10.0kV 5.0mm x4.49k   10.0um

Fundamentals and Applications of Nanophotonics

Edited by Joseph W. Haus

Fundamentals and Applications of Nanophotonics

Related titles

Directed Self-assembly of Block Co-polymers for Nano-manufacturing
(ISBN 978-0-08-100250-6)

Modeling, Characterization, and Production of Nanomaterials
(ISBN 978-1-78242-228-0)

Optofluidics, Sensors and Actuators in Microstructured Optical Fibers
(ISBN 978-1-78242-329-4)

Woodhead Publishing Series in Electronic and Optical Materials: Number 85

Fundamentals and Applications of Nanophotonics

Edited by

Joseph W. Haus

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

For information on all Woodhead Publishing publications
visit our website at http://store.elsevier.com/

Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Contents

This page intentionally left blank

List of contributors

D. de Ceglia   National Research Council, AMRDEC, Redstone Arsenal, AL, USA

J.W. Haus   University of Dayton, Dayton, OH, USA

N.M. Litchinitser   University at Buffalo, The State University of New York, NY, USA

A. Sarangan   University of Dayton, Dayton, OH, USA

M. Scalora   Charles M. Bowden Research Center, AMRDEC, RDECOM, AL, USA

J. Sun   University at Buffalo, The State University of New York, NY, USA

M.A. Vincenti   National Research Council, AMRDEC, Redstone Arsenal, AL, USA

This page intentionally left blank

Woodhead Publishing Series in Electronic and Optical Materials

This page intentionally left blank

Preface

Life is like riding a bicycle. To keep your balance you have to keep moving.
Albert Einstein Letter to his son Eduard (February 5, 1930)

ἡ γὰρ νοῦ ἐνέρεια ζωή

The energy or active exercise of the mind constitutes life.
From The Metaphysic, translated from the Greek by Rev. John H. M'Mahon in The
Metaphysics of Aristotle (1857), Book XI, p. 332

The genesis of this book, *Fundamentals and Applications of Nanophotonics*, was a graduate course that was taught for several years at the University of Dayton. It was once taught at Harbin Institute of Technology in China and reconstituted as a short course at many professional meetings. My colleagues at the University of Dayton, Andrew Sarangan and Qiwen Zhan, who are experts in the field of nanophotonics, originally co-taught and brought to the course their complementary knowledge and expertise in subfields of nanophotonics. The reception that we had with our courses persuaded me that a book incorporating the basic concepts of nanophotonics would be a valuable guide for students seeking to understand the field. This book is the product of my deliberations on the essential topics. Unfortunately, completing the book was delayed by a myriad of distractions over the past decade. It was only when I let go of all administrative duties that I was able to really devote my time and energy to completing this book.

Even so, writing this book in a reasonable time frame turned out to be a task that required more effort than I could provide, so I solicited support from experts in the field. I am grateful that my colleagues have produced chapters that add to the essential subject matter in the field and will contribute to students' understanding of the subject. Each chapter is written in a tutorial fashion and the concepts are treated in depth. Therefore, we do not try to cover everything in the field to keep the material confined to a subset of topics.

The book is written for engineers and scientists at a first-year graduate student level. It contains adequate mathematical detail to give the reader an understanding of how specific physical concepts are applied to nanotechnology and more detailed references are provided for the reader interested in delving into the real depths of the subject. The chapters contain problems that can be assigned for homework.

The field we call nanophotonics (i.e., the study of optical devices with nanometer-scale features) is driven by developments in three different areas: fabrication, characterization, and materials. It is a vast field that cuts across many disciplines: biology, chemistry,

physics, materials science, electrical and mechanical engineering, and many other disciplines not mentioned. The story behind the development of nanophotonics is a fascinating one with many contributors across the globe reporting novel results and sometimes the ideas have remained nascent for decades before a researcher has a keen new insight that propels the field into the limelight. For instance, this is the case with photonic crystals and metamaterials.

This book provides the reader with the background knowledge of the foundation of nanophotonics and then it applies the basic principles to advance new concepts in designing materials with desired physical properties. The book will prepare the diligent reader with the tools to read and understand the research literature. The field is still emerging with advancements in research tools that are rapidly developing. It is the hope of the nanophotonics community that this initial development phase will evolve one day to useful technological applications that harness the unique physical phenomena at the nanoscale. Of course, applications can come from any field applying photonics devices, such as medical and environmental sensors, energy harvesting, and information technology.

The reader should not be afraid to break out of the traditional mold and follow their own instincts to discover an unexplored path. All contributors to this book share a deep conviction that students should learn skills in all subfields and become facile enough with the concepts to contribute their own ideas as they embark on their journey of discovery. Each author is an expert who has made significant research contributions to progress in the field. Together we are attempting to meet the requirement of a broad base of knowledge and experiences with a mathematical foundation to understand physical phenomena on the nanoscale.

In our approach to the subject, there is a balance of the two requirements of mathematical rigor with physical understanding of the phenomena; moreover, the student will have exposure to experimental fabrication tools and characterization instrumentation and techniques. I generally believe that real progress in the field demands knowledge of widely different skills; the student should have the capability to perform simulations and have knowledge of the suite of fabrication tools available with their limitations and be aware of techniques to characterize the materials and devices.

I want to thank my colleagues, Domenico de Ceglia, Natalia M. Litchinitser, Andrew Sarangan, Michael Scalora, Jingbo Sun, and Maria Antonietta Vincenti, who kindly gave of their time and talent to write chapters and work together to make the book an organic whole. I am indebted to the many students who have taken the nanophotonics course over the years and who have given feedback about their judgment on various lectures and topics.

Because of the constraints on my time, there are undoubtedly errors in the text. This occurs despite readings of the chapters by colleagues. The reader who finds errors or misconceptions is encouraged to email me so that I can publish amendments to the book. The amendments can be obtained by emailing a request to me. My email address is widely available on the Web.

J.W. Haus
Dayton, OH, USA
August 2015

Introduction to nanophotonics

1

J.W. Haus
University of Dayton, Dayton, OH, USA

> *If I have seen further it is by standing on the shoulders of Giants.*
> **Isaac Newton, from the Correspondence of Isaac Newton.**

> *Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.*
> **Marie Skłodowska-Curie, quoted in** Our Precarious Habitat *(1973) by*
> **Melvin A. Benarde, p. v.**

1.1   Introduction

We are embarking on a journey across the broad subject of nanophotonics. Nanophotonics activities engage researchers in many disciplines: optics; physics; chemistry; electrical, chemical, and mechanical engineering; materials science; biology; and mathematics. The range of activities is far too large to be captured by a single tome. Therefore, the number of topics covered in this book is narrowed by our personal choices. We divide the field of nanophotonics into three broad categories, which are distinguished by their functionality: materials, fabrication, and characterization. Of course, the prefix *nano* can be inserted in front of all three words for emphasis, but let us not overuse the term at the outset. The three aspects of nanophotonics are analogous to the legs of a three-legged stool with nanophotonics resting on top of them, as illustrated in Figure 1.1. Researchers and engineers apply all three of these areas to solve technological challenges. The image conveys our bias that removing one of the topics would leave out essential knowledge that we believe every student of nanophotonics should be familiar with.

The required breadth of the field is also conducive to the formation of collaborations within and across institutions; the members of the collaboration contribute their expertise in a specific field to the completion of the project. The student or new researcher in the field will find it advantageous to possess knowledge of the basic principles for all three aspects to effectively communicate with colleagues and to understand the main issues from different perspectives. A solid foundation in all three aspects provides access to a broader range of literature and gives deeper insights into physical limitations. This book was conceived to provide readers with a working knowledge of the principles in several fields.
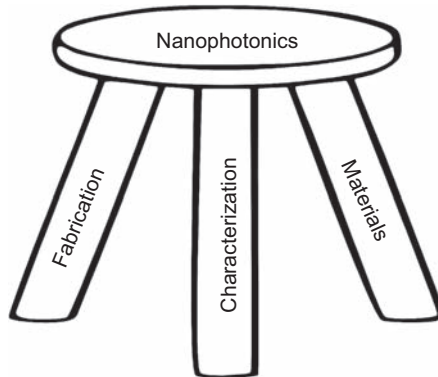
Figure 1.1 Three legs of a stool illustration to emphasize the three aspects of nanophotonics.

1.2   Materials

Technological progress is often based on the development of new materials. To highlight that point in a historical context, recall that two of the three named archaeological periods of man's ascent are named after materials technology—the bronze and iron ages. The further development of new materials has proceeded through the millennia and today, we are guided by a deeper quantitative understanding of materials' composition and properties.

Modern-age materials are much more sophisticated, but the main technology drivers are the same. In the modern era, silicon has emerged as one of the most important materials with applications ranging from electronics to micro-electro-mechanical systems (MEMS) to electronic sensors to photonics. There is market pressure to fabricate devices that have new functionality and can be produced using materials that are widely available and inexpensive. Silicon is one of the most abundant elements on Earth, second only to oxygen, making it an ideal material for many technological applications.

Today nanophotonics research is devoted to creating new materials with functionality that is not ordinarily available by processing natural elements, alloys, or compounds. The elements of the periodic table are cast into a form that is not normally found in nature. The purpose is to elicit a new response to electronic current or optical waves. The new response may be potentially useful for fabricating new devices or for making a new instrument with improved resolution.

