ph} is the number of photons present in the system. Equation (6.42) is a statement that the excitation of each mode of the system is that of a quantum mechanical harmonic oscillator.

At zero temperature each mode of the electromagnetic fields is in its lowest energy state. Since in this limit there are no photons in the system, in terms of (6.42) this occurs when $n_{ph} = 0$. However, the mode still has a net energy arising from the factor of one-half in (6.42). This remaining modal energy in (6.42) is known as the zero point energy.

Considering all of the modes of the form in (6.41) for the region between the two parallel plates gives a total zero point energy of the form

$$E_{total} = \frac{1}{2}\sum_{m,n,p} \hbar\omega_{m,n,p} \tag{6.43}$$

In the limit that $L_y, L_z \to \infty$, (6.43) can be rewritten in terms of integrals over the wave vectors in the *y*-*z* plane. This yields an expression for the energy per area measured in the *y*-*z* plane given by [21]

$$\begin{aligned}\frac{E_{total}}{L_yL_z} &= \frac{\hbar c}{2}\sum_m \frac{2}{(2\pi)^2}\iint dk_y dk_z \sqrt{\left(\frac{m\pi}{a}\right)^2 + k_y^2 + k_z^2} \\ &= \frac{\hbar}{2}\frac{c}{2\pi}\sum_m 2\int_0^\infty dk_{||}k_{||}\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}\end{aligned} \tag{6.44}$$

where

$$k_{||}^2 = k_y^2 + k_z^2. \tag{6.45}$$

Notice that an account has been made in (6.44) for the two polarizations of light, and in the second line of (6.44) a change to polar coordinates has been made in the integration in the y-z plane. This simplifies the considerations of the zero point energy in the following discussions.

In (6.44) the integral in $k_{||}$ is readily seen to be infinite so that the zero point energy per area between the plates is infinite. This infinity arises from the extreme short wavelength modes (i.e., the $k_{||} \to \infty$ modes) that occur in the mathematics. The divergence is well known to the study of field theories where it is termed as an ultraviolet catastrophe. It is a property of many field theories and standard renormalization procedures exist for dealing with these types of infinities. In the following one of these approaches shall be employed to understand the properties of the zero point energy given by (6.44).

The idea of the renormalization procedure is to introduce a parameterized factor into the integrand in (6.44) as a multiplicative term. The factor is chosen such that for most wavelengths it is unity but at very small wavelengths it goes rapidly to zero. This essentially cuts out of the system short wavelength modes in a controlled parameterized way. Varying the control parameter allows one to understand the nature and importance of the small wavelength modes to the behavior of the system.

Applying these ideas to the system in (6.44), the integral in (6.44) can be rewritten as [21]

$$\begin{aligned}\frac{E_{total}}{L_yL_z} &= \frac{\hbar}{2}\frac{c}{\pi}\sum_m \int_0^\infty dk_{||}k_{||}\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}\,e^{-\xi\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}} \\ &= \frac{\hbar}{2}\frac{c}{\pi}\sum_m -\frac{d}{d\xi}\int_0^\infty dk_{||}k_{||}e^{-\xi\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}}.\end{aligned} \tag{6.46}$$

Here a factor of $e^{-\xi\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}}$ has been introduced into the integrand for $\xi > 0$ a small parameter which in the limit that $\xi \to 0$ reduces (6.46) to (6.44).

In (6.46) the main effect of the exponential factor is that it is zero for the short wavelength modes. As a consequence, it removes these from contributing to the integral in (6.46). The factor ξ is also chosen so that the exponential has no effect on the intermediate to long range modes. The exponential form then introduces a cutoff parameter into the problem which can be varied at will.

The introduction of a cutoff factor in (6.46) is reasonable from a physical standpoint. Modes of arbitrarily large wave vector are also modes of arbitrarily large energy. At some point on the energy scale it seems physically unreasonable that such modes should be included in the considerations of the system.

Generally, applying the ideas of renormalization, (6.46) is evaluated for arbitrary $\xi > 0$. The resulting expression for (6.46) is separated into terms containing divergent factors in $\xi \to 0$ and terms in which such divergent factors in $\xi \to 0$ are absent. It is usually found that the terms involving divergent factors in $\xi \to 0$ do not enter into the physically measureable properties obtained from (6.46).

The measurable physics of the system, consequently, resides in the terms in which divergent factors in $\xi \to 0$ are absent. In the following this will be seen, specifically, to be the case for the changes in the zero point energy associated with changes in the boundary conditions of the electromagnetic fields of the system studied.

In order to obtain a closed algebraic form for (6.46) it is necessary to evaluate the integral [21]

$$\int_0^\infty dk_{||} k_{||} e^{-\xi\sqrt{\left(\frac{m\pi}{a}\right)^2 + k_{||}^2}}. \tag{6.47}$$

This can be done based on the application of the following two identities [21]

$$\frac{d}{dk} e^{-\xi\sqrt{b^2+k^2}} = -\xi \frac{k}{\sqrt{b^2+k^2}} e^{-\xi\sqrt{b^2+k^2}} \tag{6.48a}$$

and

$$\frac{d}{dk}\left[\sqrt{b^2+k^2} e^{-\xi\sqrt{b^2+k^2}}\right] = -\xi k e^{-\xi\sqrt{b^2+k^2}} + \frac{k}{\sqrt{b^2+k^2}} e^{-\xi\sqrt{b^2+k^2}}. \tag{6.48b}$$

From these two it is found that

$$k e^{-\xi\sqrt{b^2+k^2}} = -\frac{d}{dk}\left[\left(\frac{1}{\xi}\sqrt{b^2+k^2} + \frac{1}{\xi^2}\right) e^{-\xi\sqrt{b^2+k^2}}\right]. \tag{6.49}$$

This identity relates the integrand in (6.47) to a total derivative and, consequently, greatly simplifies the evaluation of the integral.

Applying (6.49) in the evaluation of (6.47) then gives [21]

$$\int_0^\infty dk_{||}k_{||}e^{-\xi\sqrt{\left(\frac{m\pi}{a}\right)^2+k_{||}^2}} = \left(\frac{1+\xi\left(\frac{m\pi}{a}\right)}{\xi^2}\right)e^{-\xi\left(\frac{m\pi}{a}\right)}. \tag{6.50}$$

This expresses the integral in terms of a closed form algebraic expression involving ξ.

The results for the zero point energy can now be assembled in terms of the above relationships. From (6.50) and the far right hand expression in (6.46) the zero point energy per area in the y-z plane becomes

$$\begin{aligned}\frac{E_{total}}{L_yL_z} &= \frac{\hbar}{2}\frac{c}{\pi}\sum_m -\frac{d}{d\xi}\left(\frac{1+\xi\left(\frac{m\pi}{a}\right)}{\xi^2}e^{-\xi\frac{m\pi}{a}}\right)\\ &= \frac{\hbar c}{\pi}\sum_m\left(\frac{1}{\xi^3}+\frac{1}{\xi^2}\frac{m\pi}{a}+\frac{1}{2\xi}\left(\frac{m\pi}{a}\right)^2\right)e^{-\xi\frac{m\pi}{a}}.\end{aligned} \tag{6.51}$$

This gives an expression for the zero point energy in terms of a number of sums over integers and exponentials.

The sums in (6.51), however, are much easier to handle if (6.51) is rewritten using a derivative notation. In this way it can be put into the form [21]

$$\frac{E_{total}}{L_yL_z} = \frac{\hbar c}{\pi}\left(\frac{1}{\xi^3}-\frac{1}{\xi^2}\frac{d}{d\xi}+\frac{1}{2\xi}\frac{d^2}{d\xi^2}\right)\sum_m e^{-\xi\frac{m\pi}{a}}. \tag{6.52}$$

Now only one infinite series need be summed.

From the identity [21]

$$\sum_m e^{-mc} = \frac{1}{1-e^{-c}} \tag{6.53}$$

where m runs over zero and the positive integers. Using this identity in (6.52), the zero point energy in (6.52) is given by [21]

$$\frac{E_{total}}{L_yL_z} = \frac{\hbar c}{2\pi}\left(\frac{1}{\xi^3}-\frac{1}{\xi^2}\frac{d}{d\xi}+\frac{1}{2\xi}\frac{d^2}{d\xi^2}\right)\left[e^{\frac{1}{2}\frac{\xi\pi}{a}}\cos ech\left(\frac{1}{2}\frac{\xi\pi}{a}\right)\right] \tag{6.54}$$

now expressed in terms of closed form well known functions.

In the limit that $\xi \to 0$ some terms in (6.54) will be divergent and some will not. To make this separation, it is useful to use the following expansion in small $\xi > 0$

$$e^{\frac{\xi\pi}{2a}}\cos ech\left(\frac{\xi\pi}{2a}\right) \approx \frac{2a}{\xi\pi}+1+\frac{1}{3}\frac{\xi\pi}{2a}-\frac{1}{45}\left(\frac{\xi\pi}{2a}\right)^3+\cdots. \tag{6.55}$$

From this it follows that [21]

$$\begin{aligned}&\left(\frac{1}{\xi^3}-\frac{1}{\xi^2}\frac{d}{d\xi}+\frac{1}{2\xi}\frac{d^2}{d\xi^2}\right)\left[e^{\frac{1}{2}\frac{\xi\pi}{a}}\cos ech\left(\frac{1}{2}\frac{\xi\pi}{a}\right)\right]\\&\approx\frac{6a}{\pi\xi^4}+\frac{1}{\zeta^3}-\frac{1}{360}\frac{\pi^3}{a^3}+O(\xi).\end{aligned} \tag{6.56}$$

Combining (6.55) and (6.56) in (6.54), the zero point energy per area between the two parallel plates takes the form of a Laurent series expansion in ξ. It is given by the specific form [21]

$$\frac{E_{total}}{L_yL_z}=\frac{\hbar c}{2\pi}\left[\frac{6a}{\pi\xi^4}+\frac{1}{\zeta^3}-\frac{1}{360}\frac{\pi^3}{a^3}+\cdots\right]. \tag{6.57}$$

For the limit $\xi\rightarrow 0$ of the series in (6.57) the first two terms are found to diverge, and the resulting divergence adds an infinite contribution to the zero point energy. This should be expected from the original considerations of the divergent integral in (6.44). Now, however, the nature of the divergence can be studied as a function of the cutoff parameter ξ.

In the limit that $\xi\rightarrow 0$, however, the third term in (6.57) is a constant independent of $\xi\rightarrow 0$, and the remaining terms of the series all go to zero. It shall be seen later that the third term in (6.57), which is a constant independent of ξ, is the important term in determining the Casimir force acting between the two plates. Before this conclusion can be reached, considerations of the system outside of the region between the plates must be made. In particular, the regions of vacuum outside that contained between the plates also have an important effect on the forces experienced by the parallel plates, and considerations must be extended to these.

To determine the force on the right hand plate in Fig. 6.4a it is important to realize that the plate interacts with two regions of vacuum. One is to the left of the plate and the other is to the right of the plate. The net force on the right hand plate is then obtained as the total contribution from these two regions of vacuum.

To simplify the treatment of determining the force on the right hand plate in Fig. 6.4a, consider the system in Fig. 6.4b. Here two parallel plates are again located at $x=0$ and $x=a$, but a third perfect conducting plate is now introduced at $x=L$. In the limit that $L\rightarrow\infty$, however, the system in Fig. 6.4b is seen to reduce to that in Fig. 6.4a.

A study of the system in Fig. 6.4b facilitates arriving at a result for the problem in Fig. 6.4a. Specifically, it allows for the determination of the force on the plate at $x=a$ based on the earlier presented calculations for the zero point energy between the plates in Fig. 6.4a. In the discussions of the force on the plate at $x=a$ it is, consequently, useful to determine the force in the context of the system in Fig. 6.4b.

For these considerations the following approach is taken: First a calculation of the force on the plate at $x=a$ are made for finite L. The result for the system in

Fig. 6.4a is then found by taking the $L \longrightarrow \infty$ limit at the end of the calculations made in the system of Fig. 6.4b.

Consider the vacuum zero point energy for the system in Fig. 6.4b. Of particular importance for determining the force on the plate at $x = a$ is the treatment of the net zero point energy in the regions of vacuum to the right and to the left of the plate at $x = a$. The net zero point energy in these two regions is the sum of that in the region $0 < x < a$ and that in the region $a < x < L$.

Between the plates at $x = 0$ and at $x = a$ the zero point energy of this region is still given by (6.57). This follows from the nature of the perfect conducting boundary conditions at the plates. It is a consequence of the separation of space in four isolated regions by the three plates in Fig. 6.1b and their perfect conducting boundary conditions [21].

Following this reasoning, in the region of vacuum between $x = a$ and $x = L$ the zero point energy can be obtained applying the exact same arguments as used in obtaining (6.57). In this manner it is found that in $a<x<L$ the zero point energy is [21]

$$\frac{E'_{total}}{L_y L_z} = \frac{\hbar c}{2\pi}\left[\frac{6(L-a)}{\pi\xi^4} + \frac{1}{\xi^3} - \frac{1}{360}\frac{\pi^3}{(L-a)^3} + \cdots\right]. \tag{6.58}$$

Alternatively, (6.58) can be directly obtained from (6.57) by replacing the variable a in (6.57) with $L - a$.

Summing (6.57) and (6.58) gives the total zero point energy within the region $0<x<a$ and $a<x<L$. Doing this the resultant sum is given by [21]

$$\frac{E^T_{total}}{L_y L_z} = \frac{\hbar c}{2\pi}\left[\frac{6L}{\pi\xi^4} + \frac{2}{\xi^3} - \frac{1}{360}\frac{\pi^3}{a^3} - \frac{1}{360}\frac{\pi^3}{(L-a)^3} + \cdots\right]. \tag{6.59}$$

From (6.59) the total zero point energy is found to be represented in terms of a series in the variables a and L.

An important point to note in (6.59) is that the terms that become infinite as $\xi \longrightarrow 0$ do not depend on the plate separation, a. This allows for a determination of the force on the plate at $x = a$ in terms of the derivative of the zero point energy with respect to a.

Upon doing this to determine the force on the plate no infinity is found to occur in the result. Taking the derivative, in the limit that $L \longrightarrow \infty$, the pressure on the $x = a$ plate is [21]

$$P = -\frac{d}{da}\frac{E^T_{total}}{L_y L_z} = -\frac{\hbar c}{2\pi}\frac{1}{120}\frac{\pi^3}{a^4} = -\frac{\pi^2 \hbar c}{240 a^4}. \tag{6.60}$$

From (6.60) it is seen that the pressure on the plate is negative so that the plates at $x = 0$ and $x = a$ are attracted to one another. In addition, the $\frac{1}{a^4}$ dependence on the plate separation indicates that the force is very short ranged. This signifies that the interaction is essentially limited to the length scales of nanoscience phenomena.

As an illustration of the restrictive nature of the Casimir interaction. Consider a particular numerical example. For two 1 × 1 cm plates separated from one another by 1 μm the Casimir force from (6.60) is 0.013 dynes. This is a small interaction which would quickly decay away with increasing distance. Another way of looking at this is that for this separation the pressure on the plates is of the order of atmospheric pressure [21].

The earlier calculations all deal with zero temperature systems. This is an extreme limit which highlights the unusual nature of the Casimir interaction. Some increase in the force associated with field fluctuations between the two plates is observed with an enhancement of the electromagnetic fluctuations in the system. For example, such an increase can be accomplished by introducing a non-zero temperature to the problem. The thermal fluctuations generated in the system themselves have a total energy. Like the energy of the zero point fluctuations, the energy of the thermal fluctuations depends on the boundary conditions of the system. These type of temperature effects and some related effects arising in the presence of dielectrics are now discussed.

Effects of Temperature and Dielectrics

The force generated by the thermal fluctuations of the system in Fig. 6.4a are now addressed. This is followed by additional examples of modifications to the physics of the system that can affect the Casimir interaction. Finally, some applications to nano-science are discussed.

As the temperature of the system is increased, thermal fluctuations begin to enter into the regions of vacuum in Fig. 6.4. In particular, the photon occupancy, n_{ph}, in (6.42) is no longer zero but increases to have an average positive value. For this the occupancy of the photons is determined by the Planck distribution and the boundary condition dependent frequencies of the electromagnetic modes in the system. Consequently, the energy of the photonic fields is changed from that of the zero-point energy as well as are the forces they exert on the perfect conducting plates in Fig. 6.4.

At finite temperatures, the fluctuations in the vacuum contributing to the total vacuum energy are now of two types. The first type is the zero-point fluctuations of the earlier considerations in (6.38) through (6.60). To these, upon introduction of temperature effects, are added the fluctuations contributing to a non-zero average of the photon occupation, n_{ph}. The calculations for the force between two plates now proceed similar to those of the zero temperature Casimir effect but with the introduction of the Planck distribution of photon modes in the additional energy terms arising from the n_{ph} in (6.42).

The pressure on the plates follows similarly from the derivative of the total energy with respect to a. In this way, considered to lowest order in the temperature, it is found that the first temperature correction to (6.60) is of the form [21–31]

$$P = -\frac{\pi^2 \hbar c}{240 a^4}\left[1 + \frac{16}{3}\frac{a^4}{\beta^4}\frac{1}{(\hbar c)^4}\right]. \tag{6.61}$$

This is valid at low temperature or for short plate separation distances a.

From (6.61) it is seen in this limit that as the temperature increases the pressure increases rapidly as the fourth power of the temperature. This is to be expected as the energy in the fluctuating fields increases rapidly with increasing temperature. Consequently, the thermal fluctuations should quickly overwhelm the fluctuations of the zero-point motion. This must be the case as the zero-point fluctuations, themselves, remain constant with changes in the temperature.

In the opposite limit of high temperatures or large plate separation distances the pressure on the plates is given by [21, 32]

$$P = -\frac{\zeta(3)}{4\pi\beta a^3}. \tag{6.62}$$

which is expressed in terms of the zeta function $\zeta(3)$. The pressure in (6.62) is found to be proportional to the temperature. This is reasonable as in the classical limit the field energy is proportional to the temperature.

Another question of interest in the study of Casimir problems is the effect on the Casimir force brought about by the introduction of a dielectric medium between the perfect conducting plates. In the idealized case of a medium with a constant frequency independent dielectric constant, $\varepsilon > 0$, the generalization is simple. The force between perfect conducting plates separated by a distance a is generalized from (6.61) to have the form [21–31]

$$P = -\frac{\pi^2 \hbar c}{240\sqrt{\frac{\varepsilon}{\varepsilon_0}}a^4}. \tag{6.63}$$

As in the case of vacuum between the plates, the force in the presence of a dielectric is attractive with the same rapid variation in the separation distance of the plates.

From (6.63) it is seen that for dielectric constants greater than one, the force on the plates is found to decrease with increasing values of the dielectric constant. This dependence of the force on the dielectric and its ability to decrease the Casimir force can be of significance in the study of nano- and micro-machines and other such electromechanical systems. In these type of devices, which are of interested to nanoscience, the Casimir force can become a factor affecting device performance.

In another consideration, the case has been treated in which the metal plates in Fig. 6.4a are characterized by a frequency dependent dielectric constant of the form of a Drude dielectric response. This type of response is found in systems with free conduction electrons or ions. It characterizes the response found in metals, ionic media, and plasmas.

The Drude dielectric response has a general form [21–31]

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \tag{6.64}$$

where ω_p is the plasma frequency of the free carriers given by

$$\omega_p^2 = \frac{4\pi e^2 N}{m} \tag{6.65}$$

in which N is the number density of free charged carriers in the medium, and m is the carrier mass.

For plates with the Drude response of the form in (6.64) an application of the ideas in the derivation of (6.60) yields a pressure on the two plates separated by a. In this way the pressure on the plates in Fig. 6.4a is given for $\frac{2\pi c}{\omega_p a} \ll 1$ by [29]

$$P = -\frac{\pi^2 \hbar c}{240 a^4}\left[1 - \frac{16}{3}\hbar c\frac{\delta}{a} + 120(\hbar c)^2\frac{\delta^2}{a^2}\right]. \tag{6.67}$$

where $\delta = \frac{1}{\hbar\omega_p}$. It is seen that the frequency dependent dielectric (similar to the results for the constant frequency dielectric media) decreases the attractive pressure acting between the two plates.

In all of the systems consider above the Casimir force has been found to exhibit an attraction between the two perfect conducting plates. A recent interesting study, however, has shown that in some cases the Casimir force between two parallel planar surfaces can be repulsive rather than attractive [26–28]. In particular, Kenneth et al. [28] have shown that the zero temperature Casimir force between a planar perfect conducting surface and a parallel planar surface of infinite permeability material experiences a repulsive Casimir interaction. Similarly, a repulse Casimir interaction is also found between parallel surfaces of two infinite permeability materials. These examples of Casimir repulsion have important implications for the design of metamaterials which exhibit the appropriate surface characteristics.

In line with earlier remarks, the adjustment and reversal of the Casimir force has important applications in the nanoscience of mechanical devices. At nanoscales the functioning of mechanical mechanisms can be affected by the Casimir force. As an example, attractive forces between the surfaces of such devices result in the phenomena known as ‘stiction’. This is the sticking together of the nano-surfaces arising from their attraction and can act to jam their mechanical operations.

The decrease or reversal of the Casimir force interaction can be of benefit in the design of such nanoscale devices. For the details of nanoelectromechanical systems (NEMS) and microelectromechanical systems (MEMS) and other mechanical properties associated with nanoscience the reader is referred to the literature [21–31].

References

1. R.S.M. Rikken, R.J.M. Nolte, J.C. Maan, J.C.M. van Hest, D.A. Wilson, P.C. Christianen, Manipulation of micro- and nanostructure motion with magnetic fields. Soft Matter **10**, 1295–1308 (2014)
2. R.F. Ismagilov, A. Schwartz, N. Bowden, G.M. Whitesides, A. Chem, Autonomous movement and self-assembly. Int. Ed. **41**, 652–654 (2002)
3. W. Gao, K.M. Manesh, J. Hua, S. Sattayasamitsathit, J. Wang, Hybrid nanomotor: a catalytically/magnetically powered adaptive nanowire swimmer. Small **7**, 2047–2051 (2011)
4. P. Tierno, R. Albalat, F. Sagues, Autonomously moving catalytic microellipsoids dynamically guided by external magnetic fields. Small **6**, 1749 (2010)
5. K. Guevorkian, J.M. Valles, Aligning Paramecium caudatum with static magnetic fields. Biophys. J. **90**, 3004–3011 (2006)
6. P. Dhar, Y. Cao, T. Kline, P. Pal, C. Swayne, T.M. Fischer, B. Miller, T.E. Mallouk, A. Sen, T.H. Johansen, Autonomously moving local nanoprobes in heterogeneous magnetic fields. J. Phys. Chem. C **111**, 3607–3613 (2007)
7. I.O. Shkilyarevski, P. Jonkheijm, P.C.M. Christianen, A.P.H.J. Schenning, E.W. Meijer, O. Henze, A.F.M. Kilbinger, W.J. Feast, A. Del Guerzo, J.-P. Desvergne, J.C. Maan, Magnetic deformation of self-assembled sexithiophene spherical nanocapsules. J. Am. Chem. Soc. **127**, 1112 (2005)
8. Y. Lui, K. Oh, J.G. Bai, C.-L. Chang, W. Yeo, J.-H. Chung, K.-L. Lee, W.K. Liu, Manipulation of nanoparticles and biolecules by electric field and surface tension. Comput. Methods Appl. Mech. Eng. **197**, 2156–2172 (2008)
9. Q. Chen, H. Huang, L. Chen, X. Ge, T. Chen, Z. Yang, L. Sun, Dielectrophoresis for Bioparticle Manipulation. Int. J. Mol. Sci. **15**, 18281–18309 (2014)
10. T.B. Jones, Basic theory of dielectrophoresis and electrorotation. IEEE Eng. Bio. Med. Mag. **22**, 33–42 (2003)
11. R.E. March, An introduction to quadrupole ion trap mass spectrometry. J. Mass. Spectro. **32**, 351–369 (1997)
12. J.R.C. Pita, Design, development and operation of novel ion trap geometries. Ph.D. thesis, Blackett Laboratory, Imperial College (2007)
13. M.S. Rocha, Optical tweezers for undergraduates: theoretical analysis and experiments. Am. J. Phys. **77**, 704–712 (2000)
14. A. Ashkin, Acceleration and trapping of particles by radiation pressure. Phys. Rev. Lett. **24**, 156–159 (1970)
15. A. Ashkin, J.M. Dziedzic, Optical trapping and manipulation of viruses and bacteria. Science **235**, 1517–1529 (1987)
16. A. Ashkin, Forces of a single-beam gradient laser trap oon a dielectric sphere in the ray optics regime. Biophys. J. **61**, 569–582 (1992)
17. K. Svoboda, S.M. Block, Biological applications of optical forces. Annu. Rev. Biophys. Biomol. Struct. **23**, 247–285 (1994)
18. A. Ashkin, Optical trapping and manipulation of neutral particles using lasers. Proc. Natl. Acad. Sci. U.S.A. **94**, 4853–4860 (1997)
19. H.-L. Guo, Z.-Y. Li, Optical tweezers technique and its applications. Sci. China Phys. Mech. Astronomy **56**, 2351–2360 (2013)
20. J.E. Molloy, M.J. Padgett, Lights, action: optical tweezers. Contemp. Phys. **43**, 241–258 (2002)
21. W.M.R. Simpson, *Surprises in Theoretical Casimir Physics: Quantum Forces in Inhomogeneous Media* (Springer, Heidelburg, 2015)
22. P. Ball, Feel the Force. Nature **447**, 772–774 (2007)
23. K.A. Milton, Recent developments in the Casimir effect. J. Phys.: Conf. Ser. **161**, 012001-1–012001-29 (2009)

24. H. De Los Santos, Nanoelectromechanical quantum circuits and systems. Proc. IEEE **91**, 1907–1922 (2003)
25. M. Sedighi Ghozotkhar, The Casimir for control in nano and micro electomechnical systems. Ph.D. thesis, University of Groningen (2016)
26. F.S.S. Rosa, On the possibility of Casimir repulsion using metamaterials. J. Phys. Conf. Ser. **161**, 012039-1–012039-8 (2009)
27. E. Buks, M.L. Roukes, Casimir force changes sign. Nature **419**, 119 (2002)
28. O. Kenneth, I. Klich, A. Mann, M. Rezen, Repulsive Casimir forces. Phys. Rev. Lett. **89**, 033001 (2002)
29. S.K. Lamoreaux, The Casimir force: background, experiments, and applications. Rep. Prog. Phys. **68**, 201–236 (2005)
30. G.L. Klimchitskaya, U. Mohideen, V.M. Mostepanenko, The Casimir force between real materials: experiment and theory. Rev. Mod. Phys. **81**, 1827–1880 (2009)
31. I. Bresvik, A. Ellingsen, A. Milton, Thermal corrections to the Casimir effect. New J. Phys. **8**, 236–256 (2006)
32. K.A. Milton, *Casimir Effect: Physical Manifestations of Zero-Point Energy* (World Scientific Publishing Co., Singapore, 2001)

Chapter 7
Lasers

In this chapter some basic discussion is presented about principles of laser operations and various types of lasers that are of interest in nanoscience applications [1–4]. Only an outline of the fundamentals of laser operation are presented as a quick review, and for more details the reader should consult the literature. In line with the focus of this book on systems commonly studied in nanoscience, discussions of vertical column lasers, spasers, and other types of nanoscience based lasers are presented. These considerations again are not meant to be comprehensive.

First a model is presented which explains many of the aspects of laser operation in terms of principles of nonlinear dynamics [1, 2]. This should act as a review, offering a simplified presentation of some of the basic aspects of laser operation. In this approach, the model chosen for consideration represents laser operation as a phase transition which is dependent on the power supplied to a media acting as a light source. The source media converts the power supplied into the generation of light.

Below a certain input power threshold, the light outputted from the media of the laser system is emitted from the laser as a regular incoherent light source. It displays no phase coherence and no amplification properties.

Above the laser power threshold, however, the nonlinearity of the system allows for a transformation causing the system to act like a laser. In this region of operation, an amplified coherent light is outputted from the laser.

While the phase transition model illustrations many of the basic feature encountered in the physics of lasers, it ignores quantum mechanical considerations which are a foundation for a complete understanding of the laser system. For a complete treatment of lasers, a full quantum field theory treatment is needed. This would be too much of a diversion at this point so that for such a treatment the reader is referred to one of the many texts available on the subject.

Following the discussions of the basic operation of the laser the focus will turn to the medium acting as a light source [3, 4]. Two basic sources of great importance to nanoscience applications are the heterojunction light source and surface plasmon light sources. Both of these types of sources will be discussed as well as some of

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_7

their basic applications. In particular, a brief review will be presented to understand the principles of operation of vertical column lasers and of the spaser. Some considerations of the laser threshold will also be made as well as discussions of the development of so-called zero threshold lasers.

7.1 A Simplified Model of Laser Operation

In this section, a general outline will be given of some of the basic features of laser operation. An extremely simplified discussion is presented of a model displaying many of the elementary features of a lasing system. It is meant as a brief introduction or review of some of the most salient features of the topic. The presentation does this within the context of a grossly simplified system which nevertheless provides an illustration of the mechanics of general laser operation. In most applications, however, the model must be generalized with the introduction of the detailed properties of the specific system being considered. In addition, the methods of quantum field theory must be used in a proper theoretical development of the topic.

A simple model that is an aid in understanding the basic operation of lasers is provided by considering a system of *N* identical atoms interacting with external electromagnetic fields [1, 2]. To facilitate the treatment, it is also assumed in the discussions that, in the absence of the external fields, the atoms do not interact with one another. As an additional simplifying point, the atoms themselves are each regarded as essentially having three energy levels with a single electron transitioning between these various atomic energy levels.

Such considerations of the proposed model can ultimately be extended to treat an arbitrarily defined electronic media. The assumptions outlined, however, form quite general considerations, illustrating the essential processes in many types of lasing systems. The complexities of the interactions in real systems are needed, however, to determine in detail the precise features of the lasing in a given system, but these do not change the basic ideas of the simple lasing process.

In the following the simple laser model is treated in order to obtain a general feel for laser operation. After this treatment, the discussions of lasers are applied to a qualitative treatment of the operation of semiconductor lasers and spasers. Semiconductor lasers are of great importance, forming a basis of many optoelectronic technological applications. They are a focus of many experimental investigations of the properties of the nanophotonic systems discussed in this book. Spasers are a more recent development which allows for the coherent generation of surface plasmon-polaritons into plasmonic circuits. This is another important development in nanoscience technology.

7.1.1 Statistical Properties of the N Atom System

In order to develop an understand the statistical nature of the energy level occupancy and the electromagnetic transitions between excited states of the atoms, first consider a system composed of two level atoms. Each atom has a ground state energy, E_0, and an excited state energy, E_1. (See Fig. 7.1 for a schematic of this system.)

A single atomic electron on each atom is shuttled between the energy levels of the atom through interaction with the electromagnetic fields. For this system of atoms and electromagnetic waves in thermal equilibrium, the number of atoms in excited states is N_1 and the number of ground state atoms is N_0 so that the total number of atoms $N = N_0 + N_1$.

From statistical physics [1–3], the Boltzmann weight

$$p_0 = \mathrm{e}^{-\beta E_0} \tag{7.1a}$$

is the relative probability that one of the atoms of the system is in its ground state, and

$$p_1 = \mathrm{e}^{-\beta E_1} \tag{7.1b}$$

is the relative probability of the same atom being in its excited state. In terms of these weights it then follows that the absolute probabilities for a given atom to be in the ground or the excited state are

$$P_0 = \frac{\mathrm{e}^{-\beta E_0}}{\mathrm{e}^{-\beta E_0} + \mathrm{e}^{-\beta E_1}}, \tag{7.2a}$$

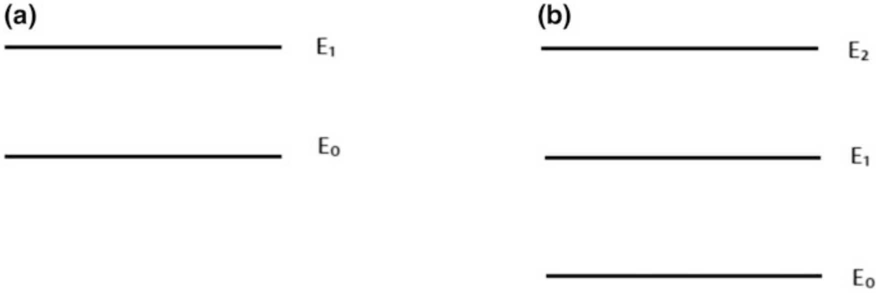


Fig. 7.1 Schematic of: **a** two level atom with ground state energy, E_0, and excited state, E_1, and **b** three level atom with ground state energy, E_0, and excited states, E_1, E_2 with $E_2 > E_1$

and

$$P_1 = \frac{e^{-\beta E_1}}{e^{-\beta E_0} + e^{-\beta E_1}}, \tag{7.2b}$$

respectively.

Using the probabilities in (7.2), it is found that for a system of N atoms that the number of atoms N_0 and N_1, respectively, in their ground and excited states are related to one another through the temperature of the system. In particular, their ratio is given by [1–3]

$$\frac{N_1}{N_0} = e^{-\beta(E_1 - E_0)}. \tag{7.3}$$

In statistical equilibrium the electron in a given atom is transferred between the ground and excited state through an interaction with the electromagnetic fields. The electrons in all of the atoms do this in such a way that the probability distributions in (7.1) and (7.2) are maintained. This means that in thermal equilibrium the rate at which an electron transitions from the excited state to the ground state must equal the rate at which it transitions form the ground state to the excited state. Otherwise the average occupancy of the two states would change with time, and this would not represent an equilibrium situation. The rates of these two types of transitions are well known form kinetic theory.

In kinetic theory it is shown that for a fixed volume system the rate of transition, $R_{0\to1}$, of the electrons in a system of N atoms from the ground states to the excited states is given by [1, 2]

$$R_{0\to1} = BN_0 n(\nu_{01}), \tag{7.4}$$

where $n(\nu_{01})$ is the number of photons in the system having the energy of the transition (i.e., $\nu_{01} = \frac{E_1 - E_0}{h}$), $B > 0$ is a frequency dependent constant of proportionality, and $R_{0\to1}$ is in units of $(\mathrm{s})^{-1}$.

The expression is a common type of rate expression for binary collision processes, and it should be noted that the expression for the rate, $R_{0\to1}$, is essential the same as that used in chemistry for the study of chemical reactions in which two reactants combine to form a single final product.

Going in the other direction, for a fixed volume system the rate of transition, $R_{1\to0}$, of the system electrons from the excited states to the ground states of the N atoms is given by

$$R_{1\to0} = N_1 f[n(\nu_{01})]. \tag{7.5}$$

where $f[n(\nu)]$ is a function of $n(\nu)$. The determination of the function $f[n(\nu)]$ is next addressed.

Applying the conditions for equilibrium, the transition rates in (7.4) and (7.5) are equated to yield the relationship

$$BN_0 n(\nu_{01}) = N_1 f[n(\nu_{01})]. \tag{7.6}$$

From (7.6), upon using (7.3), it is found that a solution for $f[n(\nu)]$ is of the form

$$f[n(\nu_{01})] = B\mathrm{e}^{\beta(E_1 - E_0)} n(\nu_{01}) = B\mathrm{e}^{\beta h \nu_{01}} n(\nu_{01}). \tag{7.7}$$

In the result in (7.7) the number of electromagnetic modes, $n(\nu)$, are related to the density of electromagnetic states, $\rho(\nu)$, by

$$n(\nu) = \rho(\nu) d\nu. \tag{7.8}$$

The Planck distribution of the density of electromagnetic modes in statistical equilibrium, $\rho(\nu)$, entering into (7.7) and (7.8) is a standard result of the statistical physics of the electromagnetic fields. There it is shown that the Planck distribution has the form [1, 2]

$$\rho(\nu) = C \frac{1}{\exp(\beta h \nu) - 1}, \tag{7.9}$$

where in (7.9) the coefficient $C > 0$ depends on the geometry of the system, e.g., $C = \frac{8\pi h \nu^3}{c^3}$ in three dimensions.

From (7.5), (7.7) and (7.9) it then follows that [1, 3]

$$\begin{aligned} f[n(\nu)] &= CB\left[1 + \frac{1}{\exp(\beta h \nu) - 1}\right] d\nu \\ &= B[Cd\nu + n(\nu)]. \end{aligned} \tag{7.10}$$

and, consequently,

$$R_{1\to 0} = BN_1 C d\nu + BN_1 n(\nu_{01}). \tag{7.11}$$

As with (7.4), (7.11) is similar to a transition rate found in chemistry for the study of chemical reactions in which a single compound dissociates into two compounds. The equality of (7.4) and (7.11) is similar to the condition used in the determination of the chemical equilibrium for two chemical reactants combining into a single product.

Unlike the rate of transition in (7.4) from the ground state to the excited state which consists of a single term, the rate of transition in (7.11) from the excited state to the ground state consists of a sum of two different terms. The terms in the sum in (7.11) represent two distinctly different types of processes. These differences will now be discussed.

The single process in $R_{0\to1}$ involves the product of the number of ground state atoms and the number of photons present in the system. It describes processes in which the transitions of the ground states to the excited states are induced by an absorptive interaction with the photons present.

The two processes involved in $R_{1\to0}$, however, are distinctly different from one another. The second term on the right of (7.11) depends on both the number of atoms in the excited state and the number of photons in the system so that the photons are actively involved in mediating the transition. These processes are known as stimulated emission processes and involve the addition of a photon to the system. Mathematically they are similar to the absorption processes represented in (7.4) as they do not occur if photons are not already present in the system.

The first term on the right of (7.11), on the other hand, depends only on the number of atoms in the excited states so that the transitions to the ground state can occur even in the absence of photons in the system. These processes are known as spontaneous emission processes. They add a photon to the system but do not require the assistance of a photon already in the system in order to make the transition.

One of the basic ideas of the operation of a laser is to use the dependence of the $R_{0\to1}$ and $R_{1\to0}$ on N_0 and N_1, respectively, to generate a surge of photons from the system. Specifically, if the ratio $\frac{N_1}{N_0}$ in the system is increased from its equilibrium value, $\frac{N_1}{N_0} = \mathrm{e}^{-\beta(E_1-E_0)}$, the excess excited states will decay to the ground state and, in the process, dump an excess of photons into the system. This is due to the increase in $\frac{R_{1\to0}}{R_{0\to1}}$ arising from the increase in $\frac{N_1}{N_0}$. A consequence of the increase in these ratios is that the rate of photon emission in the system is increased over the rate of photon absorption. The basic laser mechanism arising from these considerations will now be discussed [1, 3].

7.1.2 Laser Mechanism

To use these population related rate changes most effectively in the discussion of lasers it is best to treat a system of N isolated atoms with three energy levels $E_2 > E_1 > E_0$. Again, in this model each atom is considered to have a single electron that can transition between the three energy levels by means of interactions with the electromagnetic fields [1, 3]. The probabilities of finding the electrons in the various energy levels is given by the Boltzmann weights.

The idea for the operation of this system as a laser is to pump the individual atoms away from their equilibrium configuration. For example, pumping the atoms of the system with radiation for which $h\nu_{02} = E_2 - E_0$ causes the electrons in the ground states to transition to the E_2 excited states. This creates a population imbalance in the system of N atoms.

In the pumped system, the number of atoms in their E_2 excited states is increased over that in the equilibrium system. Once the system is pumped into a nonequilibrium configuration the next step is to apply radiation for which $h\nu_{12} = E_2 - E_1$. By means of this radiation, the pumped atoms can by stimulated emission be sent to the state, E_1.

This last process dumps photons of frequency ν_{12} into the system, enhancing the number of ν_{12} modes present in the system. The photons dumped in the stimulated emission process eventually show up in the laser output. This is one of a number of configurations of pumping and stimulated emission that can be used to generate a laser output [1, 3].

For example, it should be noted that in some systems electrons have very short lifetimes for the decay from the E_2 to the E_1 states. In these systems $h\nu_{02} = E_2 - E_0$ radiation is used to populate the E_2 level which then rapidly decays to the E_1 state. If E_1 has a slow transition rate it can act as the pumped state which develops an enhanced population from that in the system at equilibrium. Applying $h\nu_{01} = E_1 - E_0$ can then stimulate the transition of E_1 to E_0 with the consequent dump of ν_{01} photons into the system. This is another process in which the photons dumped in the stimulated emission process eventually show up in the laser output.

The development of a pumped state with an increased population over its equilibrium population is not the only condition necessary for laser operation. There is also a condition on the electromagnetic field used to stimulate the emission of the coherent laser light that is outputted by the laser.

The generation of a coherent output is done by putting the N pumped atoms in a Fabry-Perot resonator. This is a cavity resonator which has a series of resonant electromagnetic modes formed by its reflective walls and which surrounds the N atoms of the lasing medium. One of the walls of the cavity must be partially transparent so that the coherent light generated in the cavity is outputted as the output light from the laser.

The function of the cavity is to support the coherent mode generated from the N atoms as one of its resonant modes. This mode is used to stimulate the coherent dumping from the excess excited states of the atoms from their pumped excited states, creating the intense coherent output of light emitted from the partially transmitting wall of the resonant cavity.

In the following some simple considerations based on nonlinear dynamics are used to give a basic operational understanding of laser functions. These considerations are focused on modeling the rate of changes in the atomic energy level occupancies and the photons generated in the stimulated generation of coherent light, using the discussion provided earlier.

The details of the nature of the coherent field generated by the laser and the relationship between the coherent fields and atoms used to generate them in the system are not treated in this approach. The development of a coherent output from the resonant cavity is a topic of quantum electrodynamics and cannot be fully treated in an approach based on classical electrodynamics. For this the reader is referred to more advanced treatment of the subject.

Transition Processes in the Three Level System

To develop a simple nonlinear model of the laser based on atomic and photon transition and generation rates, begin with the various transition rates, $R_{a\rightarrow b}$, between states a and b for the atom-photon system in the earlier discussions. This is done for the three-level model of the single electron atom with $E_2 > E_1 > E_0$ [1, 3]. (See Fig. 7.1b for a schematic of the three-level atom.)

Following the earlier discussions of the three-level system, E_0 is the ground state of the system and E_2 is the level to be pumped [1, 3]. The rate of transition from the ground state to the excited state E_2 is

$$R_{0\rightarrow 2} = B_p N_0 n_p \tag{7.12}$$

where N_0 is the number of atoms with energy E_0, n_p is the number of pumping photons of energy $hv = E_2 - E_0$, and $B_p > 0$ is the rate coefficient. As in (7.4) the transition rate is proportional to the product of the numbers of the reactants, N_0 and n_p, that combine to form the final pumped states.

Once the atoms have been pumped, the stimulated emission of the excited state E_2 can be treated. This is the focus of the next considerations for the system.

Consider the system of N_2 pumped atoms having their electrons in the energy level E_2. At the time the system is in this pumped state the Fabry-Perot resonator has within it a number, n_s, of photons of energy $hv_s = E_2 - E_1$ interacting with the electrons in the pumped level E_2. These photons are available to act as agents of stimulated emission, inducing transitions of the electron from the E_2 to the E_1 levels.

During the transition process of the simulated emission, the n_s photons generate more $hv_s = E_2 - E_1$ photons causing n_s to increase. From simple kinetics, the manner in which n_s changes in time is described by the rate equation [1, 3]

$$\frac{dn_s}{dt} = B_s n_s N_2 - \alpha n_s. \tag{7.13}$$

The first term on the right of the equality describes the stimulated emission transition rate from E_2 to E_1 generating additional $hv_s = E_2 - E_1$ photons and increasing n_s. The coefficient $B_s > 0$ is the rate coefficient for this reaction path. In addition, to the stimulated emission processes, the laser also has losses due to the photons emitted in the laser beam and due to dissipation. The second term on the right of (7.13) represents these losses in the $hv_s = E_2 - E_1$ photon population of the cavity.

The effects on the photon population of the second term in (7.13) are seen by considering the case of (7.13) in which $B_s = 0$. In this limit the solution of (7.13) is in the form of an exponential decay given by

$$n_s(t) = \mathrm{e}^{-\alpha t}. \tag{7.14}$$

Equation (7.14) is the standard form associated with changes in a population made by simple dissipative processes. The losses represented by these processes contribute to the rate coefficient, α, and include: the removal of radiation from the cavity as it passes through the partially transmitting wall of the Fabry-Perot cavity, energy losses due to joule dissipation in the mirror and lasing medium, and spontaneous emission processes which are also in the system and are described by the transition rates

$$R_{2\to 1} = B_s N_2 C_s dv \tag{7.15a}$$

and

$$R_{2\to 0} = B_{s'} N_2 C_{s'} dv \tag{7.15b}$$

Equations (7.15a) and (7.15b) are based on the form for the spontaneous emission transition rate (i.e., transitions per second) described in the second term on the right of (7.11). In these expressions he coefficients $B_s, B_{s'}, C_s, C_{s'} > 0$.

From (7.13) it is seen that in the case that $N_2 = 0$ the number of photons in n_s only decreases in time, i.e., there is no lasing or amplification in the resonator cavity. It is known, however, that as N_2 increases from zero the N atom system eventually does, at some point, exhibit lasing. The point at which the power pumping into the E_2 states starts the system to lase is known as the threshold of the laser. It is a very important aspect of laser operations and the factors determining are now addressed [1, 3].

Lasing Equation and Fix-Point Solutions

Let N_2^* be the number of atoms required to be in the pumped state for lasing to begin. In the following N_2^* and the behavior of the system in the neighborhood of N_2^* are estimated. The approach used is reminiscent of the Landau theory of a second order phase transition, and, indeed, the beginning of lasing in the system is a second order phase transition. In these considerations, the object is to understand the nonlinear dynamics of (7.13) in the vicinity of the lasing transition [1].

Consider the system in the absence of stimulated emission and let the number of pumped atoms for this case be N_{20}. Once stimulated emission is introduced into the system the number of N_2 modes in the system is decrease from N_{20}. To model this decrease, assume that the pumped modes are approximated by

$$N_2(t) \approx N_{20} - \beta n_s. \tag{7.16}$$

where n_s are the number of photons available to generated spontaneous emission processes.

In (7.16) the second term on the right represents a decrease in the pumped modes due to decay through stimulated transitions, and the coefficient $\beta > 0$ relates this loss to the number of modes n_s or the intensity of the of the fields stimulating the transitions. Under these considerations (7.13) becomes [1]

$$\begin{aligned}\frac{dn_s}{dt} &= B_s n_s [N_{20} - \beta n_s] - \alpha n_s \\ &= [B_s N_{20} - \alpha] n_s - \beta B_s n_s^2.\end{aligned} \tag{7.17}$$

The dynamics of the system in (7.17) contains the description of the lasing process. To understand this, the solutions of (7.17) are now studied. The nature of these solutions are qualitatively discussed in terms of the functioning of the laser operations which they describe.

Equation (7.17) is a well know equation of nonlinear dynamics. To begin the study of its solutions, it is good to determine the $\frac{dn_s}{dt} = 0$ fixed points of the system. At these points the system does not change in time.

Denoting the fixed-point solutions of (7.17) by n_s^0, the values of n_s^0 are determined as solutions of

$$[B_s N_{20} - \alpha] n_s^0 - \beta B_s \left(n_s^0\right)^2 = 0. \tag{7.18}$$

The stationary solutions are then found to be [1]

$$n_s^0 = 0 \tag{7.19}$$

and

$$n_s^0 = \frac{B_s N_{20} - \alpha}{\beta B_s}. \tag{7.20}$$

Examining the results in (7.19) and (7.20), a variety of behaviors are found in the system. In particular, (7.19) is the case in which there are no photons available to induce stimulated emissions. As shall be seen in the following, this is not of interest for laser operation.

On the other hand, for positive n_s^0 (7.20) represents the case in which photons are available to induce stimulated emissions in the system, and it appears at this point that the system is self-sustaining. This is the condition for laser operation.

These three points are now described and discussed in detail as well as the behavior of the system in the neighborhood of three fix-points. At the end of these consideration an expression for the laser threshold will be obtained in terms of the various transition rates discussed earlier.

Behaviors Near the Fix-Points

The three configurations of the solutions in (7.19) and (7.20) are the following [1]:

1. When $B_s N_{20} - \alpha = 0$ both solutions of (7.19) and (7.20) for the fixed point of n_s^0 coalesce and are located at $n_s^0 = 0$. From (7.17) the time derivative at general n_s is then given by

$$\frac{dn_s}{dt} = -\beta B_s n_s^2. \tag{7.21a}$$

Considering this equation there are three possible initial conditions of n_s to be treated. These are initial states of $n_s^0 = 0$, $n_s^0 > 0$, and $n_s^0 < 0$.

If the system is initially at the two coalesced $n_s(t) = 0$ fixed points, then from (7.21a) it follows that $\frac{dn_s}{dt} = 0$ always. The solutions for $n_s^0(t)$ are zero and constant in time, with the consequence that no light is ever present in the system.

If $n_s(t) > 0$ it follows from (7.21a) that $\frac{dn_s}{dt} < 0$ so that as time progresses $n_s(t)$ always decreases as it travels towards the $n_s^0 = 0$ fixed point. Consequently, the system undergoes a relaxation to the fixed point of the system, and in the end state there are no photons in the system.

In the case of $n_s(t) < 0$ the system would have a negative occupancy of photons which is not physically of interest. It is not possible to have a negative photonic occupancy in the system so that this case will not be considered.

It is seen from these discussions of (7.21a) that the system always approaches the $n_s^0 = 0$ fixed point. This is the state in which no photons are available to induce spontaneous emissions. In this configuration, the system is acting as a light source with an intensity of light which is being dissipated away through photons leaving the system and by losses within the materials of the cavity and lasing medium.

2. When $B_s N_{20} - \alpha = D < 0$ only the $n_2^0 = 0$ fixed point in (7.19) is of interest. This follows as the non-zero fixed point in (7.20) is negative. This is an unphysical condition, and, consequently, not of interest in the following considerations.

   From (7.17) it follows that

   $$\frac{dn_s}{dt} = Dn_s - \beta B_s n_s^2 \tag{7.21b}$$

   and for $n_s(t) > 0$ the righthand side is always less than zero. It follows that as time progresses $n_s(t) > 0$ always decreases as it travels towards the $n_s^0 = 0$ fixed point. The $n_s(t) < 0$ case, again, is not physically interesting.

   In this case of the system, as earlier with that in case 1, the $n_s^0 = 0$ fixed point is always approached. No photons are available to induce spontaneous emissions. In this way, the system is acting as a light source with an intensity of light dissipated away through photons leaving the system and by losses within the materials of the cavity and lasing medium.
3. When $B_s N_{20} - \alpha = D > 0$ both (7.19) and (7.20) are physical fixed points of the system. From a treatment of the fix point in (7.20) in this limit follows the lasing transition.

From (7.17) it follows that [1]

$$\frac{dn_s}{dt} = Dn_s - \beta B_s n_s^2. \tag{7.21c}$$

This can be rewritten into the form

$$\frac{dn_s}{dt} = \beta B_s \left[n_s^0 - n_s\right] n_s. \tag{7.21d}$$

where $n_s^0 = \frac{D}{\beta B_s}$ is the nonzero fixed point.

It is readily seen that for $n_s > n_s^0$, $\frac{dn_s}{dt} < 0$ and $n_s(t)$ travels uniformly to the $n_s^0 = \frac{D}{\beta B_s}$ fixed point. In addition, for $0 < n_s < n_s^0$, $\frac{dn_s}{dt} > 0$ and again $n_s(t)$ travels uniformly to the $n_s^0 = \frac{D}{\beta B_s}$ fixed point. Initial conditions in both of these regions are found to display a relaxation to the non-zero fixed point.

On the other hand, the case $0 > n_s, \frac{dn_s}{dt} < 0$ is not of physical interest. It will not be considered further here.

In the physical cases of the system, the $n_s^0 = \frac{D}{\beta B_s}$ fixed point is always approached. This is the state in which photons are available to induce spontaneous emissions, and the system is maintaining itself as a steady state source of intense light. That is, it is acting as a steady state lasing source.

The systems first begins to lase when $D = B_s N_{20} - \alpha$ passes through zero, heading along its positive $D > 0$ trajectory. This condition occurs for

$$N_{20} = \frac{\alpha}{B_s} \tag{7.22}$$

which is the population of the pumped E_2 level at the threshold of the laser.

To develop a more concise understanding of the solutions discussed above and their relation to the fix-points of (7.17), it is helpful to rewrite the equation in a nicer format. In particular, (7.17) can be rewritten in the form [1, 2]

$$\frac{dn_s}{dt_r} = [n_0 - n_s] n_s \tag{7.23a}$$

where

$$n_0 = \frac{N_{20}}{\beta} - \frac{\alpha}{\beta B_s} \tag{7.23b}$$

and the time variable is renormalized to the form $t_r = \beta B_s t$. This allows for a detailed illustration of the properties of the system solely in terms of the reduced variable n_0.

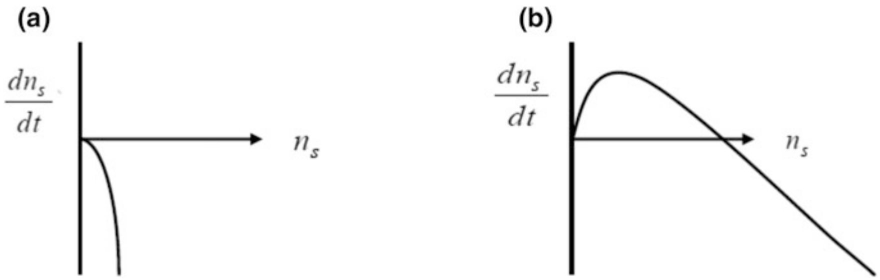


Fig. 7.2 Plot of $\frac{dn_s}{dt_r}$ versus n_s for: **a** cases 1 and 2 in which there is only a physical solution at the fixed point $n_s = 0$ and **b** case 3 in which there also is a non-zero physical fix-point that is an attractor

In this regard, in Fig. 7.2 some plots of $\frac{dn_s}{dt_r}$ as a function of n_s from (7.23) are presented for the three cases of nonlinear laser dynamics treated earlier. For case 1 the parabolic function only intersects the abscissa at $n_s = 0$, and no lasing occurs. Similarly, for case 2 the only non-negative intersection of the function with the abscissa is at $n_s = 0$ and again no lasing occurs. In both of these cases the system has not attained the lasing threshold and the power feed into the system to pump the atoms does not sustain the system.

For case 3, the physical solutions of the system approach the stable fixed point of the system at $n_s^0 = \frac{D}{\beta B_s}$. The laser threshold is first attained at $D = 0$ for which at this point $N_{20} = \frac{\alpha}{B_s}$. In addition, at the nonzero fixed-point, (7.23) yields an expression for N_{20} given by

$$N_{20} = \beta n_0 + \frac{\alpha}{B_s}. \tag{7.24}$$

The laser threshold in this simple model is found in (7.24) to depend on the loss coefficient, α, and the gain coefficient, B_s, in (7.13). Much technological effort in laser engineering has gone into lowering the power needed to create the N_{20} threshold level of pumping in the laser medium.

7.2 Semiconductor Lasers

An important light source in nanoscience applications is based on heterojunctions formed between two different types of semiconducting materials [3, 4]. Heterojunctions have become popular in engineering applications as they are based on widely studied semiconductor technologies and offer many degrees of freedom for potential design purposes. In this regard, semiconductors have formed the basis of an extensive industry where they have been developed in the designs of diodes, transistors, and a variety of other devices and electronic circuitry. They also have been shown to function with high efficiency and exhibit the durable properties needed in effective component formulations. Of a great significance, they operate effectively at frequencies commonly used in many technological and fiber optics applications.

In the following, some of the basics of semiconductor junctions operated as light sources in laser applications will be presented [3, 4]. These topics will be addressed at a qualitative level to give an idea of the functioning of heterojunctions, and the emphasis will be on basic operating principles. More advanced treatments can readily be encountered in the extensive existing literature on these topics.

After this brief outline, the discussions will culminate in a treatment of the application of heterojunctions to interesting recent applications in the design of vertical column heterojunction lasers. These types of lasers have been a focus in the design of efficient low threshold lasers and have been offered for a number of other applications in nanoscience.

In the study of semiconductor junctions there are two types of junctions. The first type is a homojunction and the second type is a heterojunction. The homojunction is formed by doping an otherwise homogeneous material with small concentrations of two different types of impurity atoms. This is done so that a planar interface is formed between the regions of two different types of doping. On one side of the interface the material contains p-type (acceptor) impurities, and on the other side the material contains n-type (donor) impurities. In the p material the current carried by the medium is due to the motion of holes, while in the n material the current carried by the medium is due to the motion of electrons [3].

The second type of junction is a heterojunction. This is formed as an interface arrangement involving two different materials. Typically, in laser designs, two planar interfaces are formed to make a sandwich of the two different materials. At the center of the sandwich is one type of semiconductor medium and on the two sides of the center material is a second type of material. The second type of material on one side of the sandwich is p doped and the second type of material on the other side of the sandwich is n doped. The material at the center of the sandwich need not be doped.

For the heterojunction design the two different types of materials used must have similar lattice parameters. This is so that a relaxed interface can be created between the three layers forming the junction and for the structural stability and uniformity of the electrical properties of the junction. In this arrangement, the p and n doping is small and has little effect on the lattice parameters of the materials being doped.

7.2.1 Homojunctions

To understand the basis of the operation of a homojunction consider the n-p junction shown schematically in Fig. 7.3. The homojunction portrayed is the simplest type of junction, and it is formed at the interface developed between

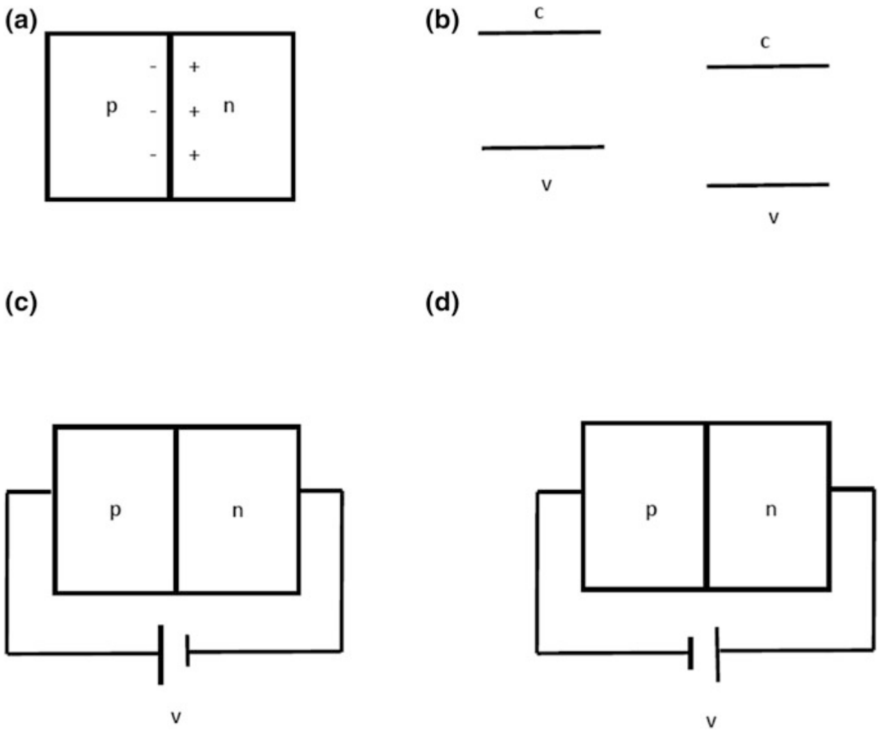


Fig. 7.3 **a** n-p homojunction, **b** relation of bulk band structures, **c** forward biased circuit, and **d** reversed biased circuit

differently doped versions of the same material. Since the dopants are generally at very small concentrations the junction can be created by diffusion of the impurities into the materials or by deposition techniques [3].

In the figure, such a junction is illustrated for the case in which on one side of the interface the medium has p type impurities and on the other side is n type impurities. Ultimately, in discussions of the junctions in the context of circuit applications, the interest is on electric currents traveling perpendicular to the interfaces. For these applications, the junction is found to exhibit a variety of rectification and light generating properties associated with the flow of electrical currents.

To the left side of the interface in Fig. 7.3 is the p material in which the current is carried by holes and to the right side of the interface is the n material in which the current is carried by electrons. Due to osmosis effects acting on the carriers in the two materials, the holes on the left of the interface will tend to diffuse into the n material on the right of the interface and the electrons on the right of the interface will tend to diffuse into the p material on the left of the interface [3].

A consequence of this is that a plane of positive charge develops to the right of the interface and a plane of negative charge is developed on the left of the interface. These planes are indicated on the figure by the + and − signs positioned on the respective sides of the interface plane. For the moment, a focus will be on considerations of the factors responsible for the development of the two charged planes. These planes are important factors which are responsible for the technological applications of the junction so that the mechanisms affecting them must be understood.

The reason that planes of charge are developed on the adjacent sides of the interface is due to a balance of osmotic and electrostatic forces. As the charges are driven across the interface by the process of osmosis, an electrostatic interaction develops between the planes of + and − charges formed at the sides of the interface. The electrostatic interactions oppose the osmotic pressures so that eventually a balance is reached between these two factors.

An equilibrium is ultimately established between the charge densities at the two planes of charge, and this distribution of charges generates a change in the electrical potential between the n and p doped materials. Alteration of the opposing forces of osmotic pressure and electrostatic forces by the means of external interactions with the system then becomes the source of a variety of important applications.

In addition, the proximity of electrons and holes generated near the interface between the p and n media allows for the generation of light through the mechanism of recombination. This is the processes in which the electron and hole charges neutralize some of the donor or acceptor ions in their respective media around the vicinity of the p-n junction interface. Essentially the holes in the valence band are filled by electrons from the conduction and in the transition a photon is created. A photon of characteristic radiation from the recombination is emitted in the course of the process [3].

The process of electron hole recombination is the basis of LED’s that are developed through the application of junction technology. In these applications, the light originates in the neighborhood of the interfaces where populations of both electrons and holes are present and injected into the p and n media, respectively.

The electric potential associated with the two planes can be calculated because these planes represent the only net accumulation of charge within the system. All other regions of the n and p materials are charge neutral. Consequently, the change in the electric potential in going from the negative charged plane on the p side of the interface to the positively charged plane on the n side of the interface is a constant $\phi_0 > 0$. This means that the difference in electric potential in going from the bulk of the p material to the bulk of the n material is ϕ_0.

The change in electric potential at the interface, represented by ϕ_0, is seen to shift the potential energy of the electrons in the bulk of the n material by [3]

$$U_0 = -e\phi_0. \tag{7.25}$$

(Note that here the charged of the negatively charged electron is represented by $-e$.) Specifically, the energies of the electrons in the n material for $\phi_0 = 0$ are for $\phi_0 \neq 0$ shifted in potential energy by $U_0 = -e\phi_0$. This has important consequences for the electron and hole states in the n and p type materials of the junction materials. As shall be seen the potential energy in (7.25) shifts the energy bands of the n and p materials relative to one another.

The value of ϕ_0 is related to an important difference in the band structure of the media separated by the junction interface. In this regard, it is known form statistical physics that at equilibrium the chemical potential of the junction system must be constant over the junction materials. However, in the p material the chemical potential is close to the lower edge of the semiconductor stop band, while in the n material the chemical potential is close to the upper edge of the semiconductor stop band. For the chemical potential to be a constant over the junction, the bulk band structure of the n type of material must be shifted downward relative to that of the p type materials.

Consequently, for the chemical potential to be constant over the media of the junction, the arrangement of Fig. 7.3b is required between the bulk of the two materials. The bands of the n material are shifted by $U_0 = -e\phi_0$ relative to those in the p material. In this regard, the electrostatics and statistical physics of the media match together.

Now consider connecting the junction to a battery. How will the addition of the external potential change the equilibrium at the junction interface? Since the n-p junction can be connected to the battery in two different way, there are two different cases of the battery-junction system to be treated.

First consider the case that a battery of potential V is connected with the positive terminal attached to the p material and the negative terminal attached to the n material. (This is schematically shown in Fig. 7.3c.) In this arrangement, the potential of the n material relative to the p material changes from ϕ_0 to $\phi_0 - V$, and the potential energy changes from (7.25) to [3]

$$U_0 = -e\phi_0 + eV. \tag{7.26}$$

Here it is assumed that the drop of the batteries potential across the junction occurs solely at the interface between the p and n materials. This is generally a very good assumption.

The shift in (7.26) between the band structures of the p and n materials is seen to decrease the offset of these band structures. The net effect is to readjust the offset between the band structure of the two media, making them look the same. This readjustment, pushing the bands back into alignment, upsets the flows of osmotic and electrostatically induced current providing for the equilibrium configuration of the system. Changes to these processes now are such as to facilitate a net flow of electrons in the n bulk to pass to the p material. As a result, a complete net steady state flow of charge is established through the circuit.

It should be noted in this regard that the conduction electrons are distributed over the energy levels of the conduction band due to thermal effects. An upward shift of the bands of the n material facilitates the passage of thermally excited electrons to the p materials. Likewise, thermally excited holes from the p material are facilitated in flowing through the circuit. The effect on both the electron and hole flows are dramatically changed in this configuration, known as forward biasing of the junction.

In the reversed configuration, the battery is now connected with the negative terminal applied to the p material and the positive terminal applied to the n material. (This is shown schematically in Fig. 7.3d.) Here the relative potential between the bulk p and n materials is given by

$$U_0 = -e\phi_0 - eV, \tag{7.27}$$

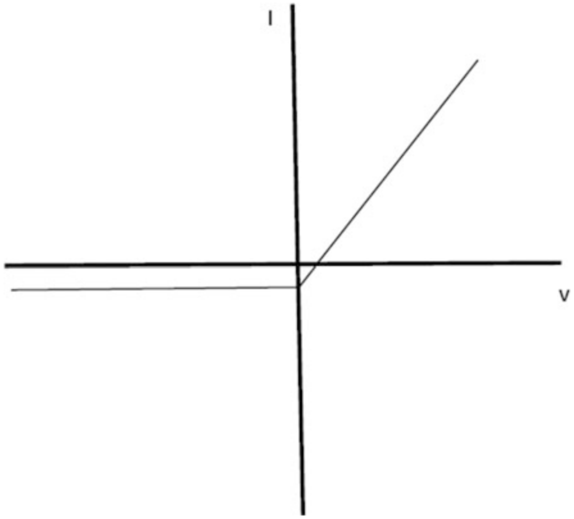
and the shift widens the separation offset between the band edges in the p and n. Now the offset of the band structures within the two media is enhanced. However, the net effects on the currents arising from osmosis and electrostatic effects is much smaller than in the case of forward biasing. The resulting current generated in the circuit is much less than in the forward biased system.

The difference between the forward and reverse biasing is shown in the typical nonlinear current versus voltage curve shown in Fig. 7.4. Applications of the curve are found in various diode and transistor technologies where the difference in the forward and reversed biased currents are a source of a variety of switching and rectification application.

7.2.2 Heterojunctions

The function of the heterojunction is in many ways similar to that of the homojunction [3]. There are, however, some important differences. The most import of these differences involves the band structure of the two types of materials used in

Fig. 7.4 Qualitative current versus voltage relationship for the p-n junction



the heterojunction design. This difference is very important in the development of laser applications, particularly in the generation of pumped systems. In the following brief discussions, the focus will be on the commonly used heterojunction design materials of GaAs and GaAlAs.

In application of the heterojunction as a source of light in laser designs, it is useful to consider a system of three materials separated by two planar interfaces. Unlike the homojunction interface discussed earlier, this is a type of semiconductor sandwich. However, as with the homojunction the current flow in the system is again perpendicular to the interfaces. Of the two materials utilized in the sandwich design, the two outer layers of the junction are made of n doped GaAlAs and p doped GaAlAs, while the material located between these is formed of GaAs.

The band structure of the three materials are shown schematically in Fig. 7.5. In these band structures, the doping of the GaAlAs again has little effect on the relative band structure so that these are represented as being the same on both sides of the sandwiched GaAs layer. As with the homojunction, the doping offsets the band structure of the n and p materials but otherwise has little effect on the differences between the energy levels in the two systems. This offset property of the band structures of the n and p materials is again a result of the uniformity of the chemical potentials over the two materials forming the junction [3].

In the middle layer, the stop band of the GaAs is seen to be much less than that of either of the GaAlAs layers. This is an important point as it allows for the development of a region to trap electrons and holes in the GaAs layer. For example, in the forward biased configuration of the junction (i.e., positive battery terminal connected to the p material and negative battery terminal connected to the n material), electrons will enter the GaAs layer and collect there. This is because many electrons will not have sufficient energy to enter into the p medium and will become localized to the GaAs layer. In a short time, the collected electrons will thermalize into lower energy level conduction band states of the GaAs layer. Similarly, holes enter the GaAs layer from the p material on the left of the GaAs sandwich and will also become thermalized within the GaAs layer [3].

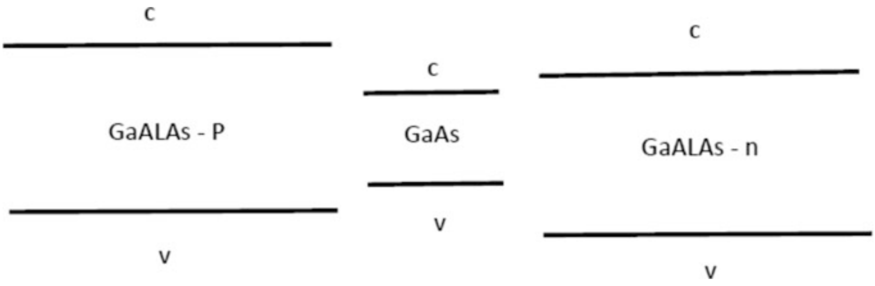


Fig. 7.5 Band structure in the bulk of the regions of GaAlAs-p, GaAs, and GaAlAs-n regions of a heterojunction. In an application the current would flow in the direction normal to the planar interfaces

The two thermalized populations of electrons and holes in the GaAs layer are now set to recombine and generate emitted photons. This recombination is the mechanism of light generation in the heterojunction laser, and the motion of the electrons and holes into this region sets the junction up for an emission process as the thermalization into two populations are configured for recombination.

In the lasing process, the function of the GaAs layer is to collect and hold the electrons and holes and to facilitate their recombination with the emission of a photon. The GaAs also has an important property which allows the radiation generated in the electron hole recombination processes to be tuned in frequency. This comes from the quantum well nature of the GaAs layer.

In particular, the energy levels of the electrons and holed confined within the GaAa layer are solutions of the Schrodinger equation subject to appropriate boundary conditions at the layer surfaces. These boundary conditions affect the energy levels of the electron and hole modes confined within the layer and introduces a dependence of these energies on the width of the layer.

Due to the dependence of the energy density of electron states within the GaAs layer on the width of the GaAs layer, the energy of the light emitted can be tuned to generate a desired laser output. This is an aid in many design applications. In addition, if a number of GaAs-GaAlAs layers are used to create a series of heterojunction wells, the output efficiency of the sequence of heterojunctions can be managed for an increased performance of the laser. The ability of tuning the light emitted and controlling the efficiency of its generation are two great successes of the heterojunction light source [3].

7.2.3 Vertical Cavity Surface Emitting Laser

A recent example of a laser system based on the heterojunction or arrays of heterojunctions is the vertical cavity surface emitting laser [3]. In this type of device, a sequence of layers is formed on a substrate by various successive depositions. Some of the depositions are made to form a heterojunction light source and others are made to form a Bragg reflector to confine and control the emitted light traveling through the layers [3]. These form the vertical column of the laser.

The first layers of the deposition are those of a distributed Bragg reflector. These are part of the confining mechanism for light emitted through the interfaces of the layered system. These are then followed by the heterojunction or heterojunctions which generates the light in the system. This is the region of the active media of the laser. A final arrangement of layers is another sequence forming a second region of a distributed Bragg reflector. The totality of the layers represents a heterojunction light source surrounded by Bragg layers which confine the light passing through the layer interfaces.

In addition to the structure made normal to the substrate surfaces, a photonic crystal patterning can also be applied parallel to the plane of the substrate surfaces. The photonic crystal pattern is designed to help modulate the light flow in the planes parallel to the substrate surfaces. In regard of the layering and photonic crystal, the distributed Bragg reflectors modulate the flow of light normal to the substrate surface and the photonic crystal controls motion in the plane of the layers. In this arrangement, the vertical column of layered materials forming the heterojunction and Bragg reflectors is surrounded by the confining media forming the photonic crystal.

In the laser design, the layering is arranged into a column sitting on the substrate, and the light emitted from the laser is outputted in one direction normal to the surfaces of the layers. Consequently, the important component of light emitted from the junctions is that generated normal to the junction interface. In most of the systems studied to date, the layers are generally formed of materials so that the light generated in the laser is at frequencies useful in fiber optics technologies. As a result, in the majority of applications, these types of laser devices have found uses in fiber optics and in communication systems for data transmission over fiber optics [3].

As an important point, the vertical cavity surface emitting laser designs facilitate the lowering of the threshold of operation of the laser. Aside from their applications in fiber optics technology, this is one of their features of current interest. Particularly, the incorporation of photonic crystal in the laser design is intended to be an aid in achieving this lowering of the lasing threshold and the increase of device efficiency. These techniques are helpful in the goal to create highly efficient so-called zero threshold lasers, and the reader is referred to the literature for further details of these applications [3].

7.3 Spasers

Another type of laser device that can be important to nanoscience technologies is the spaser [4]. Spaser stands for surface plasmon amplification by stimulated emission of radiation, and it involves the creation of an output of amplified coherent surface plasmon-polaritons. In contrast, the laser involves the creation of an output of amplified coherent light. Both of these devices have applications in nanoscience, but the spaser has a focus on plasmonic technologies.

The direct generation of surface plasmon-polaritons on a surface can be of importance in study of plasmonic circuits and in the operation of plasmonic devices. In such systems, the indirect introduction of surface plasmons onto a surface by means of the coupling of a laser output into the surface plasmon-polariton modes can often be inefficient. In most cases, it more effective to avoid intermediary process and to introduce surface plasmons by means of spacers [4].

The principles for the operation of a spaser are very close to those of a laser. An example of a simple spaser geometry is a nano-spherical shell which is either coated with or encloses a layer of quantum dots. In this arrangement, the nano-shell supports surface plasmon-polaritons modes and acts as a resonator for these modes.

The quantum dots are the active media for the generation of the surface plasmon-polaritons. Consequently, the two components central to the spaser operation are the nano-resonator and the quantum dots.

First consider the operation of the quantum dots. In the spaser operation an excited state of independent electrons and holes is introduced into the quantum dots. These then relax to form bound excitons with energies near those of the surface plasmon-polariton modes of the nano-resonators. As the excited exciton states of the quantum dots return to the ground state they radiate into the surface plasmon-polariton modes of the nano-resonators [4].

The second component of the spaser is the nano-resonator. This receives the energy from the excitons formed in the quantum dot and performs the stimulated emission function of the resonator in the laser. In this case, however, the amplified radiation is that generated in the surface plasmon-polaritons of the resonator modes.

As is seen, the spaser differs from the laser in the nature of the modes outputted from the device. The surface plasmon-polaritons outputted from the spaser are a combination of electromagnetic modes and modes of the dielectric. In addition, the surface plasmon-polaritons are based on different type of resonator principles from those of the optical laser resonator. Particularly, the wavelengths of the surface plasmon-polariton resonators involve smaller wavelength excitations than those typically handled by laser resonators. Laser resonators require length along the axis of the emission of radiation that are multiples of half-wavelengths of the light being emitted. This means that nano-particles can perform the resonator function in spaser designs [4].

References

1. S.H. Strogatz, *Nonlinear Dynamics and Chaos* (Perseus Publishing, LLC, 1994)
2. H. Haken, *Synergetic*, 3rd edn. (Springer, Berlin, 1983)
3. L. Solymar, D. Walsh, *Electrical Properties of Materials*, 7th edn. (Oxford University Press, Oxford, 2004) and C. Kittle, *Introduction to Solid State Physics*, 7th edn. (Willey, New York, 1996), and B. Hitz, J.J. Ewing, J. Hecht *Introduction to Laser Technology*, 3rd edn. (IEEE Press, New York, 2001)
4. M.I. Stockman, Spaser, plasmonic amplification, and loss compensation, in *Active Plasmonics and Tuneable Plasmonic Metamaterials*, ed. by A.V. Zayats, S. Maier (Willey, New York, 2013)

Chapter 8
Near Field Microscopy

In the following, some of the basic considerations at the foundation of near field microscopy are presented [1–5]. Near field microscopy is a new form of microscopy which was first proposed in 1928 as a means of increasing microscope resolution over that available by traditional techniques of far field microscopes. In this regard, it offers great potential applications both in nanoscience and in the study of biological materials. It has taken many years to realize the ideas of near field microscopy into functioning devices, but since the 1980s near field microscopy has developed into a recognized laboratory technique. Recently, a Nobel Prize has been awarded based on the applications of near field microscope techniques.

Microscopy can be roughly divided into two basic methodologies [1–10]. These include far field and near field microscopy. The older technique of far field microscopy is formulated on systems involving designs based primarily on focusing arrays of lenses. In far field microscopy, light from an object, located several wavelengths away from the collecting aperture of the microscopy, is focused by the microscope into an image. This is accomplished through the interaction of light generated at the object with the lenses of the device. In this way, ultimately the light is steered to form an image of the original object which approximates the features found within the object.

In the study of systems based on far field techniques several factors contribute to the successful application of the microscopy. Specific questions affecting the usefulness of the microscopy involve the quality and the magnification of the images generated by the system. Regarding these considerations, it is found that to successfully relate information about the nature of the object, the microscope must form a clear, detailed, image of the system. This must be done on length scales relevant to the level of detail at which the object needs to be understood.

An important aspect of the image formation properties of the lenses in far field systems is, then, the resolution of the features of the image created by the system. To understand in detail the factors affecting the image resolution, it is helpful to treat an object as a collection of points in space. The question of the resolution of the system is then one of how close together two points can be separated in the

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_8

object and still show up as distinct point in the image generated by the microscope. As shall be seen, in the case of the far field microscope this property is essentially limited by the diffractive nature of light.

The object of a lens system can be regarded as formed as a collection of points sources of radiation, but in terms of a far field image of the optical system these points are really spread out in space. That is they are not resolved as points but as localized distributions of intensity. Due to the lens apertures of the system and the nature of the modes of light that can be collected by the microscope, a point sources of radiation is imaged by the system in the form of little diffraction patterns. To understand in more detail how this difference comes about consider the difference between geometric and wave optics.

In the theory of geometric optics, the wave nature of light is ignored. Consequently, under the considerations of geometric optics each point of an object is ultimately projected by the lens system into a point of the image generated by the optics. This is done through the geometric mapping of the object into the image on a point by point basis [11].

In a more precise treatment, however, light travels through the far field system by Huygen's principle as a sum of phase generated processes [11]. This accounts for the diffractive effects of light. It is ultimately the reason why the imagen formed by a point in the object shows up in the image as a little diffraction pattern rather than a perfect point. The transfer of light through the microscope is a process of Fourier transfer which is only approximated by the transfer entailed in geometric optics.