The abundance of the elements is another important consideration in pursuing new technologies. Figure 1.2 is a plot of the elements' relative abundance found in the earth. The high abundance of elements such as hydrogen, silicon, oxygen, iron, carbon, and nitrogen (abundant in the atmosphere) make them inexpensive and easy to procure. Other in-demand elements may have a supply that is controlled by a single country; thus, they are subject to political crises or pressures. For instance, the rare-earth metals have found important uses in electronics, lighting, displays, and communications technologies. Fortunately, the amount of rare-earth elements used
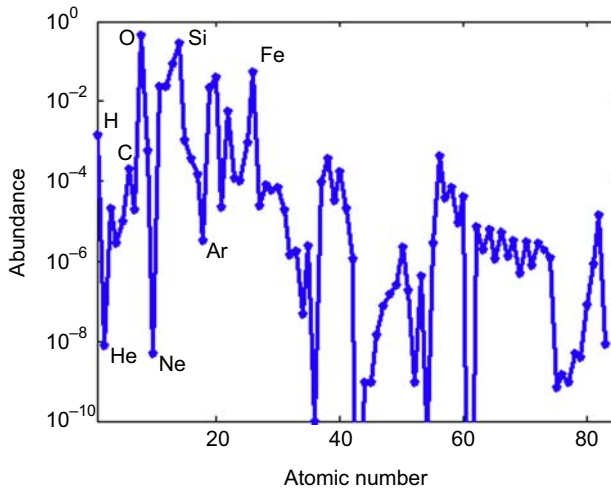
Figure 1.2 Abundance (by mass) of the elements in the Earth's upper continental crust up to element 82.
Plotted from data available from http://en.wikipedia.org/wiki/Abundance_of_the_chemical_elements.

in many of these applications is tiny, and global-scale use is not affected by their market availability. In developing new materials with desired functionality, one may be driven by the need to reduce or eliminate expensive or geopolitical elements. A commercial product that uses scarce, expensive, or geographically concentrated elements will not be sustainable on a massive scale over a long period of time.

Chemical and physical properties of materials are deeply connected to the quantum mechanics of the hydrogen atom. The framework for understanding atomic spectra is based on the Bohr model of electrons in orbits around a small, positively charged nucleus. His quantization of the angular momentum of the electrons prevented them from collapsing to the nucleus, as would be expected from classical Newtonian dynamics. The essence of the Bohr model was captured in quantum mechanics, which describes electrons in their orbital states. For the quantum solution of the hydrogen atom, the Coulomb attraction of the electron to the proton yields a set of states that are labeled by a set of quantum numbers describing the energy, angular momentum, and the spin momentum of the electrons. Surprisingly, the simple model of the hydrogen atom extends to the electronic states of multielectron atoms; the occupation of electronic states conforms to state designations that are identical to the hydrogen atom.

The treatment of molecules and solids is organized around the atomic orbital states of constituent atoms. However, to understand the electron bonding energies and geometry, such as bond angles and lattice symmetry, multiple electronic states are combined. The bonding length and direction between atoms, although closely connected to the atomic states, are based on simple electronic principles; namely, paired electrons with opposite intrinsic spin are shared between atoms to form a *covalent bond*.

The covalent bond angles are based on electron pairs shared between atoms and can be understood by using a linear combination of atomic orbitals (LCAOs). In chemistry, the LCAO is termed hybridization and is discussed in Chapter 4. The LCAO is a simplification that uses valence electron wave functions as a set of basis states to calculate a wide range of physical properties. The technique provides a simple physical description of the bond angles in molecular and lattice systems.

Electron interactions manifest different bonding mechanisms that may be identified from their electronic density distributions. *Metallic bonding* has electron density delocalized throughout the solid, whereas covalent bonds have valence electron density largely localized between the atomic nearest-neighbor pairs. The malleability of many metals is an indication that the metallic bonds are relatively weak. Similar to covalent bonding, *ionic bonding* is based on electron density displacement between the pair of atoms to form a cohesive crystal; the atoms called cations give up the electrons, leaving a net-positive charge and the electron has a density displaced more to the other atoms called anions. Ionic and covalent bonds can both be present to some degree in solids. The covalent bonds are highly dependent on orientation because of the bonding orbitals, whereas the ionic bonds depend on the difference in affinity of the atoms to attract an electron, so-called electronegativity. Finally, there is a fourth type of bonding that is commonly mentioned called *van der Waals bonds*, which is dominant for atoms with closed electron shells (e.g., inert gas atoms). The electrons in inert gas atoms are tightly bound to the nuclei and they are not shared with neighboring atoms. For inert atoms, their bonding in the solid state is promoted by induced atomic dipoles on the atoms due to quantum fluctuations, which are *van der Waals interactions*. This form of bonding is weak.

Scientists and researchers refuse to be confined by the materials refined from naturally occurring compounds. Chemists and materials scientists have tweaked the growth and synthesis processes to create new materials or nanostructured materials that exhibit unusual electronic and optical properties. For instance, there has been enormous progress made in designing and fabricating quantum-confined structures that are now commercially available in semiconductor lasers and photodetectors. Recent activities have led to the development of novel, fabricated materials' classifications—called photonic crystals, plasmonics, and metamaterials—that are designed to accentuate specific properties of light. These specific materials are introduced and discussed in subsequent chapters of this book. Because each class of new materials have exposed the community to novel physical concepts, they have spawned worldwide research efforts to discover new applications for the novel laboratory materials. For instance, metals, also called plasmonic materials, have taken center stage with traditional applications in biosensing and potential applications in energy harvesting and photonic surfaces.

1.3   Fabrication and characterization

The techniques to create new photonic devices and optical materials have been divided into two areas—*top-down* and *bottom-up* approaches—and their general features are illustrated in Figure 1.3. The top-down approach is exemplified by the fabrication
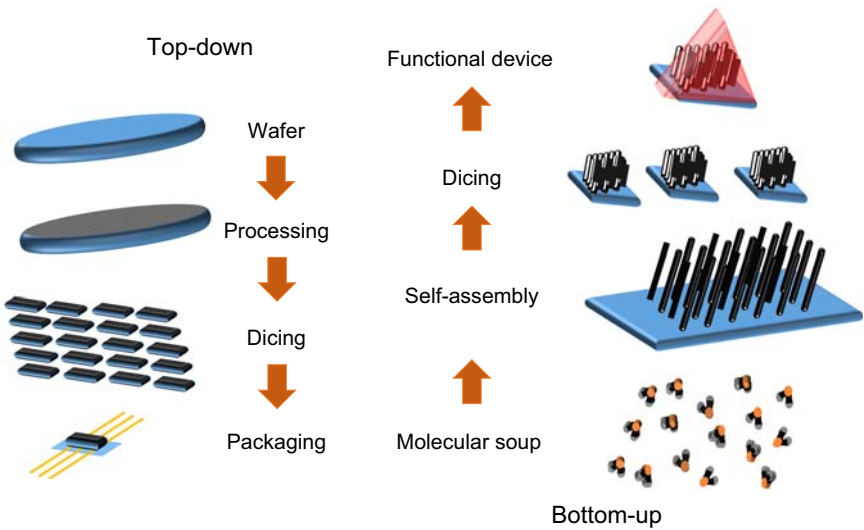
Figure 1.3 Top-down and bottom-up approaches.

techniques used to make electronic circuitry and MEMS devices, which have become ubiquitous elements in commercial devices from phones and iPads to automobiles. The starting point is a wafer substrate material that is a building block with required characteristics. The wafer has material removed, added, or doped using the available suite of fabrication tools. A high yield of devices demands that the environment be controlled using a cleanroom. After completing the fabrication process, the wafer may contain thousands of functional chips. By dicing the wafer, the chips are available for packaging to produce the final product. The available processing tools have seen several generations of improvements because of the electronics industry's relentless drive to make smaller and smaller devices.

Bottom-up approaches are characterized by controlled chemical synthesis or by physically directing growth of an initial "molecular soup" with all of the ingredients to build the final structure. The processes that organize the constituents into a final product are broadly called self-assembly techniques. Many of the approaches are bio-inspired following nature's ability to build complex systems from molecular building blocks. In self-assembly, the molecules are collected into nanosized structures that can be termed superatoms. An example of nature's self-assembled bottom-up approach is found in opaline materials. The beautiful colors of opals are the result of light interference and diffraction from a periodic arrangement of silica spheres. In modern terminology, the structures would be called photonic crystals. Chemists and materials scientists have harnessed nature's ability to form opals and have improved on it using methods to synthesize complex one-, two-, and three-dimensional photonic crystals using block copolymers. The bottom-up creation of inverse opals (i.e., synthesizing a scaffolding around the silica spheres before removing the spheres by chemical etching) is an example of stepping beyond nature to create a material with properties not naturally found.

We mention as a historical reference a 1959 lecture by the iconic Nobel Laureate Richard Feynman. The talk went largely unnoticed for several decades, but it was recently popularized in several publications. In December 1959, Feynman presented a talk entitled "There's Plenty of Room at the Bottom." As the title suggests, he saw no physical limitations to writing information on a much smaller scale (e.g., writing the entire *Encyclopedia Britannica* on the head of a pin). In Feynman's words, "there is plenty of room to make them (computers) smaller. There is nothing that I can see in the physical laws that says the computer elements cannot be made enormously smaller than they are now. In fact, there may be certain advantages." Although this article does not have a historical position in the development of nanotechnology, it still is relevant for the physical foresight Feynman showed as sizes are scaled down.

The top-down fabrication techniques used to make nanophotonic devices and materials leverage on decades of progress and investments made in the electronics industry to develop higher precision tools and control of nanometer-sized critical dimensions sustained over wafer-sized areas. The drive toward nanofabrication is most evident in the electronics industry. After the initial ideas for designing and fabricating integrated circuitry (IC) were conceived, the technology was continually refined and improved in continuous cycles to reduce the dimensions of electronic devices on a chip. A process that continues to the present day and a comprehensive, industry-driven is found in the International Technology Roadmap for Semiconductors. Its origins go back to the mid-1960s and the prognostication by Gordon Moore that the number of components integrated on a chip would grow exponentially (i.e., double approximately every 2 years). This prediction has been dubbed Moore's law, and it is remarkable that exponential growth has been adhered to for more than 4 decades.

To keep pace with Moore's law, the industry developed new techniques and tools that improved the production and yield of the products. Modern nanotechnology requires a suite of instruments performing specific tasks that are needed to make the final product. In general, making high-performance devices requires lithography and patterning, materials etching techniques, and materials deposition to fabricate. Table 1.1 lists the major nanofabrication technology steps. Semiconductor lithography is defined by several techniques to write a pattern on a wafer surface. Knowledge about each fabrication step is invaluable for understanding the obstacles that need to be overcome in fabricating a new material or device.