Due to the diffraction effects, whether two object points can be distinguished from one another on the image formed by the microscope depends on the diffraction patterns they create on that image. After being mapped through the system the two points show up as two diffraction patterns which must be distinguished on the image. It is generally found in optical systems that the separation distance needed to distinguish between two points in the image is half of the wavelength of the radiation being used in the imaging. Consequently, this is a natural limitation of the ability of a far field microscope to form a detailed image. This limitation is removed in the consideration of a near field microscope which shall next be discussed.

The technique of near field microscopy, unlike that of far field microscopy, is intimately tied to the study of surface and evanescent electromagnetic waves [1–10]. It is a method in which the most important fields in the measurement process include the evanescent waves near the surface of the object. Typically, the probe forming the collection device of the microscope must be located at a distance from the object surface which is less than a wavelength. This is different from the case of the far field microscope which collects light that has traveled a distance of many wavelengths from the object.

In the following, after a discussion of evanescent waves, the basic ideas of near field microscopy will be introduced. These involve the applications of probes to determine the localized field intensities at a measuring surface. In practical implementations, many variations on this type of light collection system are encountered. In the presentation, a focus is on how near field techniques allow for the increased resolution over that obtained in far field microscopes. A brief review is then presented of some recent experiments based on the near field microscope.

8.1 Evanescent Waves

In order to understand the importance of evanescent waves in the subwavelength imaging of the surface properties, consider the propagation of electromagnetic waves in a vacuum region above a surface which is planar on average [1–5]. (See the schematic is Fig. 8.1a.) The average surface is the x-y plane and separates a region of vacuum above the interface from a region of dielectric media located below the interface. The idea is to collect information about the surface properties from the radiation received from the surface.

The waves in the vacuum region, used in the imaging of the surface properties, have two basic forms. These include waves that propagate away for the surface and waves that decay evanescently away from the surface. The propagating waves are the waves that are studied in far field optics while the evanescent waves generally decay to zero before entering the far field microscope. Both the propagating and evanescent waves carry information about the surface properties, and this accounts for the less than perfect resolution of far field microscopes.

The difference in the propagating and evanescent waves is now discussed. In particular, it is shown that these two types of waves carry different information about the Fourier components of the surface properties. Consequently, to develop a complete understand of the surface both components of information need to be gathered in an effective microscopy of the surface. The near field microscopy techniques used to develop a more complete set of information about the surface are the topic in the remainder of this chapter.

The first form of radiation from the surface is that of a propagating plane wave. This is given by

$$\vec{E}(x,y,z,t) = \vec{E}_0 \exp\left[i\left(k_x x + k_y y + k_z z - \frac{\omega}{c} t\right)\right]. \tag{8.1}$$

Here the wave vector component for propagation away from the surface is characterized by $k_z > 0$ which is written in terms of the frequency and other wave vector components as

$$k_z = \sqrt{\frac{\omega^2}{c^2} - k_x^2 - k_y^2}. \tag{8.2a}$$

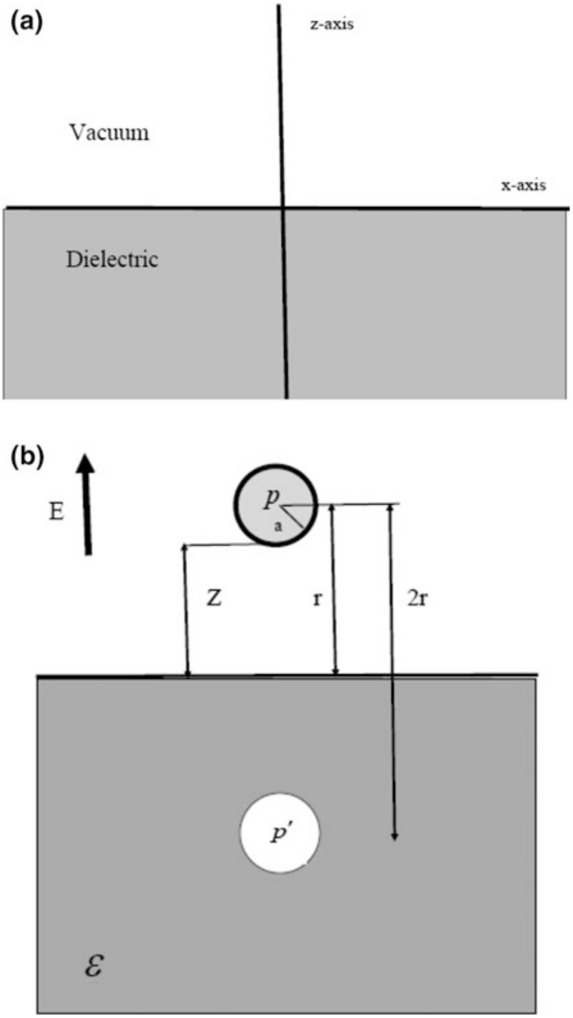


Fig. 8.1 Schematic figures: **a** the interface between the semi-infinite vacuum-dielectric media, **b** the scattering of surface waves from a dielectric sphere in close proximity to a planar vacuum-dielectric interface. In **b** the dielectric sphere above the surface is of radius, a, and has a dipole moment, p. An image of the dipole sphere above the surface is located below the surface and has an image dipole moment p'. Evanescent electromagnetic surface waves on the planar surface are scattered by the dielectric sphere and its image into bulk waves in the two media

A consequence of (8.2a) is that the inequality

$$k_x^2 + k_y^2 < \frac{\omega^2}{c^2} \tag{8.2b}$$

must hold in order to have a real z-component of the wave vector. It is important to note from this that the requirements of a propagating wave are seen to have a fundamental limitation on the x-y Fourier components that can be radiate from the surface. This restriction is posed by the inequality in (8.2b).

To understand this, consider the imaging properties of the waves propagating away from the surface. Specifically of interest is the ability of the propagating waves to care information about the surface and to relate this to the image formed by the microscope. For these considerations, the surface structure properties can be described by functions of the form $f(x, y)$ and their representation by Fourier series. In this format, the surface properties take the general two-dimensional form

$$f(x,y) = \int d^2k\hat{f}\left(k_x, k_y\right)\mathrm{e}^{i\left(k_x x + k_y y\right)}, \tag{8.3}$$

and, for an accurate representation of the property $f(x, y)$, all its Fourier components must be used by the microscope in the creation of the image.

It should be noted, from (8.1) through (8.3), that if the propagating electromagnetic wave is to efficiently carry off information regarding the $\mathrm{e}^{i\left(k_x x + k_y y\right)}$ component in (8.3), the condition

$$k_x^2 + k_y^2 < \frac{\omega^2}{c^2} \tag{8.4}$$

must be satisfied. This is a restriction on the ability of the propagating waves to carry information from the surface in far field microscopy.

In the case that the condition in (8.4) is not satisfied, the surface information remains localized about the interface. As a result, surface features with lengths less than

$$L_s = \frac{2\pi}{\omega/c} = \lambda \tag{8.5}$$

will not be accurately represented in the image generated by a far field microscope. This is a consequence of the fact that the far field microscope only images the information carried to it over distances of many wavelengths by waves propagating from the object surface.

The second type of electromagnetic solution at the surface is the evanescent waves. These are of a form given by

$$\vec{E}(x,y,z,t) = \vec{E}_0 \exp\left[i\left(k_x x + k_y y - \frac{\omega}{c}t\right) - k_z z\right]. \tag{8.6}$$

They represent fields moving along the interface but with components that decay away from the surface. In (8.6) the exponential decay of the fields away from the surface is characterized by a real k_z given by

$$k_z = \sqrt{k_x^2 + k_y^2 - \frac{\omega^2}{c^2}}. \tag{8.7}$$

Consequently, the condition on k_z in (8.6) and (8.7) so that the wave is localized on the interface is that $k_x^2 + k_y^2 > \frac{\omega^2}{c^2}$. If this condition does not hold, (8.6) reverts to the propagating wave solutions in (8.1) and (8.2). While the solution in (8.1) and (8.2) is the focus of far-field microscopy, the second type of solution in (8.6) and (8.7) is the focus of near field microscopy.

The reason for the need of a microscopy based on the evanescent waves in (8.6) and (8.7) is that they are waves containing the short length scale information about the interface. These evanescent waves uniquely contain information which characterizes surface features on lengths, l, satisfying

$$l < L_s = \lambda. \tag{8.8}$$

Schemes of near field microscopy, then, focus on the extraction of the information contained in the evanescent waves localized about the interface as well as the standardly treated long wavelengths of far field optics.

An important way of accessing the information in the surface modes is based on the scattering of evanescent waves by a feature placed on or near the surface to be imaged. This is the basis of near field microscopy.

In this regard, in earlier discussions of plasmonics it was shown that surface plasmon-polaritons were bound electromagnetic modes at the planar interface between two media. There they were shown to propagate along and to be localized at the interface. It was also discussed how a perturbation on the interface could scatter the surface plasmon-polaritons into waves propagating away from the interface. The scattering converts the evanescent surface waves into propagating waves which move away from the interface. These propagating waves, created from the scattering of evanescent waves, can carry the information contained within the evanescent waves so as to be received in the far field.

These type of studies of the scattering of surface waves will now be revisited as an introduction to techniques of near field optics. They provide a basis of the scanning probe technology upon which near field microscopy is based, the details of which will be addressed later.

8.2 Model of a Surface Probe

In near field microscopy, the idea is to place some type of probe or scattering feature within a wavelength of the surface. The probe or feature causes a transition of the evanescent fields on the surface into scattered propagating fields which are then picked up in the far field. There are a variety of different arrangement that can be made to accomplish this scattering and to collect the resulting propagating fields.

As the probe or feature is scanned across the surface the variation in the collected scattering must be interpreted to arrive at an understanding of the nature of the surface properties generating the collected fields.

A simple model of this process can be given in terms of a sphere of dielectric placed above a surface. The dielectric sphere is intended to model a probe or scattering feature which is meant to transform the information in the evanescent fields to propagating modes in the far field. This model will now be discussed in terms of it scattering properties.

In Fig. 8.1b a schematic of the scattering problem is considered [3]. The planar interface is between vacuum (above the interface) and a dielectric medium of dielectric constant ε (below the interface). Above the interface in the region of vacuum is a dielectric sphere of radius, a, and polarizability, α. The center of the sphere is located a distance $r > a$ from the surface so that $z = r - a$ is the distance from the bottom of the sphere to the surface. In the following considerations, the scattering interaction in the system of an electric field E applied normal to the interface is treated in the quasi-static approximation [3].

The problem is solved based on the method of images. The applied electric field, in its leading order interaction, induces a polarization in the spherical dielectric which is given by [3]

$$p = \alpha E. \tag{8.9}$$

The induced dipole in turn generates a dipole field given at the nearest point of the adjacent surface by the expression

$$E_{dipole}(r) = \frac{p}{2\pi r^3}. \tag{8.10}$$

In the quasi-static limit an image dipole is induced by the dipole fields from the generated dipole in (8.9). The form of the image dipole is given in terms of the induced dipole in the vacuum by [3]

$$p' = \frac{\varepsilon - 1}{\varepsilon + 1} p = \beta p \tag{8.11}$$

and is located with its center at a perpendicular distance r below the surface. In the formation of the imaging dipole, the vectors of the dipole above the surface and its image below the surface are opposite to one another. Furthermore, the image is chosen so that the fields of the dipoles above the surface and the semi-infinite media below the interface are given as the sum of the fields from the dipole above the surface and its image below the surface.

This process of induction can be continued on as the fields from the image dipole contribute to the dipole moment induce in the sphere located in vacuum, and the subsequently affected moment of the sphere in vacuum changes the fields at the image. A whole series of perturbations generated in this way then needs to be summed and taken into account. In this manner, it is found that the total induce moment of the sphere in vacuum becomes [3]

$$p = \frac{\alpha}{1 - \frac{\alpha\beta}{16\pi(z+a)^3}} E. \tag{8.12}$$

An electric field is found to arise from the induced dipole above the surface and its imagine in the dielectric medium below the surface. It is found that the effective fields from these two dipole sources can be represented by a field arising from an effective dipole located on the interface. In this way, the sum of the two dipoles represented by spheres in Fig. 8.1b is, to leading order, an effective dipole at the surface given by

$$p^{eff} = \alpha^{eff} E \tag{8.13}$$

where the effective polarizability is given by [3]

$$\alpha^{eff} = \frac{\alpha(1+\beta)}{1 - \frac{\alpha\beta}{16\pi(z+a)^3}}. \tag{8.14}$$

The effective polarization, p^{eff}, generated as a response to the applied electric field, E, is the origin of the far field dipole radiation entering the system in the form of bulk propagating scattered fields. In this generation, the electric field E driving the effective polarization can be composed of both incident propagating fields and surface evanescent waves. As a result, the effects from evanescent waves are only part of the source of the generated fields.

In an ideal microscopy, however, the evanescent waves would be a predominant source of the effective polarization. Consequently, a number of techniques have been introduced to enhance the components of the scattered field that originate in the surface evanescent waves. This is to improve the performance of the near field microscope, separating out from it effects not related to the evanescent waves. These techniques will be mentioned later.

Some additional considerations regarding the coupling between surface and the incident fields are also needed in understanding the application of the above results. Two types of scattering processes occur as the incident wave encounters the surface and dielectric sphere. In a first type of scattering the incident fields of the light couple to a larger region of the surface surrounding the sub-region containing the induced dipole. Consequently, a component of the scattering is from the dielectric mismatch at the interface along with surface structure that is present at the interface.

A second scattering component is related to the smaller region of the effective dipole composed from the sphere in vacuum and the image of the induced dipole. A number of techniques have been developed to make a separation of these two different contributions in the generation of bulk waves leaving the interface.

This lets the features with small length scales to become easier to identify from those features arising from the scattering of propagating waves.

In the development of a near field microscope an effective design must be employed which emphasized the processes involving evanescent waves in the generation of the bulk waves scattered from the surface. This allows for the identification of the subwavelength features of the surface. In addition, a focus must be on these interactions at the position of the surface probe. The variety of other phenomena mention in the earlier paragraphs must be minimized for an effective result in the determination of surface properties [3].

8.3 First Proposal by Synge

The earliest proposed system of near field microscopy, based on some of the features of the scattering problem in Fig. 8.1, was put forth by Synge in 1928 [1–5, 12, 13]. The idea of Synge was to make a system for the measurement of surface properties that involved the surface to be measure, a light source, and an opaque screen pierced by a subwavelength hole. A schematic of the measuring configuration is shown in Fig. 8.2.

During the measurement process the screen is kept at a constant separation from the mean surface and is sequentially moved over the plane of the surface in order to make a series of measurements. A fundamental requirement of these

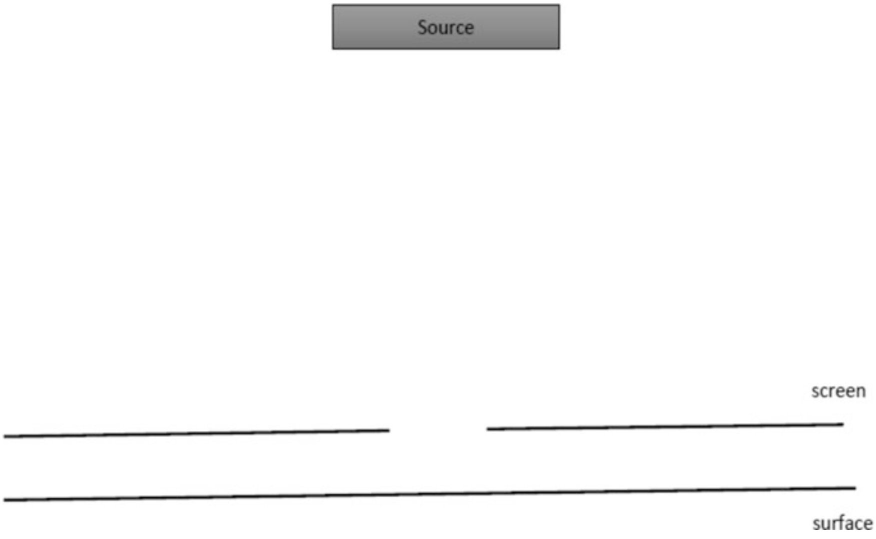


Fig. 8.2 Synge's method: The screen and the surface are separated by a distance which is less than a wavelength of the radiation emitted by the source, and the aperture in the screen has a subwavelength diameter

measurements is that the separation between the screen and the mean surface must be less than a wavelength of the measuring light originating in the light source. This is needed so that evanescent waves emitted from the subwavelength hole in the screen can be efficiently coupled into the surface interaction required for the microscopy of the surface. It represented a major problem in the implementation of Synge's ideas [12, 13].

The smaller the separation between the screen and surface, the greater is the concentration of evanescent waves in the region between the screen and surface. As shall be discussed, this enhancement of evanescent waves is a fundamental basis effecting the quality of the near field microscopy. The development of means for the subwavelength positioning of the screen and surface is, in fact, the primary reason near field microscope took fifty years from its initial proposal to develop.

In the course of the sequential processes, a measurement is first made at one positioning of the screen. Following this the screen is moved to another position over the mean surface at which another measurement is performed. Results from each of these measurements is recorded and correlated with the positions on the mean surface at which each of the measurements was made. This generates a map of recorded measurement made over the plane of the mean surface.

For each of the measurement, light is incident on the screen from a source at the top of the schematic figure. Upon encountering the screen most of the light is reflected and absorbed, but a small portion is passed by the screen and sent on to the surface of the sample. This component of light is composed of a mixture of waves that are propagating waves and/or evanescent waves that leave the screen to interact with the surface being measured.

The light passed by the screen interacts with the surface, and the light resulting from this interaction is ultimately collected by a device which records its intensity. The measured intensity creates what is essentially a pixel of light with an intensity determined by the properties of the surface being measured. The surface properties recorded through their effect on the light detected include: the surface profile, the physical and chemical properties of the materials forming the surface, and the properties of the surface plasmon-polariton modes at the surface.

Following an intensity measurement of the pixel the screen is moved parallel to the mean surface and another pixel measurement is made. This is done throughout the entire surface and the pixel intensities are recorded, forming an intensity map over the region of the sample that is being investigated. It then remains to interpret the map of pixel intensities and correlate these with the physical property of the surface that is of interest.

In the Synge approach the conversion of the evanescent waves passing the subwavelength aperture to bulk far field modes provides the subwavelength resolution of the microscopy. These evanescent waves contain the small wavelength components need for the accurate image the surface features. However, as with the considerations of the model in Fig. 8.1b, the signal received in the formation of the pixels has origins in both evanescent and propagating components [1–5, 12].

As noted earlier, the emphasis of the microscope should be on intensities with origins in the evanescent components. These contain the measurement of the small wavelengths features and must somehow be separated from those of the longer wavelengths which are a type of noise in the microscopy.

Another important feature of the outlined technique is that the resolution of the surface features is governed by the size of the aperture in the screen. In this regard, the ability to make apertures which can be much less than the half-wavelength limitation of far field microscopy is very important. In the near field system diffraction is not so much a problem as that the operation is based on the presence of evanescent waves, and the resolution of the near field microscope is essentially the diameter of the subwavelength aperture it is based upon.

8.4 Subsequent Realizations

In the realizations of the near field technique made since its proposal, the movable screen is often replaced by a scattering tip or a tapered fiber optics waveguide with a metallic coating applied over its length. (See Fig. 8.3 for a schematic of these components.) The idea is to pass the scattering tip or tapered waveguide over the surface as a probe of the surface, and the difference between the probes is that the tip is a scattering site whereas the tapered guide is a light pipe which emits or collects radiation at its tip. Similar to the screen with an aperture, the probes are positioned to be within less than a wavelength of light from the surface being imaged. The readings from the probes are then make sequentially to form an intensity map over the surface [1–10, 12].

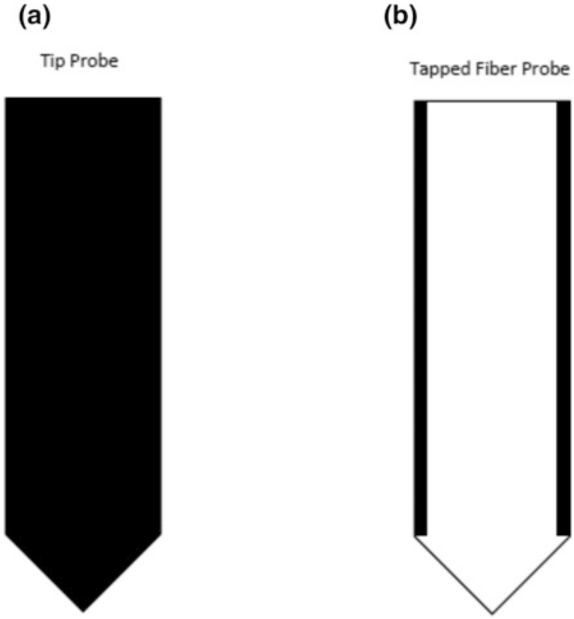


Fig. 8.3 Two types of probes. In **a** the probe is a scattering tip and in **b** the probe is a tapered fiber waveguide with a metal sheath. In some cases fiber waveguide probes are made without the metal cladding

In the case of the scattering tip an arrangement similar to that of an atomic force microscope is often made in guiding the tip over the surface. Specifically, an atomic force microscope is designed to measure the profile of the surface by using an electronic feedback arrangement. This mechanism maintains a constant separation distance between the probe and the surface being measured. In the near field microscope, the same feedback mechanism can be used in scanning the microscope tip of the near field system so that it maintains a fix height from the surface. The light scattered by the tip is then collected in the formation of the pixel profile of the surface properties [1–5].

For the measurements made with the scattering tip, light can be incident on the surface-tip system from above or below the surface. The collection of the scattered light can also be made from above or below the surface. No matter the configuration chosen, as with the original proposal made by Synge, the resolution of the tapered waveguide tips is fixed by the aperture of the tip. Such types of system have been used in a variety of near field arrangements and systems and have achieved resolutions even at the level of molecular scales [1–5, 10].

As another example, a collection device is based on waveguides. In one waveguide configuration, the tapered waveguide is a source of radiation passed within a wavelength of the surface. This is the case employing a waveguide without the metal cladding. In this mode of operation, the tapered waveguide replaces the subwavelength hole within the screen in the original proposal of Synge. The pixels are then formed by this method in the same way as in Synge's screen system. The system, however, can also be reversed and operated so that light form the surface is collected by the tapered waveguide. This collected light is then used to make a pixel and the resolution of the microscope is again set by the subwavelength aperture.

In the case of the tapered waveguide with the metal coating, the waveguide can act either as a source or collector of the scattered light in the system. In some cases, however, it can be both the source and collector of the light interacting with the surface. In all the examples considered, the emission and collection of the tapered waveguide is made at the tip of the waveguide which is left without a metal covering. Only the cylindrical sides of the waveguide fiber are cover with metal.

Another method proposed by Synge that can be of application is a near field microscopy developed by placing a dielectric particle on the surface. In this method, the particle would act as a scattering site for the formation of a pixel. By moving the particle, a pixel map of the surface is then formed. The physics operating here is very similar to the scattering problem treated in Fig. 8.1b.

As a final important note about the near field techniques, it should be pointed out that they can also be applied to a variety of spectroscopic methodologies. These include the inelastic events arising in various Raman and fluorescent spectroscopies. The near field application enables for the local study of structures which emit signals based on the inelastic transitions arising from these types of spectroscopic transitions. Such features can then be mapped over the surface of the sample just as the structure and compositional features of the surface are revealed by the elastic counterpart of near field microscopy [1–10].

8.5 An Experimental Result

As an example of some experimental results, in Fig. 8.4 results are shown from [1] for the imaging of a grating using atomic force microscopy and three variations of the near field optical techniques.

In Fig. 8.4a, b the atomic force microscopy of the grating surface is shown. Specifically, Fig. 8.4b shows a plot of the profile height of the surfaces as a function of distance along the line indicated in white in Fig. 8.4a. The grating was etched in a glass surface and was formed with a period of 383 nm and a step height of 8 nm.

A scanning near field microscope results for the imaging of the grating are presented in Fig. 8.4c [1]. In this figure the imaging is made with a probe configuration at a constant height of 1 nm. Similar results are presented in Fig. 8.4d, e, shown for constant probe height at a few nm and at 100 nm, respectively. From these figures, it is found that the resolution of the scanning near field microscope is strongly dependent on the separation of the probe from the surface. The closer the probe is positioned to the surface, the better is the image formed [1].

An example from biology of the interesting application of near field microscopy is in recent development of fluorescence microscopes for the study of biological materials [14]. These rely on the fluorescent response of biological materials and/or stains applied to these materials to develop the images of subwavelength features in these materials. In Fig. 8.5 are presented some microscopic studies of human cells and those of other organisms studied by these means. The focus of the study was to apply various processing techniques to enhance the resolution of images obtained by near-field techniques. For the details the reader is referred to the original literature [14].

Further results are available in the literature for both physical [1–8, 10] and biological samples [1–5, 9]. These include microscopy result based on the elastic scattering of light as well as applications of florescent and Raman spectroscopy effects in these systems.

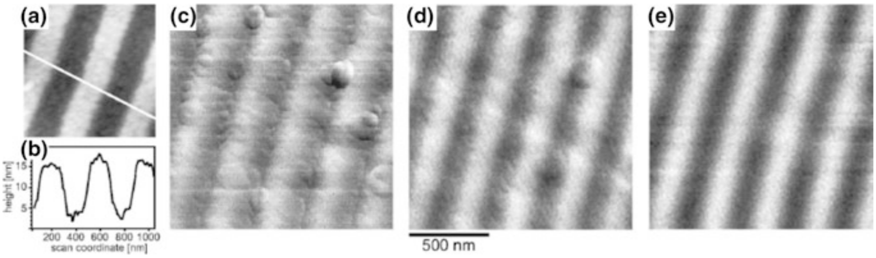


Fig. 8.4 Scanning near field microscope results compared to results from a study of atomic force microscopy of a grating etched on a glass surface. The results in: **a** and **b** are for the atomic force microscope. The results in **c**–**e** are from a near field microscope with heights of 1 nm, several nm, and 100 nm, respectively. The results are taken from [1]. Reproduced from [1] with permission of AIP Publishing

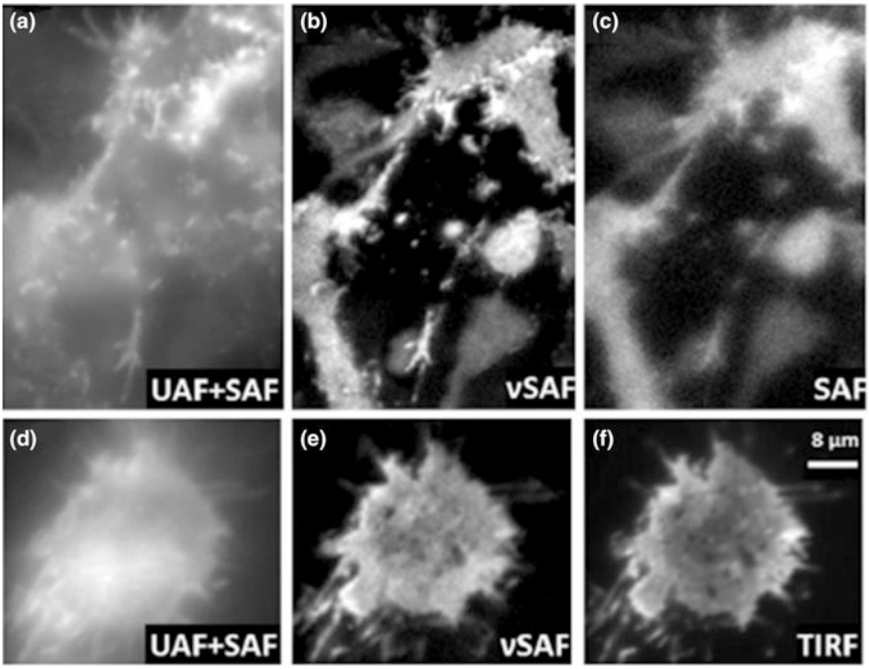


Fig. 8.5 Scanning near field fluorescence microscope results of biological cells. A variety of imaging processing techniques have been used on these samples of human embryo kidney cells and cells from mice. The focus was to develop a better imaging. For the details of these results the reader is referred to the original paper. Reprinted with permission from [14]. Copyright 2012 American Physical Society

As a final point it should be noted that subwavelength resolution has also been achieved using techniques based on photonic crystals [15] and metamaterials [16, 17]. Some of these techniques have been discussed in earlier chapters. In addition, the techniques of near field microscopy have recently been extended to acoustic microscopy where they have achieved resolutions on the scale of nanometers [18].

References

1. B. Hecht, B. Sick, U.P. Wild, V. Deckert, R. Zenobi, O.J.F. Martin, D.W. Pohl, Scanning near-field optical microsopy with aperture probes: fundamentals and applications. J. Chem. Phys. **112**, 7761–7774 (2000)
2. A. Dereux, C. Girard, J.-C. Weeber, Theoretical principles of near-field optical microscopies and spectroscopies. J. Chem. Phys. **112**, 7775–7789 (2000)
3. R. Hillenbrand, B. Knoll, F. Keilmann, Pure optical contrast in scattering-type scanning near-field microscopy. J. Microsc. **202**, 77–83 (2001)

4. M.A. Paesler, P.J. Meyer, *Near-Field Optics: Theory, Instrumentation, and Applications* (Wiley, 1996)
5. D. Courjon, *Near-Field Microscopy and Near-Field Optics* (Imperial College Press, 2003)
6. E. Betzig, J.K. Trautman, Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. Science **257**, 189–195 (1992)
7. E. Betzig, P.L. Finn, J.S. Weiner, Combined shear force and near-field scanning optical microscopy. Appl. Phys. Lett. **60**, 2484–2486 (1992)
8. E. Betzig, J.K. Trautman, T.D. Harris, J.S. Weiner, R.L. Kostelak, Breaking the diffraction barrier: optical microscopy of a nanometric scale. Science **251**, 1468–1470 (1991)
9. F. de Lange et al., Cell biology beyond the diffraction limit: near-field scanning optical microscopy. J. Cell Sci. **114**, 4153–4160 (2001)
10. E. Betzig, R.J. Chicester, Single molecules observed by near-field scanning optical microscopy. Science **262**, 1422–1425 (1993)
11. G.R. Fowles, *Introduction to Modern Optics* (Dover Books on Physics, Holt Rienhart and Weston, New York, 1975)
12. E.H. Synge, Suggested method for extending microscopic resolution into the ultra-microscopic region. Phil. Mag. **6**, 356–362 (1928)
13. E.H. Synge, An application of piezo-electricity to microscopy. Phil. Mag. **13**, 297–300 (1932)
14. T. Barroca, K. Balaa, S. Leveque-Fort, E. Fort, Full-field near-field optical microscope for cell imaging. Phys. Rev. Lett. **108**, 218101 (2012)
15. E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopoulou, C.M. Soukoulis, Subwavelength resolution in a two-dimensional photonic crystal-based superlens. Phys. Rev. Lett. **91**, 207401 (2003)
16. D. Lu, Z. Lui, Hyperlenses and metalenses for far-field super-resolution imaging. Nat. Commun. (2012). https://doi.org/10.1038/ncomms2176
17. S. Xu, Y. Jiang, H. Xu, J. Wang, S. Lin, H. Chen, B. Zhang, Realization of deep subwavelength resolution with singular media. Sci. Rep. **4**, 5212 (2014)
18. P. Xu, W. Cai, R.M. Wang, Scanning near-field acoustic microscope and its application. Sci. China Technol. Sci. **54**, 126–130 (2011)

Chapter 9
Nonlinear Optics

In this chapter some of the basic ideas of nonlinear optics are presented as they apply to nanoscience [1–24]. The presentation is aimed at giving an introduction to a few important phenomena that are found to be useful to technology. In this regard, a particular focus is on photonic crystal applications as these systems have been most addressed in the recent literature. Nevertheless, the topic is much broader than this, and, as an aid to the interested reader, some applications of nonlinearity in other areas of nanoscience are briefly mentioned.

Nonlinear optics covers a great variety of topics and is responsible for the explanation of a large number of different optical phenomena [25–28]. There are, however, two general types of phenomena that have drawn most of the recent attention in the study of nano-systems. These are phenomena arising from Kerr nonlinearity and phenomena arising from the generation of second harmonics of radiation.

Kerr nonlinearity is a property, found in many nonlinear optical materials, in which the material exhibits an index of refraction that is dependent on the intensity of the electric field applied to it [1, 25–28]. The effect is small and usually requires the application of intense light such as that generated in lasers for its observation. Depending on the particular medium under consideration, increasing the intensity of the electric field applied to a material can either increase or decrease its refractive index.

By changing the index of refraction through the application of an external field intensity, Kerr nonlinearity has been a focus in the design of various switching devices. The idea is based on configuring an optical device in which small changes of the refractive index of a component can redirect or turn on and off a particular response of the system. Proposals of various optical diodes and transistors have been made based on this type of switching effect.

Another effect of Kerr nonlinear systems is that they can support soliton-like excitation modes [1, 28]. These are new types of modes that exist solely due to the nonlinearity of the system, and they are not present in linear systems. The basic type of solitons are bright solitons, dark solitons, and grey solitons.

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_9

A soliton is formed in a system of nonlinear media as an excitation having a field intensity pattern that causes a particular change in the index of refraction of the media [1, 28]. The change of the index of refraction is made in such a way as to self-consistently support the field pattern of the soliton excitation. In this scheme, then, the field modifies the index of refraction which in turn leads to a variation of the index of refraction of the system supporting the modifying field.

A consequence of this is that a soliton intensity pulse can be supported in a nonlinear system. This type of pulse excitation is known as a bright soliton. Similarly, under the proper conditions an intensity dip can be supported in a nonlinear system. This type of soliton excitation then travels in the system as a pulse of decrease in field intensity. If the dip gives a drop in intensity going to zero, the soliton is a dark soliton. In the case that the dip is a drop in intensity going to a non-zero intensity minimum, the soliton is a grey soliton [1, 28].

These types of intensity pulses and dips can have a variety of technological applications. They have been a focus of recent considerations in the context of fiber optics and in photonic crystal wave guides.

The second phenomenon of nonlinear optics to be treated here is second harmonic generation of light [1–28]. This is a nonlinear phenomenon in which a fundamental of light inputted into the nonlinear media has an outputted component generated at twice the frequency of the fundamental. It has a number of technological applications and a number of technological problems to be overcome in its generation.

As an elementary focus on these two types of nonlinearity, they shall both be discussed in the context of one-dimensional optical systems. This allows for a simple analytic treatment which illustrates the basic theory of the systems being considered. Higher dimensional systems usually required computer simulation methods which give a poorer illustration of the physics involved. A brief review will be made of some of the literature regarding higher dimensional systems treated by simulation methods.

In the following, first a treatment will be given of a one-dimensional (layered) photonic crystal composed of Kerr nonlinear optical media. A theory of the properties of the excitations in the system will be discussed with a particular focus on the new soliton modes. This will be followed by considerations of the generation of second harmonic of radiation within a nonlinear media and some indications of how photonic crystal and other nanoscience systems can be used as an aid in the efficient generation of second harmonics of radiation.

9.1 Photonic Crystal Composed of Kerr Media

In the following section discussions are presented of the physical properties of photonic crystals formed of Kerr nonlinear media. This is done using one-dimensional models of photonic crystals in which the system is formed as a periodic array of dielectric layers. The one-dimensional models are simple enough that they can be treated analytically. At the same time, they exhibit many of the important properties of photonic crystals in higher dimensions. In addition, they are often of interest from a technological standpoint in the design of optical diodes, transistors, and optical coatings.

As in the case of one-dimensional photonic crystals composed of linear media (see Chap. 2), the solutions of the electrodynamics of one-dimensional photonic crystals composed of Kerr nonlinear media can be written in terms of the individual solutions within each of the slabs of the layering. In this process, the individual solutions obtained in each of the layers are matched to one another by applying appropriate electromagnetic boundary conditions at the interfaces between adjacent slabs. The total solution for the electrodynamics of the photonic crystal is then given by the piecewise matching of these individual slab solutions [1, 28–33].

Applying this procedure along the complete layering generates a system of algebraic equations relating the field amplitude coefficients between adjacent interfaces of the dielectric slabs to one another. The resulting set of difference equations is often easily solved to obtain a variety of scattering and standing wave solutions for the electrodynamics of the system. In many systems the relation between the coefficients of adjacent interfaces are expressed as matrix equations so that, consequently, the procedure is often referred to as the transfer matrix method.

In the following discussions the transfer matrix method will be used to study the electrodynamics of one-dimensional photonic crystals of Kerr nonlinear media for electromagnetic waves that are propagating in the system incident normal to the slab surfaces [29, 30, 32, 34]. Due to the dependence of the Kerr index of refraction on the intensity of the electric field, the electromagnetic dispersion relation and scattering interactions in the layered system are found to be dependent on the amplitude of the electromagnetic waves. This, however, does not affect the energy conservation of the waves as it is easily shown that the energy is conserved under refractive interactions at the interfaces of Kerr media and other Kerr or linear media.

The transfer matrix method arises from a consideration of a single slab or layer of dielectric medium. Consequently, in the following, single slab treatments will be given which are appropriate for both linear media or Kerr media slabs [1, 29, 30, 32–34]. For these discussions, the slab surfaces are taken to be in the x-y plane and the direction of normal incidence on the slab surfaces is the z-axis. (See Fig. 9.1 for a schematic of the slab surrounded by vacuum.) The electric and magnetic field vectors of the wave solutions are taken to be polarized in the x-y plane. To begin the discussions, consider a single slab with surfaces located at $z = z_m$ and $z = z_{m+1}$. The matrix is considered to have an unspecified index of refraction which is allowed to be position dependent.

Within the slab the electric field amplitudes obey a Helmholtz equation of the form [1, 29, 30, 34]

$$\frac{d^2E_m}{dz^2} + \frac{\omega^2}{c^2}n^2(z)E_m = 0 \tag{9.1}$$

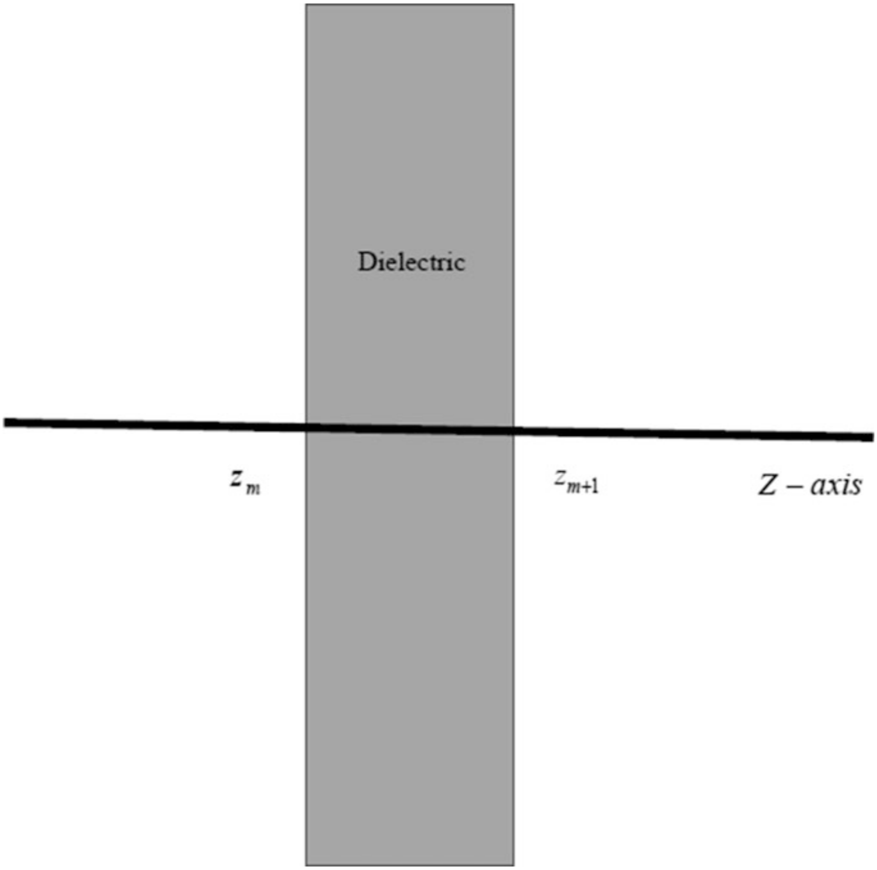


Fig. 9.1 A single dielectric slab with surfaces at $z = z_m$ and $z = z_{m+1}$

where $n(z)$ is the index of refraction and ω is the mode frequency. The Helmholtz equation is a second order differential equation so that it has a solution with the general form

$$E_m = E_m(z, A_{m1}, A_{m2}) \tag{9.2}$$

where A_{m1} and A_{m2} are two integration constant which are to be fixed by the boundary conditions.

As an example of (9.1) and (9.2), consider the case in which the slab is composed of linear medium of constant refractive index n. In this limit the field within the slab is given by the form [1]

$$E_m = A_{m,1}\mathrm{e}^{ikz} + A_{m,2}\mathrm{e}^{-ikz}, \tag{9.3}$$

where $k = n\frac{\omega}{c}$. Similarly, in the vacuum to the left of the slab the fields are of the form

$$E_{m-1} = A_{m-1,1}\mathrm{e}^{ik_0 z} + A_{m-1,2}\mathrm{e}^{-ik_0 z} \tag{9.4}$$

where $k_0 = \frac{\omega}{c}$, and in the region of vacuum to the right of the slab the field is given by

$$E_{m+1} = A_{m+1,1}\mathrm{e}^{ik_0 z} + A_{m+1,2}\mathrm{e}^{-ik_0 z}. \tag{9.5}$$

Each of these solutions exhibit the general form of (9.2).

Now consider the case of the linear medium slab with solutions given in (9.3)–(9.5) and determine how the amplitude coefficients in the two vacuum regions are related to one another. The boundary conditions at the interface between the slab and each of the vacuum regions are that the electric field and the z-derivative of the electric field are continuous at the interface. Matching the solutions in both vacuum regions with the solutions within the slab, the coefficients in the vacuum to the left of the slab can be related to those in the vacuum on the right of the slab. In this way a matrix equation form [18] is generated.

The resulting matrix relation can be written as [1]

$$\begin{vmatrix} M(m)_{11} & M(m)_{12} \\ M(m)_{21} & M(m)_{22} \end{vmatrix} \begin{vmatrix} A_{m-1,1} \\ A_{m-1,2} \end{vmatrix} = \begin{vmatrix} A_{m+1,1} \\ A_{m+1,2} \end{vmatrix} \tag{9.6}$$

where the matrix elements $M(m)_{i,j}$ for $i,j = 1,2$ are association with the slab labeled m and are written in terms of the dielectric parameters and the positions of the two surfaces of the slab [11, 12, 18, 35, 36]. For the linear dielectric medium the $M_{i,j}$ are expressed in terms of algebras involving plane wave forms.

In the case of the general system in (9.1) and (9.2) the problem is a little more complicated. Now the coefficients $\{A_{m-1,1}, A_{m-1,2}\}$ are related to $\{A_{m+1,1}, A_{m+1,2}\}$ by two relationships obtained from the boundary conditions. Given these two relationships, however, a complete solution of the electrodynamics of many interesting problems can be easily written down. The procedure is essentially the same as that used in the treatment of systems based on (9.6).

As an important example of such a transfer matrix problem, consider the scattering of an electromagnetic plane wave incident on a finite photonic crystal layering of slabs. In this example, the solution for a plane wave at normal incidence from the left on a finite array of slabs N is studied. The object of the discussions is to obtain the scattered and reflected wave components of the system.

To generate the solution start at the right hand edge of the layering, considering the slab on the far right of the array. In the region on the right of this slab there is only a transmitted wave. This transmitted wave is given by the form [1]

$$E_{m+1} = t\mathrm{e}^{ik_0 z}. \tag{9.7}$$

where t is the transmission amplitude of the wave. In particular, it is seen from (9.7) that in the region of vacuum to the right of the array the field only propagates away from the layers and to the right.

Given the form of the scattered solution in (9.7) it only remains to apply the transfer matrix method to obtain the other scattering amplitudes along the chain. In this process one moves towards the left on the array of slabs, relating the sets of amplitudes $\left\{A_{m-1,1}, A_{m-1,2}\right\}$, $\left\{A_{m+1,1}, A_{m+1,2}\right\}$ to each other and to the amplitude t. In this way, eventually all of the coefficients in the layering can be given in terms of t.

At the far left of the array, to the left of the left most slab in the finite layering, the vacuum fields are of the form [1]

$$E_{m-1} = i\mathrm{e}^{ik_0 z} + r\mathrm{e}^{-ik_0 z} \tag{9.8}$$

where i and r are the amplitudes of the incident and reflected waves, respectively. Again, applying the boundary conditions expresses i and r in terms of t.

From (9.7) and (9.8) and field amplitudes the reflection coefficient is given by $R = \left|\frac{r}{i}\right|^2$ and the transmission coefficient is given by $T = \left|\frac{t}{i}\right|^2$. As an expression of energy conservation, it can be generally shown that for a real index of refraction that $R + T = 1$. This is true for both linear and Kerr nonlinear media and the processes outline above are essentially the same for both linear and nonlinear models.

A variety of problems have been handled using one-dimensional models of linear and nonlinear media based on the discussions outlined above [12, 18]. Not only do many of the solutions have practical applications, but often they have exact solutions illustrating physical principles found in the qualitative properties of much more complex, higher dimensional systems.

Examples of such nonlinear properties illustrated by layered media models include: (a) band structures [1, 25–28] with renormalized band structures and gap soliton modes, (b) bound state impurity problems at stop band frequencies [1, 27], reminiscent of those observed in semi-conductor electronics [1, 25–28], displaying impurity energy levels dependent on the intensity of the fields at the impurity site, (c) optical properties which exhibit bistability behaviors, e.g., bistability of some of the transmission and dispersion properties of the modes of the system [1, 25–28], and (d) various types of properties related to disorders in the one-dimensional array [1]. In addition, Kerr nonlinearity in one-dimensional systems has also, in its own right, been of technological application in the development of optical switches [1, 25–28] and optical diodes [28] for opto-electric circuits. For the details of these treatments the reader is referred to the literature.

In the following the transfer matrix techniques outline above will be applied to problems involving finite layered media. In particular, some discussion of the transmission and reflection properties of finite layers and coatings at dielectric interfaces and mirror surfaces will be presented. These provide important illustrations of the general features of photonic crystals involving nonlinear dielectric media, including the dispersive properties of the electromagnetic excitations and the nature of soliton solutions. In addition, many of the properties of infinite photonic crystals are seen to be approximated by these finite structures which are in their own right of significant technological interest.

9.1.1 Model of Finite Kerr Nonlinear Layers: Scattering Properties

The problem of the scattering of a normal incident plane wave from a finite layering of a one-dimensional photonic crystal composed of Kerr nonlinear media illustrates many of the basic feature of general photonic crystals formed of Kerr nonlinear media [1, 29, 30, 34]. These include the basic properties of the band structure of the system and of the different classes of new types of soliton modes present in nonlinear media.

In this regard, it is often found that even a system with a relatively small number of layers displays the essential features of an infinite photonic crystal and, consequently, provides a good study of these properties. Aside from the illustrative use of one-dimensional models, they also have potential as models of technologically important finite systems, i.e., the properties of small number of layerings can in their own right have important technological applications in the design of coatings at the interface of two different optical media.

In the following, finite layerings of photonic crystal will be investigated in the context of coatings. First a treatment of a coating that transmits the incident radiation from a region on one side of the coating to a transmitted wave in a propagating medium on the other side of the coating will be treated. Following this a second type of coating which is placed on a perfect conducting mirror will be treated. In both systems a focus will be on the wave functions of soliton-like excitations within the coating materials and the conditions on the Kerr media that are required for the soliton modes to exist.

In the first model, a finite one-dimensional photonic crystal is composed as a system of five identical Kerr medium slabs that are separated by four vacuum slabs. Each of the separating vacuum slabs is of the same width as one of the Kerr slabs, and the entire finite array is surrounded by vacuum. (An illustration of this geometry is shown in the schematic diagram in Fig. 9.2a.)

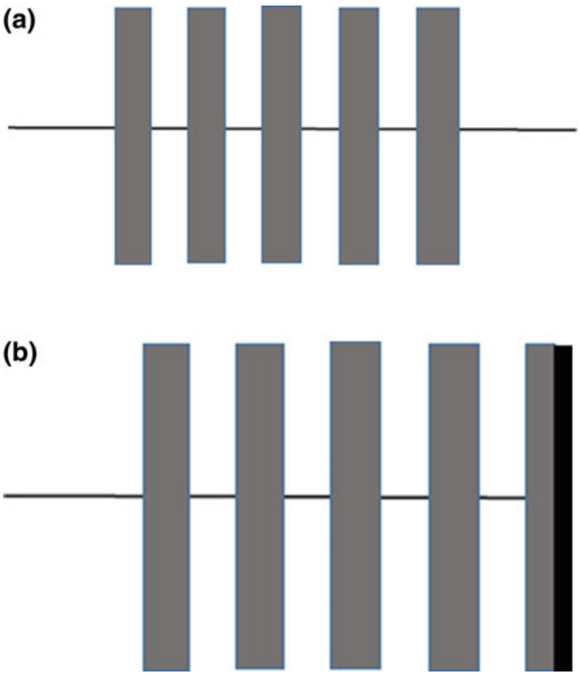


Fig. 9.2 Schematic drawing of a coating consisting of five Kerr slabs each of thickness *d* separated by four vacuum layers each of thickness *d*. Two coating models are illustrated: **a** a free standing coating surrounded by vacuum and **b** a coating applied on its right to a perfect conducting mirror and interfaced to the left of the coating with vacuum. The horizontal line in both **a** and **b** is the z-axis and the slab surfaces are parallel to the x-y plane

In terms of the slab thickness, *d*, the slabs of Kerr nonlinear medium are located in the regions [34]

$$(2m-1)d<z<2md \tag{9.9}$$

for $m = 1, 2, 3, 4, 5$. The vacuum slabs separating the Kerr slabs are located in the regions

$$2md<z<(2m+1)d \tag{9.10a}$$

for $m = 0, 1, 2, 3, 4$. Outside the finite layering, within the regions $z < d$ and, $z > 10d$, are two semi-infinite regions of vacuum. The layering is infinite and translationally invariant in the x-y plane.

In the second model, the layering of five Kerr medium slabs separated by vacuum slabs is interfaced on the right with a perfect conducting mirror. The right most Kerr slab is chosen to be half the width of the other Kerr slabs of the coating and shares and interface with the perfect conducting mirror. This arrangement is shown schematically in Fig. 9.2b and mathematically the positions of the Kerr medium slabs are given by [34]

$$9d < z < 9.5d, \tag{9.10b}$$

$$(2m-1)d < z < 2md \tag{9.10c}$$

for $m = 1, 2, 3, 4$, with the vacuum slabs located in the regions

$$2md < z < (2m+1)d \tag{9.10d}$$

for $m = 1, 2, 3, 4$. The region $z < d$ is vacuum and the perfect conducting mirror is placed at $z = 9.5d$.

As seen from the above geometries, the first model has a full scattering solution with incident, reflected, and transmitted waves. The second model, however, has only incident and reflected waves. While the transmission and reflection coefficients and the associated wave functions excited within the barrier are a focus of the first model, the discussions of the second model will focus on the wave functions excited within the coating media and how these correlate with the band structure of the photonic crystal coating.

In the scattering problems now considered for both systems, a plane wave of light at infinity is normal incident from the left of the coatings. In the first model, transmission anomalies are found to be associated with the resonant excitation of soliton modes within the coatings at stop band frequencies. These types of modes are referred to as gap soliton modes. In the second model, resonantly excited gap soliton modes can be excited by the incident fields within the coating medium. These modes show up physically as intense field enhancements associated with bright solitons. Such field enhancements can have important technological applications [34].

The electric fields in both models are waves propagating along the z-axis at normal incidence to the slab interfaces. They are solutions of the Helmholtz equations for propagation in vacuum and in the Kerr medium. These are solved and matched together by boundary conditions at the interfaces of the two different types of media.

The electromagnetic plane waves in the regions of vacuum have a dispersion relation $k_0 = \omega/c$, with electric fields that are solutions of the vacuum Helmholtz equation given by [1, 29, 30, 34]

$$\frac{d^2E}{dz^2} + k_0^2 E = 0. \tag{9.11}$$

The solutions of (9.11) are then expressed in standard from as a linear combination of $\sin(k_0 z)$ and $\cos(k_0 z)$ functions.

In the region of Kerr dielectric, the index of refraction is dependent on the intensity of the electric field. This introduces the problem of determining a correct form for modeling the field dependence of the nonlinear refractive media. Such models can be complex, involving index of refraction tensors and tensor relations between the field components. Nevertheless, in a general treatment a scalar index of refraction, involving a simple second order relationship in the field amplitude, often suffices to provide a semi-quantitative understand of Kerr media systems.

The nature of the field dependence of Kerr media has been studied in many experimental systems and has been explained by theoretical considerations based on first principles treatments of the field interactions with materials. From these considerations a standard expression to represent the square of the index of refraction of Kerr media in discussions of the Helmholtz equation is often taken to be given by the form [1, 29, 30, 34]

$$n_{Kerr}^{2} = n^{2}\left[1 + \lambda |E|^{2}\right]. \tag{9.12a}$$

Here λ is the Kerr parameter and n is the zero field (linear medium) limit of the Kerr index of refraction. In experimental systems, the Kerr parameter is very small so that the Kerr nonlinearity is generally a perturbation to the electrodynamics of problems involving Kerr medium.

Equation (9.12a) is the simplest form of a field dependent refractive index which models the properties of a Kerr nonlinear medium. In general, real systems tend to be more complicated, having dielectrics represented by tensors as well as tensor Kerr parameters. Saturation effects may also come to play. For a more detained consideration of these aspects and for a theoretical discussion of the dielectric model in (9.12a), the reader is referred to the literature. Here (9.12a) will be studied as a representation of a Kerr medium which yields a solvable Helmholtz problem.

The electromagnetic solutions in the region of Kerr media described by (9.12a) have electric fields obtained from the Helmholtz equation of the form [1, 29, 30, 34]

$$\frac{d^{2}E}{dz^{2}} + n^{2}k_{0}^{2}\left[1 + \lambda |E|^{2}\right]E = 0. \tag{9.12b}$$

In the $\lambda \to 0$ limit of the Kerr refractive index, the wavenumber $k = n\omega/c = nk_{0}$, where n is the linear part (i.e., the low power limit) of the Kerr index of refraction. This wavenumber then characterizes the weak field behavior of the Kerr medium, and outside this region, nonlinear effects become important. As with (9.11) an exact solution of (9.12b) can be obtained, and it shall be shown later that the solutions of (9.12b) are explicitly written in terms of Jacobi elliptic functions.

For the treatment of the boundary conditions in the two problems under consideration, it should be noted that: At the dielectric-vacuum interfaces in Fig. 9.2a, b, the boundary conditions connecting the solutions of (9.11) and (9.12b) are that both E and $\frac{dE}{dz}$ are continuous functions. However, in the case of the system in Fig. 9.2b one interface is between a Kerr medium slab and a perfect conducting mirror. At this interface the electric field solution obtained from (9.12b) vanishes at the perfect conducting mirror.

First consider the case in Fig. 9.2a of a transmitting coating. In this geometry the layered coating is surrounded by vacuum so that to the left of the array the incident and reflected waves can be written in the form [1, 34]

$$E = E_0\left[\mathrm{e}^{ik_0 z} + r_0 \mathrm{e}^{-ik_0 z}\right], \tag{9.13}$$

and to the right of the array the transmitted wave is given by

$$E = t_0 E_0 \mathrm{e}^{ik_0 z}. \tag{9.14}$$

In the general solution, these fields must be matched by boundary conditions to the left most and right most Kerr medium slabs of the coating.

The solutions for the electric fields within the slabs forming the layers of the coating are obtained from (9.11) and (9.12). Inside both types of slabs of the coating media the electric field solutions obtained from (9.11) and (9.12) are found to take the general form [1, 29, 30, 34]

$$E = E_0 \varepsilon(z) \mathrm{e}^{i\phi(z)}. \tag{9.15}$$

In (9.15) both of $\varepsilon(z)$ and $\phi(z)$ are real and E_0 in (9.13) through (9.15) is the amplitude of the electric field component of the incident electromagnetic wave. For the general solution, the slab solutions must be connected with one another and with the field solutions in the surrounding vacuum.

To obtain the solution of the coating problem, the forms of the fields in (9.13) through (9.15) must be substituted into (9.11) and (9.12). (See [29, 30] for the detains of this.) Upon applying the boundary conditions, three equations for $\varepsilon(z)$ and $\phi(z)$ are obtained for each of the slabs of the array.

In both the vacuum and Kerr media slabs the following relationships hold [1, 29, 30, 34]

$$\frac{d\phi}{dz} = \frac{W}{I} \tag{9.16a}$$

where

$$I(z) = \varepsilon^2(z) \tag{9.16b}$$

and

$$W = k_0 |t_0|^2. \tag{9.16c}$$

Note in (9.16) that W is a constant throughout the layering of the coating and is fixed by the amplitude of the transmitted wave. In addition, it is seen that (9.16a) determines the phase variation in each of the slabs of the coating.

In addition to the variation of the phases in each slab of the coating, the amplitude $\varepsilon(z)$ variation in each slab must be obtained. From the same substitution used to obtain (9.16) it is found for the case of the vacuum slabs that $I(z) = \varepsilon^2(z)$ is obtained as a solution of [1, 29, 30, 34]

$$\frac{1}{4}\left(\frac{dI}{dz}\right)^2 + W^2 + k_0^2 I^2 = A_l I. \tag{9.17a}$$

Here A_l is an integration constant which is used to match the boundary conditions at the slab interfaces. In general, the constants $\{A_l\}$ are different for each vacuum slab of the finite array but are related to each other between the various different slabs of the coating.

The differential equation in (9.17a) for the vacuum slabs can be rewritten in terms of an equation involving an integral. The resulting integral relationship is given by the indefinite integral form [1, 29, 30, 34]

$$\int \frac{1}{\sqrt{A_l I - W^2 - k_0^2 I^2}} dI = \pm 2z + C_l, \tag{9.17b}$$

where C_l is an integration constant. The integral in (9.17b) can be evaluated in terms of elementary functions giving the functional relationship

$$-\frac{1}{k_0}\sin^{-1}\left(\frac{A_l - 2k_0^2 I}{\sqrt{A_l^2 - 4W^2 k_0^2}}\right) = \pm 2z + C_l. \tag{9.17c}$$

In the case of the Kerr medium slabs, the earlier substitutions used to obtain (9.16) yield the Kerr media Helmholtz equations of the form [1, 29, 30, 34]

$$\frac{1}{4}\left(\frac{dI}{dz}\right)^2 + W^2 + n^2 k_0^2 I^2 + \frac{1}{2} n^2 k_0^2 \tilde{\lambda} I^3 = A_{nl} I. \tag{9.18a}$$

where A_{nl} is an integration constant and $\tilde{\lambda} = \lambda |E_0|^2$ measures the strength of the Kerr nonlinearity. The constants $\{A_{nl}\}$ are used to match the boundary conditions and, consequently, are generally different for each Kerr slab of the coating array.

Again, the differential equation in (9.18a) can be converted into an integral relationship expressed as

$$\int \frac{1}{\sqrt{A_{nl} I - k^2 I^2 - \frac{1}{2} k^2 \tilde{\lambda} I^3 - W^2}} dI = \pm 2z + C_{nl}, \tag{9.18b}$$

where C_{nl} is an integration constant. The indefinite integral in (9.18b) is evaluated in terms of Jacobian elliptic functions. For this rendering, the reader is referred to the literature for the details [29–34].

The constant W, defined in (9.16c), is set by matching boundary conditions across the coating array and, consequently, this also sets the transmitted wave amplitude. In this way the transmission coefficient, $T = |t_0|^2$, for the transfer of light through the of the array in Fig. 9.1a is ultimately determined. In addition, it is seen from (9.15) that the field intensity $|E(z)|^2$ is given by

$$|E(z)|^2 \propto I(z) = \varepsilon^2(z), \tag{9.19}$$

so that $I(z)$ will be used in the later discussions as a dimensionless indicator of the intensity of the fields within the slab.

In the treatment of the problem in Fig. 9.2b for the perfect mirror the solutions generally follow as above with an exception to one of the boundary conditions. In particular, the electromagnetic wave is reflected from the perfect reflecting mirror. Consequently, there is not transmitted wave component so that $W = 0$, and at the perfect conducting mirror $E = 0$. As a result of these changes $\frac{dE}{dz}$ at the mirror surface is now used as a parameter which is set to match boundary conditions over the array of slabs.

Numerical Examples of the Two Coating Models

As an illustration of the two models, the above theoretical results have been evaluated for some numerical examples [34]. In these investigations, both models in Fig. 9.2a, b have been treated for a system of five Kerr nonlinear media slabs to determine their properties of transmission, reflection, and the nature of the soliton wave functions resonantly excited in the coating media. First some results are presented for the transmission coating model in Fig. 9.2a. These are followed by a presentation of results for the mirror coating model in Fig. 9.2b.

In Fig. 9.3 results are presented for the transmission properties of the model in Fig. 9.2a. Specifically, the plot in Fig. 9.3 is of the transmission coefficient of the coating versus the linear part of the Kerr refractive index, n, for an incident wave with a fixed $k_0 d = 1.5$ [34]. Results are shown for two cases of the transmission coating: In Fig. 9.3a the plots are for the linear media limit of the system, in which $\tilde{\lambda} = 0.0$. Consequently, all of the medium in the slabs is linear dielectric media. For a comparison with these results, in Fig. 9.3b a similar plot is presented for a system in the case of a Kerr nonlinear medium with $\tilde{\lambda} = 0.008$. For the plot in Fig. 9.3b the transmission coefficient is again presented as a function of the linear part of the Kerr refractive index, n, for an incident wave with a fixed $k_0 d = 1.5$. Many of the feature found in the two plots are similar, but there are some important differences.

In these plots it is seen that, as n is varied, the transmission passes through a series of stop and pass bands, i.e., regions of near zero and near unit transmission [34]. The stop and pass bands of the finite layering compare well with the stop and pass bands in the photonic crystal composed of an infinite number of slabs. In the infinite photonic crystal, the stop bands occur in the regions of: $1.0540 < n < 1.2340$, $2.6110 < n < 3.8755$, and for $4.4965 < n < 6.0805$. Outside these bands are the pass band regions.

It is also generally found that, in a comparison of Fig. 9.3a, b, the nonlinearity has a small effect on the pass and stop bands observed. In addition to this, within the stop band regions of both the linear and nonlinear coatings the wave functions

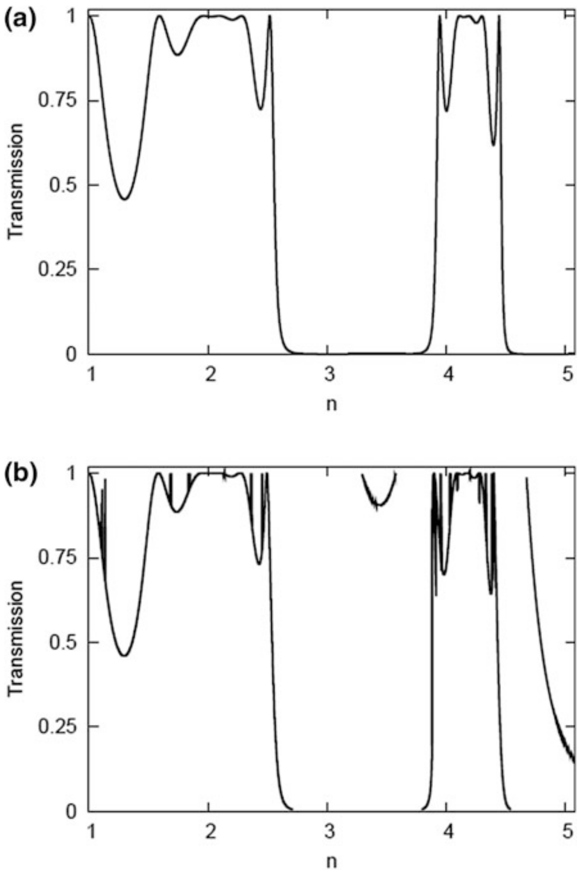


Fig. 9.3 The Transmission Coefficient versus n for: **a** an array of linear dielectric media and vacuum slabs for the geometry in Fig. 9.2a and **b** an array for the geometry in Fig. 9.2a with dielectric slabs formed of Kerr medium characterized by $\tilde{\lambda} = 0.008$. In both plots $k_0 d = 1.5$ and there are five dielectric slabs forming the coating

located in the coating media are much smaller in amplitude than those found in the pass band solutions.

A difference in the pass band structure is, however, observed between Fig. 9.3a, b. This is found in a band of transmission states that is roughly located within the regions $3.29 < n < 3.52$ and $4.67 < n$. The transmission in these bands is associated with soliton modes that are excited within stop bands. These soliton modes are only present in the system in Fig. 9.3b which has a Kerr nonlinearity and are absent from the linear media system in Fig. 9.3a. The soliton solutions in the nonlinear coating are located within the stop band of the linear system and are found to have larger wavefunction amplitudes than their corresponding linear media wavefunctions counterparts.

In order to see the nature of the pulse soliton modes in the new bands found within stop bands of the Kerr coating, it is necessary to plot the wave functions of the excitations associated with the observed transmission enhancements. Results for the field intensities of gap soliton excitations in both of the regions $3.25 < n < 3.52$ and $4.67 < n$ are now discussed [34].

In Fig. 9.4 results for the wave function field intensity, $I(z)$, versus position within the barrier are presented for representative solitons in the system shown in Fig. 9.3b [34]. The results shown in the figures are for solitons that exist in the system for the parameters: (a) $\tilde{\lambda} = 0.008$ and $n = 3.2932$ in the lowest stop band shown in Fig. 9.3b and (b) $\tilde{\lambda} = 0.008$ and $n = 4.6748$ in the next higher stop band shown in Fig. 9.3b. The values of $\tilde{\lambda}$ and n in both of the plots in Fig. 9.4 were selected at the point of maximum transmission in the new soliton bands. Both sets of parameter values are located in the two regions, mentioned earlier in the discussions of Fig. 9.3b, of enhanced transmission anomaly arising from the Kerr nonlinear medium, and these modes are absent from the linear media system in Fig. 9.3a.

For the plots in Fig. 9.4 of $I(z)$ versus position inside the barrier of five Kerr slabs separated by vacuum, the slabs forming the coatings are located within the

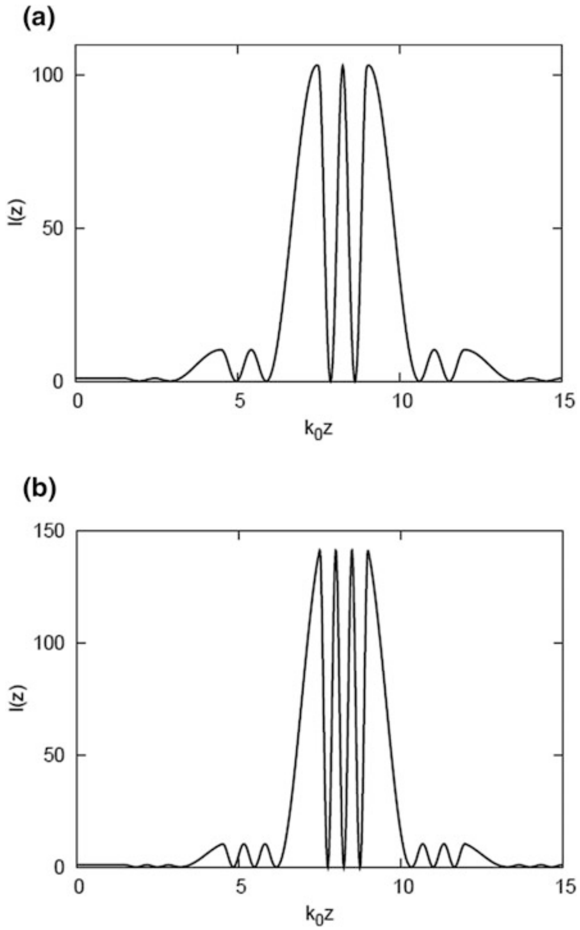


Fig. 9.4 Plot of $I(z)$ versus $k_0 z$ for a coating of five Kerr slabs for: **a** $\tilde{\lambda} = 0.008$ and $n = 3.2932$ in the lowest stop band, **b** $\tilde{\lambda} = 0.008$ and $n = 4.6748$ in the next higher stop band. In these plots the dielectric slabs of the coating are located between $1.5 \leq k_0 z \leq 15.0$ [34]

region $1.5 \leq k_0 z \leq 15$, and the middle of the coating region is located at $(k_0 z)_{middle} = 8.25$. In the case of both of the wave functions in Fig. 9.4, it is seen that the fields of the pulses are highly concentration within the center of the slabs [34].

Compared to the incident and transmitted field intensities the intensity at the center of the coating is much greater. In this regard, it is of interest to note that the presence of such highly concentrated fields has a number of possible technological applications which will be discussed later. First, however, some discussion of the wave functions associated with the mirror problem will be given.

The mirror system in Fig. 9.2b has only a reflected wave, but the presence of gaps soliton modes can still be found in the physical properties observed in the mirror coating. In particular, the band structure effects of the periodic coating are such that the fields penetrating the coating and reflected from the mirror are small within the coating for stop band modes and large within the coating for pass band excitations. Consequently, the field intensities of the modes in the coating are significantly affected by the stop and pass bands of the photonic crystal structure. This is also seen to be the case with the new pass band of soliton modes [34]. The soliton solution wave functions are found to exhibit an intensity increase over the corresponding stop band modes found in the linear limit of the coating.

The enhancement of the fields is particular prominent for the resonant excitation of soliton modes within the coating. As an illustration of this, consider the five Kerr layer coating in Fig. 9.2b at the condition for a gap soliton to be resonantly excited within the system. An example of a soliton intensity profiles under these conditions is presented in Fig. 9.5.

In Fig. 9.5 the soliton wave function intensity, $I(z)$, versus $k_0 z$ are shown for a gap soliton resonantly excited by an incident field for which $k_0 d = 1.5$ and for an array with the parameters $n = 3.2997$ and $\tilde{\lambda} = 0.008$. (See Fig. 9.3b.) To make a spatial reference frame for the plot, the five Kerr dielectric slabs are located within the region $1.5 \leq k_0 z \leq 14.25$ in the figure.

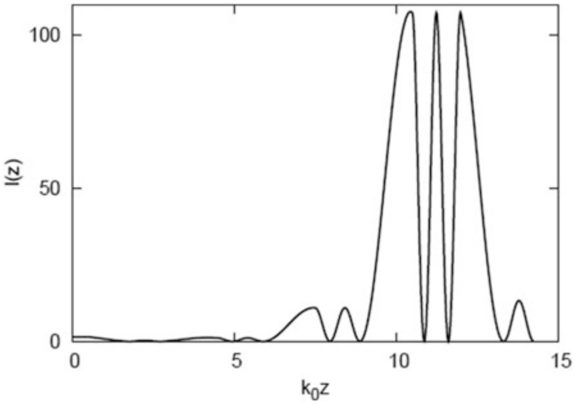


Fig. 9.5 Plot of the soliton intensity fields $I(z)$ versus $k_0 z$ for $n = 3.2997$, $\tilde{\lambda} = 0.008$, and for an incident field with $k_0 d = 1.5$. The figure show the gap soliton pulse excited in the band gap of the photonic crystal coating on a perfect conducting mirror [34]

It is found in Fig. 9.5 that the intensity of the soliton is concentrated within the region next to the mirror surface, where the mirror surface is located at $k_0 d = 14.25$. In general, the Kerr medium photonic crystal coating, consequently, acts to enhance the fields in the coating from those of the incident amplitude at the far left of the plot. Since the enhancement is found to be within a region close to the mirror surface, it can have consequences for the development of surface enhanced Raman spectroscopy. In such a scheme, molecules residing within the enhanced fields will experience an increase spectroscopic coupling to the incident fields [1, 29, 30, 34].

9.2 Generation of Second Harmonics

The second type of nonlinear effect that will be treated here is the generation of second harmonics of radiation [1, 16, 22–28, 35–41]. This is an optical effect in which light at a fundamental frequency is applied to a nonlinear medium and in response the medium generates an additional weak component of radiation at twice the frequency of the fundamental. In the previous discussions, the Kerr nonlinearly left the frequency of the light unchanged while altering the refractive properties of the medium in a way which depended on the intensity of the light. Now in second harmonic generation the frequency of light is doubled by its interaction with the optical medium via mechanisms which also introduce complications in the dynamics of the light generated within the nonlinear medium.

Second harmonic generation is a more problematic effect than the Kerr effect. It occurs at a lower order of nonlinearity than the Kerr effect but involves more crystalline asymmetry in the generating media than is required for the Kerr effect. For a medium to exhibit the property of the generation of second harmonics it is necessary that the crystal structure of the medium lacks inversion symmetry. In addition, there are a variety of other symmetry limitations that come into play on the generated radiation. These restrictions depend on the details of the crystal structure and will not be discussed here so that the reader is referred to the literature for a complete discussion of such symmetry considerations [25–28]. Here only the basic properties are presented of a medium which is assumed to generate second harmonics.

Second harmonic generation has a number of important applications in technology [1, 16, 22–24, 35–41]. In particular, there are a number of laser applications where it is applied to generate second harmonics from the fundamental of a laser output or from the laser itself. It is also important in a variety of optical diagnostics technologies. In this regard, techniques of microscopy have been developed which image biological systems [19, 23] and the properties of surfaces [24] based on the use of their ability to generate second harmonics of radiation applied to them.

The origin of the effect is the nonlinearity of the optical medium in response to an incident beam. Theoretically this is most often discussed in terms of the medium's time-dependent polarization response to the applied field. In this way, the polarization generated in the medium by the applied field can be represented as a Fourier series composed of many different frequency harmonic components, i.e., many different frequency outputs which are multiples of the fundamental driving frequency [1, 25–28]. The various multiples of the fundamental frequency of the incident beam arise directly from the nonlinearity of the dynamics of the positive and negative charges of the material, and it is generally found that the weak second harmonic response of the system is the most dominant of the generated harmonics.

In this process of generation, the amplitude of the polarization response of a frequency harmonic is generally found to decrease with the increasing difference of the harmonic frequency from that of the applied fundamental frequency. As a result of the vector nature of the polarization response and that of the applied fields, it follows for the generated response that there are various symmetry considerations that need to be taken into account in order to understand the polarization tensor and response properties of the medium [25–28]. The asymmetry required for the second harmonic response has already been noted, but restrictions apply for the generation of all of the other harmonics. These in general are dependent on the tensor nature of the polarization involved in their generation.

As a result, it is often found for nonlinear optical materials that aside from the second harmonic terms there are many other possible frequency responses involving other different mixings of the radiation fields applied to the material. These higher harmonic responses are generally weaker than those of the second harmonic fields. This is particularly evident in the application to the material of a number of different radiation fields which often involve high orders of tensor interactions between the various fields. In the introductory treatment of this chapter, however, only second harmonic generation will be a consideration.

Aside from various crystal symmetry considerations that are important to the efficient generation of second harmonics in an optical medium, there are a number of engineering problems that naturally arise in the design of devices based on second harmonic generating media. These difficulties are found in the basic physics considerations involved in the generation of a harmonic waves throughout the spatial extent of a generating material.

In particular, the waves generated throughout the medium have phases which can add constructively and destructively to provide the total wave outputted by the generating device. A problem then arises as to how to extract the maximum intensity of the second harmonic fields from the sum of these phase additive processes.

Some of the problems associated with these phases questions can be solved through the applications of photonic crystals and metamaterials. Consequently, a technological interest in the application of engineered materials in second harmonic generation involves the application of photonics crystals and metamaterials to enhance the generation of second harmonics [22–24, 37–41]. These artificial materials are the focus here along with how they are employed to offer reliable, intense, sources of such radiations.

In this regard, both photonic crystals and metamaterials provide useful tools in the solution of problems involving spatial generation of waves throughout a medium. This problem is often referred to as the phase matching problem in the generation of second harmonics. It is a basic design consideration in devices meant to provide for the generation of second harmonics and has had a long history involving a variety considerations based in classical optics [1, 25–28].

In the following, first some elementary considerations of second harmonic generation will be presented along with a basic mathematical statement of the phase matching problem. Following these discussions a basic presentation is given of the applications of photonic crystals and metamaterials which are used to circumvent the phase matching problem [1, 22–24, 37–41].

9.2.1 Basics of Second Harmonic Generation and the Phase Matching Problem

To begin with, a general discussion of the equations for the generation of second harmonics in a uniform nonlinear medium is outlined [1, 25–28]. These equations are then developed to understand the second harmonic response of a uniform homogeneous medium to an applied radiation field, and to show how the phase matching problem arises.

For these considerations the Maxwell curl equations in a uniform polarizable medium are written in the form [1]

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \tag{9.20a}$$

$$\nabla \times \vec{B} = \mu_0 \frac{\partial \vec{D}}{\partial t}. \tag{9.20b}$$

Here the electric displacement field $\vec{D} = \varepsilon_0 \vec{E} + \vec{P}$ is expressed in terms of the electric field and the electric polarization, $\vec{P}$.

In terms of the displacement field (9.20b) can, consequently, be rewritten as [1, 25–28]

$$\nabla \times \vec{B} = \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t} + \mu_0 \frac{\partial \vec{P}}{\partial t}, \tag{9.20c}$$

where the polarization response of the medium is separated out to the righthand side of the equation. It describes the material science of how the dielectric medium interacts with the applied electric field.

In (9.20c) the polarization vector $\vec{P}$ contains the response of the dielectric medium to an applied electric field. This response consists both of linear response terms, having a linear dependence on the electric field, and of nonlinear response terms, which depend on higher order tensor powers of the electric field.

The responses of the system driven by the linear components of the polarization, then, represent the topic of standard treatments in traditional classical optical systems. It only involves frequency terms at the fundamental frequency of the incident applied field. These are topics that were a primary interest before the development of intense laser fields which could generate field intensities sufficient to excite a nonlinear dielectric response from the system involving frequencies different from the fundamental.

The nonlinear parts of $\vec{P}$ include terms which are responsible for the generation of the frequency responses of the system that differ from the fundamental frequency of the applied electric field. In particular, some of these terms are responsible for the second harmonic response of the generating medium as well as some of the weaker effects related to higher order harmonics developed in the medium.

In addition, some of the nonlinear terms, however, also renormalize the dielectric response of the system at the fundamental frequency applied to the medium. These are, for example, responsible for the Kerr effect in which the index of refraction is found to be dependent on the intensity of the applied electric field.

The linear terms in $\vec{P}$ contribute to the renormalization of the optical response of the medium at the fundamental frequency of the system. This is familiar from the optics of linear media. In addition, the linear response also affects the propagation of the fields of the second harmonic generated as well as those of any higher harmonics that may be present in the medium. The mathematics involved in the study of the linear and nonlinear polarization are now addressed.