Table 1.1 Major fabrication technology processes

Process	Examples
Lithography and patterning	Ultraviolet or extreme ultraviolet light, X rays, electron beams, dip-pen, and nanoimprint lithographies
Deposition techniques	Physical vapor deposition, chemical vapor deposition, molecular beam epitaxy
Etching	Wet and dry (plasma) techniques

As the origin of the name suggests, lithography is the process of transferring a pattern onto a substrate. The pattern was traditionally made on a mask, and the pattern was transferred to a photoresist by optical illumination using ultraviolet (UV) light. As feature sizes were reduced, new processing techniques were worked out and new tools, such as extreme UV and electron beam writing, were developed. Additional tools deposited materials with nanometer accuracy or embossed a nanometer feature in a thermoplastic polymer. The pattern once transferred to a thin film is used to either add material to the substrate by deposition techniques or remove material using etching techniques. The fabrication processes and techniques are treated in detail in the fabrication chapter. There are also instances in which the fabrication of new devices is a blend of top-down and bottom-up approaches. Intermediate steps in the process of making a final device can incorporate new bottom-up techniques that are compatible with traditional cleanroom fabrication requirements.

In Feynman's 1959 lecture "There's Plenty of Room at the Bottom," he also challenged the community to improve metrology by developing higher resolution instruments to image the written features. Writing the *Encyclopedia Britannica* on the head of a pin is a simple fabrication challenge demonstration that Feynman proposed; however, writing information is only one part of the process. To be complete and useful, an instrument is needed that can faithfully read the written content, for example, on the head of a pin. Without the reader, the fabrication process would be incomplete without a process to verify the faithful writing of the data.

The same need carries over to the design of functional nanostructures on substrates. After designing, fabricating, or chemically synthesizing the nanostructures, the samples need to be characterized with an instrument that resolves to a scale less than the critical dimensions of the nanostructures. Thus, the development of nanoresolution imaging instrumentation is indispensable for examining the features created during the fabrication process. Fabrication is not complete until the process has been monitored and verified at all stages. Characterization is a necessary and important step in the development of our understanding of nanostructured materials and the processing steps so that the fabrication can be cyclically improved.

The list of characterization techniques in Table 1.2 reflects traditional and recent metrology developments. Optical microscopes are well understood and have been developed over several centuries with new techniques to increase resolution discovered mainly in the twentieth century. Added to traditional optical microscopy

Table 1.2 **Characterization techniques**

Imaging technique	Examples
Optical techniques	Optical microscopy, ellipsometry, optical spectroscopies, laser pulse-probe techniques
Electron microscopes	Scanning electron and transmission electron microscopy
Scanning probe techniques	Atomic force, scanning tunneling and near-field optical microscopies, etc.

is a slew of other techniques as a result of the development of coherent sources and of our understanding of incoherence. Nonlinear optical techniques have also added to our repertoire of instruments that enable submicron-feature resolution. The 2014 Nobel Prize in Chemistry shared by Stefan Hell, William Moerner, and Eric Betzig acknowledged the remarkable resolution improvements that can be achieved by developing novel applications of advanced laser technology. Electron imaging techniques based on the quantum wave nature of the electron (i.e., its de Broglie wavelength) were first demonstrated in the 1930s, and the technology was greatly advanced after World War II. There is a plethora of modern scanning techniques that have appeared since the invention of the scanning tunneling microscope by Gerd Binnig and Heinrich Rohrer in 1981 (1986 Nobel Prize in Physics); the tunneling instrument has a resolution extending to the scale of Angstroms or less. Since then, the list of scanning probe methods has expanded, which adds to the list of extremely high-resolution techniques. Many of the metrology breakthroughs have also been internationally acknowledged by conferring of Nobel Prizes.

1.4   Devices

Although the pursuit of new knowledge that confers deeper insights into atom−photon interactions on the nanoscale is an important quest that shapes our fundamental understanding of nature, it is the development of new photonic devices applying the new principles that is affecting rapid change in our world. The pace of technological advancement has accelerated on a scale that was scarcely comprehensible only a century ago. The march of technology is apparent in the constant evolution of electronic devices developed for entertainment. For example, recorded music is widely enjoyed by everyone, and the electronics technology we use to enjoy it has gone through a rapid evolution over the past six decades. The storage of music saw a progression from analog technology using vinyl records; then tapes gave way to digital technology with CDs, which were used in compact players largely because of the introduction of semiconductor lasers; and, subsequently, to electronic as electronic memory technology (e.g., flash memory storage) matured and they were incorporated into music devices. The need to lug around dozens of CDs to play in a bulky player on long flights was obviated after solid-state storage devices were introduced in the marketplace; the devices are so compact that they fit in the palm of our hands and contain hundreds of albums (try carrying that many CDs on a long flight!).

The Nobel Laureate and nanotechnologist Richard Smalley was a passionate advocate for science education and used his talents to espouse a list of the top global problems facing humanity in the next half century. Smalley's list in Table 1.3 (modified from the original by combining environment and poverty) includes challenging problems that are open ended and closely interrelated. Incremental technology improvements make the problems more manageable. Nanotechnology researchers continue working toward solutions to Smalley's challenges, especially if it has special potential applications to the fields of energy, water, disease, food, and environment.

Table 1.3 **Smalley's list of top global problems facing humanity in the next half century**

Humanities top problems
Energy
Water
Food
Environment and poverty
Terrorism and war
Disease
Education
Democracy
Population

Tackling problems in these fields engages engineers and scientists from across the globe who practice the concept of "science without borders." They not only attend international conferences, but they develop collaborative research partnerships with personnel exchange and division of labor among fabrication, characterization, and device testing to advance photonics technology.

To illustrate the complexity and connectivity of the problems, consider the energy challenge. It is a problem with many facets that can be addressed. An energy solution should be sustainable and use renewable resources, suggesting alternative approaches that also affect water, environment, poverty, and disease. Among several approaches, there are solar cells or photovoltaics. They are a clean source of energy production, but they have relatively low energy harvesting efficiency. Photovoltaic device research is spurred to discover continual improvements by incorporating novel materials, such as perovskites and nanoparticles, and designing tandem layers to enhance their energy conversion efficiency. Furthermore, solid-state lighting, such as light-emitting diodes, and electronics have substantially improved humans' standard of living, and optical communication has enabled the dissemination of information with unprecedented speed around the globe. Both technologies impact Smalley's education and democracy challenges, while at the same time managing energy consumption. Other areas of the energy challenge in which nanotechnology can play a role are reducing energy losses in transmission and improving energy storage devices.

In the field of telecommunications, the development of low-loss optical fibers and photonics technology with its enormous bandwidth has outpaced other means of communication and ushered in a new era of the information age in which photonics is the dominant technology. Driving these systems are photonic devices, semiconductor lasers, and photodetectors that are products of nanophotonics research. There are also breakthroughs in the channel carrying the information. For example, the introduction of optical amplifiers in long-haul fiber-optic systems sustained the optical

signals over distances of tens of thousands of kilometers, thus removing the electronics bottleneck from fiber-optic systems. This enabled fiber-optic systems to be transparent to the bit-rate of the signal, and many channels could be multiplexed together and transmitted over distances extending between and around continents. The fiber communication systems' reliance on photonic devices extends beyond semiconductor lasers and detectors to include modulators to encode data and optical multiplexers and demultiplexers to route the data from one point to another.

Semiconductor materials grown with nanometer-scale dimensions elicit quantum properties because of the confinement of carriers (i.e., electrons and holes). Thin films that confine carriers are common in optoelectronic devices. For instance, semiconductor lasers fabricated with the quantum confinement of carriers have lower power consumption than their predecessors to meet the growing demands of data centers to increase the speed and density of information transport. Photodetectors designed with quantum confinement properties can be used to design imagers for long infrared wavelengths.

In 1958, Jack Kilby at Texas Instruments received an IC patent that marks the beginning of the electronics revolution. He was recognized for this invention by a Nobel Prize in 2000. Within 6 months after Kilby's patent, Robert Noyce at Fairchild Semiconductor patented an IC concept that proved to be practical and formed the basis for the development of complementary metal-oxide semiconductor (CMOS) technology.

Research in the field called silicon photonics devices has piggybacked on recent CMOS developments by applying the same tools used in cleanroom environments. In addition, a silicon wafer is prepared by burying a silicon oxide layer electrically insulating the silicon surface layer; this technique is called silicon-on-insulator (SOI) technology, and it is useful for designing photonic waveguides using the top layer of silicon using the high index of silicon relative to silicon oxide to confine the electromagnetic field. Silicon technology would introduce functionality derived from fiber communications to the silicon wafer scale. Silicon waveguides fabricated on a silicon wafer replace optical fibers as the transmission medium. Silicon wafers will contain monolithically created functional photonic devices, such as multiplexers/demultiplexers, modulators, and detectors, that can encode the information in light, route the signals for transmitting them to neighboring chips, and finally decode the signal. Silicon photonics technology is envisioned to eliminate communication bandwidth bottlenecks on the chip scale and significantly reduce power consumption in data centers, server farms, and supercomputers. However, photonic sources, such as lasers and light-emitting diodes, is one essential function that silicon does not naturally accommodate. The inability to create sources in silicon is due to its electronic band structure characteristics, which are treated in detail in Chapter 4. To overcome the photonic source limitation in silicon, researchers have explored hybrid integration of suitable semiconductor materials on silicon and nonlinear optical effects in tightly confined electromagnetic modes in silicon waveguides.

This chapter touched on the scope of developments in nanophotonics using new materials, fabrication, metrology, and devices. The field's scope is broad, with journals reporting new findings with applications to information technology, medicine,

communications, renewable energy technology, and computing. It embraces the efforts of international scientists and engineers across many traditional disciplines, such as biology; chemistry; physics; mathematics; and electrical, mechanical, and materials engineering. The problem is that the number of articles published in the field is rapidly growing; for the reader to gain an appreciation of the content of specific articles, important conceptual notions are required. It is not our intention to cover all aspects of nanophotonics in this book, which is truly a fool's errand; instead, the subjects provide students and researchers access to journal articles by presenting a solid foundation in essential topics.