Consider the evaluation of (9.20c) for the fundamental frequency fields and for the second harmonic fields generated by the action of the nonlinearity on the fundamental fields. For the discussions, an incident fundamental harmonic wave will be considered to interact with a uniform homogeneous isotropic medium to generate a second harmonic wave. The nonlinear interaction of the medium will be considered small so that only the leading order nonlinear corrections on the fields will be considered. This is found to be a good approximation in the study of most nonlinear optical systems.

Under these considerations, (9.20c) is found to decouple into two equations. The first equation describes the behavior of the fundamental frequency applied to the medium. The second equation accounts for the fields generated at twice the fundamental frequency.

In this way from (9.20a) and (9.20c) it is found that the fundamental fields, $\vec{E}_1$, occurring at frequency ω satisfy the homogeneous Helmholtz equation

$$\nabla^2\vec{E}_1 + \mu_0\varepsilon(\omega)\omega^2\vec{E}_1 = 0. \tag{9.21a}$$

Here $\varepsilon(\omega)$ represents the response of the medium at the frequency of the fundamental applied field. It ultimately is related to the linear component of the polarization of the medium. In particular, all of the waves propagating in the medium, to leading order, will satisfy a linear medium homogeneous Helmholtz equation similar to that in (9.21a).

The small second harmonic fields, $\vec{E}_2$, that are generated from the fundamental wave occur at frequency 2ω and from (9.20) are found to satisfy an inhomogeneous Helmholtz equation of the form

$$\nabla^2 \vec{E}_2 + \mu_0 \varepsilon(2\omega)(2\omega)^2 \vec{E}_2 = -\mu_0 (2\omega)^2 \vec{P}_{2\omega}. \tag{9.21b}$$

In (9.21b), the $\vec{P}_{2\omega}$ term on the right of the equation is the nonlinear component of the polarization. It is the component of the polarization response generated in the nonlinear medium and is the source of the second harmonic radiation that is generated within the medium.

The second harmonic response represented in $\vec{P}_{2\omega}$ is dependent on the presence of the fundamental applied field, and in the absence of the fundamental field no second harmonic is present in the system. In this regard, the $\vec{P}_{2\omega}$ response can ultimately be shown to be related to the square of the fundamental field. Such a dependence on the fundamental fields, in fact, accounts for the generation of a wave at twice the fundamental frequency. In addition, it is found to also explain the dependence of the intensity of the generated second harmonic wave on the intensity of the fundamental wave.

Regarding the terms on the left hand side of (9.21b), the factor of $\varepsilon(2\omega)$ in the second term of the sum renormalizes the free space permittivity at frequency 2ω. It comes from the linear part of the polarization vector response, $\vec{P}$, at the fundamental driving frequency 2ω. Consequently, the righthand side of (9.21b) gives the Helmholtz equation for the leading order response of the system to an applied polarization at the frequency 2ω.

Experimentally as well as theoretical it can be shown that, for a nonlinear system which generates a second harmonic, a simple model for the nonlinear polarization is given by the form [1, 25–28]

$$\vec{P}_{2\omega} = \vec{E}_1 \cdot \tilde{d}(2\omega, \omega, \omega) \cdot \vec{E}_1. \tag{9.22}$$

In this equation $\tilde{d}(2\omega, \omega, \omega)$ is a third rank tensor which is subject to certain important symmetry considerations that have been extensively discussed in the literature of nonlinear optics.

The formulations in (9.21) and (9.22) are now set to treat the dynamics of the fundamental and the generation of second harmonics from the fundamental in a second harmonic generating medium. In a simplified model based on these equations the origins of the problem of phase matching associated with the generation of second harmonics can be explained.

The Problem of Phase Matching

In the following presentation, an aim of the discussions is to give a simplified treatment of second harmonic generation, illustrating the origins of the ideas of phase matching. Consequently, a simple one-dimensional geometry will be considered which provides the easies example of the mechanism of second harmonic generation. An idealized version of polarization terms in (9.22) will also ultimately be used in the development of the theory based on (9.21). The reader is referred to the literature for the details of treatments focused on models based on (9.22) which represent specific materials for purposes of technological applications and for more complex geometries of the generating medium.

As an illustration of the phase matching problem associated with the generation of second harmonic within a uniform medium, a simplified model of a system for the generation and propagation treated in the context of a one-dimension system is now discussed. Essentially what this model is meant to show is that as the fundamental wave propagates through and interacts with the nonlinear generating medium, it gives rise at each point of the medium to an outputted second harmonic wave.

As a consequence, the outputted wave causes the fundamental wave to decrease in intensity. In addition, as the fundamental waves moves through the medium, its energy loss is found to reappear in the generated second harmonic output of the medium. This second harmonic generation is made throughout the extent of the nonlinear medium and leads to various phase effects.

Since the generated second harmonic waves not only involve an amplitude but also a phase, various processes of phase coherent addition of the second harmonic fields occur throughout the generating medium. These phase additions are both constructive and destructive in nature over the entire medium. Consequently, the final wave emitted from the medium is found to be very sensitive to the totality of the phase related processes throughout the entire medium. Only under very special considerations do the constructive and destructive processes result in a total generated second harmonic wave of good intensity exiting the source device.

For a simple treatment illustrating these points, consider a planar interface between vacuum and a second harmonic generating medium. The interface is taken in the y-z plane with the x-axis perpendicular to the interface, located at $x = 0$, between the two semi-infinite media. (See Fig. 9.6 for a schematic.) On the left of the interface is vacuum and a uniform nonlinear second harmonic generating medium is to the right of the interface.

The incident electromagnetic plane wave of frequency ω propagates along the x-axis in the vacuum, moving towards the right. This is the fundamental wave which enters the nonlinear medium and generates a second harmonic response. Upon entering the nonlinear medium it is assumed that the nonlinearity of the medium is very small so that it has little effect on the fundamental. Consequently, the depletion of energy from the fundamental wave of frequency ω is negligible as it propagates into the nonlinear medium.

Under the assumption that there is no significant energy loss in the fundamental wave, it follows from (9.21a) that the form of the solution for the fundamental wave in the nonlinear medium is given by [1, 25–28]

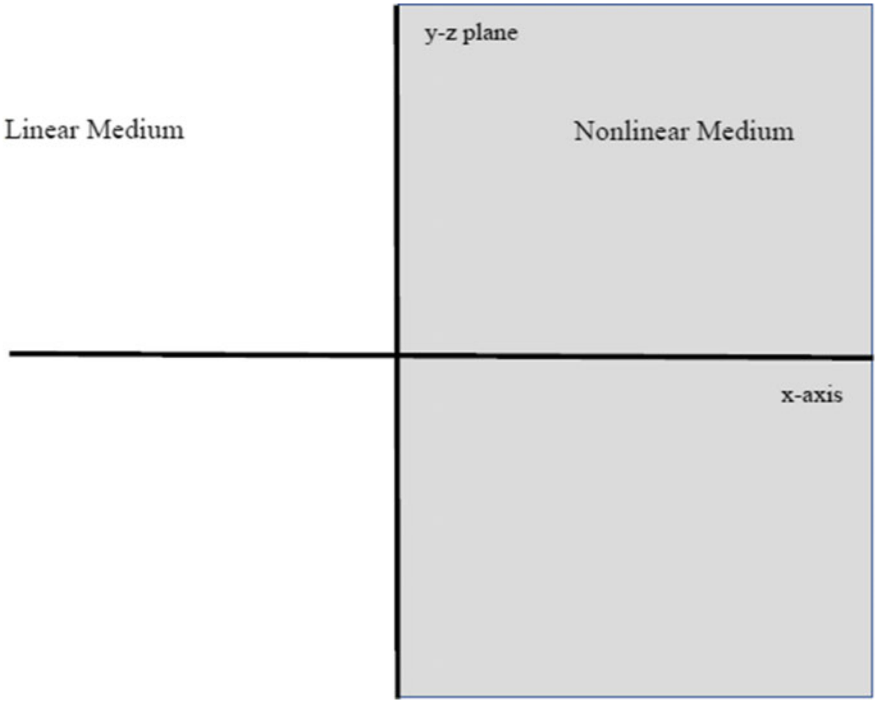


Fig. 9.6 Planar interface between a linear media and a nonlinear second harmonic generating media

$$E_1 = E_{10} \mathrm{e}^{i[k(\omega)x - \omega t]}. \tag{9.23}$$

In (9.23) the wave number $k(\omega) = \sqrt{\mu_0 \varepsilon(\omega)}\omega$ is that of the fundamental wave of frequency ω, and the constant amplitude E_{10} in (9.23) is determined by the boundary conditions for the dielectric mismatch at the interface. The field in (9.23) is then a vector field which lies in the y-z plane.

In terms of the form of the solution in (9.23) for the fundamental harmonic wave, (9.21b) for the generated second harmonic wave is written as [1, 25–28, 42]

$$\frac{\partial^2 E_2}{\partial x^2} + \mu_0 \varepsilon(2\omega)(2\omega)^2 E_2 = -\mu_0 (2\omega)^2 d(2\omega)(E_{10})^2 \mathrm{e}^{2i[k(\omega)x - \omega t]}. \tag{9.24}$$

In obtaining (9.24), the polarization generating the second harmonic radiation is written in terms of the fundamental wave and expressed by the simple form $P_{2\omega} = d(2\omega)(E_{10})^2 \mathrm{e}^{2i[k(\omega)x - \omega t]}$. The vector of second harmonic polarization is also taken parallel to the y-z plane and parallel to the polarization of the generated second harmonic wave, E_2. The coefficient of the polarization coupling for the second harmonic generation has also been assumed to be of a simple form which is a constant throughout the nonlinear medium [42].

An approximate solution of (9.24) is obtained by assuming that E_2 is of the form

$$E_2 = E_{20}(x)\mathrm{e}^{i[k(2\omega)x-2\omega t]}. \tag{9.25}$$

where $k(2\omega) = \sqrt{\mu_0\varepsilon(2\omega)}2\omega$. Here the field envelop function $E_{20}(x)$ is treated as slowly varying in space over the wavelength of the plane wave which it multiplies. In particular, in this regard a necessary condition on the spatial variation of the envelope function is that $\left|\frac{\partial^2 E_{20}}{\partial x^2}\right| \ll \left|k(\omega)\frac{\partial E_{20}}{\partial x}\right|$.

Under these conditions a substitution of (9.25) into the second spatial derivative in (9.24) is approximately given by the form [1, 25–28]

$$\frac{\partial^2 E_2}{\partial x^2} \approx \left[2ik(2\omega)\frac{\partial E_{20}}{\partial x} - k^2(2\omega)E_{20}\right]\mathrm{e}^{i[k(2\omega)-2\omega t]}. \tag{9.26}$$

Applying this in (9.24), it is found that [1, 25–28, 42]

$$\frac{\partial E_{20}}{\partial x} = -\frac{\mu_0 2\omega d(2\omega)(E_{10})^2}{2i\sqrt{\mu_0\varepsilon(2\omega)}}\mathrm{e}^{i[2k(\omega)-k(2\omega)]x}. \tag{9.27}$$

The problem of determining the spatial variation of the amplitude of the generated second harmonic wave in the semi-infinite nonlinear medium has now been reduced to the integration of a first order differential equation. In the course of the following discussion, the phase matching problem will be seen to arise on the righthand side of (9.27) from the rapid variation of the complex exponential in the x variable.

A direct integration of (9.27), subject to the boundary condition that $E_{20}(x=0) = 0$ (i.e., the second harmonic wave only begins to be generated when the fundamental wave first encounter the surface of the semi-infinite nonlinear medium), yields the position dependence of the amplitude of the generated second harmonic wave. In general, it is seen from (9.27) that the relevant integral involved is of the form [42]

$$\int_0^x \mathrm{e}^{i[2k(\omega)-k(2\omega)]x}dx. \tag{9.28}$$

For the case in which $2k(\omega) - k(2\omega) \neq 0$ the integral in (9.28) is seen to display an oscillatory behavior. In this behavior the integral exhibits successive waves of constructive and destructive phase additions, resulting in an oscillating intensity of the second harmonic wave output between two bounded limits. This is the heart of the phase matching problem and represents a fundamental limitation on the intensity of the second harmonic generated within the nonlinear medium. The focus of many optical designs for second harmonic systems is to maximize the results from integrals of this form.

Notice, however, that in the case that

$$2k(\omega) = k(2\omega) \tag{9.29a}$$

a good phase matching result can be achieved [1, 42]. Under this condition, solving (9.27) for $x > 0$ results in a linear growth of the amplitude of the second harmonic wave. In particular, the integration of (9.27) yields [1, 25–28, 42]

$$E_{20}(x) = -\frac{\mu_0 2\omega d(2\omega)(E_{10})^2}{2i\sqrt{\mu_0 \varepsilon(2\omega)}} x, \tag{9.29b}$$

with the generated wave amplitude proportional to the distance away from the interface.

In principle, the intensity of the second harmonic fields in (9.29) can increase indefinitely. The result in (9.29), nevertheless, is based on the assumption that the amplitude of the fundamental is a constant, and this is not the case as the energy from the fundamental is transferred to the second harmonic field. A proper accounting of this energy transfer provides a second limitation on the intensity of the generated second harmonic that can ideally be achieved.

The preferred condition for generating an intense second harmonic source output is to have a system which displays the linear growth with position obtained in (9.29) as opposed to the oscillatory behavior in (9.28). In general, however, such a phase matching condition requires a very precise set of conditions be placed on the dielectric properties and dispersion relations of the fundamental and second harmonics within the nonlinear system.

Many technological considerations are needed to fix the problems associated with these type of phase matching considerations [1, 22, 24, 38–41]. Such treatments of the phase problem involve a number of general geometric and material science solutions, and the reader is referred to the literature for discussions based on designs not involving photonic crystals and metamaterials. In the following solutions focused on photonic crystals and metamaterials will be briefly reviewed.

9.2.2 Applications of Photonic Crystal and Metamaterials to Second Harmonic Generation

Metamaterials [1, 24, 38–41] and photonic crystals [1, 16, 22, 35, 36] offer a variety of methods which can be exploited in the design of enhancement mechanisms for the generation of second harmonics. These technologies allow light to be manipulated in new ways that are not available using traditional methods of optics. At nanoscience length scales these features are found to reduce the effects of the phase matching problem or to enhance the generating fields at the fundamental frequency. A brief outline of some of the more recent developments is given.

In this regard, meta-material mirrors have recently found applications [16, 22–24, 35–41] providing for a great enhancement of the second harmonics generated at certain types of nano-technology based mirror surfaces. In this technology the surface of the mirror is formed in part from a layering of nanostructures. The nanostructures employed are designed to be of smaller length scales than the wavelengths of the fundamental and second harmonic of radiation generated at the mirror surface and are arranged on the surface in the form of a metamaterial functioning as a quantum well heterostructure [38].

In the second harmonic generation from this system, the mirror surface is composed of a multi-quantum well semi-conductor heterostructure [38]. The heterostructure is formed upon a metal (gold) substrate and has a periodic metal (gold) patterning of facets opposite the substrate on the reflecting side of the mirror surface. In its operation, a fundamental frequency of light is sent into the mirror at normal incidence to the mirror surface.

A second harmonic is then generated in the heterostructure and subsequently is reflected by the mirror in a direction normal to surface of the mirror. The mechanism for enhancing the optical nonlinearity, which generates the enhanced second harmonic, is based on the excitation of surface plasmons in the heterostructure forming the mirror surface. In this regard, the mechanism is reminiscent of surface enhanced spectroscopy.

In addition to the study of mirror surfaces, the use of metamaterials in the design of bulk media exhibiting second harmonic generation properties has also been pursued. A number of different designs based on split ring resonators and on various types of nano-particles have been investigated [39–41].

Photonic crystals [1, 16, 35, 36] are another class of artificial materials that can be engineered to facilitate the generation of second harmonics. It is often found that correctly choosing the periodicity of the photonic crystal can be used to enhance the second harmonic effects of the materials forming the photonic crystal. Unlike metamaterials, photonic crystal systems operate on radiation for which the wavelengths of the fundamental and/or second harmonics are of the order of the smallest translations of the photonic crystal lattice into itself.

Consequently, strong interactions of these radiations within the photonic crystal can be used to guide and enhance their effects in the generation process [1, 16, 35, 36]. Mechanisms at play in the generation of second harmonics in the photonic crystals arise from [1, 16, 35, 36]: (a) creating a periodic nonlinear response in the system by periodically positioning the media generating the second harmonic fields to enhance phase coherent additions of the generated fields, (b) formulating designs involving a periodic dielectric in which the fundamental and/or second harmonic strongly interact with the band structure effects of the photonic crystal to concentrate their interactions, or (c) designing photonic crystals to modify the density of states via the Purcell effect in such a manner as to increase the efficiency of second harmonic generation.

Two ideas form an operative basis in the application of photonic crystals to second harmonic generation. In the first idea, it is realized that for second harmonic generation in a uniform medium there are regions of the medium from which the generated radiation adds constructively, and there are region of the medium from which the radiation generated adds destructively. In this regard, a rough solution of the problem is made by arranging a segmented array of second harmonic generating slabs positioned in space so as to support an optimally efficient second harmonic generation.

In one-dimension such an array is formed by slicing the uniform semi-infinite slab up into a series of slabs of finite thickness positioned along the *y*-*z* plane. Removing slabs at the proper periodic intervals, so that only a finite number of slabs remain, can create an efficient generating structure. The slabs in the array are chosen to remain in the array so that their generated fields add constructive. Such a finite periodic array has been shown to easily be formulated to meet the above conditions [1, 16, 35, 36].

A second approach is based on the ability of photonic crystals to modify the electromagnetic density of states. This modification of the density of states of the electromagnetic modes is known as the Purcell effect [1], and it has great effects on the electrodynamic properties exhibited by a system. The importance of the Purcell effect in this regard arises from the fact that transition processes which generate electromagnetic waves are described by the Fermi Golden Rule. Processes described by this rule are shown to be closely tied to the density of electromagnetic states.

One of the consequences of the Fermi Golden Rule is that the rate of transition of processes generating electromagnetic fields are related to the electromagnetic density of states available into which the generated electromagnetic waves can make a transition. Consequently, enhancing the density of states available to transition into increases the rate of transition. For example, states immediately above and below the stop bands of a photonic crystal have density of states that are increased over the density of states in a uniform medium. This would increase the rate of transition into these states. On the other hand, the density of states within a stop band of the photonic crystal are zero and so in this case transitions into these states are suppressed.

The idea of applying the ideas in the second method is to arrange for a system in which the density of states is enhanced at both the fundamental and harmonic frequencies. These enhancements would facilitate the generation of second harmonics over that which occurs in the fields generated in a uniform medium [1, 16, 39–41] and for which the corresponding density of states are not enhanced.

References

1. A.R. McGurn, *Nonlinear Optics of Photonic Crystals and Meta-Materials* (Claypool & Morgan, San Rafael, 2015)
2. T.F. Krauss, O. Painter, A. Scherer, J.S. Roberts, R.M. De La Rue, Photonic microstructures as laser mirrors. Opt. Eng. **37**(4), 1143–1148 (1998)
3. S.A. Moore, L. O'Faolain, T.P. White, T.F. Krauss, Photonic crystal laser with mode selective mirrors. Opt. Express **16**(2), 1365–1370 (2008)
4. T.D. Happ, M. Kamp, F. Klopf, J.P. Reithmaier, A. Forchel, Two-dimensional photonic crystal laser mirrors. Semicond. Sci. Tech. **16**(4), 227–233 (2001)
5. C. Shemelya, D.F. DeMeo, T.E. Vandervelde, Two dimensional metallic photonic crystals for light trapping and anti-reflective coatings in thermophotovoltaic applications. Appl. Phys. Lett. **104**, 021115 (2014)
6. Y. Ohtera, D. Kurniatan, H. Yamada, Antireflection coating for multilayer-type photonic. Opt. Express **18**(12), 12249–12261 (2010)
7. D.M. Callahan, K.A.W. Horowitz, M.F. Atwater, Light trapping in ultrathin silicon photonic crystal supperlattices with randomly-textured dielectric couplers. Opt. Express **21**(25), 30315–30326 (2001)
8. L. Zhu, A.P. Raman, S. Fan, Radiative cooling of solar absorbers using a visibly transparent photonic crystal thermal blackbody. PNAS **112**(40), 12223–12224 (2015)
9. P. Bermel, C. Luo, L. Zeng, L.C. Kimerling, J.D. Joannopoulos, Improving thin-film crystalline silicon solar cell efficiencies with photonic crystals. Opt. Express **32**(1), 934–937 (2007). Complete this
10. D. Zhou, R. Biswas, Photonic crystal enhanced light-trapping in thin film solar cells. J. Appl. Phys. **103**, 093102 (2008)
11. S. Zanotto, M. Liscidini, C. Andreani, Light trapping regimes in thin-film silicon colar cells with a photonic crystal pattern. Opt. Express **18**(5), 4260–4274 (2010)
12. M. Florescu, H. Lee, I. Puscasu, M. Pralle, L. Florescu, D.Z. Ting, J.P. Dowling, Improving solar cell efficiency using photonic band gap materials. Sol. Energy Mater. Sol. Cells **91**(1), 1599–1610 (2007)
13. A. Knapitsch, P. Lecoq, Review on photonic crystal coatings for scintillators. Int. J. Mod. Phys. A **29**(30), 1430070–1430101 (2014)
14. Z. Zhu, B. Liu, H. Zhang, W. Ren, C. Cheng, S. Wu, M. Gu, H. Chen, Improvement of light extraction of LYSO scintillator by using a combination of self-assembly of nanospheres and atomic layer deposition. Opt. Express **23**(6), 7085–7093 (2015)
15. J.L. Lee, M. Tymchenko, C. Argyropoulos, P.-Y. Chen, F. Lu, F. Demmerie, G. Boehm, M.-C. Amann, A. Alu, M.A. Belkin, Giant nonlinear response from plasmonic metasurfaces coupled to interband transitions. Nature **511**, 65–69 (2014)
16. J. Matroell, R. Corbalan, Enhancement of second harmonic generation in a periodic structure with a defect. Opt. Commun. **108**, 319–323 (1994)
17. M. Scalora, M.J. Bloemer, A.S. Manka, J.P. Dowling, C.M. Bowden, R. Viswanathan, J.W. Haus, Pulsed second-harmonic generation in nonlinear, one-dimensional, periodic structure. Phys. Rev. A **56**, 3166–3174 (1997)
18. G. D'Aguanno, M. Centini, M. Scalara, C. Sibilla, Y. Dumeige, P. Vidakovic, J.A. Levenson, M.J. Bleomer, C.M. Bowden, J.W. Haus, M. Bertolotti, Photonic band edge effects in finite structures and applications to $\chi^{(2)}$ interactions. Phys. Rev. E **64**, 016609 (2001)
19. X. Zhao, J. Xue, Z. Mu, Y. Hyang, M. Lu, Z. Gu, Gold nanoparticle incorporated inverse opal photonic crystal capillaries for optofluidic surface enhanced Raman spectroscopy. Diosens. Bioelectron. **72**, 268–274 (2015)

20. S.-M. Kim, W. Zhang, B.T. Cunningham, SiO_2-Ag post-cap nanostructure coating for surface enhanced Raman spectroscopy. Appl. Phys. Lett. **93**, 142112 (2008)
21. L.D. Tuyen, A.C. Liu, C.-C. Huang, P.-C. Tsai, J.H. Lin, C.-W. Wu, L.-K. Chau, T.S. Yang, L.D. Minh, H.-C. Kan, C.C. Hsu, Doubly resonant surface-enhanced Raman scattering on gold nanorod decorated inverse opal photonic crystals. Opt. Express **20**(28), 29266–29275 (2012)
22. L.-P. Peng, C.-C. Hsu, Y.C. Shih, Second-harmonic green generation from two-dimensional nonlinear photonic crystal with orthorhombic lattice structure. Appl. Phys. Lett. **83**, 3447 (2003)
23. B.E. Cohen, Biological imaging: beyond fluorescence. Nature **467**, 407 (2010)
24. Y.R. Shen, Surface properties probed by second-harmonic and sum-frequency generation. Nature **337**, 519 (1989)
25. R.W. Boyd, *Nonlinear Optics*, 2nd edn. (Academic, Amsterdam, 2003)
26. D.L. Mills, *Nonlinear Optics* (Springer, Berlin, 1998)
27. P.P. Banerjee, *Nonlinear Optics* (Dekker, New York, 2004)
28. Y.S. Kivshar, G.P. Agrawal, *Optical Solitons* (Academic, Amsterdam, 2003)
29. W. Chen, D.L. Mills, Optical response of a nonlinear dielectric film. Phys. Rev. B **35**, 524–532 (1987)
30. W. Chen, D.L. Mills, Optical response of nonlinear multilayer structures: bilayers and supperlattices. Phys. Rev. B **36**, 6269–6278 (1987)
31. V.M. Agranovich, S.A. Kiselev, D.L. Mills, Optical multistability in nonlinear superlattices with very thin layers. Phys. Rev. B **44**, 10917–10920 (1991)
32. W. Chen, D.L. Mills, Gap solitons and the nonlinear optical response of superlattices. Phys. Rev. Lett. **58**, 160–163 (1987)
33. R. McGurn, A.A. Maradudin, Localization of light in linear and nonlinear random layered dielectric systems that are periodic on average. Phys. A **207**, 435–439 (1994)
34. R. McGurn, Kerr nonlinear layered photonic crystal coatings, in *Proceeding of SPIE on Photonic Properties of Engineered Nanostructures, VII*, vol. 10112, 101120H (2017), 18pp. https://doi.org/10.1117/12.2250040 (20 Feb 2017)
35. V. Andreev, O.A. Andreeva, A.V. Balakin, D. Boucher, P. Masselin, I.A. Ozheredov, I.R. Prudnikov, A.P. Shkurinov, Mechanisms of second-harmonic generation in one dimensional periodic media. Quantum Electron. **29**, 632 (1999). M. Scalora, M.J. Bloemer, A.S. Manka, J. P. Dowling, C.M. Bowden, R. Viswanathan, J.W. Haus, Pulsed second-harmonic generation in nonlinear, one-dimensional, periodic structures. Phys Rev. A **56**, 3166 (1997)
36. S. Buckley, M. Radulaski, J. Petykiewica, K.G. Lagoudakis, J.-H. Kang, M. Brongersma, K. Biermann, J. Vuckovic, Second harmonic generation in GaAs photonic crystal cavities in (111)B and (001) crystal orientations. ACS Photonics **1**, 516–523 (2014). G. D'Aguanno, M. Centini, M. Scalora, C. Sibilia, Y. Dumeige, P. Vidakovic, J.A. Levenson, M.J. Bloemer, C. M. Bowden, J.W. Haus, M. Bertolotti, Photonic band edge effects in finite structures and applications to $\chi^{(2)}$ interactions. Phys. Rev. E **64**, 016609 (2001)
37. Y.B. Band, *Light and Matter: Electromagnetism, Optics, Spectroscopy, and Lasers* (Wiley, Chichester, 2006)
38. P.N. Melentiev, A.E. Afanasev, V.I. Balykin, Giant optical nonlinearity of plasmonic nanostructures. Quantum Electron. **44**, 547 (2014). J.L. Lee, M. Tymchenko, C. Argyropoulos, P.-Y. Chen, F. Lu, F. Demmerie, G. Boehm, M.-C. Amann, A. Alu, M.A. Belkin, Giant nonlinear response from plasmonic metasurfaces coupled to intersubband transitions. Nature **511**, 65 (2014)
39. M.W. Klein, C. Enkrich, M. Wegener, S. Linden, Second-harmonic generation from magnetic metamaterials. Science **313**, 502–504 (2006)

40. Z.A. Kudyshev, I.R. Gabitov, A.I. Maimistov, R.Z. Sagdeev, N.M. Litchintser, Second harmonic generation in transiton metamaterials. J. Opt. **16**, 114011 (2014)
41. G. Biris, N.C. Panoiu, Second harmonic generation in metamaterials based on homogeneous centrosymmetric nanowires. Phys. Rev. B **81**, 195102 (2010)
42. A. Yariv, *Introduction to Optical Electronics* (Holt, Rinehart, and Winston, Inc., New York, 1971)

Chapter 10
Quantum Computers

A fundamental distinction between classical and quantum mechanical systems is how the two theories treat the idea of probability [1–10]. Quantum mechanics is a theory which has a fundamental basis in the ideas of probability and probability distributions. This, however, is not so much the case with classical mechanics. It is not necessary in classical discussions to introduce the ideas of probability distributions for many important applications of the theory, nevertheless, the idea of probability distribution can be introduced into the theory in a natural way. Since probability is known to be an essential element in the understanding of quantum theory, for a comparison of classical and quantum ideas it is necessary to view both in the context of probability.

In classical mechanics the particles of a system are described by a set of generalized coordinates which are uniquely developed in time by Newton's laws of motion and the particle force laws. Even though in this formulation the coordinates of the particles can be known at all times, it is often still useful to describe the dynamics by a probability function. This can be done for any system of classical particles and is often important as it allows for the treatment of the averaged properties of the particles. For example, such a probability centered approach is a basis in classical treatments of statistical mechanics and kinetic theory.

In these types of statistical formulations of classical mechanics, the probability of finding a particle with a particular set of generalized coordinates is described by a probability distribution. The time development of the system can be directly related to changes in the particle distribution function that evolves in time according to the classical equations of motion. In such a treatment of classical mechanics, then, the focus of the study of mechanical systems is on the equations of motion and the probability distribution of the particles. These represent the complete description of the properties of the system in time and space.

A quantum mechanical treatment, on the other hand, represents the system from a completely different viewpoint [1–10]. In quantum mechanics, the particles of a system are described by a probability amplitude commonly known as a wave function. In this view, the statistical properties of the system are related to a

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7_10

probability distribution which is the modulus squared of the probability amplitude. Unlike classical mechanics, it is, however, the probability amplitude and not the probability distribution that evolves in time in accord with the equations of motion of the system.

The quantum probability amplitude evolves in time by a Schrodinger equation based on a set of potentials representing the particle interactions [7, 8]. Unlike in the classical theory the dynamics of the quantum mechanical probability distribution is a manifestation of the dynamics of the probability amplitude. This introduces a fundamental distinction in the nature of classical and quantum probability distributions. Both theories treat properties that can be represented by probability distributions but the development of these distributions in time is quite different between the two theories.

As a result, the behaviors of a system allowed in a description based on a quantum probability distribution are more general than those of the same system based on a classical mechanics description. Solely with the introduction of the probability amplitude as an intermediary to the probability distribution of the systems, the distribution displays a richer variety of possibilities and new types of phenomena from those found in classical theories [1–10].

In the following, discussions will be presented focused on two new phenomena found in quantum systems. These phenomena arise in regard to the difference in quantum and classical probabilities as revealed by the Bell inequalities [1–3, 7] and the properties of quantum entanglement [1–7].

The Bell inequalities are a set of relationships satisfied by classical probability distributions [1–3]. The interesting aspect of these relationships is that they are not always satisfied by quantum mechanical probability distributions. In the early theoretical development of quantum theory, the inequalities provided one of the first quantitative measures of the difference between the classical and quantum theories. Since they were put forth it has been shown in a number of experiments that the probabilities of naturally occurring systems do not in fact satisfy the Bell inequalities. Consequently, systems in nature are not based on classical mechanical descriptions.

In addition, discussions are presented of another important aspect of quantum systems. This is the property of entanglement which is directly related to the expanded set of wave functions and probabilities accessed by quantum mechanical theories. Entanglement is a new property arising in quantum systems and was originally closely related to a problem in the foundations of quantum theory.

Specifically, entanglement originally was the source for the discussions of the Einstein-Podolsky-Rosen paradox [1–7]. This paradox was presented early on in the development of quantum theory as a potential problem in the completeness of the quantum description of nature and was related to the proposal of the need for various hidden variables. In this viewpoint, quantum theory was not considered to be a complete theory, and the hidden variables would enter into a more complete theory of physical systems.

The entanglement of quantum systems is an important new outlook in the quantum mechanics of many body problems [4–7]. It was originally presented as a paradoxical feature in, for example, cases of particle decay. In decays of some systems an angular momentum conserving process occurs in which two identical entangled particles of spin one-half are generated, traveling off in different directions.

Consider such a decay into a system where the two particles are generated in a singlet spin state. For such a state the measurement of the spin angular momentum of one particle fixes the spin of the second particle as being opposite that of the first particle, even though the two may be separated by a great distance. This appeared to some physicists as a violation of relativistic mechanics and an indication that quantum theory may be an incomplete description of nature.

Aside from its importance to the foundations of quantum theory, the entanglement feature of quantum mechanical systems is at the heart of the ideas upon which quantum computation is based. Entangled systems offer design applications for parallel processing computational mechanisms and dense storage of information. Entangled states are a recent focus in the attempt to design new types of computer algorithms and mechanisms of information storage that are more efficient and compact than those currently available based on conventional technologies [7–10].

Both the Bell inequality and entanglement aspects of quantum theory contribute to the basis of quantum computation. They must be understood in order to grasp the advantages of quantum mechanics in the design of more effective computers. After a discussion of these foundations of quantum theory, the discussion of quantum computation will begin.

In the presentation of the chapter, the focus will be on developing quantum computers as they offer a new method of doing calculations that are based on the difference in the probabilistic natures of classical and quantum mechanical systems and offer many possible advantages over traditional classical mechanical techniques of computation. At the basis of this are the entanglement features of quantum mechanical systems which can be developed for performing massively parallel computations run in novel nontraditional forms of computational algorithms. Such computational techniques employ mechanisms designed specifically to function on nano-systems. Furthermore, these types of parallelized calculations offer a potential for the rapid calculation of things that are not approachable using the techniques of traditional computers based on classical mechanic systems. The possibility of the applications of these types of calculational features of quantum systems has been an incentive for the investigation of a variety of systems for the development of quantum computers.

In the following, after a treatment of the properties of quantum mechanical systems that form a foundation for quantum computers, quantum computational methods will be discussed. The chapter will end in a presentation of some illustrative examples of quantum computing algorithms.

10.1 Bell's Inequality

To understand the nature of the difference between classical and quantum probability distribution functions consider the conditional probabilities involving three different random variables [1–3]. This is the smallest number of random variables that can exhibit the nature of the difference between probabilities in classical and quantum systems. First some discussions of the probabilities involving three different random variables will be treated for a classical system. This will be followed by a discussion of three random variables modeled in a quantum system.

Consider three random variables $\{S_1, S_2, S_3\}$ representing different properties exhibited by a classical mechanical system but that are similar in their nature [1–3]. In addition, as a further simplification which facilitates a comparison with quantum mechanics systems, assume that the variables are binary taking on values from the set $\{0, 1\}$. Examples of such variables would be: rotations taken as clockwise (denoted by 1) or anticlockwise (denoted by 0) about the $i = 1, 2, 3$ axes, motions in the positive (denoted by 1) or negative (denoted by 0) direction on the $i = 1, 2, 3$ axes, etc. In the statistical description of the dynamics, the variables are random so that their occurrence in the system is found to be distributed by a probability function $P(S_1, S_2, S_3)$ which is generated by the physics contained within the dynamics.

As this is a problem in classical dynamics the fundamental properties of the system are contained within the probability function and its time development. The distribution function of the variables is then computed directly from the application of Newton laws applied to the knowledge of the initial dynamical configuration. Consequently, all of the average properties of the system are related to averages weighted by $P(S_1, S_2, S_3)$ as developed by the dynamics [1].

A particularly important aspect of the dynamics of the classical system is the nature of the correlations between the random variables. In particular, $\{S_1, S_2, S_3\}$ are random variables, but due to the interactions within the system they develop statistical correlations between one another. These correlations represent the nature of the dynamics and the degree to which the initial configuration of the system is understood.

For example, from $P(S_1, S_2, S_3)$ one can compute the probability that $S_1 = S_2$ or $S_1 = S_3$, i.e., that the indicated variables have the same numerical value. Note that this equality holds for the numerical values of the two different variables regardless of the fact that the variables are defined about different axes in space. For example, there may be two clockwise rotations in the system which occur about different axes of rotations or two axis along which particles move in the positive direction. These would have the same binary values while representing processes about different axes.

For the discussions presented later, denote these two probabilities as [1–3]

$$P_0(S_1 = S_2) \tag{10.1}$$

and

$$P_0(S_1 = S_3), \tag{10.2}$$

respectively. These two defined functions then, specifically represent the respective probability that $S_1 = S_2$ and the probability that $S_1 = S_3$ for the statistical distribution $P(S_1, S_2, S_3)$. It shall now be shown in the following that a complete understanding of the correlations of the random variables in the statistical distribution can be developed in terms of two variable distributions of the type in (10.1) and (10.2).

Some particular correlations of the random variables become important in understanding the difference between classical and quantum correlations. The focus in the following is to develop an inequality based on $P_0(S_1 = S_2)$, $P_0(S_1 = S_3)$, and $P_0(S_2 = S_3)$. This inequality will be used to develop and understand the fundamental difference between classical mechanical and quantum mechanical probability distributions.

To this end begin by considering the correlations of the type presented in (10.1) and (10.2), using them to determine the probability of finding systems satisfying $S_1 = S_2$ or $S_1 = S_3$. Applying basic probability reasoning, this probability is given by [1–3]

$$P_0(S_1 = S_2\, or\, S_1 = S_3) = P_0(S_1 = S_2) + P(S_1 = S_3) - P(S_1 = S_2 = S_3), \tag{10.3}$$

where $P_0(S_1 = S_2 = S_3)$ is the probability that $S_1 = S_2 = S_3$ in the system. The last term on the right enters into (10.3) so as to remove the double counting of the $S_1 = S_2 = S_3$ erms present in both of the preceding two terms on the right. Consequently, the probability $P_0(S_1 = S_2\, or\, S_1 = S_3)$ does not involve a double or over counting of states.

It then follows, denoting by $P_0(S_1 \neq S_2\ and\ S_1 \neq S_3)$ the probability that $S_1 \neq S_2$ and $S_1 \neq S_3$ in the system, that [1–3]

$$\begin{aligned} &P_0(S_1 = S_2 or S_1 = S_3) + P_0(S_1 \neq S_2 and S_1 \neq S_3) \\ &\quad = P_0(S_1 = S_2) + P(S_1 = S_3) - P(S_1 = S_2 = S_3) \\ &\qquad + P_0(S_1 \neq S_2 and S_1 \neq S_3) = 1, \end{aligned} \tag{10.4}$$

Equation (10.4) represents the statement that all configurations of the system are in one or the other distributions summed on its far left hand side. Consequently, it must follow from (10.4) that [1]

$$P_0(S_1 = S_2) + P_0(S_1 = S_3) + P_0(S_1 \neq S_2 \textit{ and } S_1 \neq S_3) \geq 1. \tag{10.5}$$

The first two terms involve the correlations $P_0(S_1 = S_2)$ and $P_0(S_1 = S_3)$ that are of interest, and it will now be shown that the third term on the left hand side of the inequality in (10.5) can be rewritten in terms of $P_0(S_2 = S_3)$.

To arrive at this further reduction in the inequality, note that the binary variables $\{S_1, S_2, S_3\}$ can only take the values $\{0, 1\}$. Consequently,

$$P_0(S_1 \neq S_2 \textit{ and } S_1 \neq S_3) = P_0(S_2 = S_3 \textit{ and } S_1 \neq S_2). \tag{10.6}$$

This is the statement that since neither S_2 nor S_3 can equal S_1 and the variables in the system can only take the values $\{0, 1\}$, it must follow that $S_2 = S_3$. In addition it follows from (10.6) that [1–3]

$$P_0(S_2 = S_3) \geq P_0(S_2 = S_3 \textit{ and } S_1 \neq S_2). \tag{10.7}$$

Here (10.7) follows from relaxing the $S_1 \neq S_2$ restriction in going from the right to the left side of the inequality.