Further reading

Books

[1]   P. Prasad, Nanophotonics, Wiley, NY, 2004.
[2]   P. Prasad, Biophotonics, Wiley, NY, 2004.
[3]   A. Lakhtakia, Handbook of Nanotechnology, SPIE Press, NY, 2004.
[4]   L. Novotny, B. Hecht, Principles of Nano-optics, second ed., Cambridge Univ., Cambridge, 2012.
[5]   S.V. Gaponenko, Introduction to Nanophotonics, Cambridge University Press, Cambridge, 2010.
[6]   M. Ohtsu, K. Kobayashi, T. Kawazoe, T. Yatsui, M. Naruse, Principles of Nanophotonics, Taylor and Francis Group, Boca Raton, 2008.

Review articles

[7]   M. Liu, Z. Ji, L. Shang, Top-down fabrication of nanostructures, Nanotechnology 8 (2010) 1−47.
[8]   Y. Shen, C.S. Friend, Y. Jiang, D. Jakubczyk, J. Swiatkiewicz, P.N. Prasad, Nanophotonics: interactions, materials, and applications, J. Phys. Chem. B 140 (2000) 7577−7877.
[9]   B.D. Gates, Q. Xu, M. Stewart, D. Ryan, C.G. Willson, G.M. Whitesides, New approaches to nanofabrication: molding, printing, and other techniques, Chem. Rev. 105 (2005) 1171−1196.
[10]   R.P. Feynman, There's plenty of room at the bottom (data storage), J. Microelectromech. Syst. 1 (1) (1992) 60−66.
[11]   C. Mack, The multiple lives of Moore's law, IEEE Spectrum 52 (2015) 31.

This page intentionally left blank

Electrodynamics for nanophotonics

2

J.W. Haus
University of Dayton, Dayton, OH, USA

> *And God said, "Let there be light," and there was light. God saw that the light was good, and he separated the light from the darkness.*
>
> **Bible: Genesis 1:3–4**

> *For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled.*
>
> **Richard Feynman "What Do You Care What Other People Think?"**

2.1   Introduction

Electrodynamics is a foundational theory that is well known to every student of physics and of electrical engineering. The subject begins with writing Maxwell's equations and the constitutive relations between the different fields, but even here there are a multitude of materials in which the theory can be applied, and there are surprising new phenomena that can be found within the confines of ordinary classical electrodynamics. Maxwell's equations are easy enough to write down and to establish their physical meaning; after all, Maxwell's celebrated treatise on electrodynamics was published just 150 years ago. Despite the lapse of time from its first publication, we continue to discover new phenomena that can be extracted from Maxwell's equations.

Maxwell's theory is the basis for advanced electromagnetic technology and has enjoyed phenomenal success in the twentieth century. In addition to encompassing all electrostatic and magnetostatic phenomena, it predicted new dynamical phenomena from radiofrequency (RF) to X rays with exquisite accuracy. The modern development in electrodynamics has been a steady migration of calculations from analytic techniques to more numerically intensive calculations. It is a clear sign of the times that as systems become more complex, researchers will have to heavily rely on computational results.

2.2   Maxwell's equations

Maxwell's equations are the foundation for how electromagnetic systems behave. They provide physical insights about the scattering properties and local field distributions of a system, and quantitative results can be experimentally validated. Simulations of

a system using Maxwell's equations guide us to determine the best design that will optimize the electromagnetic performance of our system. Indeed, with some caveats and exceptions, the electromagnetic calculations that we perform for macroscopic systems are the same calculations that we perform on the nanometer scale.

In this chapter, we will tailor the presentation to cover only those aspects of electrodynamics that will be important for our understanding of nanophotonics topics covered in this book. There are plenty of books on the subject of electrodynamics, and the interested reader is encouraged to study, in depth, all classical aspects of the theory. We mention a small subset of electrodynamics books among the references that provide a good foundation for advanced study of the subject. Now, let us begin by listing Maxwell's equations side by side in differential and integral form in Table 2.1. The physical interpretation of the equations is more apparent in the integral form. The reader who may be unfamiliar with the forms should refer to several excellent tomes of the subject.

Each equation is briefly discussed separately first to provide the reader with their essential physical meaning. The first line in Table 2.1 is Coulomb's law, and it expresses the connection between the displacement field and the free change density, ρ_f. In integral form, the displacement field is visualized by space-filling lines that are tangent to the displacement vector field at each point; the field lines start on positive charges (the sources) and end on negative charges (the sinks). The free charges are sources and sinks for the displacement field lines. In the differential form, $\vec{\nabla} \cdot \ldots$ is the divergence operator; when it operates on the displacement vector field, the charge density is recovered. It is physically visualized as an operation related to fluid-like flow through space. In the case of Coulomb's law, the flow through space is the flux of D-field lines spreading through space. In succinct terms, Coulomb's law expresses the fact that D-field lines begin or end on free charges. In the region of space empty of charges, one should not conclude from the expression that it is also a field-free region. The fields pass through the space with the same number of lines passing into and out of the surface. Equation (2.2) expresses that the lack of magnetic monopoles (no sources or sinks) can be physically visualized in this picture as B-field lines closing in on themselves.

Table 2.1 Maxwell's equations in differential and integral form

Description	Differential form	Integral form	
Coulomb's law	$\vec{\nabla} \cdot \vec{D} = \rho_f$	$\oint \vec{D} \cdot d\vec{S} = \iiint \rho_f \, dV$	(2.1)
No magnetic monopoles	$\vec{\nabla} \cdot \vec{B} = 0$	$\oint \vec{B} \cdot d\vec{S} = 0$	(2.2)
Faraday's law	$\vec{\nabla} \times \vec{E} = -\partial \vec{B}/\partial t$	$\oint \vec{E} \cdot d\vec{l} = -\dfrac{d}{dt} \iint \vec{B} \cdot d\vec{S}$	(2.3)
Ampere's law	$\vec{\nabla} \times \vec{H} = \vec{j}_f + \partial \vec{D}/\partial t$	$\oint \vec{H} \cdot d\vec{l} = \iint \vec{j}_f \cdot d\vec{S} + \dfrac{d}{dt} \iint \vec{D} \cdot d\vec{S}$	(2.4)

Faraday's law of induction in integral form shows the close relationship between a time-varying magnetic flux (magnetic flux is defined by $\Phi_B = \iint \vec{B} \cdot d\vec{S}$) through a closed loop and an induced potential around the loop called the electromotive force (emf: $\mathscr{E}_e = \oint \vec{E} \cdot d\vec{l}$). The curl operator $\vec{\nabla} \times \ldots$ is connected with rotational flow of a "fluid" in space. The emf is not a conserved quantity in general as is the case for the electrostatic potential; a charged particle under the influence of this electric field may gain or lose energy by going around the closed path. Likewise, Ampere's law expresses a deep connection between a quantity called the magnetomotive force (mmf: $\mathscr{E}_m = \oint \vec{H} \cdot d\vec{l}$) and the current flowing through a surface. In the equation, the current has two contributions: one from the free charge current ($I = \iint \vec{j}_f \cdot d\vec{S}$) flowing through the materials and one from the displacement current $\left(\frac{d}{dt} \iint \vec{D} \cdot d\vec{S} \right)$ that may exist in space outside of electrical conductors. The displacement current that Maxwell added to prior existing electromagnetic laws ensures that charges are conserved. The divergence of the differential form of Ampere's law reduces to

$$\frac{\partial \rho_f}{\partial t} = \vec{\nabla} \cdot \vec{j}_f. \tag{2.5}$$

As simple as Maxwell's equations are to write and explain, the applications are as vast as the human intellect can devise. In the past quarter century, nanophotonics has created several paradigm shifts that have changed our understanding of an established field and injected new vigor to our study of electromagnetic phenomena.

2.2.1 Boundary conditions

Applying the integral form of Maxwell's equations, the boundary conditions connecting the electromagnetic field between two regions can be extracted. The surface integrals are evaluated using an infinitesimal Gaussian pillbox that straddles the surface under study (Figure 2.1). As the pillbox volume is reduced to zero, the vector
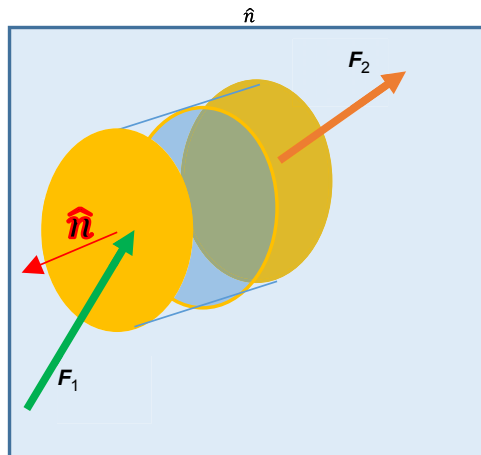
Figure 2.1 Gaussian pillbox construction for the divergence-related boundary conditions.
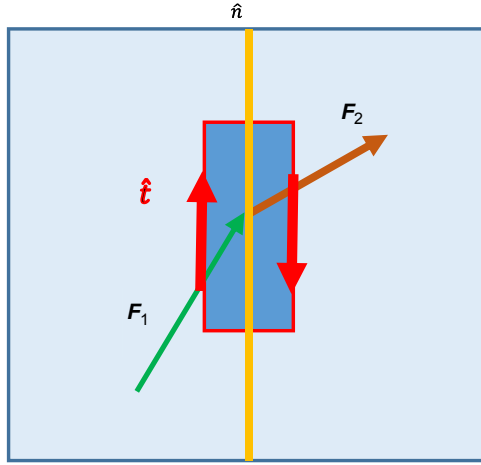
Figure 2.2 Circuit loop construct straddling the interface for the curl boundary conditions.

fields are nearly constant throughout the pillbox; thus, the cylindrical surface integrals containing the transverse field components vanish and the surface integrals over the end caps containing field components normal to the boundary satisfy the following equations:

$$\widehat{\mathbf{n}} \cdot \left(\vec{D}_2 - \vec{D}_1\right) \; = \; \sigma_{\mathrm{f}} \quad \text{and} \quad \widehat{\mathbf{n}} \cdot \vec{B}_1 \; = \; \widehat{\mathbf{n}} \cdot \vec{B}_2. \tag{2.6}$$

The surface charge density, σ_{f}, may be nonzero under ideal conditions of perfect conductivity, but it normally vanishes in the limit of small volume and the normal component of the displacement field is continuous across the interface.