Finally, from (10.4)–(10.7) it follows that [1–3]

$$P_0(S_1 = S_2) + P_0(S_1 = S_3) + P_0(S_2 = S_3) \geq 1, \tag{10.8}$$

which represents a statement of the Bell inequality for the system discussed in (10.1)–(10.8). This equation is fundamental to classical systems with binary coordinates. It is, however, an inequality which is not satisfied by a similar two level quantum system in which the probability distribution of the system is obtained in terms of a probability amplitude.

As an example of a classical mechanical system that obeys the inequality in (10.8), consider the problem of three coins. By tossing each of the three coins a distribution of heads and tails is generated for each coin. In particular, for tosses of the first coin of the system of three $S_1 = 1$ (heads) or 0 (tails). For tosses of the second coin $S_2 = 1$ (heads) or 0 (tails), and from tosses of the third coin again $S_3 = 1$ (heads) and 0 (tails). Computing the probabilities in (10.8) one finds for a large sequence of coin tosses that $P_0(S_1 = S_2) = P_0(S_1 = S_3) = P_0(S_2 = S_3) = \frac{1}{2}$ so that $P_0(S_1 = S_2) + P_0(S_1 = S_3) + P_0(S_2 = S_3) = \frac{3}{2} > 1$. This demonstrates (10.8) for the particular system of measurements made in tosses of the three coins. Next these considerations will be made for a system involving three different measurements on a quantum mechanical wave function.

The focus will now be towards demonstrating that the inequality in (10.8) fails in a quantum system involving measurements on spin one-half particles. The comparison will provide a distinction between classical and quantum systems.

To understand the difference between classical and quantum probabilities, consider the particular example of a set of two quantum mechanical spin one-half particles. This provides the simplest case of a wave function illustrating the difference between classical and quantum mechanical probabilities. For the set of two particles the measurement of three different spin related properties, denoted by the set of outcome values $\{S_1, S_2, S_3\}$, will be defined. It will then be shown that the wave function of the particles generates a probability distribution for the values $\{S_1, S_2, S_3\}$ which does not obey the inequality in (10.8). This will provide a contradiction between quantum and classical dynamics which distinguishes between the two theories.

First some of the properties of spin will be reviewed. These will then be applied directly in a study of two spin one-half particle wave functions as a test of (10.8) for a quantum system.

First consider some properties of the wave functions of a single spin one-half particle. For the particle the z-component of the spin one-half can be represented in Dirac notation as $|1\rangle$ for spin up and $|0\rangle$ for spin down. The set $\{|1\rangle, |0\rangle\}$ forms a complete orthonormal basis in the spin space of the particle. This, however, is not the only basis available for the study of the particle as any set of functions related to $\{|1\rangle, |0\rangle\}$ by a unitary transformation also is an acceptable basis for the system.

Consequently, an equally good basis is provide by the set $\{|1'\rangle, |0'\rangle\}$ where the primed basis are given by [1–3]

$$\begin{aligned} |0'\rangle &= \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle \\ |1'\rangle &= \frac{\sqrt{3}}{2}|0\rangle - \frac{1}{2}|1\rangle \end{aligned} \tag{10.9a}$$

with

$$\begin{aligned} |0\rangle &= \tfrac{1}{2}|0'\rangle + \tfrac{\sqrt{3}}{2}|1'\rangle \\ |1\rangle &= \tfrac{\sqrt{3}}{2}|0'\rangle - \tfrac{1}{2}|1'\rangle \end{aligned}. \tag{10.9b}$$

These basis states are orthogonal and normalized to unity, and they are related to the unprimed basis by a unitary transformation.

Similarly, another possible set of basis states is offered by $\{|1''\rangle, |0''\rangle\}$ where the double primed basis is related to the unprimed basis by

$$\begin{aligned} |0''\rangle &= \frac{1}{2}|0\rangle - \frac{\sqrt{3}}{2}|1\rangle \\ |1''\rangle &= \frac{\sqrt{3}}{2}|0\rangle + \frac{1}{2}|1\rangle \end{aligned} \tag{10.10a}$$

with

$$\begin{aligned} |0\rangle &= \tfrac{1}{2}|0''\rangle + \tfrac{\sqrt{3}}{2}|1''\rangle \\ |1\rangle &= -\tfrac{\sqrt{3}}{2}|0''\rangle + \tfrac{1}{2}|1''\rangle \end{aligned}. \tag{10.10b}$$

These basis states again are orthogonal and normalized to unity and related to the unprimed basis by a unitary transformation. The dynamics of the spin can be treated equally well in any of the three given basis choices.

Now in the unprimed, primed, and double primed bases it is possible to study the measurements of the binary variables $\{|1\rangle, |0\rangle\}$, $\{|1'\rangle, |0'\rangle\}$, and $\{|1''\rangle, |0''\rangle\}$ as they are statistically related to one another. A determination of the relative probability distribution of these measurements can then be made within the context of (10.8). This allows for an understanding of the correlated probabilities, providing for a comparison with the identity in (10.8) for the classical probabilities. For this comparison it is convenient to treat a two-particle wave function.

The two particle wave function that is of interest for this study is given by [1–7]

$$\begin{aligned}|\psi\rangle &= \frac{1}{\sqrt{2}}[|0\rangle|0\rangle + |1\rangle|1\rangle] = \frac{1}{\sqrt{2}}[|0'\rangle|0'\rangle + |1'\rangle|1'\rangle] \\ &= \frac{1}{\sqrt{2}}[|0''\rangle|0''\rangle + |1''\rangle|1''\rangle].\end{aligned} \tag{10.11}$$

This involves two identical particles in a combination of the $\{|1\rangle, |0\rangle\}$, $\{|1'\rangle, |0'\rangle\}$, or $\{|1''\rangle, |0''\rangle\}$ basis states. The second and third equalities in (10.11) can be shown by using (10.9) and (10.10) to rewrite all of the wave functions in terms of the $\{|1\rangle, |0\rangle\}$ states.

Consequently, the form of the wave functions in (10.11) is invariant under transformations between the three bases, i.e., it retains its form in all three bases. This is very useful in computing the probabilities of finding the basis states as components of the wave function.

The focus is next on looking at the nature of the probability distributions generated from the probability amplitudes in (10.11). A search will be made of three random variables in the system which are distributed in the system so as to violate the inequality in (10.8). This would represent a fundamental departure of the nature of the quantum system from that found in classical mechanical systems.

Considering (10.8) for the description of the statistics represented by the wave function in (10.11), define the random variable S_1 by [1–7]

$$\begin{matrix} S_1 = 1 \\ S_1 = 0 \end{matrix} \quad \text{if} \quad \begin{matrix} |1\rangle \\ |0\rangle \end{matrix}. \tag{10.12}$$

This represents a measurement in the unprimed basis. Similarly for the other two variables in (10.8) define [1–7]

$$\begin{matrix} S_2 = 1 \\ S_2 = 0 \end{matrix} \quad \text{if} \quad \begin{matrix} |1'\rangle \\ |0'\rangle \end{matrix}, \tag{10.13}$$

and

$$\begin{matrix} S_3 = 1 \\ S_3 = 0 \end{matrix} \quad \text{if} \quad \begin{matrix} |1''\rangle \\ |0''\rangle \end{matrix}. \tag{10.14}$$

where these variables are set to measure in the primed and double primed bases.

It is seen comparing (10.11)–(10.14) that the set of binary variables $\{S_1, S_2, S_3\}$ represents a complete description of the wave functions in (10.11) as measured in each of the three bases. This allows for the development of a probabilistic theory of the three random binary variables that can then be compared within the context of the inequality in (10.8).

The wave functions in (10.11) can be rewritten into mixed forms involving the $\{|1\rangle, |0\rangle\}$, $\{|1'\rangle, |0'\rangle\}$, or $\{|1''\rangle, |0''\rangle\}$ bases. For instance, in this way applying (10.9b) and (10.10b) gives

$$\begin{aligned} |\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|0\rangle + |1\rangle|1\rangle] = \frac{1}{\sqrt{2}}\Big[&|0\rangle\frac{1}{2}\left(|0'\rangle + \sqrt{3}|1'\rangle\right) \\ &+ |1\rangle\frac{1}{2}\left(\sqrt{3}|0'\rangle - |1'\rangle\right)\Big], \end{aligned} \tag{10.15a}$$

$$\begin{aligned} |\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|0\rangle + |1\rangle|1\rangle] = \frac{1}{\sqrt{2}}\Big[&|0\rangle\frac{1}{2}\left(|0''\rangle + \sqrt{3}|1''\rangle\right) \\ &+ |1\rangle\frac{1}{2}\left(-\sqrt{3}|0''\rangle + |1''\rangle\right)\Big], \end{aligned} \tag{10.15b}$$

and

$$|\psi\rangle = \frac{1}{4\sqrt{2}}\left[\left(|0'\rangle + \sqrt{3}|1'\rangle\right)\left(|0''\rangle + \sqrt{3}|1''\rangle\right) + \left(\sqrt{3}|0'\rangle - |1'\rangle\right)\left(-\sqrt{3}|0''\rangle + |1''\rangle\right)\right]. \tag{10.15c}$$

With these representations, the probability of finding the states $S_1 = 1$ and $S_2 = 1$ in the system is obtained by projecting out $|1\rangle|1'\rangle$ from (10.15a). This gives from (10.15a) [1–3]

$$P(S_1 = S_2 = 1) = |\langle 1|\langle 1'||\psi\rangle|^2 = \frac{1}{8}. \tag{10.16a}$$

Similarly, it follows from (10.15a) that [1–3]

$$P(S_1 = S_2 = 0) = |\langle 0|\langle 0'||\psi\rangle|^2 = \frac{1}{8}. \tag{10.16b}$$

and, consequently,

$$P(S_1 = S_2) = \frac{1}{4}. \tag{10.17}$$

By the same argument it follows from (10.15b) that

$$P(S_1 = S_3) = \frac{1}{4}, \tag{10.18}$$

and from (10.15c) that

$$P(S_2 = S_3) = \frac{1}{4}. \tag{10.19}$$

As a result in the quantum system

$$P_0(S_1 = S_2) + P(S_1 = S_3) + P_0(S_2 = S_3) = \frac{3}{4}. \tag{10.20}$$

Comparing this with the result in (10.8) it is seen that the classical mechanical relationship in (10.8) is violated in the particular spin system studied. This is an essential difference between the classical and quantum mechanical probabilities, arising from the fact that the fundamental element in quantum probabilities is the wave function or probability amplitude whereas classical mechanics deals directly with the probability distribution.

10.2 Entanglement and the Einstein-Podolsky-Rosen Paper

Another important aspect of quantum probabilities as they relate to the wavefunction is the idea of entanglement and its connection to an early paper of Einstein, Podolsky, and Rosen [1–7]. Entanglement in quantum mechanics arises from the unusual nature of many-body wavefunctions as they are developed in quantum theory. The ideas of the Einstein, Podolsky, and Rosen paper come from a conflict with the qualitative properties found between classical mechanical and quantum mechanical systems studied in many-body theory. These differences were originally regarded as paradoxical in early treatments of quantum theory. In the following, after a brief description of the features of entangled wavefunctions, the ideas of Einstein, Podolsky, Rosen will be summarized. This will be followed by a discussion of some of the current theoretical measures and problems associated with the measurement of entanglement.

10.2.1 Nature of Entangled and Non-entangled State Wavefunctions

Entangled states first occur in the theory of two identical particles. However, it is in fact found that entangled states can exist in the wavefunctions of all systems of more than one particle. The entanglement properties of two particle states are the easiest to study and to offer an effective measurement procedure for quantifying the degree of entanglement. That is why they will be the focus here. The study of the entanglement of systems of higher numbers of particles and the measure of the degree of entanglement in these wavefunctions is still problematic.

For these discussions, useful examples of multiple particle wavefunctions can be composed from spinor states. In this formulation, a two-particle entangled wavefunction is composed from the inner product of two single spin one-halves. Consequently, if the single spin states are denoted $|1\rangle$ for spin up and $|0\rangle$ for spin down, the two particle states are composed as linear combinations of the inner product states $|1\rangle|1\rangle$, $|1\rangle|0\rangle$, $|0\rangle|1\rangle$, and $|0\rangle|0\rangle$.

First consider the difference between an entangled and a non-entangle wave function for the case of a system of two spins. A two particle wave function $|\psi\rangle$ for particles with single spin coordinates $|1\rangle$ for spin up and $|0\rangle$ for spin down is said to be a non-entangled wave function provided that [1–7]

$$|\psi\rangle = |\alpha\rangle|\beta\rangle, \tag{10.21a}$$

where $\alpha = 0, 1$ and $\beta = 0, 1$ denote the spin of the wave function for the single particle states or

$$|\psi\rangle = [a_1|1\rangle + a_0|0\rangle][b_1|1\rangle + b_0|0\rangle], \tag{10.21b}$$

where $|a_1|^2 + |a_0|^2 = 1$ and $|b_1|^2 + |b_0|^2 = 1$. In this representation, then, in (10.21a) α labels the first particle and β labels the second particle, respectively, and, similarly, the first and second brackets on the left in (10.21b) represent the first and second particles, respectively. In the second example, it is seen that the wavefunctions of the first and second particles are separately linear combinations of different states of the same particle. In both cases the two particle wave function separates into a product of a wavefunction of the first particle with that of the second particle.

It is seen in the non-entangled state that the two particle wavefunction can be factorized into the product of two single spin wavefunctions for each of the particles composing the two particle wavefunction. As a result, the probability density obtained from (10.21a) has the form

$$|\langle\vec{r}_1, \vec{r}_2|\psi\rangle|^2 = |\langle\vec{r}_1|\alpha\rangle|^2|\langle\vec{r}_2|\beta\rangle|^2 \tag{10.22}$$

so that the two particle probability density is the product of two separate single particle probability densities. Here $\langle \vec{r}_1|$ and $\langle \vec{r}_2|$ are the position states for the first and second particle, and $\langle \vec{r}_1, \vec{r}_2|$ is the position of the two particles taken together. It should also be noted that the wavefunction in (10.21b) displays a similar separation. The two particle wavefunctions in (10.21) are not, however, the most general form of the two-particle wavefunction.

Consider for example the following two particle wavefunctions:

$$|\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|0\rangle + |1\rangle|1\rangle], \tag{10.23a}$$

$$|\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|0\rangle - |1\rangle|1\rangle], \tag{10.23b}$$

$$|\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|1\rangle + |1\rangle|0\rangle], \tag{10.23c}$$

and

$$|\psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle|1\rangle - |1\rangle|0\rangle]. \tag{10.23d}$$

These four two particle wavefunctions are known as the Bell states [1–7] for two spin one-halves and are examples of entangled wavefunctions. The reason that they are termed entangled is that they cannot be factorized into the product of two wave functions each of which separately account for the probabilities of only the first or only the second particle of the pair. Only non-entangled wavefunctions can be rewritten so that their probability densities can be expressed as the product of a probability distribution for the first spin times a probability distribution for the second spin. This is the form developed earlier in (10.22).

As shall be seen later, the Bell wavefunctions involve a state of maximal entanglement. The sense of the degree to which the Bell wavefunctions are entangled can be seen from the following observations: For each spin in the Bell wavefunction the probability of observing that spin in a spin up state is $\frac{1}{2}$, and the probability of observing the other spin in a spin down state is again $\frac{1}{2}$. Consequently, a measurement of the spin state of either of the two spins generates a maximum information regarding the spins of the Bell states. Much more information is arrived at on determining a measurement of a Bell state than, for example, a non-entangled state such as $|\psi\rangle = |1\rangle|1\rangle$.

10.2.2 Einstein, Podolsky, and Rosen

An interesting property associated with wavefunctions of the form given in (10.23) comes from the measurement properties of the entangled states that they represent. Consider, as an example, a measurement made on the wavefunction in (10.23c).

If the first particle of the pair is measured and found to be in a state of spin up, the second particle must be in a state of spin down. Likewise, if the first particle is in a state of spin down, the second particle must be in a state of spin up. In this way, a measurement of one particle sets the value displayed by the other particle, and the measurement of the first particle is a random probabilistic event.

This was viewed as paradoxical because two particles in different locations could be viewed as acting instantaneously with one another in the determination of their properties. However, as Bell's work indicates the paradoxical behavior is, in fact consistent with physical reality. Through many years the underlying entanglement properties of the quantum mechanical wavefunctions have been proven experimentally, and the reader is referred to the literature for a further discussion.

The degree to which wavefunctions represent non-entangled or entangled states can be quantified. This quantification shall now be discussed in terms of the von Neumann entropy and the entanglement entropy.

10.2.3 Measurements of Entanglement

An important measure of the information in a probabilistic system is the entropy of the probability distribution [4–10]. This also applies to quantum mechanical systems where it is found that entropy can be used to characterize the nature of many particle quantum wavefunctions. In particular, the entanglement entropy of the probability distribution of non-entangled states and, consequently, the information content of their wavefunctions are less than those of entangled states. The entanglement entropy then acts as a measure of the degree to which a system is entangled. The focus in the subsequent discussions will be on the development of the details of the application of entropy as a measure of entanglement.

The idea of entropy is very important in the study of probability and is a topic of the general mathematics of probability and statistics. In the following it shall be seen that the ideas of entropy can be applied to a variety of measures of statistically distributed systems. These include both the von Neumann entropy and the entanglement entropy. For these applications, the formulation of the entropy of a system is given in terms of the density matrix description of the system. The ideas of the density matrix, entropy, and their applications in the determination of entanglement are now reviewed.

Density Matrix

An important construction in understanding the quantum mechanical properties of systems is the density matrix. This plays an important part in characterizing the nature of wavefunctions of a quantum mechanical system as well as the degree to which a wavefunction is entangled. In the following a focus will be on exploring these ideas for a two-particle or bipartite system.

To begin with, remember that the density matrix is written in terms of the complete set of eigenfunctions of the system, $\{|n\rangle\}$. In the context of the later discussions, it can be defined to be of the form [1–7]

$$\tilde{\rho} = \sum_n \rho_{n,n}|n\rangle\langle n| \tag{10.24}$$

where the $\{\rho_{n,n}\}$ weight the $\{|n\rangle\langle n|\}$ and are the probabilities of occurrence of the $\{|n\rangle\}$ in the system. The weights are chosen so that the average of an operator $A = \sum_{i,j} a_{i,j}|i\rangle\langle j|$ in the system is written in terms of the density matrix as

$$\langle A\rangle = \mathrm{tr}(Ap) = \sum_n a_{n,n}\rho_{n,n} \tag{10.25}$$

where tr represents the trace over $\{|n\rangle\}$. Note that the $\{\rho_{n,n}\}$ may be chosen to represent purely quantum mechanical averages, statistical mechanics averages, averages related to instrumental measurements, etc.

Written in terms of a general orthonormal basis set $\{|n'\rangle\}$ the density matrix becomes [1–7]

$$\begin{aligned}\tilde{\rho} &= \sum_{n',m'}\sum_n |n'\rangle\langle n'|\rho_{n,n}|n\rangle\langle n||m'\rangle\langle m'| \\ &= \sum_{n',m'} |n'\rangle\rho'_{n',m'}\langle m'|\end{aligned} \tag{10.26}$$

where $\rho'_{n',m'} = \sum_n \langle n'|n\rangle\rho_{n,n}\langle n|m'\rangle$. Consequently, the density matrix is only diagonal in the basis of eigenvalues of the Hamiltonian of the system. It assumes a more complex form in other bases.

For a system of two non-interacting spin one-half particles a complete set of basis eigenstates states of the free particle Hamiltonian is

$$\left\{|\phi_1\rangle = |1\rangle|1\rangle, |\phi_2\rangle = |0\rangle|0\rangle, |\phi_3\rangle = \frac{1}{\sqrt{2}}(|0\rangle|1\rangle + |1\rangle|0\rangle), |\phi_4\rangle = \frac{1}{\sqrt{2}}(|0\rangle|1\rangle - |1\rangle|0\rangle)\right\}. \tag{10.27}$$

Each of these four wave functions is a state of fixed quantum numbers of the total spin and total z-component of the two spins. Consequently, any of these four states can be taken as a pure state of the system with unique values of the quantum numbers of the system. The density matrix for the system in any of these pure system eigenstates would then be, from (10.24), written in the form

$$\rho_i = |\phi_i\rangle\langle\phi_i| \tag{10.28}$$

where $i = 1, 2, 3, 4$.

A different type of state of the quantum system from pure states are mixed states. These are states that arise including considerations apart from those of quantum theory. For example, the system may involve statistical mechanics in its description so that a statistical mechanical probability is assigned to the possible quantum states in which the system may be found. Another example arises from the uncertainty in setting the initial configuration of the quantum mechanical system being studied. These states are, consequently, not simple pure wavefunctions of the system. As a result of these additional consideration, the mixed states require more complex density matrices than those encounter in pure states of the system.

As an example, consider a two-spin system that can occur in the state $|\phi_1\rangle$ with a statistical probability p_1 and in a state $|\phi_2\rangle$ with a statistical probability p_2. Here the two probabilities $p_1 + p_2 = 1$ are classical probabilities not related to the quantum mechanics of the system. They arise as the state of the system is not well defined by the quantum mechanics alone but represents a mix of states with different sets of quantum numbers.

The density matrix for the mixed mode of the example is then written as [1–7]

$$\rho = p_1|\phi_1\rangle\langle\phi_1| + p_2|\phi_2\rangle\langle\phi_2| \tag{10.29a}$$

which can alternately be rewritten in the notation of the full four by four matrix form

$$\rho = \begin{vmatrix} p_1 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix} \tag{10.29b}$$

Note that as a further example of the matrix formulation: For the case of the pure state $\rho_2 = |\phi_2\rangle\langle\phi_2|$, $p_1 = 0$ and $p_2 = 1$ in (10.29b), and for the case of the pure state $\rho_1 = |\phi_1\rangle\langle\phi_1|$, $p_1 = 1$ and $p_2 = 0$ in (10.29b).

It should be noted that for the pure state it follows from (10.28) that $\rho_i^2 = \rho_i$. On the other hand, however, from (10.29) it is found that for mixed states $\rho^2 \neq \rho$. This is a general relationship that allows for a distinction between pure states and mixed states. Nevertheless, a more useful relationship for quantifying the difference between pure and mixed state wave functions follows from the mathematical study of probability distributions. Specifically, this quantification involves the introduction of the concept of statistical entropy and it relationship to mixed and pure states.

The states encountered in the description of a physical systems then fall into three different types. In pure states, the states can be classified into entangled and non-entangled states. Here the entangled states are those new states introduced in the study of dynamics by the nature of quantum theory and are quantified by the entanglement entropy. In addition, states of the system can be classified as mixed states. In these states, other concepts of statistical physics enter aside from those of quantum theory. These involve the von Neumann entropy.

Entropy

In the mathematical study of probability distributions an important question is how much information is learned about a system upon making a measurement which determines its state [4–6]. If the probability distribution of the system studied ranges over a very narrow group of states exhibiting very similar properties, the information derived from a measurement is not very great. On the other hand, if the probability distribution of the system studied ranges over a very broad group of states exhibiting very dissimilar properties, the information derived from a measurement is very great.

A means of quantifying the information obtained from a measurement made on a system characterized by a probability distribution is the entropy. The entropy of a probability distribution $P(s_i)$ generated over the distributed random variables $\{s_i\}$ is defined by [1–7]

$$S = -\sum_i P(s_i) \ln P(s_i) \tag{10.30}$$

where the sum is over all of the different realizations of the random system. Here the set $\{s_i\}$ can represent the spin values of a collection of particles or some other array of values used to characterize each realized configuration possible to the system. The sum is, then, over each of the i labeled different configurations. In terms of the density matrices in (10.28) and (10.29) of the two-particle spin one-halves this entropy becomes

$$S = -\mathrm{tr}[\rho \ln \rho] \tag{10.31}$$

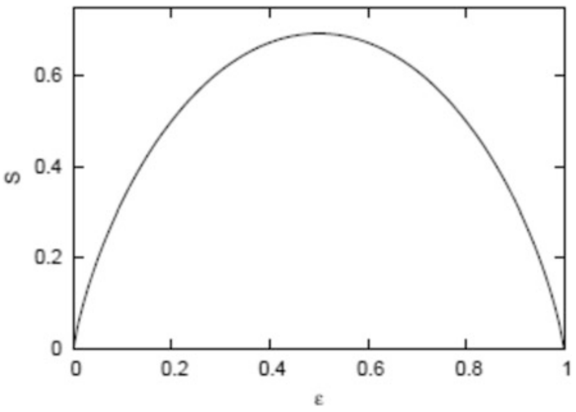
As an example of the properties of the entropy of the two-spin system in (10.31) note that for the pure state in (10.28) $S = 0$. The zero entropy indicates that there is no information content in the pure system. This makes sense as if the system is in an eigenmode of the system the result of a measurement on the mode yields no new information about the system. Since the properties of the eigenmode are completely known, the lack of information output is indicated by the zero entropy.

For the mixed state wave function described in (10.29b), however, the situation is quite different. Evaluation (10.31) for the density matrix in (10.29b) gives an entropy of the form

$$S = -p_1 \ln p_1 - (1 - p_1) \ln(1 - p_1). \tag{10.32}$$

In Fig. 10.1 a plot is presented of S as a function of $\varepsilon = p_1$.

Fig. 10.1 Plot of the von Neumann entropy versus $\varepsilon = p_1$ for the system in (10.29)



From the plot in Fig. 10.1 it is seen that, as expected, the entropy of the pure states at $\varepsilon = 0$ and $\varepsilon = 1$ are zero. No information is gathered by measurements on these states. Between these states are a range of mixed states. The information gathered from measurements on these states varies with the degree of mixing of the states.

At $\varepsilon = 0.5$ the wavefunctions contains equal amounts of the two eigenmodes composing the wave function, and it is for this wavefunction that the most information is gleaned upon performing a measurement on the wave function. At other fractional values of ε the mixture contains one of the eigenmodes of the system to a greater or lesser extent that the other. Consequently, the result of a spin measurement on the mixed state offers less of a surprise or less information regarding the result of the measurement.

The entropy measured on the total density function of the system is the von Neumann entropy, and it is essentially involved in measuring the amount of mixing of the wavefunction represented by the density matrix. The larger the non-zero entropy of the system the greater is the mixing of eigenstates in the wavefunction representing the state of the system.

The next type of entropy that is important in characterizing the system is the entanglement entropy. This type of entropy or information measurement quantifies the degree to which a state of the system is entangled. This is a different type of information than that from the von Neumann entropy which relates to the extent that the wavefunction is formed as a mixture of eigenstates.

Remember that a non-entangled eigenstate of a system of multiple particles is one that generates a probability distribution for the particles which is a product of independent single particle probability distributions. For this type of eigenstate the probability amplitude describing the state of the system also factorizes into a product of independent probability amplitudes for each of the particles of the system. The important new properties of quantum systems arise from the existence of entangled wavefunctions which do not exhibit these types of factorization properties.

The entanglement entropy is designed to measure the degree of entanglement of the wavefunctions of the system. In the following, the considerations will be for a two-particle spin one-half system. The question of how to quantify the entanglement of systems with higher number of particles is still an open question.

One can see straightaway that if a non-entangled eigenstate of the form $|\psi\rangle = |\varphi\rangle|\lambda\rangle$ is treated, it has a density matrix given by [1–7]

$$\rho = |\varphi\rangle|\lambda\rangle\langle\lambda|\langle\varphi|. \tag{10.33}$$

On tracing out the $|\lambda\rangle$ states of the second particle it is found that

$$\rho_1 = |\varphi\rangle\langle\varphi|. \tag{10.34}$$

which is the density matrix of the pure state of the first particle. The reason for this reduction is that the first and second particles obey distinct and independent probability distributions.

The information content of the reduced density matrix in (10.34) can be computed using the entropy defined in (10.31). This is done by replacing the density matrix in (10.31) by the identification $\rho = \rho_1$ and taking the trace over the states of the first particle of the pair. Computed in this way the entropy is known as the entropy of entanglement or entanglement entropy.

In this way, the entropy of ρ_1 is found to be $S = 0$. Consequently, no new information is gathered on the state as it is one of the two states described by an isolated independent particle distribution. Similarly, all of the same information can be extracted from (10.33) by tracing over the first particle instead of the second particle.

Performing this trace one finds for the first particle that the reduce density matrix is [1–7]

$$\rho_2 = |\lambda\rangle\langle\lambda|. \tag{10.35}$$

Using (10.31) with $\rho = \rho_2$ again gives $S = 0$. No new information is obtained from the reduced density matrix of the non-entangled system. This, however, is not the case if the two-particle wavefunction is entangled.

The reduction of the two-particle density matrix to a single particle density matrix by tracing over the basis for one of the particles does not occur for two particle density matrices involving entangled states. For example, consider an eigenstate with a two-particle entangled wavefunction of the form [1–7]

$$|\psi\rangle = \frac{1}{\sqrt{(1-\varepsilon)^2 + \varepsilon^2}}[(1-\varepsilon)|\varphi\rangle|\lambda\rangle + \varepsilon|\lambda\rangle|\varphi\rangle] \tag{10.36}$$

where $0 \le \varepsilon \le 1$ and $\langle\lambda|\varphi\rangle = 0$. Now the density matrix is

$$\rho = |\psi\rangle\langle\psi|. \tag{10.37}$$

so that tracing over the second particle variables gives

$$\rho_1 = \frac{1}{(1-\varepsilon)^2 + \varepsilon^2}\left[(1-\varepsilon)^2|\varphi\rangle\langle\varphi| + \varepsilon^2|\lambda\rangle\langle\lambda|\right] \tag{10.38}$$

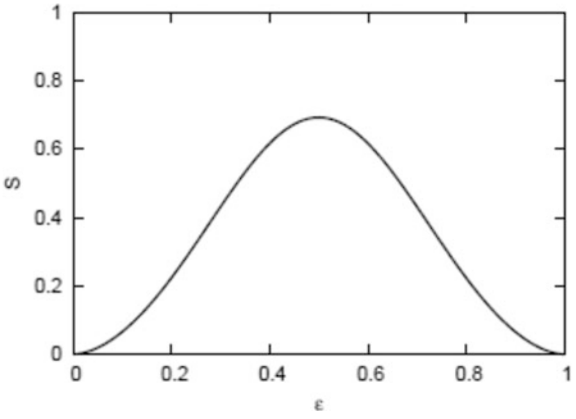
Applying the standard statement of entropy given in (10.31) but now on the reduced density in (10.38) obtained by tracing over the second of the two particles gives [1–7]

$$\begin{aligned} S &= -\mathrm{tr}[\rho_1 \ln \rho_1] \\ &= -\frac{(1-\varepsilon)^2}{(1-\varepsilon)^2+\varepsilon^2}\ln\left[\frac{(1-\varepsilon)^2}{(1-\varepsilon)^2+\varepsilon^2}\right] \\ &\quad -\frac{\varepsilon^2}{(1-\varepsilon)^2+\varepsilon^2}\ln\left[\frac{\varepsilon^2}{(1-\varepsilon)^2+\varepsilon^2}\right] \end{aligned} \tag{10.39}$$

Notice that for $\varepsilon = 0$ and $\varepsilon = 1$ (10.38) is a non-entangled state and the entropy is zero. Otherwise, however, the entropy is non-zero and the state is entangled. Specifically, for $\varepsilon = \frac{1}{2}$ the system is maximally entangled. This is seen from Fig. 10.2 where the entropy is plotted as a function of ε.

In the case of the Bell states given in (10.23), each of the entangled states is a maximum of the entanglement entropy. This is obtained from an application of the methods used in (10.36)–(10.39) to the states in (10.23) and is at the point of interest in the study of the Bell states. The entanglement properties of wavefunctions involving more than two-particles is currently still a focus of research interests and as such remains an open problem.

Fig. 10.2 Plot of the entropy S in (10.39) versus ε



10.3 Quantum Information and Computing

In quantum information and quantum computing, information is stored and acted upon in physical structures for which quantum effects are dominant [7–10]. Such physical realizations are often taken to be composed of two level features, i.e., they are systems formed as a collection of physical units that can each exist in two different states. Examples of such materials are a collection of spin one-half particles, a collection of atoms or ions that can exist in a ground state-excited state complex, light of various polarizations, Josephson junction arrays, etc.

The basic idea in forming these collects of binary units is that information can be encoded into the array of physical units. This is done by reading the information into the level occupancy distributed throughout the two-level units forming the system. In this assignment, the two levels of the array are denoted by 1 and 0, allowing them to represent the binary encoding of information that is the basis for information handling in computer science. The read-in information can at some later time be readout of the array or acted upon by a physical process so as to change the information content within the array.

Not only can information be stored in arrays of two levels but it can also be changed. This is accomplished through the application of various physical processes which interaction with the two-level states. Such processes can be designed to change the occupancies of the two-level states of the system and reconfigure them into another collection of 1's and 0's from that of the initial read-in state. Consequently, an important fundamental element in the design and implementation of the quantum computer programs is the formulation of processes which give rise to specific changes of the information content of the array of two levels.

In order to understand these basic ideas of quantum computing in more detail, it is first necessary to consider some of the aspects of information in computers based on classical mechanics [7–10]. The ideas of quantum computers are then seen as an extension of classical computer ideas to take advantage of the phase coherence and the unitary time evolution of quantum systems.

These two quantum properties allow for the development of algorithms for quantum computing. In many instances quantum algorithms exhibit the great advantage of quantum computers over those of classical computers. To this end, in the following, some ideas of classical computers will be developed. After this a development and comparison of the ideas of classical computation will be made with quantum computer methods. At the end, some examples of problems which benefit from quantum algorithms are presented.

10.3.1 Ideas of Classical Computers

The handling of information in classical computers is based on storing the information in a pattern of 0's and 1's. In this design, a mechanism described by classical physics is employed as a basic unit of the computer and the computer is composed as an array of such basic units. In its formulation the basic unit consists of two states signifying the presence of 1 or 0 held in the memory of the computer [7–10].

Each such unit for information storage in the classical computer is referred to as a bit, and the computer must have many bits to effectively hold significant information. In particular, a bit represents a bit of the total information to be stored, and the information that can be held in a single bit is only one of 1 or 0. As shall be seen later the classical bit is quite different from the unit of information storage (termed the qubit) in a quantum computer [7–10].

A collection of complex information is stored in the classical computer by assigning values to a large array of bits. This information can be either read out later or subsequently used as input for a computation.

In the case that the information is input data for a calculation, the computer can go further and act to modify the input data. For this modification, it uses classical mechanical based processes to arrive at a final state of the information. This is then the act of computation which is governed by an algorithm specifying how the computer should change the inputted data.

As with the input data, the final state of information after a computation is stored as an array of 1's and 0's. This information can then be outputted from the computer as the answer of the particular processing task assigned on the initial data.

For the purposes of processing the inputted data only a limited number of operations are available to the classical computer. These operations take the sequence of 1's and 0's in the data and change them into other patterns of 1's and 0's. For processing of the data it can be shown that only two classes of basic types of operations are needed. These involve operations which change the value of a single entry of the input or operations which take two entries of the input and use them to generate a resultant third value as an output. A vast sequential assembly of such single and binary operations then constitute an algorithm for generating the computer output.

Elaborate algorithms are written for the processing of the input into output data. It can, however, be shown that the most complicated algorithm is only based on a limited number of different basic standardized logic operations or gates. A general algorithm is then the repetition over and over of these basic gates.

This is similar to the idea of biological processes based on the molecules of DNA and RNA. These molecules are both formed as the repetition of a long sequence of chemical units involving a limited number of chemical building blocks, i.e., the amino acids. When put together as a sequenced chain both molecules function as an algorithm for a molecular assembly process in the biological cell. The resulting molecules act as a computer in the generation of other molecular forms.

In the formulation of computer algorithms it can be shown that all calculations that can be performed by a computer can be composed from three logic gates [7–10]. These three important operations are the N-gate, the AND-gate, and the OR-gate. Their function is similar to that of the molecular units of DNA and RNA in the biological system, specifically, to decompose the input data and reassemble it into a useful output. In the following these three gates will be explained. After this some examples of more complicated classical gates that can be expressed in terms of them will be given.

The simplest logic gate of a classical computer is the N-gate [7–10]. It involves an operation which deals with only a single value or bit of the inputted data. In this sense the functioning of the N-gate is very simple. It just reverses the value of the bit (which is either 1 or 0) that it is operating on (i.e., replacing it by 0 or 1). This can be expressed in tabular form by the rule [8] (Table 10.1). If a is the input 0 or 1, then $b = \bar{a}$ is the output 1 or 0, respectively. Note that in this formulation the bar over the a indicates that the value of a is replace by its opposite value.

The next two gates of interest for classical computing involve operations between two values or bits of the inputted data. These bits are then used to produce a third bit as a contribution to the output. The two basic binary logic gates for these types of operations are the AND- and the OR-gates [8].

The binary operation of the AND-gate is given in tabular form by [8] (Table 10.2). Here the product is essentially the logic multiplication defined in Boolean algebra. The table can alternately be viewed as a defined tabular operation which functions as part of a scheme to generate more complex processes.

The final binary gate in the set of algorithm generating processes is the OR-gate. This gate is represented by the logic operation given in tabular form by [8] (Table 10.3). Here the addition is essentially the binary addition of Boolean algebra. Again the table can be regarded as a rule yielding an essential component of the formation of more complex algorithms.

Table 10.1 N-gate

a	$b = \bar{a}$
0	1
1	0

Table 10.2 AND-gate

a	b	$c = ab$
0	0	0
0	1	0
1	0	0
1	1	1

Table 10.3 OR-gate

a	b	$c = a + b$
0	0	0
0	1	1
1	0	1
1	1	1

Table 10.4 XOR-gate

a	b	$c = a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

As an example of an algorithm composition from these three logic gates, consider another important logic-gate operation which can be written in terms of the N-, AND-, and OR-gates. A gate that shall be important later is the exclusive-or-gate. This is often denoted as the XOR-gate. The table representing the XOR-gate is given by [8] (Table 10.4). In terms of the Boolean addition and multiplication processes and the N-gate defined earlier, the XOR table can be expressed as an algebraic form in these gates. Specifically, it is found that

$$c = a \oplus b = \bar{a}b + a\bar{b}. \tag{10.40}$$

The reproduction of the $a \oplus b$-table then follows from an application of the earlier discussed tabular rules.

The three logic-gates that form the basic units in algorithm composition display an important characteristic of classical computation. Specifically, the execution of a classical algorithm is inherently not reversible. Not all calculations can be reversed in the sense that one can go from the output of the computer back to the input data by taking a reversed sequence of the logic-gates.

In this regard, though the N-gate is a reversible operation, the other two AND and OR-gates are not reversible. In the case of the N-gate, its reversibility follows from the fact that [8]

$$a = \bar{\bar{a}}. \tag{10.41a}$$

However from an inspection of the logic-gate tables it is seen that in the case of the AND-gate [8]

$$c = ab = 0 \tag{10.41b}$$

and in the case of the OR-gate [8]

$$c = a + b = 1 \tag{10.41c}$$

do not allow for a unique determination of a and b for the indicated values of c. These two gates cannot be operated in reverse so that the operations of general algorithms tend to be irreversible processes. As shall be seen later, this has consequences in the compositional and thermodynamic properties of algorithms implemented using classical gates.

Returning to the composition of algorithms that transform input data to a set of output data, consider such a general process in a little more detail. This will later be used to reveal the essential differences between classical and quantum computation.

In the process of classical computation an input array of 1's and 0's is acted upon by the computer to generate an output array of 1's and 0's. As a simple example of such a process consider the reversible transformational process [8]

$$|0,1,0,0,0,1| \Leftrightarrow |1,0,1,1,1,0|. \tag{10.42}$$

Here the input data is on the left of the ⇔ sign and the outputted data is on its right. The sign ⇔ indicates all of the various algorithmic steps done by the computer to change the input data into the output data.

In the present example one can see that each bit in the input data has been acted upon by an N-gate to reverse its value. Consequently, the processes is reversible as each N-gate application is reversible. As the net transformation is reversible and can go either way, it is denoted by the ⇔.

As an example of an irreversible process consider the transformation [7, 8]

$$|0,1,0| \Rightarrow |0,0|. \tag{10.43a}$$

Here the binary operation AND-gate has operated on the first two bits of the input to produce the first entry of the output. The AND-gate is known to be irreversible so that the operation in (10.43a) cannot be inverted. In this regard, for example, the transformation [7, 8]

$$|1,0,0| \Rightarrow |0,0|. \tag{10.43b}$$

leads through the application of the AND-gate on the first two bits to the same output data. The inputs in (10.43a) and (10.43b) are completely different but, nevertheless, yield the same output.

Another interesting feature of (10.43) is that the final 0 in the input and output is left unchanged by the transformation. This can occur in certain computations which use a value of one of the inputs as an indicator or control of the operation to be performed. For example, the 0 in the last place of the input in (10.43) may be used to indicate that the output of the binary action on the first two bits of the input is computed using the AND-gate.

Assume then that if a 1 occurs in the last entry of the input, the AND-gate operation in (10.43a) and (10.43b) is replaced by the OR.-gate. Here again the 1 is acting as a control to indicate the action to be performed. In this case one would find [7, 8]

$$|0,1,1| \Rightarrow |1,1| \tag{10.43c}$$

and

$$|1,0,1| \Rightarrow |1,1|. \tag{10.43d}$$

where the first two bits on the left determine the first bit on the right through the application of the OR-gate. The last bit remains unchanged as a control.

The results in (10.43) indicate how the application of a binary process may differ depending on the nature of a control statement. In addition, it should be noted that all of the processes in (10.43), due to the irreversible nature of the AND and OR-gates, are irreversible.

As another example of a computational process that will be helpful in understanding the potential power of quantum computation, consider computations that have a final output or answer that is in the form of a yes or no statement about the data inputted into the computer algorithm. For example, given a function of the form [7]

$$y = f(x), \tag{10.44}$$

an interest may be in determining if x is a value for which $f(x) > 0$, whether or not $f(x)$ is an even or odd function, whether or not $f(x)$ is a periodic function, etc. All of these questions have yes or no answers and, consequently, 1 or 0 data outputs.

Since the values of x can be represented in the base 2 by a string of 1 and 0, it becomes necessary to design an algorithm in which a string of 1 and 0 are inputted, and the values 1 and 0 for yes and no are outputted. This is represented in a typical irreversible process of the form [7, 8]

$$|0,1,0,1,1,0| \Rightarrow |1| \tag{10.45a}$$

or

$$|0,1,0,0,1,0| \Rightarrow |0| \tag{10.45b}$$

In classical computation the algorithm generally functions by going through a sequence of x's, computing their associated values of $f(x)$, and testing whether or not the conditions on $f(x)$ are met by the calculated value. The determination is done in a loop type of process which is not cleaver but relies on brute sequential computations to find an answer. On the whole, this can be very inefficient to realize on a classical computer, even to the extent that it may not be practical to implement with a finite CPU time.

Turning to quantum computation, it will be seen that quantum computations are quite different than those in classical systems. To begin with, ideally quantum computations are based on unitary, reversible, processes. Irreversibility, in fact, is not part of the design of the quantum computer logic-gates but is a problem to be overcome in quantum computers. In addition, due to the entanglement properties of quantum mechanics, the nature of the quantum bit in quantum computing (known as the qubit) is different than the bit in classical computing. These feature will now enter into the discussions as the treatment turns to an explanation of the basics of quantum computing.

10.3.2 Quantum Computing

In quantum computing the basic unit of information storage is not the classical bit but the quantum mechanical qubit [7–10]. Similar to the case of the classical system, the 1's and 0's of the quantum computer data are denoted by the states of two level units, and the computer medium is composed of an array of many such individual units. The basic unit composing the array may be, for example, the spin up and spin down state of a spin one-half particle, an excited state and ground state of a trapped ion, the polarization of light, or any other component that can exist in two different states. Due to the nature of quantum mechanics, however, the two-level units in the quantum mechanical array exhibit completely different properties from their classical mechanical counterparts.

Unlike classical mechanical systems, quantum mechanical systems can exist in a linear combination of eigenstates. This is known as a superposition of basis states. For example, if $|1\rangle$ and $|0\rangle$ are the wave functions of an orthonormal basis of a two level quantum unit of the system, a wavefunction composed as a linear combination of basis states of the unit may be composed as [7–10]

$$|\alpha\rangle = \alpha_0|0\rangle + \alpha_1|1\rangle \tag{10.46}$$

where α_0 and α_1 are complex numbers satisfying $|\alpha_0|^2 + |\alpha_1|^2 = 1$.

An array of the quantum system composed solely of basis states would look like, e.g.,

$$|0,0,1,1,0,1,0\rangle = |0\rangle|0\rangle|1\rangle|1\rangle|0\rangle|1\rangle|0\rangle \tag{10.47}$$

where the righthand side of the equation represents a direct product of the wave functions of an array of seven different quantum units, and the left-hand side is a compact notation for the direct product defined on the right. This particular array is essentially representing the sequence of 1's and 0's given in the classical data portrayed by the data set [7–10]

$$|0,0,1,1,0,1,0|. \tag{10.48}$$

From (10.47) and (10.48) it is found that the classical representation of data by 1's and 0's is contained as a subset of the wave functions of an array of quantum mechanical two level states. The particular subset of states must always be represented as a direct product involving only the orthonormal set $\{|0\rangle, |1\rangle\}$ of each of the two-level units of the array. Furthermore, it must not involve linear combination of such direct product representations. Due to the possibility of forming wave functions as linear combinations of quantum states, however, the available states for storing data is much increased in the full quantum system from that of the classical represented data just discussed.

When using the full set of quantum states of the two-level units forming the array of the quantum system, the data array of the system is much extended from the subset of classical-like states. In particular, the state in (10.47) can be generalized to wave functions formed as linear combinations of basis states [7–10]. These may be written in the general form

$$|\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4, \tilde{\alpha}_5, \tilde{\alpha}_6, \tilde{\alpha}_7\rangle = |\tilde{\alpha}_1\rangle|\tilde{\alpha}_2\rangle|\tilde{\alpha}_3\rangle|\tilde{\alpha}_4\rangle|\tilde{\alpha}_5\rangle|\tilde{\alpha}_6\rangle|\tilde{\alpha}_7\rangle. \tag{10.49a}$$

Here

$$|\tilde{\alpha}_i\rangle = \tilde{\alpha}_{i,0}|0_i\rangle + \tilde{\alpha}_{i,1}|1_i\rangle \quad \text{for } i = 1, 2, \ldots, 7 \tag{10.49b}$$

where $\left|\tilde{\alpha}_{i,0}\right|^2 + \left|\tilde{\alpha}_{i,1}\right|^2 = 1$, $\{|0_i\rangle, |1_i\rangle\}$ is the orthonormal set of the *i*th two-level unit of the array, and the righthand side of (10.49a) again represents a direct product of the quantum wave function of the seven different units forming the array. This state is essentially a linear combination or superposition of many classical-like states of the form in (10.47).

In addition, due to the nature of quantum dynamics, it is, in general, not possible to change the data states of the quantum system [e.g., a data set such as the classical-like state in (10.47)] without generating or dealing with wave functions involving linear combinations of basis states such as those in (10.49). This is because, in the ideal world of theoretical physics, the time evolution of the quantum states is generated by unitary transformations.

The nature of these transformations constitutes another essential difference between classical and quantum computers as unitary transforms only generate reversible changes in the system. Consequently, quantum computations must be accomplished through the application of a set of reversible processes, and the algorithms of such computations must be reversible [7–10].

In this regard, however, note that in the real world the dynamical transformations of quantum systems are generally a little less than unitary. This is because systems in the real world exhibit losses arising from their interactions with the universe in which they find themselves, including effects of thermodynamic fluctuations. In the study of quantum computers, this difficulty (termed decoherence) is generally regarded as a design problem to be overcome.

In particular, all computations made using algorithms generated by unitary processes, must act over a time which is much shorter than the decoherence time [7, 8]. This is the time scale characterizing the length of the period of time over which the system no longer appears to be unitary. The isolation of the quantum computation medium and the absolute characterization of the Hamiltonian of that medium are then fundamental necessities for the study of such systems and the development of computer algorithms to operate within them.

Changes made to the input data during an ideal quantum computational process involve the application of a set of unitary transformations to the wave function characterizing the initial read in data. The unitary transforms for such time evolutions are of a general form which, from basic quantum theory, is given by [7, 8]

$$U\left(t_f, t_i\right) = \exp\left(-i\int_{t_i}^{t_f} Hdt\right). \tag{10.50}$$

Here t_i and t_f are the initial and final times of the operation of the unitary transformation and H is the Hamiltonian governing the dynamics of the system during this period of time.

The time evolution of a state such as that given in (10.47) or (10.49) is then obtained by acting on these initial value wave functions with $U\left(t_f, t_i\right)$ where H in (10.50) is a Hamiltonian of the computing medium. This evolution is set up to create the appropriate wave function changes during the $t_i \rightarrow t_f$ time interval of its application to represent various logic-gate operations of the system. As shall be discussed later these logic-gate operations must be reversible as $U\left(t_f, t_i\right)$ is unitary.

The form of the Hamilton for the time evolution can and often, of necessity, changes during the time evolution of the logic-gate operation. These changes are brought about by the application to the computing medium of a series of various external fields which interact with the medium. As an example, in a simple system composed of a trapped ion, the occupancy of a ground state and an excited state has the general qubit form given in (10.46). The trapped ion is an often studied system in quantum optics, and its manipulation through the application of an external light source is common knowledge, used as a basis of many aspects of optics and their applications in technology.

By applying a beam of monochromatic light to the trapped ion system, a unitary transform of the form [7, 8]

$$(\alpha_0, \alpha_1) \rightarrow \left(\alpha'_0, \alpha'_1\right) \tag{10.51}$$

can be made on the qubit. During the process, described in (10.51), the original coefficients α_0 and α_1 are changed by the unitary transform to turn them into the primed coordinates α'_0 and α'_1 where $\left|\alpha'_0\right|^2 + \left|\alpha'_1\right|^2 = 1$. Using various pulses of light applied during specific time intervals a wide range of unitary transformations of the single qubit can be made.

For an array of N trapped and isolated ions which can be separately or pairwise interacted with by external light sources, it is possible to perform adjustments on the state occupancy of the ion qubits composing the array. Under these processes, the adjustments can be formulated so as to generate basic logic-gate operations on the qubit wave functions of the array. Algorithms are then created as a sequential application of these logic-gate operations to the array.

In the following it will be shown that computations based on the applications of unitary transforms to data expressed in the general state format of (10.49) display a number of advantages over classical computations such as those discussed in the previous subsection. Before these advantages are presented, along with examples of algorithms of quantum computation, a treatment of the logic-gates that enter into quantum computation will be given. Due to the difference in the nature of reversible and irreversible computation, these quantum gates are quite different from the classical logic gates upon which classical computation is based.

Quantum Logic-Gates

A primary difference between the logic-gates of classical and quantum computers is that the logical operations of quantum computers must be invertible, i.e., it should be possible to determine the input of the gate given its output [7–10]. This requires that the number of qubits of the input data should equal the number of qubits in the output data. In addition, the number of qubits acted upon by a quantum logic-gate should be the same as the number of qubits output by the logic-gate. This has consequences in the table of these reversible logic-gates.

To understand the operation of some of the fundamental logic-gates arising in quantum computation, the operation of these gates will be considered on a data array of the form [7–10]

$$|\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_{N-1}, \tilde{\alpha}_N\rangle = |\tilde{\alpha}_1\rangle|\tilde{\alpha}_2\rangle\ldots|\tilde{\alpha}_{N-1}\rangle|\tilde{\alpha}_N\rangle. \tag{10.52}$$

This is a generalization of the direct product defined in (10.49) to the case of N qubits.

N-Gate

The simplest unitary logic-gate operation to treat is the N-gate which is described by the logic table [7, 8] (Table 10.5).

This is the same N-gate Table as mentioned in the context of classical computation. Nevertheless, the N-gate is reversible so that it can be described by a quantum mechanical process where it ultimately takes the form of a unitary operator.

Consider the N-gate operation as it is applied to the ith ket in the direct product in (10.52). The ket it operates on has the form $|\tilde{\alpha}_i\rangle = \tilde{\alpha}_{i,0}|0_i\rangle + \tilde{\alpha}_{i,1}|1_i\rangle$ which was originally given in (10.49b). The unitary operator generating the N-gate transformation on $|\tilde{\alpha}_i\rangle$ takes $|1_i\rangle$ into $|0_i\rangle$ and $|0_i\rangle$ into $|1_i\rangle$ and is given by $\tilde{N}_i$ defined as [7, 8]

Table 10.5 N-gate

a	$b = \bar{a}$
0	1
1	0

$$\tilde{N}_i = |1_i\rangle\langle 0_i| + |0_i\rangle\langle 1_i|. \tag{10.53}$$

It is readily seen that $\tilde{N}_i^+\tilde{N}_i = \tilde{N}_i\tilde{N}_i^+ = 1$ so that the operator is unitary and only operates on one of the kets in the direct product wave function.

Consequently, in terms of the *i*th ket

$$\tilde{N}_i|\tilde{\alpha}_i\rangle = \tilde{\alpha}_{i,0}|1_i\rangle + \tilde{\alpha}_{i,1}|0_i\rangle \tag{10.54}$$

where the other kets in (10.52) have been ignored as the N-gate does not affect them. In the case that $\left(\tilde{\alpha}_{i,0}, \tilde{\alpha}_{i,1}\right) = (1, 0)$ or $\left(\tilde{\alpha}_{i,0}, \tilde{\alpha}_{i,1}\right) = (0, 1)$ the *i*th ket reduces to a classical-like state, and the N-gate operation in (10.53) and (10.54) takes the form familiar from the discussions of the classical computation N-gate.

CN-Gate

Next consider a logic-gate that involves two of the kets in (10.52). Let these two kets be the *i*th and *j*th kets in the direct product wave function. For these two kets consider as an example an application of the control-not or CN-gate. This has the logic table [7, 8] (Table 10.6).

In the application of the table, a_i will refer to the initial value of the *i*th ket and b_i will refer to the initial value of the *j*th ket. Similarly, a_f will refer to the final value of the *i*th ket and b_f will refer to the final value of the *j*th ket.

The unitary operator representing the CN-gate logic table is then given by the operator [8]

$$CN_{i,j} = \left|0_i, 0_j\right\rangle\left\langle 0_i, 0_j\right| + \left|0_i, 1_j\right\rangle\left\langle 0_i, 1_j\right| + \left|1_i, 1_j\right\rangle\left\langle 1_i, 0_j\right| + \left|1_i, 0_j\right\rangle\left\langle 1_i, 1_j\right|. \tag{10.55}$$

In this notation, proceeding from left to right on the righthand side of (10.55) represents the operations of the rows of the logic table going from the top to the bottom of the table. Appling the CN operator to the $|\tilde{\alpha}_i\rangle\left|\tilde{\alpha}_j\right\rangle$ kets in (10.52) it is found that [8]

$$\begin{aligned} CN_{i,j}|\tilde{\alpha}_{i,}\rangle\left|\tilde{\alpha}_j\right\rangle &= CN_{i,j}\left|\tilde{\alpha}_i, \tilde{\alpha}_j\right\rangle \\ &= \alpha_{i,0}\alpha_{j,0}\left|0_i, 0_j\right\rangle + \tilde{\alpha}_{i,0}\tilde{\alpha}_{j,1}\left|0_i, 1_j\right\rangle \\ &\quad + \tilde{\alpha}_{i,1}\tilde{\alpha}_{j,0}\left|1_i, 1_j\right\rangle + \tilde{\alpha}_{i,1}\tilde{\alpha}_{j,1}\left|1_i, 0_j\right\rangle. \end{aligned} \tag{10.56}$$

Table 10.6 CN-gate

a_i	b_i	a_f	b_f
0	0	0	0
0	1	0	1
1	0	1	1
1	1	1	0

Table 10.7 CNN-gate

a_i	b_i	c_i	a_f	b_f	c_f
0	0	0	0	0	0
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	1	0	1
1	1	0	1	1	1
1	1	1	1	1	0

Note that the other kets have been omitted in (10.56) as the CN has no effect on them. In addition, the transformation is seen to be unitary as $CN_{i,j}(CN_{i,j})^+ = (CN_{i,j})^+ CN_{i,j} = 1$, and it is also Hermitian as $(CN)^+ = CN$.

CNN-Gate

An example of a logic-gate that operates on three qubits is the control-not-not or CNN-gate. The logic table of the CNN-gate is given by [8] (Table 10.7).

Consider now the effects of the CNN-gate as it operates on the ith, jth, and kth kets in (10.52). As in the previous discussions the other kets forming the state in (10.52) will be ignored in the following as they remain unchanged by the proposed application of the CNN-gate. For the proposed application of the table, a_i will refer to the initial value of the ith ket, b_i will refer to the initial value of the jth ket, and c_i will refer to the initial value of the kth ket. Similarly, a_f will refer to the final value of the ith ket, b_f will refer to the final value of the jth ket, and c_f will refer to the final value of the kth ket.

The unitary operator that represents the changes in the ith, jth, and kth kets in (10.52) coming from the application of the CNN-gate Table is given by the form [8]

$$\begin{aligned} CNN_{i,j,k} = {} & |0_i, 0_j, 0_k\rangle\langle 0_i, 0_j, 0_k| + |0_i, 0_j, 1_k\rangle\langle 0_i, 0_j, 1_k| \\ & + |0_i, 1_j, 0_k\rangle\langle 0_i, 1_j, 0_k| + |0_i, 1_j, 1_k\rangle\langle 0_i, 1_j, 1_k| \\ & + |1_i, 0_j, 0_k\rangle\langle 1_i, 0_j, 0_k| + |1_i, 0_j, 1_k\rangle\langle 1_i, 0_j, 1_k| \\ & + |1_i, 1_j, 1_k\rangle\langle 1_i, 1_j, 0_k| + |1_i, 1_j, 0_k\rangle\langle 1_i, 1_j, 1_k|. \end{aligned} \tag{10.57}$$

The CNN-gate is, as with the earlier examples of reversible logic-gates, found to be unitary and Hermitian.

Universal-Gates

In addition, the CNN-gate is an example of a universal gate. This means that all of the other logic gates involving two or less qubit operations can be obtained from it. For example, in the case that it is specified that $a_i = 1$ always, the CNN table reduces to the logic table [8] (Table 10.8).

Table 10.8 Reduced CNN-gate

b_i	c_i	a_f	b_f	c_f
0	0	1	0	0
0	1	1	0	1
1	0	1	1	1
1	1	1	1	0

Table 10.9 CN-gate formed from reduction of a CNN-gate

b_i	c_i	b_f	c_f
0	0	0	0
0	1	0	1
1	0	1	1
1	1	1	0

If, as $a_i = 1$ always, the a_f column in the resulting table is ignored, it is found that the table reduces to [8] (Table 10.9) which is just the table of the CN-gate. Consequently, the operator in (10.57) reduces to [8]

$$\begin{aligned} CNN_{i,j,k} = {} & \left|1_i, 0_j, 0_k\right\rangle\left\langle 1_i, 0_j, 0_k\right| + \left|1_i, 0_j, 1_k\right\rangle\left\langle 1_i, 0_j, 1_k\right| \\ & + \left|1_i, 1_j, 1_k\right\rangle\left\langle 1_1, 1_j, 0_k\right| + \left|1_i, 1_j, 0_k\right\rangle\left\langle 1_i, 1_j, 1_k\right|. \end{aligned} \tag{10.58}$$

which is the $CN_{j,k}$ operator [8].

In the case of the CNN-gate Table in which $c_i = 0$ always, the resulting table is [8] (Table 10.10). Notice that in the table $a_f = a_i$ and $b_f = b_i$ so these two states remain unchanged by the operation, but $c_f = a_i b_i$. Consequently, the table is a representation of an AND-gate. In the operator representation it follows that [8]

$$\begin{aligned} CNN_{i,j,k} = {} & \left|0_i, 0_j, 0_k\right\rangle\left\langle 0_i, 0_j, 0_k\right| + \left|0_i, 1_j, 0_k\right\rangle\left\langle 0_i, 1_j, 0_k\right| \\ & + \left|1_i, 0_j, 0_k\right\rangle\left\langle 1_i, 0_j, 0_k\right| + \left|1_i, 1_j, 1_k\right\rangle\left\langle 1_1, 1_j, 0_k\right|, \end{aligned} \tag{10.59}$$

where it is seen that the values of the *i*th and *j*th kets are always set to remain unchanged by the operator.

Universal F-Gate

Another example of a universal gate is the Fredkin or F-gate which is a control-exchange-gate. The table for the F-gate is given by [8] (Table 10.11)

Table 10.10 Reduced CNN-gate

a_i	b_i	a_f	b_f	c_f
0	0	0	0	0
0	1	0	1	0
1	0	1	0	0
1	1	1	1	1

Table 10.11 F-gate

a_i	b_i	c_i	a_f	b_f	c_f
0	0	0	0	0	0
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	1	1	0
1	1	0	1	0	1
1	1	1	1	1	1

and the F-gate operator is [8]

$$\begin{aligned} F_{i,j,k} = {} & \left|0_i, 0_j, 0_k\right\rangle\left\langle 0_i, 0_j, 0_k\right| + \left|0_i, 0_j, 1_k\right\rangle\left\langle 0_i, 0_j, 1_k\right| \\ & + \left|0_i, 1_j, 0_k\right\rangle\left\langle 0_i, 1_j, 0_k\right| + \left|0_i, 1_j, 1_k\right\rangle\left\langle 0_i, 1_j, 1_k\right| \\ & + \left|1_i, 0_j, 0_k\right\rangle\left\langle 1_i, 0_j, 0_k\right| + \left|1_i, 0_j, 1_k\right\rangle\left\langle 1_i, 1_j, 0_k\right| \\ & + \left|1_i, 1_j, 0_k\right\rangle\left\langle 1_1, 0_j, 1_k\right| + \left|1_i, 1_j, 1_k\right\rangle\left\langle 1_i, 1_j, 1_k\right|. \end{aligned} \tag{10.60}$$

As with the CNN gate, it is found that by making certain fixed input assignments to the F-gate, the F-gate can be readily converted to exhibit a variety of one and two qubit gates [8].

Walsh-Hadamard-Gate

A final important unitary transformation for the implementation of quantum computer algorithms is the Walsh-Hadamard or H-gate. This is a unitary transformation which operates upon a single qubit and is best expressed in operator notation. The Walsh-Hadamard operator applied on the *i*th qubit in (10.52) has the form [8]

$$\tilde{H}_i = \frac{1}{\sqrt{2}}\{|0_i\rangle\langle 0_i| + |0_i\rangle\langle 1_i| + |1_i\rangle\langle 0_i| - |1_i\rangle\langle 1_i|\}. \tag{10.61}$$

It has the interesting and very useful property that applied to the members of the orthonormal basis $\{|0_i\rangle, |1_i\rangle\}$ it yields [8]

$$\tilde{H}_i|0_i\rangle = \frac{1}{\sqrt{2}}[|0_i\rangle + |1_i\rangle] \tag{10.62a}$$

and

$$\tilde{H}_i|1_i\rangle = \frac{1}{\sqrt{2}}[|0_i\rangle - |1_i\rangle]. \tag{10.62b}$$

The H-operator is seen to transform the elements of the orthogonal basis into wave functions which are equally weighted linear combinations of basis states. The result of the operation is a superposition wave function composed of equal weights of the different possible basis states of the system. It can be implemented experimental, for example, in a system of trapped ions by applying a pulse of light which mixes the eigenstates of the ions, taking each ion from a pure eigenstate to an equally weighted linear combination of its two eigenstates.

This can be useful in creating an initial set of linear combinations of states of the system, starting from an ordered ground state of the array of qubits. In this process, the ground state of the array of ions can be set through the operation to be composed as a superposition wave function representing an array of uniformly combined possible eigenstates of the system [8]

$$\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|0_1,0_2,\ldots,0_N\rangle = |h_1\rangle|h_2\rangle\ldots|h_N\rangle. \tag{10.63}$$

where $|h_i\rangle = \frac{1}{\sqrt{2}}[|0_i\rangle + |1_i\rangle]$. The state created in (10.63) contains every possible configuration in the set $\{|0_1,0_2,\ldots,0_N\rangle, |1_1,0_2,\ldots,0_N\rangle,\ |0_1,1_2,\ldots,0_N\rangle,\ldots,|1_1,1_2,\ldots,1_N\rangle\}$ that are available involving the various $\{|0_i\rangle,|1_i\rangle\}$ ground state-excited state basis of the N qubit array.

As an illustration of how the various logic gates developed in this section can be applied to perform an actual quantum calculation, two important algorithms for quantum computing will now be discussed. These involve preparing an initial state to be used in the computation, acting on it with an algorithmic sequence of logical operations, and finally reading out the final answer arising from the data represented by the initially prepared state. The best way to understand the basic ideas of quantum computation is to see how such algorithms operate and to develop an idea of how they can be more efficient than algorithms which are designed for classical computers.

Quantum Computer Algorithms

To understand how such a linear combination of quantum states might facilitate quantum computation, an example of a quantum algorithm based on unitary logic-gates will be given. The problem to be solved by the algorithm is a contrived problem meant to illustrate the advantages of quantum computation over classical computation. It does this by displaying some of the basic techniques of quantum computing that facilitate computation and which are not available in classical computing.

Deutsch-Jozsa Algorithm

The problem to be considered was proposed early on in the study of quantum computing and is known as the Deutsch-Jozsa problem. It involves a classification of the values of a particular function defined over the complete set of N component kets in the set [8]

$$\{|0_1, 0_2, \ldots, 0_N\rangle, |1_1, 0_2, \ldots, 0_N\rangle, |0_1, 1_2, \ldots, 0_N\rangle, \ldots, |1_1, 1_2, \ldots, 1_N\rangle\}. \quad (10.64)$$

Specifically, if the N-kets of the complete set in (10.64) are denoted by [7, 8]

$$|\vec{x}\rangle = |x_1, x_2, \ldots, x_N\rangle = |x_1\rangle|x_2\rangle\ldots|x_N\rangle \quad (10.65)$$

where each of the $x_i = 0$ or 1 for $i = 1, 2, 3, \ldots, N$, then a function $f(x_1, x_2, \ldots, x_N)$ is defined over the ket arguments taking either the value 0 or 1.

The function $f(x_1, x_2, \ldots, x_N)$ is essentially a black box that is contrived so that it can exhibit either of two types of behaviors when applied to each of the kets in (10.64) [7]. In the first type of behavior, $f(x_1, x_2, \ldots, x_N) = 1$ for all of the kets in (10.64), and the function is referred to as being a constant function.

For the second type of behavior, $f(x_1, x_2, \ldots, x_N) = 1$ for only half the kets in (10.64) and consequently $f(x_1, x_2, \ldots, x_N) = 0$ for the other half of the kets in (10.64) [7]. Furthermore, for the second type of function it is unknown which of the N-kets give either of these two values. Functions exhibiting this second type of behavior are known as balanced functions.

In the Deutsch-Jozsa problem the reader is given the black box function but is not told which of the two types of functions the box represents. It is only known that the box must be one of the two types previously mentioned. The problem left to the reader is to apply a quantum computer to find out which of the two functions the black box represents. Is the function a constant or a balanced function?

Using methods of classical computation, the problem becomes one of successively reading in the different $(x_1, x_2, \ldots, x_N)$ sets of values and recording the values of the function $f(x_1, x_2, \ldots, x_N)$ outputted from each of the inputted values. In this process, there are 2^N different values that could be read into $f(x_1, x_2, \ldots, x_N)$. It is known, however, that only one-half plus one of the set of different $(x_1, x_2, \ldots, x_N)$ need to be read into $f(x_1, x_2, \ldots, x_N)$ to completely test the behavior of the function. This follows as for the balanced function half of the set of inputs give 0 and half give 1. Consequently, the classical computer requires $2^{N-1} + 1$ tests to make the determination between the two types of functions that could be in the black box.

The test on the black box function using classical computers is seen to be a brute force endeavor, involving an extensive computational effort. A considerable speed up of the process, however, is made available to the effort by the application of processes involving quantum effects. Essential to this speed up are the ideas of wave functions represented as mixtures of eigenstates and the development of quantum algorithms as unitary transformations. As a comparison, these ideas of quantum computation will now be applied to the study of the Deutsch-Jozsa problem.

Now try to speed up the computation up by using the ideas of quantum computing. One way to increase the rate of computation is (instead of sequentially substituting pure states of the form of (10.64) as input into the quantum computer) to introduce a data set as a wavefunction which is an equally weighted sum of all basis states. This is an equally mixed wave function of the basis states and is formed as a linear combination of the states in (10.64). The quantum computer has an important feature that it can operate on the superposition wave function input with the logic-gate operations discussed earlier. This is different from the classical computer which can only operate on one of the pure state inputs of the system, contained within the set in (10.64), at a time.

In its operation the quantum computer processes the superposition wave function input by sequentially applying logic gates of the quantum algorithm to the input. At the end of this processing the result of the calculation shows up as a superposition wave function of output answers contained in the output data from the quantum computer algorithm. Whereas the classical computer sequentially applies the logic gates of its classical algorithm to one input state to obtain one output, the quantum computer applies essential the same logic processes to the wave function mixture of many different input states to obtain a superposition wave function of many output states. In this sense, the quantum computer does many things at once while the classical system does only one thing at a time.

The beginning of the quantum computer processes is then to create a superposition wave function of input data upon which the computer is to work. In this superposition wave function it is important that each of the input states in the set in (10.64) is represented with equal probability in the input data and is equally processed by the computer algorithm. This assures an output of equally mixed answer states within the output data.

To create a superposition wave function input state that is a linear sum of equally weighted states from the set in (10.64) the H-gate in (10.61)–(10.63) can be used. The H-gate takes states that are pure basis eigenstates of the system and generates an equal weighed superposition wave function of them. This is true for any of the states in (10.64) as can be readily seen from (10.61)–(10.63). Most quantum computations, however, begin by setting up a superposition wave function based on the application of the H-gate to the ground state of the system forming the computer. This particular application of the H-gate is shown in (10.63) and will be used in the following discussions.

To generate a superposition wave function from the ground state of the system in (10.64) it follows from (10.63) that for the N-ket system [7]

$$\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|0_1,0_2,\ldots,0_N\rangle = |h_1\rangle|h_2\rangle\ldots|h_N\rangle = |h_1,h_2,\ldots,h_N\rangle \tag{10.66}$$

where $|h_i\rangle = \frac{1}{\sqrt{2}}[|0_i\rangle + |1_i\rangle]$. The application of the H-gates in (10.66) is found to assure an equal mixture of all of the states contained within the set of (10.64). This can be seen by multiplying out the right hand side of (10.66), resulting in the sum [7]

$$\begin{aligned}|h_1,h_2,\ldots,h_N\rangle = \left(\frac{1}{\sqrt{2}}\right)^N \{&|0_1,0_2,\ldots,0_N\rangle + |1_1,0_2,\ldots,0_N\rangle \\ &+\cdots+|1_1,1_2,\ldots,1_N\rangle\}\end{aligned} \tag{10.67}$$

The result in (10.67) is an equal weighted mixed sum of all of the input states in (10.64).

It is now evident from the mathematical structure of the input data that the classical computer operates on the single states in (10.64), but the quantum computer operates on each of the sum of the superposed single states in (10.67). In this sense the quantum computer is a type of parallel processor.

A discussion is now given of how the superposition wave function input data is processed for the Deutsch-Jozsa problem. This is followed by an explanation of how the answer is extracted from the superposition wave function state outputted by the quantum computer.

As with all quantum computer processing of the input data, the calculation on the initial data for the Deutsch-Jozsa problem proceeds by a sequence of unitary transformations to generate an output data set. A consequence of this is that more basis states than the N-ket states in (10.64) and (10.67) are needed for the calculation. The N-ket states in (10.64) and (10.67) are an orthonormal basis that handles the input data, but there is more data in the system than just the input data. In particular, there must also be room in the basis of orthonormal states in which to develop and store the output data. This requires an expansion of the dimension of the basis of ket states handled in the computation. The expansion is required in order to make room for the 0's and 1's of the output data.

The value of the outputted function $f(x_1, x_2, \ldots, x_N)$ is a 0 or a 1 so that at a minimum the state vectors in (10.64) and (10.67) must be increased to accommodate an additional qubit of output information. Making this adjustment the basis set for performing the calculation is increased from the complete set of N-kets in (10.64) to become a complete set of N + 1 kets. This complete set of orthonormal N + 1-kets is given by [7]

$$\begin{aligned}&\{|0_1, 0_2, \ldots, 0_N, 0_{N+1}\rangle, |1_1, 0_2, \ldots, 0_N, 0_{N+1}\rangle, |0_1, 1_2, \ldots, 0_N, 0_{N+1}\rangle,\\ &\quad \ldots, |1_1, 1_2, \ldots, 1_N, 1_{N+1}\rangle\}.\end{aligned} \tag{10.68}$$

In the following it shall be shown that the set of states in (10.68) is, in fact, the basis of complete states that is needed for the quantum computational determination of the nature of $f(x_1, x_2, \ldots, x_N)$. In particular, it will be seen that the outputted state of the system consists of a listing of N qubits of the input data plus a qubit of data containing the answer. This comes about due to fact that the quantum processes are all unitary operations. Consequently, the quantum computer operates in the space of N + 1-kets forming the basis of the quantum mechanical space including both the input and output data.

The set of unitary operations which process the input data into the output in the space of N + 1-kets are then a set of $(N+1) \times (N+1)$ unitary matrices which are used to realize the quantum computer algorithm that is used to generate the computer solution. For this processing on the basis set of (10.68), the first N left entries are loaded with input data and the right most entry is arranged to receive the output data generated by the program. The first N left entries are very important at the beginning of the calculation and the last entry is most significant at the end of the calculation.

To start the calculation a set of superposition wave function input data as well as a superposition wavefunction initial state for the placement of the output data must be developed. The generation of a superposition wavefunction was done earlier in (10.66) and (10.67) for the N ket input data states and now this must be generalized to the N + 1 ket in (10.68). The generalization is almost a direct extension of that in (10.66) but with a slight modification which helps extract the final answer from the outputted state.

In the generalization of (10.66) for an N-ket to the N + 1-ket the superposition state system is set to [7]

$$\begin{aligned}&\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N\tilde{H}_{N+1}|0_1,0_2,\ldots,0_N,1_{N+1}\rangle\\&=|h_1\rangle|h_2\rangle\ldots\ldots|h_N\rangle\left|\tilde{h}_{N+1}\right\rangle=\left|h_1,h_2,\ldots,h_N,\tilde{h}_{N+1}\right\rangle\end{aligned}\tag{10.69a}$$

where

$$|h_i\rangle=\frac{1}{\sqrt{2}}[|0_i\rangle+|1_i\rangle]\tag{10.69b}$$

for $i=1,2,\ldots,N$, and

$$\left|\tilde{h}_{N+1}\right\rangle=\frac{1}{\sqrt{2}}[|0_i\rangle-|1_i\rangle]\tag{10.69c}$$

assure the equal mixture of states from (10.68). Notice that the N + 1 entry in the ket on the left side of the equality in (10.69a) is a 1 whereas all of the other entries are 0. This choice will be seen later to facilitate the construction of the outputted state in the algorithmic processing of the inputted ket. Generated in this manner the ket on the right side of the equality in (10.69a) is still an equal weighted superposition wave function of the basis set in (10.68).

Multiplying out the result in (10.69), the sum of equal weighted input-output eigenstates of the system is then [7]

$$\begin{aligned}\left|h_1,h_2,\ldots,h_N,\tilde{h}_{N+1}\right\rangle=\left(\frac{1}{\sqrt{2}}\right)^{N+1}&\{|0_1,0_2,\ldots,0_N,0_{N+1}\rangle+|1_1,0_2,\ldots,0_N,0_{N+1}\rangle\\&+\cdots-|1_1,1_2,\ldots,1_N,1_{N+1}\rangle\}.\end{aligned}\tag{10.70}$$

In (10.70) the weighting of each of the basis states is equal so that again each of the inputs is treated equally during the calculation. This provides for a series of parallel processing calculations dealing with each of the 2^{N+1} pure input eigenstates. In addition, there is a sign difference between some of the weights. This difference gives rise to an interference between the states generated in the calculation of the superposition wave function output data. It will be useful in extracting the final answer to the question of whether or not the function $f(x_1, x_2, \ldots, x_N)$ is a constant or balanced function.

Given the initial configuration of mixed wave function states, the calculation proceeds by devising a unitary process which operates on the superposition wave function initial state of input data to convert it to a superposition wave function output state. Schematically this is represented by [7]

$$\left|h_1, h_2, \ldots, h_N, \tilde{h}_{N+1}\right\rangle \overset{Process}{\rightarrow} \left|h_1, h_2, \ldots, h_N, \tilde{h}'_{N+1}\right\rangle \tag{10.71}$$

During the development of this process each of the pure eigenstates found in the superposition wave function input is transformed by an algorithm composed from quantum logic-gates to generate an answer for that particular input. The inputted data shows up in both the inputted and outputted data sets.

To understand the general nature of the process involved in the calculation, consider both sides of (10.71) written in terms of the orthonormal basis in (10.68). From (10.70) and (10.71) it is seen that [7]

$$\begin{aligned}&\left(\frac{1}{\sqrt{2}}\right)^{N+1} \{|0_1, 0_2, \ldots\ldots, 0_N, 0_{N+1}\rangle + |1_1, 0_2, \ldots, 0_N, 0_{N+1}\rangle \\ &+ \cdots - |1_1, 1_2, \ldots, 1_N, 1_{N+1}\rangle\} \overset{Process}{\rightarrow} \left(\frac{1}{\sqrt{2}}\right)^{N+1} \left\{\gamma_1 \left|0_1, 0_2, \ldots, 0_N, 0'_{N+1}\right\rangle\right. \\ &\left.+ \gamma_2 \left|1_1, 0_2, \ldots, 0_N, 0'_{N+1}\right\rangle\right\} + \cdots + \gamma_{2^{N+1}} \left|1_1, 1_2, \ldots, 1_N, 1'_{N+1}\right\rangle\big\}.\end{aligned} \tag{10.72}$$

where $\{\gamma_i\}$ are the coefficients of a unitary transformation arising from the computational algorithm.

The transformation that allows for the determination of the nature of $f(x_1, x_2, \ldots, x_N)$ can be obtained by choosing a unitary algorithmic process to operate on the input data and generate the $\{\gamma_i\}$ [7]. This unitary process will be shown to involve the application to the input data of an exclusive-or (XOR) logic-gate similar to that discussed above (10.40). The XOR operation of interest will now be discussed, followed by a presentation which shows how it is employed in creating the unitary process that generates an answer to the question of whether $f(x_1, x_2, \ldots, x_N)$ is constant or balanced.