The second set of boundary conditions connecting the two media is derived from the integral forms of Ampere's law and Faraday's law. The line integral is traced over a closed loop that straddles the interface. As shown in Figure 2.2, the fields on each side of the interface. As the size of the loop is reduced to an infinitesimal size, the field contributions tangential to the boundary are the leading terms and the line segments perpendicular to the boundary cancel one another to first order in a local field approximation. The resulting boundary conditions can be recast into a vector form as

$$\widehat{\mathbf{n}} \times \vec{E}_2 \; = \; \widehat{\mathbf{n}} \times \vec{E}_1 \quad \text{and} \quad \widehat{\mathbf{n}} \times \left(\vec{H}_2 - \vec{H}_1\right) \; = \; \vec{K}_{\mathrm{f}}. \tag{2.7}$$

Two independent tangential directions are implicit for the vector fields. The tangential E-field is continuous across the interface and the surface current passing through the loop is denoted by $\vec{K}_{\mathrm{f}}$. For most physical situations, $\vec{K}_{\mathrm{f}} = \vec{0}$ and the tangential components of the H-field are continuous across the interface.

2.2.2 Constitutive relations

Maxwell's equations in the forms given in Table 2.1 are incomplete. There are four vector fields ($\vec{E}$, $\vec{D}$, $\vec{B}$, $\vec{H}$) and only six independent equations. The fields are

connected to one another by constitutive relations describing the electromagnetic response of the media, which may have temporal and spatial dispersion. The relationship between the D- and E-fields in a temporally dispersive medium is defined by the sum of two contributions—the electric and (electric) polarization of the medium:

$$\vec{D}(r,t) = \varepsilon_0 \vec{E}(r,t) + \vec{P}(r,t). \tag{2.8}$$

In a linear, homogeneous medium, the polarization may be expressed as

$$\vec{P}(r,t) = \varepsilon_0 \iiint \int_{-\infty}^{\infty} \overline{\overline{\chi}}(\vec{r}-\vec{r}',t-t') \cdot \vec{E}(r',t') \mathrm{d}V' \mathrm{d}t', \tag{2.9}$$

where $\overline{\overline{\chi}}(\vec{r}-\vec{r}',t-t')$ is the susceptibility tensor, which can have spatial and temporal dependence. The spatial dependence expresses the effect of nonlocal contributions to the electronic response and its effects are generally present near surfaces and are of subnanometer scale. Using a four-dimensional Fourier decomposition of the fields,

$$\vec{F}(\vec{k},\omega) = \frac{1}{(2\pi)^4} \iiint \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\left(\vec{k}\cdot\vec{r}'-\omega t'\right)} \vec{F}(\vec{r}',t') \mathrm{d}V' \mathrm{d}t', \tag{2.10}$$

ω is the angular frequency and $\vec{k}$ is the spatial frequency of the field. The corresponding inverse Fourier transform is

$$\vec{F}(\vec{r},t) = \iiint \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\left(\vec{k}\cdot\vec{r}-\omega t\right)} \vec{F}(\vec{k},\omega) \mathrm{d}^3 k \mathrm{d}\omega. \tag{2.11a}$$

It is also useful to define fields that are frequency dependent while remaining in the spatial domain:

$$\vec{F}(\vec{r},t) = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t} \vec{F}(\vec{r},\omega) \mathrm{d}\omega. \tag{2.11b}$$

Fields can be decomposed into their angular frequency components and superposed to form the corresponding time-dependent field.

Equation (2.9) is transformed to the algebraic form using the field definitions in Eqn (2.11a) as

$$\vec{P}(\vec{k},\omega) = \varepsilon_0 \overline{\overline{\chi}}(\vec{k},\omega) \cdot \vec{E}(\vec{k},\omega). \tag{2.12}$$

Extending the discussion, the displacement field can be successively written as

$$
\begin{aligned}
\vec{D}(\vec{k}, \omega) &= \varepsilon_0 \vec{E}(\vec{k}, \omega) + \vec{P}(\vec{k}, \omega) \\
&= \varepsilon_0 \left(\overline{\overline{1}} + \overline{\overline{\chi}}(\vec{k}, \omega) \right) \cdot \vec{E}(\vec{k}, \omega) \\
&= \varepsilon_0 \overline{\overline{\varepsilon}}_r(\vec{k}, \omega) \cdot \vec{E}(\vec{k}, \omega).
\end{aligned}
\tag{2.13}
$$

The relative dielectric tensor is denoted as $\overline{\overline{\varepsilon}}_r(\vec{k}, \omega)$. In most applications, the spatial dependence of the dielectric function is negligible; however, the temporal dispersion always plays an important role in the dynamical theory. In a similar fashion, magnetic phenomena are also described via a linear relation between the H- and B-fields. In four-dimensional frequency space, it is succinctly expressed as

$$
\begin{aligned}
\vec{B}(\vec{k}, \omega) &= \mu_0 \left(\vec{H}(\vec{k}, \omega) + \vec{M}(\vec{k}, \omega) \right) \\
&= \mu_0 \left(\overline{\overline{1}} + \overline{\overline{\chi}}_m(\vec{k}, \omega) \right) \cdot \vec{H}(\vec{k}, \omega) \\
&= \mu_0 \overline{\overline{\mu}}_r(\vec{k}, \omega) \cdot \vec{H}(\vec{k}, \omega).
\end{aligned}
\tag{2.14}
$$

The magnetic susceptibility tensor is $\overline{\overline{\chi}}_m(\vec{k}, \omega)$ and the relative magnetic permeability tensor is $\overline{\overline{\mu}}_r(\vec{k}, \omega)$. Equations (2.14) and (2.15), eliminating the spatial dependence from the coefficients, are written as

$$
D(\vec{k}, \omega) = \overline{\overline{\varepsilon}}(\omega) \cdot E(\vec{k}, \omega),
\tag{2.15a}
$$

and

$$
B(\vec{k}, \omega) = \overline{\overline{\mu}}(\omega) \cdot H(\vec{k}, \omega).
\tag{2.15b}
$$

The dielectric permittivity and the magnetic permeability are tensors (i.e., 3×3 matrices) that are indicative of an anisotropic material with possibly complex coefficients; they are expressed as

$$
\overline{\overline{\varepsilon}} = \begin{pmatrix} \overline{\varepsilon}_{11} & \overline{\varepsilon}_{12} & \overline{\varepsilon}_{13} \\ \overline{\varepsilon}_{21} & \overline{\varepsilon}_{22} & \overline{\varepsilon}_{23} \\ \overline{\varepsilon}_{31} & \overline{\varepsilon}_{32} & \overline{\varepsilon}_{33} \end{pmatrix} \quad \text{and} \quad \overline{\overline{\mu}} = \begin{pmatrix} \overline{\mu}_{11} & \overline{\mu}_{12} & \overline{\mu}_{13} \\ \overline{\mu}_{21} & \overline{\mu}_{22} & \overline{\mu}_{23} \\ \overline{\mu}_{31} & \overline{\mu}_{32} & \overline{\mu}_{33} \end{pmatrix},
\tag{2.16}
$$

respectively. Each of the components is dependent on frequency, and the symmetry of the tensors is normally invoked such that $\overline{\mu}_{\alpha\beta} = \overline{\mu}_{\beta\alpha}$ and $\overline{\varepsilon}_{\alpha\beta} = \overline{\varepsilon}_{\beta\alpha}$. Anisotropic materials found in nature are generally crystalline materials. However, we will encounter materials that are artificially synthesized or fabricated heterogeneous materials made of two or more constituents; when the artificial materials are small enough compared with the wavelength of light, they are treated as homogeneous anisotropic materials.

The imaginary parts are indicative of losses (or under special circumstances gain) in the material. As such, there are constraints on the real and imaginary functions; they are related by the Kramers–Kronig relations, which can be found in many textbooks on electrodynamics and will not be further discussed here. The physical connection between the two parts is an expression of causality; namely, that a field cannot respond to a signal at a point in time that would be applied in the future.

More complexities can be added to the constitutive relations. For instance, chiral media are composed of molecular constituents that induce coupling between electric and magnetic phenomena. In homogeneous chiral materials, the constitutive relations in Eqn (2.15) are written as

$$\vec{D}(\vec{k}, \omega) = \bar{\bar{\varepsilon}}(\omega) \cdot \vec{E}(\vec{k}, \omega) + i\bar{\bar{\kappa}}(\omega) \cdot \vec{H}(\vec{k}, \omega), \tag{2.17a}$$

and

$$\vec{B}(\vec{k}, \omega) = \bar{\bar{\mu}}(\omega) \cdot \vec{H}(\vec{k}, \omega) - i\bar{\bar{\kappa}}(\omega) \cdot \vec{E}(\vec{k}, \omega). \tag{2.17b}$$

The additional coupling coefficient can lead to new physical phenomena that can be examined by probing specific polarization states of the fields. The chiral parameter is invoked to describe circular dichroism, which is a difference in the absorption coefficient between right and left circularly polarized waves and rotary dispersion, which is an angular rotation of a linear polarized wave's orientation.

2.3 Microscopic dynamical models

The microscopic connection between the macroscopic coefficients and the microscopic dynamics of atomic or molecular constituents of the medium is through the polarization of the medium. The microscopic quantity of interest is the dipole moment, defined as

$$\vec{p} = -e\vec{r}. \tag{2.18}$$

The variable $\vec{r}$ is the position of an electron of charge $-e$ and the nuclei are heavy enough and the frequency sufficient so that their motion can be neglected. When the dipoles are distributed throughout the volume with a dipole density of n per unit volume, the polarization of the medium is

$$\vec{P} = n\vec{p} = -e\vec{r}n. \tag{2.19}$$

The polarization's linear relationship to the displacement field in Eqn (2.13) can be applied to connect microscopic models to macroscopic fields. The treatment for homogenizing heterogeneous material and specific effective medium expressions will be treated in Chapter 7. In isotropic media, the tensors are diagonal and the diagonal elements have the same value.