To understand the processing operation going on in (10.72) it is helpful to shift the focus of the treatment from the superposition wave function states in (10.71) and (10.72) and consider the working of the processing operation at the level of the individual basis states in (10.68). In this way, the processing operation can be discussed in terms of an operation based on XOR logic-gate applications. For these discussions, the orthonormal states of the set defined in (10.68) are denoted by the general form

$$|\vec{x}, y_{N+1}\rangle = |x_1, x_2, \ldots, x_N, y_{N+1}\rangle \tag{10.73}$$

where $\vec{x}$ represents the set of 0's and 1's for the N-kets of input data and y_{N+1} is the 0 or 1 in the ket reserved for the output data.

In terms of the notation of (10.73) the left hand side of (10.71) and (10.72) becomes [7]

$$\left|h_1, h_2, \ldots, h_N, \tilde{h}_{N+1}\right\rangle = \left(\frac{1}{\sqrt{2}}\right)^{N+1} \sum_{\vec{x}} \{|\vec{x}, 0_{N+1}\rangle - |\vec{x}, 1_{N+1}\rangle\}. \tag{10.74}$$

Here the negative sign between the two terms in the bracket comes from the H-gate operations in (10.69) which is used to create the entanglement of the input data associated with the N + 1-ket. The sum in (10.74) is over the 2^N different states of $\vec{x}$ for the N-kets of input data.

Focusing on the individual members of the basis set in (10.68) and their general form in (10.73), the process that can be used to solve the problem involves a transformation which takes the last entry,y_{N+1}, of the N + 1-kets and converts it to the output data entry $y_{N+1} \oplus f(\vec{x})$. Here $\oplus$ is the exclusive-or (XOR) operation defined in the table above (10.40). In tabular form the outlined process is given by [7] (Table 10.12).

Applying the $y_{N+1} \oplus f(\vec{x})$ transformation to the N + 1-kets of the entangled input state represented in (10.74) yields the processed output state [7]

$$\left|h_1, h_2, \ldots, h_N, \tilde{h}'_{N+1}\right\rangle = \left(\frac{1}{\sqrt{2}}\right)^{N+1} \sum_{\vec{x}} \{|\vec{x}, 0_{N+1} \oplus f(\vec{x})\rangle - |\vec{x}, 1_{N+1} \oplus f(\vec{x})\rangle\}. \tag{10.75}$$

Considering the expression in the brackets in (10.75) and applying the results for $y_{N+1} \oplus f(\vec{x})$ from the table below (10.75) to the kets in the sum on the right in (10.75), it is found that

$$|\vec{x}, 0_{N+1} \oplus f(\vec{x})\rangle - |\vec{x}, 1_{N+1} \oplus f(\vec{x})\rangle = (-1)^{f(\vec{x})}\{|\vec{x}, 0_{N+1}\rangle - |\vec{x}, 1_{N+1}\rangle\} \tag{10.76a}$$

The identity in (10.76a) can be checked using the XOR table for $y_{N+1} \oplus f(\vec{x})$ and the fact that $f(\vec{x})$ and $y_{N+1} \oplus f(\vec{x})$ only take the values 0 and 1.

Table 10.12 Multiplication table for the Deutsch-Jozsa algorithm

y_{N+1}	$f(\vec{x})$	$y_{N+1} \oplus f(\vec{x})$
0	0	0
0	1	1
1	0	1
1	1	0

To confirm the identity in (10.76a) first consider the cases for which $f(\vec{x}) = 0$. It then follows from the left hand side of (10.76a) that [7]

$$\begin{aligned}&|\vec{x}, 0_{N+1} \oplus f(\vec{x})\rangle - |\vec{x}, 1_{N+1} \oplus f(\vec{x})\rangle = |\vec{x}\rangle[|0_{N+1} \oplus 0_{N+1}\rangle - |1_{N+1} \oplus 0_{N+1}\rangle]\\ &= |\vec{x}\rangle[|0_{N+1}\rangle - |1_{N+1}\rangle] = (-1)^{f(\vec{x})}\{|x, 0_{N+1}\rangle - |x, 1_{N+1}\rangle\}\end{aligned} \tag{10.76b}$$

In the same way, for the case in which $f(\vec{x}) = 1$ it follows from the left hand side of (10.76a) that

$$\begin{aligned}&|\vec{x}, 0_{N+1} \oplus f(\vec{x})\rangle - |\vec{x}, 1_{N+1} \oplus f(\vec{x})\rangle = |\vec{x}\rangle[|0_{N+1} \oplus 1_{N+1}\rangle - |1_{N+1} \oplus 1_{N+1}\rangle]\\ &= |\vec{x}\rangle[|1_{N+1}\rangle - |0_{N+1}\rangle] = (-1)^{f(\vec{x})}\{|x, 0_{N+1}\rangle - |x, 1_{N+1}\rangle\}.\end{aligned} \tag{10.76c}$$

Applying (10.76a) in (10.75) it is found that (10.75) can be rewritten in the form [7]

$$\left|h_1, h_2, \ldots, h_N, \tilde{h}'_{N+1}\right\rangle = \left(\frac{1}{\sqrt{2}}\right)^{N+1} \sum_{\vec{x}} (-1)^{f(\vec{x})}\{|\vec{x}, 0_{N+1}\rangle - |\vec{x}, 1_{N+1}\rangle\} \tag{10.77}$$

This expresses the processed result from the application of the logic-gate in terms of the complete orthonormal basis set in (10.68).

It is interesting to note that the expressions for the wave function in (10.74) before the transformation and the wave function in (10.77) after the transformation are both normalized to one. This is a consequence of the unitary nature of the transform that was applied in going from (10.74) to (10.77). It is an indication that, since these two wave functions are related by a unitary transformation, there is a quantum mechanical process that will transform the two wave functions into each other.

In this regard, it should be noted that the physical process to be applied to make the unitary transformation is not discussed here. It is only noted that as the process is represented by a unitary transformation it should be feasible in a quantum system. The physical details of the unitary transformation process will depend on the physical system upon which the quantum computer is realized.

The processing of the state in (10.77) is not over as yet. An additional transformation must be performed on the superposition wave function before the answer to the problem can be extracted from the quantum computer. Specifically, the application of another H-gate operation is needed.

For the next step in the processing of the wave function a set of H-gate transformations is applied to the $\vec{x}$ kets on left hand side of (10.77). Specifically, H-gate transformations are applied to each of the first N kets of the N + 1-ket wave function so that [7]

$$\begin{aligned}
&\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N\left|h_1,h_2,\ldots,h_N,\tilde{h}'_{N+1}\right\rangle\\
&=\left(\frac{1}{\sqrt{2}}\right)^{N+1}\sum_{\vec{x}}(-1)^{f(\vec{x})}\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N\{|\vec{x},0_{N+1}\rangle-|\vec{x},1_{N+1}\rangle\}\\
&=\left(\frac{1}{\sqrt{2}}\right)^{N+1}\sum_{\vec{x}}(-1)^{f(\vec{x})}\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|\vec{x}\rangle\{|0_{N+1}\rangle-|1_{N+1}\rangle\}.
\end{aligned}\tag{10.78}$$

Focusing on the action of the H-gates on the N kets forming $|\vec{x}\rangle$ it follows that

$$\begin{aligned}
\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|\vec{x}\rangle&=\tilde{H}_1|x_1\rangle\tilde{H}_2|x_2\rangle\ldots\tilde{H}_N|x_n\rangle\\
&=\left(\frac{1}{\sqrt{2}}\right)^{N}[|0_1\rangle+(-1)^{x_1}|1_1\rangle][|0_2\rangle+(-1)^{x_2}|1_2\rangle]\ldots[|0_N\rangle+(-1)^{x_N}|1_N\rangle].
\end{aligned}\tag{10.79}$$

Upon multiplying out the product of kets on the far right of (10.79) and introducing the notation $z_i=0_i$ or 1_i for $i=1,2,\ldots,N$, (10.79) is rewritten as [7]

$$\begin{aligned}
\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|\vec{x}\rangle&=\left(\frac{1}{\sqrt{2}}\right)^{N}\sum_{z_1=0}^{1}\sum_{z_2=0}^{1}\cdots\sum_{z_N=0}^{1}(-1)^{\vec{x}\cdot\vec{z}}|z_1\rangle|z_2\rangle\ldots|z_N\rangle\\
&=\left(\frac{1}{\sqrt{2}}\right)^{N}\sum_{\vec{z}}(-1)^{\vec{x}\cdot\vec{z}}|\vec{z}\rangle
\end{aligned}\tag{10.80}$$

where $\vec{x}\cdot\vec{z}=x_1z_1+x_2z_2+\cdots+x_Nz_N$ and the sum over $\vec{z}$ on the far right is over the complete set of 2^N states of $\vec{z}$.

Substituting the result in (10.80) into (10.78) yields [7]

$$\begin{aligned}
&\tilde{H}_1\tilde{H}_2\ldots\ldots\tilde{H}_N\left|h_1,h_2,\ldots,h_N,\tilde{h}'_{N+1}\right\rangle=\left|h'_1,h'_2,\ldots,h'_N,h'_{N+1}\right\rangle\\
&=\left(\frac{1}{\sqrt{2}}\right)^{N+1}\sum_{\vec{x}}(-1)^{f(\vec{x})}\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|\vec{x}\rangle\{|0_{N+1}\rangle-|1_{N+1}\rangle\}\\
&=\frac{1}{2^N\sqrt{2}}\sum_{\vec{x}}(-1)^{f(\vec{x})}\sum_{\vec{z}}(-1)^{\vec{x}\cdot\vec{z}}|\vec{z}\rangle\{|0_{N+1}\rangle-|1_{N+1}\rangle\}\\
&=\frac{1}{2^N\sqrt{2}}\sum_{\vec{x},\vec{z}}(-1)^{f(\vec{x})+\vec{x}\cdot\vec{z}}|\vec{z}\rangle\{|0_{N+1}\rangle-|1_{N+1}\rangle\}.
\end{aligned}\tag{10.81}$$

Again it is seen that the normalization of the wave function is one so that the net transformation of the state is unitary. It can be performed in a quantum mechanical system.

The processed state in (10.81) can now be used to extract the answer to the problem regarding the nature of the function, $f(\vec{x})$. To determine the nature of $f(\vec{x})$ it is only needed to measure the states of the first N kets of the system, i.e., the state of $|\vec{z}\rangle$. If $|\vec{z}\rangle = |0_1, 0_2, \ldots, 0_N\rangle$ so that all of the first N kets are in the ground state, then $f(\vec{x})$ is a constant function. Otherwise, $f(\vec{x})$ is a balanced function.

This is seen from a consideration of the final state of the processed wave function generated in (10.81)

$$\begin{aligned}&\left|h'_1, h'_2, \ldots, h'_N, \tilde{h}'_{N+1}\right\rangle\\ &= \frac{1}{2^N\sqrt{2}}\sum_{\vec{x},\vec{z}}(-1)^{f(\vec{x})+\vec{x}\cdot\vec{z}}|\vec{z}\rangle\{|0_{N+1}\rangle - |1_{N+1}\rangle\}.\end{aligned} \tag{10.82}$$

The relative amplitude for measuring the state $|\vec{z}\rangle = |0_1, 0_2, \ldots, 0_N\rangle$ in the wave function in (10.82) is

$$\begin{aligned}&\langle 0_1, 0_2, \ldots\ldots\ldots, 0_N\left|h'_1, h'_2, \ldots, h'_N, \tilde{h}'_{N+1}\right\rangle\\ &\quad= \frac{1}{2^N\sqrt{2}}\sum_{\vec{x},\vec{z}}(-1)^{f(\vec{x})+\vec{x}\cdot\vec{z}}\langle 0_1, 0_2, \ldots, 0_N|\vec{z}\rangle\{|0_{N+1}\rangle - |1_{N+1}\rangle\}\\ &= \frac{1}{2^N\sqrt{2}}\sum_{\vec{x}}(-1)^{f(\vec{x})}\{|0_{N+1}\rangle - |1_{N+1}\rangle\}.\end{aligned} \tag{10.83}$$

Consequently, the amplitude for observing the state $|\vec{z}\rangle = |0_1, 0_2, \ldots, 0_N\rangle$ in the wave function reduces to [7]

$$\frac{1}{2^N\sqrt{2}}\sum_{\vec{x}}(-1)^{f(\vec{x})}. \tag{10.84}$$

In the case that $f(\vec{x})$ is a constant function (i.e., $f(\vec{x}) = 0$ or 1) the sum in (10.84) is non-zero. However, in the case that $f(\vec{x})$ is a balanced function (i.e., $f(\vec{x}) = 0$ for half the of values of $\{\vec{x}\} = 0$ and $f(\vec{x}) = 1$ for the other half of the values of $\{\vec{x}\}$) the sum in (10.84) is zero.

The problem of determining whether or not $f(\vec{x})$ is constant or balanced reduces to measuring the occupancy of the first N kets of the output. If these kets are all in the ground state, $|0_1, 0_2, \ldots, 0_N\rangle$, the function is a constant function. If the kets are not in the ground state,$|0_1, 0_2, \ldots, 0_N\rangle$, the function is a balanced function.

As a simple example, the above algorithm will be discussed for the specific case of N = 1. This provides a useful illustration of the general functioning of the algorithm [7].

In this example the function $f(x)$ involves a single variable x which can take the values 0 or 1. A constant function then has $f(0) = f(1) = 0$ or 1, and a balanced function has $f(0) = 0$ and $f(1) = 1$ or $f(0) = 1$ and $f(1) = 0$. The computation begins with the initial state in which the first ket is in the ground state and the second ket is in the excited state so that [7]

$$|0\rangle|1\rangle \tag{10.85}$$

where $|0\rangle$ holds the input data and $|1\rangle$ will contain the output data.

The state in (10.85) is first acted upon by two H-gates, one acting on the first ket and the second acting on the second ket. Consequently, under this action

$$\begin{aligned}\tilde{H}|0\rangle\tilde{H}|1\rangle &= \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle]\frac{1}{\sqrt{2}}[|0\rangle - |1\rangle] \\ &= \frac{1}{2}|0\rangle[|0\rangle - |1\rangle] + \frac{1}{2}|1\rangle[|0\rangle - |1\rangle].\end{aligned} \tag{10.86}$$

The application of the two H-gates is followed by the logic-gate operation involving the $y_{N+1} \oplus f(\vec{x})$ XOR table below (10.75).

Applying the XOR operation to (10.86) yields [7]

$$\begin{aligned}&\frac{1}{2}|0\rangle[|0\rangle - |1\rangle] + \frac{1}{2}|1\rangle[|0\rangle - |1\rangle] \overset{XOR-gate}{\rightarrow} \frac{1}{2}|0\rangle[|0 \oplus f(0)\rangle - |1 \oplus f(0)\rangle] \\ &\quad + \frac{1}{2}|1\rangle[|0 \oplus f(1)\rangle - |1 \oplus f(1)\rangle].\end{aligned} \tag{10.87}$$

Considering the first term in the sum on the righthand side: For $f(0) = 1$

$$\frac{1}{2}|0\rangle[|0 \oplus f(0)\rangle - |1 \oplus f(0)\rangle] = \frac{1}{2}|0\rangle[|1\rangle - |0\rangle] = \frac{1}{2}(-1)^{f(0)}|0\rangle[|0\rangle - |1\rangle], \tag{10.88a}$$

and for $f(0) = 0$

$$\frac{1}{2}|0\rangle[|0 \oplus f(0)\rangle - |1 \oplus f(0)\rangle] = \frac{1}{2}|0\rangle[|0\rangle - |1\rangle] = \frac{1}{2}(-1)^{f(0)}|0\rangle[|0\rangle - |1\rangle]. \tag{10.88b}$$

Similarly, considering the second term in the sum on the righthand side: For $f(1) = 1$

$$\frac{1}{2}|1\rangle[|0 \oplus f(1)\rangle - |1 \oplus f(1)\rangle] = \frac{1}{2}|1\rangle[|1\rangle - |0\rangle] = \frac{1}{2}(-1)^{f(1)}|1\rangle[|0\rangle - |1\rangle], \tag{10.88c}$$

and for $f(1) = 0$

$$\frac{1}{2}|1\rangle[|0\oplus f(1)\rangle - |1\oplus f(1)\rangle] = \frac{1}{2}|1\rangle[|0\rangle - |1\rangle] = \frac{1}{2}(-1)^{f(1)}|1\rangle[|0\rangle - |1\rangle]. \tag{10.88d}$$

Combining the results in (10.87) and (10.88), it follows that

$$\begin{aligned}&\frac{1}{2}|0\rangle[|0\rangle - |1\rangle] + \frac{1}{2}|1\rangle[|0\rangle - |1\rangle] \overset{XOR-gate}{\rightarrow} \frac{1}{2}(-1)^{f(0)}|0\rangle[|0\rangle - |1\rangle] + \frac{1}{2}(-1)^{f(1)}|1\rangle[|0\rangle - |1\rangle]\\ &= \frac{1}{\sqrt{2}}\left[(-1)^{f(0)}|0\rangle + (-1)^{f(1)}|1\rangle\right]\frac{1}{\sqrt{2}}[|0\rangle - |1\rangle].\end{aligned} \tag{10.89}$$

The calculation is finalized by applying an H-gate transformation to the left-hand qubits in the direct product in (10.87). From this application, it follows that

$$\begin{aligned}&\frac{1}{\sqrt{2}}\left[(-1)^{f(0)}|0\rangle + (-1)^{f(1)}|1\rangle\right]\frac{1}{\sqrt{2}}[|0\rangle - |1\rangle] \overset{H-gate}{\rightarrow}\\ &\quad\frac{1}{2}\left[\left((-1)^{f(0)} + (-1)^{f(1)}\right)|0\rangle + \left((-1)^{f(0)} - (-1)^{f(1)}\right)|1\rangle\right]\frac{1}{\sqrt{2}}[|0\rangle - |1\rangle].\end{aligned} \tag{10.90}$$

The probability amplitude of finding the first ket in the $|0\rangle$ state is

$$\frac{1}{2}(-1)^{f(0)} + \frac{1}{2}(-1)^{f(1)} \tag{10.91}$$

The probability of the system of two qubits to be in the $|0\rangle$ state is 1 if the system is constant and 0 if the system is balanced. By making one measurement the problem is resolved.

The preceding gives a rather contrived example of a mathematics problem which can be resolved on a quantum computer [7]. It is an example of a problem which can be solved quicker on a quantum computer than by an algorithm working on a classical computer, and this is the essential point of the exercise. All of the processes involved is the discussions are by necessity unitary processes, and how these unitary processes are to be implemented must be considered for each type of quantum computer system devised for the purposes of computation.

The unitary nature of the algorithm implies that such processes can be implemented in a quantum system as quantum systems evolve in a unitary way. In their formulations, the algorithms themselves are seen to arise from cleaver applications of the ideas of quantum mechanics and the probabilities obtained in the collapse of the final state wave functions.

The chapter will conclude with another, perhaps more realistic, problem which can be of practical importance. This will show that not all quantum algorithms focus on academic questions. It will also involve discussions of the treatment of Fourier series defined on discrete lattices [7, 8]. These types of treatments occur commonly in many branches of physics and engineering.

Period of a Periodic Function

The problem of interest is that of the determination of the period of a periodic function which is defined on a lattice. This is another example of a problem which benefits from the parallel computations that are done by applying a quantum computer algorithm on an initially superposed wave function state of input data. It also illustrates an example of the ideas of probability as they enter into the final step of collapsing the output wave function in order to determine the answer from a quantum computation. Before this type of study can be addressed, however, it is helpful to have some discussion of the properties of Fourier series.

Consider a function $f(x)$ defined over a one-dimensional lattice with sites $\{x_i = (\Delta x)i\}$ for $i = 0, 1, \ldots, 2^N - 1$ representing a set of 2^N points on the x-axis. Start initially by focusing the considerations on the set $\{x_i = (\Delta x)i\}$ of lattice positions in the direct lattice and the relationship of these lattice positions to a set of plane wave states in wave vector or k-space. In this association, the direct space set of lattice points can be expressed as a Fourier series in the orthonormal basis set of plane waves in k-space. Once this relationship is obtained a program can be written for $f(x)$ defined over a one-dimensional lattice in either the direct lattice or the k-space representation [7, 8].

Applying the theory of Fourier series to these two representations, it is found that the direct space position, x_n, is expressed as a series of plane waves in k-space given by [7, 8]

$$x_n = \frac{1}{2^{N/2}} \sum_{\tilde{m}=0}^{2^N-1} \mathrm{e}^{i\frac{2\pi}{(2^N)}n\tilde{m}} \tilde{x}_{\tilde{m}} \tag{10.92a}$$

where the k-space amplitude $\tilde{x}_{\tilde{m}}$ of each plane wave is given in terms of x_n by

$$\tilde{x}_{\tilde{m}} = \frac{1}{2^{N/2}} \sum_{n=0}^{2^N-1} \mathrm{e}^{-i\frac{2\pi}{(2^N)}\tilde{m}n} x_n \tag{10.92b}$$

In (10.92) it is seen that the direct lattice space and k-space amplitudes of the Fourier series are related to one another through unitary transformations.

As has been noted in the earlier discussions, the unitary nature of the transformation between these two different representations is ideal for a treatment of the transformations in (10.92) by means of a quantum computing algorithm. It only remains to associate the position variables and amplitudes of the wave vector sets in (10.92) with the qubits of a quantum mechanical system in order to implement quantum algorithms with which to process the position and k-space variables in the quantum system.

To make the association between these two types of variables, it is important to note that another way of representing the direct space position, x_n, is in terms of quantum kets $|n\rangle$ where $n = 0, 1, 2, \ldots, 2^N - 1$ are the integer labels of the lattice sites. In this representation, one can make the association of the direct lattice position $x_n = (\Delta x)n$ with the quantum state $\Delta x|n\rangle$, expressing the integer label of the lattice site n with the $|n\rangle$ quantum ket. It remains only to associate the $|n\rangle$ kets with the ground and excited states of the qubits of the quantum mechanical system forming the quantum computer.

For this association, in particular, the ket $|n\rangle$ can be expresses as a direct product state involving the orthonormal basis eigenstates $\{|0\rangle, |1\rangle\}$ of single qubits of the quantum mechanical system. This is done using binary arithmetic. Specifically, any decimal integer n can be realized as a sequence of 0's and 1's in a binary arithmetic so that these 0's and 1's can be realized as a direct product sequence of the $\{|0\rangle, |1\rangle\}$ basis of the quantum qubits.

To see how this works out, remember that to express a decimal integer N in terms of a binary integer representation,N must be written in the form [7]

$$N = \sum_n a_n 2^n \tag{10.93}$$

where $n = 0, 1, 2, 3, 4, \ldots$ and a_n is constrained so that $a_n = 0$ or 1. The binary representation of N is then expressed as the 0's and 1's of the array of coefficients $(\ldots, a_n, a_{n-1}, a_{n-2}, \ldots, a_0)$.

In this manner, for example,

$$7 = 2^2 + 2^1 + 2^0 \tag{10.94a}$$

so that $(a_2, a_1, a_0) = (1, 1, 1)$ is the binary representation of the decimal number 7. As another example consider the decimal number 6. To represent this in binary notation, write

$$6 = 2^2 + 2^1 \tag{10.94b}$$

so that $(a_2, a_1, a_0) = (1, 1, 0)$ is the number 6 in binary format [8]. Using these methods of translation, the resulting binary representations of the lattice site number can then be expressed in a straightforward manner as the direct product of qubits formed of ground and excited quantum states.

Consequently, in terms of a system composed of three qubits it follows that [8]

$$|7\rangle = |1\rangle|1\rangle|1\rangle = |1, 1, 1\rangle \tag{10.95a}$$

and

$$|6\rangle = |1\rangle|1\rangle|0\rangle = |1, 1, 0\rangle \tag{10.95b}$$

In (10.95) the ket on the far left represents a state of the system in the decimal representation, and the other two terms on its right are in binary representations.

The binary term in the center of the equality is written in detail in terms of the direct product of three qubits, and the binary term on the far right in the equality is in an abbreviated format. Notice that subscript site labels have been omitted on the 0's and 1's to simplify the notation.

In this way any of the decimal kets $|n\rangle$ can be expressed in terms of a direct product of binary qubit kets. This allows the problem to be performed in terms of the ground and excited states of a qubit quantum system.

Similarly, in k-space, $\tilde{x}_m$ is expressed in terms of quantum kets $|\tilde{m}\rangle$ where $\tilde{m} = 0, 1, 2, \ldots, 2^N - 1$ are the integer labels of the k-space lattice sites. One can then make the k-space association $\tilde{x}_m = (\Delta x)\tilde{m}$ with the quantum state $\Delta x|\tilde{m}\rangle$. In this representation

$$|n\rangle = \frac{1}{2^{N/2}} \sum_{\tilde{m}=0}^{2^N-1} \mathrm{e}^{i\frac{2\pi}{(2^N)}n\tilde{m}} |\tilde{m}\rangle \tag{10.96a}$$

where the associated inverse transformation is

$$|\tilde{m}\rangle = \frac{1}{2^{N/2}} \sum_{n=0}^{2^N-1} \mathrm{e}^{-i\frac{2\pi}{(2^N)}\tilde{m}n} |n\rangle. \tag{10.96b}$$

An explicit formula for the generation of the unitary transformations of the Fourier series in (10.96) by the application of logic-gate operations can be given, and the reader is referred to the literature for further discussions of these. Here it should be assumed that, due to the unitary nature of the transformations, the transformations can be realized by appropriate operations in a quantum computational system.

As an example of the unitary relationships in (10.96) consider their applications on the quantum superposition wave function state composed of 3 qubits with a total of $8 = 2^3$ possible orthonormal basis states for the total system. In direct lattice space, this superposition wave function has the form

$$\begin{aligned} &\frac{1}{\sqrt{8}}\{|0\rangle + |1\rangle + |2\rangle + |3\rangle + |4\rangle + |5\rangle + |6\rangle + |7\rangle\} \\ &\quad = \frac{1}{\sqrt{8}}\{|0,0,0\rangle + |0,0,1\rangle + |0,1,0\rangle + |0,1,1\rangle \\ &\qquad + |1,0,0\rangle + |1,0,1\rangle + |1,1,0\rangle + |1,1,1\rangle\} \end{aligned} \tag{10.97}$$

where the left-hand side of the equation is in the decimal representation and the righthand side of the equations is in the binary representation.

As in earlier discussions of the constant-balanced function problem, the superposition wave function state in (10.97) can easily be generated from the application of three H-gate operations to a binary qubit of the form $|0,0,0\rangle$. In this way [7, 8]

$$\tilde{H}_1\tilde{H}_2\tilde{H}_3|0,0,0\rangle = \frac{1}{\sqrt{8}}\{|0,0,0\rangle + |0,0,1\rangle + |0,1,0\rangle + |0,1,1\rangle + |1,0,0\rangle + |1,0,1\rangle + |1,1,0\rangle + |1,1,1\rangle\}. \tag{10.98}$$

Focusing on one particular ket in the mixed wave function in (10.98), an example of the transformation between the direct and k-space lattices can be made. For example, consider the direct space ket $|6\rangle = |1,1,0\rangle$. From (10.96) this state of the system can be expressed in terms of the k-space kets $\{|\tilde{0}\rangle, |\tilde{1}\rangle, \ldots, |\tilde{7}\rangle\}$ giving the relationship

$$|6\rangle = \frac{1}{\sqrt{8}}\sum_{\tilde{m}=0}^{7} \mathrm{e}^{i\frac{2\pi}{(8)}6\tilde{m}}|\tilde{m}\rangle. \tag{10.99}$$

Transformations of the other kets in (10.98) follow directly as in (10.99) so that (10.98) can easily be expressed either in a direct space or a k-space representation.

Given the preceding notation and transformations, the problem of determining the period of a function defined on a direct lattice can now be formulated and solved by a quantum computer algorithm. For these considerations, assume that $f(x)$ is a periodic function defined on a lattice of 2^N direct lattice sites but that its periodicity is not known. The determination of the periodicity of $f(x)$ can be shown to follow from a study involving the determination of both the direct and k-space representations of the function.

In the quantum algorithm, the solution of the periodic problem is obtained by applying unitary transformations. Consequently, the kets required for the computations must be of a highly specific form. As in the algorithm studied in the earlier computational example of the constant-balanced function problem, the kets operated on must hold both the input data needed to determine the value of the function $f(x)$ and the output form of the function obtained from that input data. The algorithm works on the input data entries of the kets to change the entries of the kets assigned to receive the function generated from that data to contain the values of the outputted functions.

In the absence of a superposition wave function, such a state ket containing a set of inputted data and the outputted value of the function computed from that input data is of the form

$$|x,f(x)\rangle = |x\rangle|f(x)\rangle. \tag{10.100}$$

Here the input data $|x\rangle$ is a ket composed as a direct product of N qubit eigenstates each of which involves the orthogonal basis $\{|0\rangle, |1\rangle\}$ and the outputted $|f(x)\rangle$ is a ket containing the value of the function obtained from the evaluation of $f(x)$. The $|f(x)\rangle$ containing the outputted function, similar to the $|x\rangle$ input data ket, can be expressed in binary notation using the ideas presented in (10.93)–(10.95) and the qubit eigenstates of the computer. Consequently, the kets in (10.100) involve only a series of 0's and 1's in their arguments and represent eigenstates of the quantum computer system.

The calculation begins by using an H-gate to generate a fully superposed wave function state of input data and output state qubits to receive the answer. In particular, this is accomplished by the following application of H-gates [7, 8]

$$\tilde{H}_1\tilde{H}_2\ldots\tilde{H}_N|0,0,0,\ldots,0\rangle\tilde{H}_{N+1}\ldots|1,1,1,1,\rangle. \tag{10.101}$$

Here the ket of initial 0's is being prepared to represent a fully superposed wave function state of the input data, x. In this mixture, all possible input data sets are represented and equally weighted in the superposition wave function input state. As shall be seen later, each of these superposed eigenstates of the input data will be used for the determination of the function value for that eigenstate value and the generated value of the function is placed in the corresponding output part of the ket corresponding to the input data.

The ket of initial 1's in (10.101) is for the reception of the output values of the function, $f(x)$. The superposition wave function of the kets reserved in (10.101) for the answer is required in order to generated a complete set of quantum states. It is necessary to have a complete orthonormal basis in which to define unitary operations on the quantum system, i.e., completeness is a necessary mathematical property of the Hilbert space.

Upon applying the H-gates to each of the single qubit kets in the direct products in (10.101), it is found that for the action on a single qubit

$$\tilde{H}|0\rangle = \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle] \tag{10.102a}$$

and

$$\tilde{H}|1\rangle = \frac{1}{\sqrt{2}}[|0\rangle - |1\rangle]. \tag{10.102b}$$

The application of the H-gates in (10.101) and (10.102) then results in a superposition wave function qubit state made ready to begin the calculation. A consequence of the mixed nature of the wave function generated in (10.101) is that the algorithms applied to the wave function state will result in a massive parallel processed computer calculation.

The next step in the calculation involves an operation on each of the pure basis states contained in the superposition wave function. A particular pure state in the sum in (10.101) is of the general form

$$|x\rangle|y\rangle. \tag{10.103a}$$

where the input data describing the direct space lattice $|x\rangle$ is a ket composed as a direct product of N qubit eigenstates each of which involves the orthogonal basis $\{|0\rangle, |1\rangle\}$ and the $|y\rangle$ ket is a direct product of single eigenstate qubits reserved to receive the value of the function obtained from the evaluation of $f(x)$. Consequently, evaluating the function $f(x)$ at the position $|x\rangle$ transforms the $|y\rangle$ ket into the ket $|f(x)\rangle$ of received function values.

This operation may be described as [7, 8]

$$|x\rangle|y\rangle \overset{Operation}{\rightarrow} |x\rangle|f(x)\rangle. \tag{10.103b}$$

where x represents a pure state composed of 0's and 1's and $f(x)$ is the value of the function evaluated for this input. Consequently, after performing the operation on each of the kets in the sum of (10.101) the system is left in as state which is a superposition wave function sum of states of the form of (10.103b) containing the positions and corresponding function values. Following this transformation, the resulting wave function is

$$\frac{1}{\sqrt{2^N}}\sum_x |x\rangle|f(x)\rangle. \tag{10.104}$$

Next the calculation switches to a focus on the state $|x\rangle$. In particular, the position state ket of the direct lattice is next transformed to the k-space representation. Under this process, from (10.92) it follows that

$$|x\rangle \overset{Process}{\rightarrow} \frac{1}{2^{N/2}}\sum_{\tilde{k}=0}^{2^N-1} \mathrm{e}^{i\frac{2\pi}{2^N}\tilde{k}x}|\tilde{k}\rangle. \tag{10.105}$$

Applying this to (10.104) it follows that the wave function is transformed to the form [8]

$$\begin{aligned}\frac{1}{\sqrt{2^N}}\sum_x |x\rangle|f(x)\rangle &= \frac{1}{2^N}\sum_{x,\tilde{k}=0}^{2^N-1} \mathrm{e}^{i\frac{2\pi}{2^N}\tilde{k}x}|\tilde{k}\rangle|f(x)\rangle \\ &= \frac{1}{2^N}\sum_{x=0}^{2^N-1}|f(x)\rangle\left(\sum_{\tilde{k}=0}^{2^N-1}\mathrm{e}^{i\frac{2\pi}{2^N}x\tilde{k}}|\tilde{k}\rangle\right).\end{aligned} \tag{10.106}$$

The form of the wave function in (10.106) is very useful in determining the period of the function $f(x)$. Of particular importance for this determination are the phase factors $\sum_{\tilde{k}=0}^{2^N-1}\mathrm{e}^{i\frac{2\pi}{2^N}x\tilde{k}}|\tilde{k}\rangle$ multiplying the states $|f(x)\rangle$. The importance of the phase factors can be seen by considering the effects of a periodic function on the evaluation of (10.106).

There are 2^n lattice sites over which the function $f(x)$ is defined so that if T is the integer number of sites over which the function is periodic, then [7, 8]

$$f(x+T)=f(x) \tag{10.107a}$$

Consequently, applying (10.107) in (10.106) the various $|f(x)\rangle$ kets for a periodic function are related by

$$|f(x+T)\rangle = |f(x)\rangle \tag{10.107b}$$

In the case of the periodic function of period T many of the function kets, $|f(x)\rangle$, in (10.106) will be equal, and the complex coefficients of these identical kets will add together.

Due to the relationship in (10.107b), the resulting phase coherence in the subsequent $\tilde{k}$ sums in (10.106) has the consequence that only terms satisfying the conditions

$$\tilde{k} = 0, \frac{2^n}{T}, 2\frac{2^N}{T}, 3\frac{2^N}{T}, \ldots, (T-1)\frac{2^N}{T} \tag{10.108}$$

are present in the final wave function in (10.106). Terms not satisfying (10.108) are absent from the wave function. From this fact a number of measurements can be made involving the output states generated by the quantum algorithm, and these measurements allow for the determination of the period of the function under study.

An example will illustrate the point. Consider, a lattice consisting of 2^3 points. On such a lattice, periodic functions defined on the lattice will only have periods of $T = 1, 2, 4, 8$. Assume that $f(x)$ is one of these periodic functions and you are asked to determine its period from the discussed algorithm.

The superposition wave function state in (10.104), resulting after the application of $f(x)$ to the entangled input state from (10.101), is given by [8]

$$\begin{aligned}\frac{1}{\sqrt{8}}\sum_{x=0}^{7}|x\rangle|f(x)\rangle = \frac{1}{\sqrt{8}}\{&|0\rangle|f(0)\rangle + |1\rangle|f(1)\rangle + |2\rangle|f(2)\rangle + |3\rangle|f(3)\rangle \\ &+ |4\rangle|f(4)\rangle + |5\rangle|f(5)\rangle + |6\rangle|f(6)\rangle + |7\rangle|f(7)\rangle\}\end{aligned} \tag{10.109}$$

Following the earlier discussions, the entangled wave function in (10.109) is next transformed to be represented in k-space.

Applying (10.106) to the system in (10.109) to make the k-space transformation gives [8]

$$\frac{1}{\sqrt{8}}\sum_{x=0}^{7}|x\rangle|f(x)\rangle = \frac{1}{8}\sum_{x=0}^{7}|f(x)\rangle\left(\sum_{\tilde{k}=0}^{7}e^{i\frac{2\pi}{8}x\tilde{k}}|\tilde{k}\rangle\right). \tag{10.110}$$

The resulting wave function in (10.110) is now in a ready form from which the periodicity of the function can be revealed. It remains only to investigate the nature of the terms composing the total wave function of the system. This can be done by collapsing the wave function through a measurement process.

As an example of determining the period of $f(x)$ from (10.110) and the conditions in (10.108), consider the form of the output wave function in (10.110) if the unknown periodicity is in fact a period $T = 2$ [8]. If this is the case, then it follows that $f(0) = f(2) = f(4) = f(6)$ and $f(1) = f(3) = f(5) = f(7)$.

From (10.108) it is seen that only $\tilde{k} = 0$ and $\tilde{k} = 4$ terms should be present in (10.110). In fact, from (10.95) and the periodicity conditions of $f(x)$ in the previous paragraph, it follows from (10.110) that

$$\frac{1}{\sqrt{8}}\sum_{x=0}^{7}|x\rangle|f(x)\rangle = \frac{1}{2}\left\{|0\rangle[|f(0)\rangle + |f(1)\rangle] + |4\rangle\left[|f(0)\rangle + \mathrm{e}^{i\pi}|f(1)\rangle\right]\right\}. \tag{10.111}$$

Consequently, the wave function in (10.111) is composed of two states of input data. These are states of $|0\rangle$ and $|4\rangle$. In the $|0\rangle$ state the three binary qubits of the input vector are all in the ground state so that

$$|0\rangle = |0,0,0\rangle. \tag{10.112a}$$

In the other $|4\rangle$ state, the three binary qubits of the input vector have one qubit in an excited state and the other two in ground states. Consequently, for the second state [8]

$$|4\rangle = |1,0,0\rangle. \tag{10.112b}$$

It is important to note that a single measurement of the wave function in (10.111) does not give the definitive answer regarding the periodicity, but multiple runs will ultimately reveal the answer. This indicates the probabilistic nature of obtaining an answer from a quantum computer. Answers with high probability, however, show up in the system to the largest extent and are often all one needs to determine that correct answer.

In quantum computing one must often devise an algorithm which gives a very high probability of presenting the correct answer to a problem when the output wave function is collapsed by a measurement. The success of the computation then relies heavily on the formulation of cleaver algorithmic processes and is only useful as it leads to much faster computations than those available from classical computers.

References

1. L. Maccone, A simple proof of Bell's inequality. Am. J. Phys. **81**, 854–859 (2013)
2. R.B. Griffiths, EPR, Bell, and quantum locality. Am. J. Phys. **79**, 954–967 (2016)
3. G. Blaylock, The EPR paradox, Bell's inequality, and the question of locality. Am. J. Phys. **78**, 111–122 (2010)

4. D. Petz, Entropy, von Neumann and the von Neumann entropy, in *John von Neumann and the Foundations of Quantum Physics*, ed. by M. Redei, M. Stoltzner (Kluwer, Dordrecht, 2001)
5. S. Rolston, Getting the measure of entanglement. Nature **528**, 48–49 (2015)
6. M.B. Plenio, S. Virmani, An introduction to entanglement measures. Quantum Inf. Comput. **7**, 1 (2007)
7. C.P. William, *Explorations in Quantum Computing* (Springer, New York, 2011)
8. G.P. Berman, G.D. Doolen, R. Mainieri, V.I. Tsifrinovich, *Introduction to Quantum Computers* (Wold Scientific, Singapore, 1998)
9. A. Steane, Quantum computing. Rep. Prog. Phys. **61**, 117–173 (1998)
10. T.D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, J.L. O'Brien, Quantum computers. Nature **464**, 45–53 (2010)

Index

A. McGurn, *Nanophotonics*, Springer Series in Optical Sciences 213,
https://doi.org/10.1007/978-3-319-77072-7