One more piece is missing from our description of electrodynamics—namely, the interaction of electromagnetic waves with a single electron with a charge, $-e$, moving with a velocity, v, expressed as the Lorentz force:

$$\vec{F}(t) = -\mathrm{e}\big(\vec{E} + \vec{v} \times \vec{B}\big). \tag{2.20}$$

The first term is the Coulomb interaction between charges and the electric field and the second term is the magnetic force or Lorentz force.

There are two prototypical classical models that have been applied throughout the literature. One model describes a point electron that is bound to a nucleus as though by a spring. This is called the Lorentz model. For an electron for which the position is labeled by $\vec{r}$ and the velocity by $\vec{v}$, the familiar damped harmonic oscillator equation of motion is

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} + \gamma\vec{v} + \omega_0^2\vec{r} = -\frac{\mathrm{e}}{m}\vec{F}(t) = -\frac{\mathrm{e}}{m}(\vec{E} + \vec{v} \times \vec{B}), \tag{2.21}$$

where γ is a phenomenological damping constant, ω_0 is the resonance frequency, and m is the mass of the electron. The result for the polarization dynamics is derived from the kinematic equation of motion by using the definition of the polarization in Eqn (2.19):

$$\ddot{\vec{P}} + \gamma\dot{\vec{P}} + \omega_0^2\vec{P} = \vec{F}(t) = \omega_p^2\varepsilon_0\vec{E} - \frac{\mathrm{e}}{m}\dot{\vec{P}} \times \vec{B}. \tag{2.22}$$

where $\omega_p^2 = \frac{e^2}{m\varepsilon_0}n$ defines the plasma frequency, which notably depends on the density of electrons in the volume and $\dot{\vec{P}} = \frac{\mathrm{d}\vec{P}}{\mathrm{d}t}$, etc. The linear susceptibilities as defined in Eqn (2.12) are explicitly written by neglecting the last term and Fourier transforming Eqn (2.22):

$$\vec{P} = \varepsilon_0\chi_{\mathrm{b}}(\omega)\vec{E} = \frac{\omega_p^2}{\omega_0^2 - \omega^2 - \mathrm{i}\gamma\omega}\varepsilon_0\vec{E}. \tag{2.23}$$

The second case, called the Drude model, is similar to the first and a simple version of it can be found by setting the restoring force equal to zero (i.e., $\omega_0 = 0$). The linear susceptibility for the free charges is identified and extracted from the following relation:

$$\vec{P} = \varepsilon_0\chi_{\mathrm{f}}(\omega)\vec{E} = -\frac{\omega_p^2}{\omega^2 + \mathrm{i}\gamma\omega}\varepsilon_0\vec{E}. \tag{2.24}$$

The free charge susceptibility has a negative real part, leading to a negative dielectric function for metals and other materials possessing free electrons.

2.4 Wave equations

In general, the wave equations are vector equations that incorporate Maxwell's equations into a single expression for a single field using the constitutive relations. For simplicity, we use the frequency space representation of Maxwell's equations and restrict our analysis to the scalar form of the constitutive equations derived earlier. However, the medium may be inhomogeneous, which means that the susceptibilities may depend on the local position coordinate. For instance, taking the curl of Eqn (2.3) and replacing the magnetic field using the constitutive relation Eqn (2.15b), and using Eqn (2.4) to eliminate the curl of the H-field, we have

$$\vec{\nabla} \times \left(\frac{\vec{\nabla} \times \vec{E}(\vec{r}, \omega)}{\overline{\mu}(\vec{r}, \omega)} \right) - \omega^2 \, \overline{\varepsilon}(\vec{r}, \omega) \vec{E}(\vec{r}, \omega) \; = \; i\omega \vec{j}_f(\vec{r}, \omega). \tag{2.25}$$

Similarly, the vector wave equation for the H-field can be derived as

$$\vec{\nabla} \times \left(\frac{\vec{\nabla} \times \vec{H}(\vec{r}, \omega)}{\overline{\varepsilon}(\vec{r}, \omega)} \right) - \omega^2 \overline{\mu}(\vec{r}, \omega) \vec{H}(\vec{r}, \omega) \; = \; \vec{\nabla} \times \left(\frac{\vec{j}_f(\vec{r}, \omega)}{\overline{\varepsilon}(\vec{r}, \omega)} \right). \tag{2.26}$$

The terms on the right-hand side are current source terms generating the field.

2.4.1 Plane-wave solutions

In a homogeneous medium, the material coefficients are independent of the spatial coordinates. We will further assume that the free charge and the current density vanish. Under these conditions, the wave equations can be written as follows:

$$\vec{\nabla} \times \vec{\nabla} \times \vec{E}(\vec{r}, \omega) - \omega^2 \overline{\mu}(\omega) \overline{\varepsilon}(\omega) \vec{E}(\vec{r}, \omega) \; = \; 0. \tag{2.27}$$

Using the identity $\vec{\nabla} \times \vec{\nabla} \times \vec{E}(\vec{r}, \omega) = \vec{\nabla}(\vec{\nabla} \cdot \vec{E}(\vec{r}, \omega)) - \nabla^2 \vec{E}(\vec{r}, \omega)$ and applying Coulomb's law $\vec{\nabla} \cdot \vec{E}(\vec{r}, \omega) = (\vec{\nabla} \cdot \vec{D}(\vec{r}, \omega))/\overline{\varepsilon}(\omega) = 0$, we arrive at a vector version of the homogeneous Helmholtz equation,

$$\nabla^2 \vec{E}(\vec{r}, \omega) + \omega^2 \overline{\mu}(\omega) \overline{\varepsilon}(\omega) \vec{E}(\vec{r}, \omega) \; = \; 0. \tag{2.28}$$

The solutions can be expressed as an expansion in plane waves:

$$\vec{E}(\vec{r}, \omega) \; = \; \vec{E}_0(\vec{k}, \omega) e^{i\vec{k} \cdot \vec{r}}. \tag{2.29}$$

The wave vector $\vec{k}$ can be chosen at any solid angle on the surface of a sphere. The wave number $k = |\vec{k}|$ is bound by the dispersion relation extracted from Eqn (2.28) as

$$k^2 \; = \; \omega^2 \mu(\omega)\varepsilon(\omega) \; = \; k_0^2 \overline{\mu}_r(\omega) \overline{\varepsilon}_r(\omega). \tag{2.30}$$

$k_0 = \omega/c$ is the free-space wave number. The wave number can be complex, and plane-wave solutions have attenuated amplitudes as the waves propagate through the lossy medium. The (complex) refractive index, defined as $k = \eta k_0$, where $\eta = n + i\kappa$, is

$$\eta^2 = \bar{\mu}_{\mathrm{r}}(\omega)\bar{\varepsilon}_{\mathrm{r}}(\omega). \tag{2.31}$$

The electric field and magnetic field amplitudes satisfy the following equations:

$$\vec{k}\cdot\vec{E}_0(\vec{k},\omega) = 0, \vec{B}_0(\vec{k},\omega) = i\vec{k} \times \vec{E}_0(\vec{k},\omega). \tag{2.32}$$

These expressions form the result that the electric and magnetic fields are orthogonal to one another and to the wave vector. The three vectors $(\vec{E}, \vec{B}, \vec{k})$ form a right-handed set of basis vectors.

2.4.2 Conservation of energy

Measurements of properties of electromagnetic waves involve quadratic forms of the fields, at least in the optical regime. One quantity of particular significance is the instantaneous Poynting vector, defined as

$$\vec{S}(\vec{r},t) = \vec{E}(\vec{r},t) \times \vec{H}(\vec{r},t). \tag{2.33}$$

It is interpreted as an energy flux density with units of Watts per square meter. The origin of this interpretation of energy flux comes from Poynting's theorem of energy conservation. The theorem is derived by invoking the following vector identity:

$$\vec{\nabla}\cdot\vec{S}(\vec{r},t) = \vec{H}(\vec{r},t)\cdot\vec{\nabla} \times \vec{E}(\vec{r},t) - \vec{E}(\vec{r},t)\cdot\vec{\nabla} \times \vec{H}(\vec{r},t). \tag{2.34}$$

Using Maxwell's equations, the equation has the form

$$\vec{\nabla}\cdot\vec{S} = -\vec{H}\cdot\frac{\partial\vec{B}}{\partial t} - \vec{E}\cdot\frac{\partial\vec{D}}{\partial t} - \vec{j}_{\mathrm{f}}\cdot\vec{E}. \tag{2.35}$$

Defining the rate of change of electric energy density,

$$\frac{\partial w_{\mathrm{e}}}{\partial t} = \vec{E}\cdot\frac{\partial\vec{D}}{\partial t}, \tag{2.36}$$

and magnetic energy density,

$$\frac{\partial w_{\mathrm{m}}}{\partial t} = \vec{H}\cdot\frac{\partial\vec{B}}{\partial t}. \tag{2.37}$$

Poynting's theorem is recast in the form ($w = w_{\mathrm{e}} + w_{\mathrm{m}}$)

$$\vec{\nabla}\cdot\vec{S} = -\vec{j}_{\mathrm{f}}\cdot\vec{E} - \frac{\partial w}{\partial t}. \tag{2.38}$$

The volume integral form of Eqn (2.38) is ($W = \iiint\limits_{V} w\mathrm{d}V$):

$$\oint_{A(V)} \widehat{\mathbf{n}} \cdot \vec{S}\mathrm{d}A = -\iiint\limits_{V} \vec{j}_{\mathrm{f}} \cdot \vec{E}\mathrm{d}V - \frac{\mathrm{d}W}{\mathrm{d}t}. \tag{2.39}$$

The left-hand side describes the flow of electromagnetic energy through the surface $A(V)$ of the volume V with $\widehat{\mathbf{n}}$ being a unit vector normal to the surface. The Poynting vector contains incident and radiation field contributions that contribute to energy flow in both directions through the surface. The first term on the right-hand side includes the dissipation of energy in the volume due to Joule heating. The second term is the rate of change of energy in the volume.

Photodetection measurements are slow compared with the oscillation period of a wave. Therefore, the measured quantity is a time average. For a continuous wave, the measured power flux or irradiance impinging on a surface can be defined by

$$I = \widehat{n} \cdot \langle \vec{S} \rangle, \tag{2.40}$$

where $\langle \vec{S} \rangle$ is the time-averaged Poynting vector defined as

$$\langle \vec{S} \rangle = \frac{1}{T} \int\limits_{0}^{T} \vec{S}\mathrm{d}t. \tag{2.41}$$

T is a time that is short compared with any changes of the instantaneous Poynting vector amplitude. For simplicity, the time may be considered as one optical cycle, but it can be an average over many optical cycles as long as the field amplitudes do not perceptibly change. For time harmonic fields,

$$\vec{E}(\vec{r},t) = \mathrm{Real}\{\vec{E}(\vec{r},\omega)\mathrm{e}^{-\mathrm{i}\omega t}\}, \quad \vec{H}(\vec{r},t) = \mathrm{Real}\{\vec{H}(\vec{r},\omega)\mathrm{e}^{-\mathrm{i}\omega t}\}, \tag{2.42}$$

and the time average in Eqn (2.41) is

$$\langle \vec{S} \rangle = \frac{1}{2}\mathrm{Real}\{\vec{E}(\vec{r},\omega) \times \vec{H}^{*}(\vec{r},\omega)\}. \tag{2.43}$$

The asterisk denotes the complex conjugate of the function. For a plane-wave field, Eqn (2.27), the Poynting vector is

$$\langle \vec{S} \rangle = \frac{1}{2}\mathrm{Real}\left\{\frac{\vec{k}}{\omega\overline{\mu}}\left|\left|E_0(\vec{k},\omega)\right|\right|^2\right\}. \tag{2.44}$$

The Poynting vector is related to the energy density of the medium. This expression for $\langle \vec{S} \rangle$ is valid for homogeneous, isotropic media with complex material parameters.

2.4.3 Dissipation and energy density in a dispersive medium

The energy densities defined in Eqns (2.36) and (2.37) can be simplified in dispersive media by an approximation. The constitutive relations given by Eqn (2.15) can be approximated to yield manageable expressions. Consider the fields decomposed into complex components called positive and negative frequencies. The integrands on the right-hand side of Eqn (2.39) are written

$$
\mathbb{Q} + \frac{\partial w}{\partial t} = \frac{1}{2}\left(\vec{j}_{\mathrm{f}}^{+}(\vec{r},t) + \vec{j}_{\mathrm{f}}^{-}(\vec{r},t)\right)\cdot\frac{1}{2}\left(\vec{E}^{+}(\vec{r},t) + \vec{E}^{-}(\vec{r},t)\right)
$$

$$
+ \frac{1}{2}\left(\vec{E}^{+}(\vec{r},t) + \vec{E}^{-}(\vec{r},t)\right)\cdot\frac{1}{2}\left(\frac{\partial \vec{D}^{+}(\vec{r},t)}{\partial t} + \frac{\partial \vec{D}^{-}(\vec{r},t)}{\partial t}\right) \tag{2.45}
$$

$$
+ \frac{1}{2}\left(\vec{H}^{+}(\vec{r},t) + \vec{H}^{-}(\vec{r},t)\right)\cdot\frac{1}{2}\left(\frac{\partial \vec{B}^{+}(\vec{r},t)}{\partial t} + \frac{\partial \vec{B}^{-}(\vec{r},t)}{\partial t}\right).
$$

The angular frequency spectrum in Eqn (2.11b) is applied to the field components as

$$
F^{+}(\vec{r},t) = \int_{0}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t} F(\vec{r},\omega)\mathrm{d}\omega. \tag{2.46}
$$

The function $\vec{F}^{-}(\vec{r},t)$ is the complex conjugate of $\vec{F}^{+}(\vec{r},t)$. The frequency-dependent dielectric function is decomposed in a Taylor series as

$$
\bar{\varepsilon}(\omega) = \sum_{n=0}^{\infty} \bar{\varepsilon}_{n}\omega^{n}. \tag{2.47}
$$

Applying the series form of the dielectric function, the positive frequency displacement field is progressively rewritten as

$$
\vec{D}^{+}(\vec{r},t) = \int_{0}^{\infty} \bar{\varepsilon}(\omega)\mathrm{e}^{-\mathrm{i}\omega t}\vec{E}^{+}(\vec{r},\omega)\mathrm{d}\omega
$$

$$
\tag{2.48}
$$

$$
= \int_{0}^{\infty} \bar{\varepsilon}\left(\mathrm{i}\frac{\partial}{\partial t}\right)\mathrm{e}^{-\mathrm{i}\omega t}\vec{E}^{+}(\vec{r},\omega)\mathrm{d}\omega = \bar{\varepsilon}\left(\mathrm{i}\frac{\partial}{\partial t}\right)\vec{E}^{+}(\vec{r},t).
$$

The final equality is useful for deriving a useful expression for the energy density. Consider a field decomposed into a carrier frequency term and a slowly varying amplitude:

$$
\vec{E}^{+}(\vec{r},t) = \vec{E}_{0}(\vec{r},t)\mathrm{e}^{-\mathrm{i}\omega_{0}t} \tag{2.49}
$$

The complex field amplitude, decomposed in Fourier components, has a spectrum for which the width, $\Delta\omega$, is much smaller than the carrier frequency (i.e., $\Delta\omega \ll \omega_0$). Returning to the time derivative of the displacement field in Eqn (2.42), one term of the series is

$$\left\{\frac{\partial \vec{D}^+(\vec{r},t)}{\partial t}\right\}_n = -i\,\bar{\varepsilon}_n \left(i\frac{\partial}{\partial t}\right)^{n+1} \vec{E}_0(\vec{r},t)e^{-i\omega_0 t}$$

$$= \left[-i\,\bar{\varepsilon}_n(\omega_0)^{n+1}\vec{E}_0(\vec{r},t) + (n+1)\bar{\varepsilon}_n(\omega_0)^n \frac{\partial \vec{E}_0(\vec{r},t)}{\partial t}\right]e^{-i\omega_0 t}.$$

$$(2.50)$$

Although series can be continued to higher order, this approximation is sufficient to provide useful results. Gathering all of the terms together, the sum of the terms in square brackets is

$$\left[-i\omega_0\bar{\varepsilon}(\omega_0)\vec{E}_0(\vec{r},t) + \frac{d(\omega_0\bar{\varepsilon}(\omega_0))}{d\omega_0}\frac{\partial \vec{E}_0(\vec{r},t)}{\partial t}\right].$$

$$(2.51)$$

The time derivative of the displacement field is

$$\frac{\partial \vec{D}^+(\vec{r},t)}{\partial t} = \frac{\partial}{\partial t}\bar{\varepsilon}\left(i\frac{\partial}{\partial t}\right)\vec{E}_0(\vec{r},t)e^{-i\omega_0 t}$$

$$= \left(-i\omega_0\bar{\varepsilon}(\omega_0)\vec{E}_0(\vec{r},t) + \frac{d(\omega_0\bar{\varepsilon}(\omega_0))}{d\omega_0}\frac{\partial \vec{E}_0(\vec{r},t)}{\partial t}\right)e^{-i\omega_0 t}.$$

$$(2.52)$$

An analogous procedure is invoked for the magnetic terms. The time average of the energy density flux in Eqn (2.42) is separated into two contributions:

$$\mathbb{Q} = \frac{1}{2}\text{Re}\{\vec{j}_{f0}(\vec{r},t)\cdot\vec{E}_0^*(\vec{r},t)\} + \frac{1}{2}\left(\omega_0\bar{\varepsilon}''(\omega_0)|\vec{E}_0(\vec{r},t)|^2\right.$$

$$\left. + \omega_0\bar{\mu}''(\omega_0)|\vec{H}_0(\vec{r},t)|^2\right),$$

$$(2.53)$$

$$\frac{\partial w}{\partial t} = \frac{1}{2}\text{Im}\{\vec{j}_{f0}(\vec{r},t)\cdot\vec{E}_0^*(\vec{r},t)\} + \frac{1}{2}\left(\frac{d(\omega_0\bar{\varepsilon}'(\omega_0))}{d\omega_0}\text{Re}\left\{\vec{E}_0(\vec{r},t)\cdot\frac{\partial \vec{E}_0^*(\vec{r},t)}{\partial t}\right\}\right.$$

$$\left. + \frac{d(\omega_0\bar{\mu}'(\omega_0))}{d\omega_0}\text{Re}\left\{\vec{H}_0(\vec{r},t)\cdot\frac{\partial \vec{H}_0^*(\vec{r},t)}{\partial t}\right\}\right).$$

$$(2.54)$$

The function $\mathbb{Q}$ describes contributions to power dissipation per unit volume. When the current density is proportional to the local field, the constitutive relation is related to Ohm's law:

$$\vec{j}_{f0}(\vec{r},t) = \sigma(\omega)\vec{E}_0(\vec{r},t).$$

$$(2.55)$$

The function $\sigma(\omega) = \sigma'(\omega) + i\sigma''(\omega)$ is the conductivity; as expressed, it is a complex function. The power dissipation simplifies

$$\mathbb{Q} = \frac{1}{2}\sigma'(\omega)\left|\vec{E}_0(\vec{r},t)\right|^2 + \frac{1}{2}\left(\omega_0\overline{\varepsilon}''(\omega_0)\left|\vec{E}_0(\vec{r},t)\right|^2 + \omega_0\overline{\mu}''(\omega_0)\left|\vec{H}_0(\vec{r},t)\right|^2\right).$$
(2.56)

The electric and magnetic terms are related to the losses in the dielectric and magnetic materials within the volume. The absorption of energy leads to a subsequent local heating of the materials.

Equation (2.54) is a reactive term leading to storage of electromagnetic energy in the volume. This can be written as

$$\frac{\partial w}{\partial t} = \frac{1}{2}\sigma''(\omega)\left|\vec{E}_0(\vec{r},t)\right|^2 + \frac{1}{2}\left(\frac{\mathrm{d}(\omega_0\overline{\varepsilon}'(\omega_0))}{\mathrm{d}\omega_0}\mathrm{Re}\left\{\vec{E}_0(\vec{r},t)\cdot\frac{\partial\vec{E}_0^*(\vec{r},t)}{\partial t}\right\}\right)$$
$$+ \frac{1}{2}\left(\frac{\mathrm{d}(\omega_0\overline{\mu}'(\omega_0))}{\mathrm{d}\omega_0}\mathrm{Re}\left\{\vec{H}_0(\vec{r},t)\cdot\frac{\partial\vec{H}_0^*(\vec{r},t)}{\partial t}\right\}\right).$$
(2.57)

The first term is a reactive contribution to the kinetic energy of the free electrons, which is related to the kinetic inductance contribution from the circuit analysis. More details on this contribution are discussed below. Leaving out the first term, the last two terms in Eqn (2.54) can be integrated in time, yielding the energy density stored in the volume,

$$w = \frac{1}{4}\left(\frac{\mathrm{d}(\omega_0\overline{\varepsilon}'(\omega_0))}{\mathrm{d}\omega_0}\left|\vec{E}_0(\vec{r},t)\right|^2 + \frac{\mathrm{d}(\omega_0\overline{\mu}'(\omega_0))}{\mathrm{d}\omega_0}\left|\vec{H}_0(\vec{r},t)\right|^2\right).$$
(2.58)

To be of use, the energy density should be a positive definite function, which implies that the coefficients are positive; that is, $\frac{\mathrm{d}(\omega_0\overline{\varepsilon}'(\omega_0))}{\mathrm{d}\omega_0}, \frac{\mathrm{d}(\omega_0\overline{\mu}'(\omega_0))}{\mathrm{d}\omega_0} > 0$. Under these conditions, the group velocity is also positive. Equation (2.58) reduces to the traditional electromagnetic energy density result for nondispersive media (i.e., $\overline{\varepsilon}'$ and $\overline{\mu}'$ are constants). The coefficients can be positive although the dielectric permittivity and magnetic permeability are negative; therefore, Eqn (2.58) applies to a wide range of materials, including plasmonic and metamaterials, which are discussed in later chapters.

2.4.4 *Fresnel equations*

At an interface between two materials, an electromagnetic field experiences reflection and refraction. The connections between the incident field amplitude and the reflected and transmitted field amplitude are called the Fresnel equations. Consider a plane at $z = 0$ separating two isotropic materials. The plane of incidence is defined by the incident, reflected, and transmitted wave vectors. For definiteness, here the plane of incidence will be the (x, z) plane.

The incident, reflected, and transmitted wave vectors are

$$\vec{k}_i = (k_{ix}, 0, k_{iz}), \quad \vec{k}_r = (k_{rx}, 0, k_{rz}), \quad \text{and} \quad \vec{k}_t = (k_{tx}, 0, k_{tz}). \tag{2.59}$$

The three wave vectors are illustrated in Figure 2.3. The angles are defined with respect to the dotted line in the figure. Referring to Eqn (2.27), the spatial variation of the plane-wave fields at a uniform interface ($z = 0$) satisfy

$$k_{ix} = k_{rx} = k_{tx}. \tag{2.60}$$

This is an expression of Snell's law, which is normally expressed in terms of the angle of incidence, reflection, and refraction. The sign of the z-component of the reflected wave vector is $k_{rz} = -k_{iz}$, and the angle of reflection is the same as the angle of incidence. A geometric interpretation of Eqn (2.60) is illustrated in Figure 2.3. The horizontal dashed line in Figure 2.3 demonstrates the equality of the component of the transmitted and reflected wave vectors expressed in Eqn (2.59). The lengths of the wave vectors satisfy the dispersion relation, Eqn (2.30), and we consider real material coefficients in this situation; in Figure 2.3, the refractive index of medium 2 is larger than that in medium 1 because the wave vector has a longer length and the angle it makes with the normal line is smaller. The reader should be aware that in media with complex refractive indices, the angles are treated as complex numbers; perhaps a less confusing approach is developing expressions using only the wave vector components, which are naturally complex valued. We derive expressions using this philosophy.

In isotropic materials, light is decomposed into two orthogonal polarizations. One is called the s-polarization (s is shorthand for *senkrecht*, meaning "perpendicular" in German) and is defined by the electric field perpendicular to the plane of incidence. The other is called p-polarization, and the electric field is parallel to the plane of incidence.
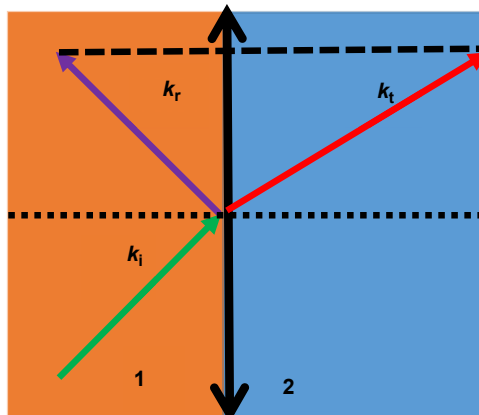
Figure 2.3 Planar interface between two media with wave vectors. Medium 1 is on the left and medium 2 is on the right.

The s-polarization electric and H-fields for each plane wave are written as

$$\vec{E}_\alpha = (0, E_{\alpha0}, 0), \ \ \vec{H}_\alpha = \mathrm{i} \frac{E_{\alpha0}}{\mu_\alpha}(-k_{\alpha z}, 0, k_{\alpha x}), \tag{2.61}$$

for $\alpha = i, r, t$, denoting the incident, reflected, and transmitted wave fields, respectively. The field amplitudes are connected through the boundary conditions. Continuity of the tangent electric and H-field provide the s-polarization Fresnel equations for the reflection and transmission amplitudes defined as $E_{r0} = r_s E_{i0}$ and $E_{t0} = t_s E_{i0}$:

$$r_{\mathrm{s}} = \frac{\left(\frac{k_{zi}}{\mu_i} - \frac{k_{zt}}{\mu_t}\right)}{\left(\frac{k_{zi}}{\mu_i} + \frac{k_{zt}}{\mu_t}\right)} \ \ \text{and} \ \ t_{\mathrm{s}} = 2 \frac{\frac{k_{zi}}{\mu_i}}{\left(\frac{k_{zi}}{\mu_i} + \frac{k_{zt}}{\mu_t}\right)}. \tag{2.62}$$

The p-polarization H- and electric field amplitudes are now

$$\vec{H}_\alpha = (0, H_{\alpha0}, 0), \ \ \vec{E}_\alpha = \mathrm{i} \frac{1}{\varepsilon_\alpha} H_{\alpha0}(-k_{\alpha z}, 0, k_{\alpha x}) \ \ \text{for} \ \ \alpha = i, r, t. \tag{2.63}$$

The p-polarization Fresnel equations, in which the reflection and transmission amplitudes are defined by $H_{r0} = r_p H_{i0}$ and $H_{t0} = t_p H_{i0}$, are

$$r_{\mathrm{p}} = \frac{(k_{zi}\varepsilon_t - k_{zt}\varepsilon_i)}{(k_{zi}\varepsilon_t + k_{zt}\varepsilon_i)} \ \ \text{and} \ \ t_{\mathrm{p}} = \frac{2k_{zi}\varepsilon_t}{(k_{zi}\varepsilon_t + k_{zt}\varepsilon_i)}. \tag{2.64}$$

The reflection and transmission coefficients are defined by

$$R_\alpha = \frac{\hat{z} \cdot \langle \vec{S}_r \rangle}{\hat{z} \cdot \langle \vec{S}_i \rangle} = |r_\alpha|^2 \ \ \text{and} \ \ T_\alpha = \frac{\hat{z} \cdot \langle \vec{S}_t \rangle}{\hat{z} \cdot \langle \vec{S}_i \rangle} = K_\alpha \frac{\mathrm{Real}\left\{\frac{k_{zt}}{\mu_t}\right\}|t_\alpha|^2}{\mathrm{Real}\left\{\frac{k_{zi}}{\mu_i}\right\}} = 1 - R_\alpha. \tag{2.65}$$

where, $\alpha = \mathrm{s}$ or p. $K_s = 1$ and $K_p = |k_i/k_t|^2$

The generalized form of the equations has some interesting consequences. The impedance of a medium is defined by

$$Z = \sqrt{\frac{\mu}{\varepsilon}}. \tag{2.66}$$

When the impedance values of the two media are identical, the reflection coefficient vanishes at a normal angle of incidence.

The graphs in Figure 2.4 cover a few cases of interest for real valued optical coefficients. For these cases, real angles of incidence and transmission are defined and applied. Figure 2.4(a) and (b) exemplify situations for commonly found dielectric

(a) s-polarization: $\varepsilon_1 = 1$, $\mu_1 = 1$ $\varepsilon_2 = 3$, $\mu_2 = 1$ (b) p-polarization: $\varepsilon_1 = 1$, $\mu_1 = 1$ $\varepsilon_2 = 3$, $\mu_2 = 1$

s-polarization: $\varepsilon_1 = 1$, $\mu_1 = 1$ $\varepsilon_2 = 3$, $\mu_2 = 3$ p-polarization: $\varepsilon_1 = 1$, $\mu_1 = 1$ $\varepsilon_2 = 1$, $\mu_2 = 3$

s-polarization: $\varepsilon_1 = 1$, $\mu_1 = 1$ $\varepsilon_2 = 2$, $\mu_2 = 2$ s-polarization: $\varepsilon_1 = 2$, $\mu_1 = 2$ $\varepsilon_2 = 1$, $\mu_2 = 1$
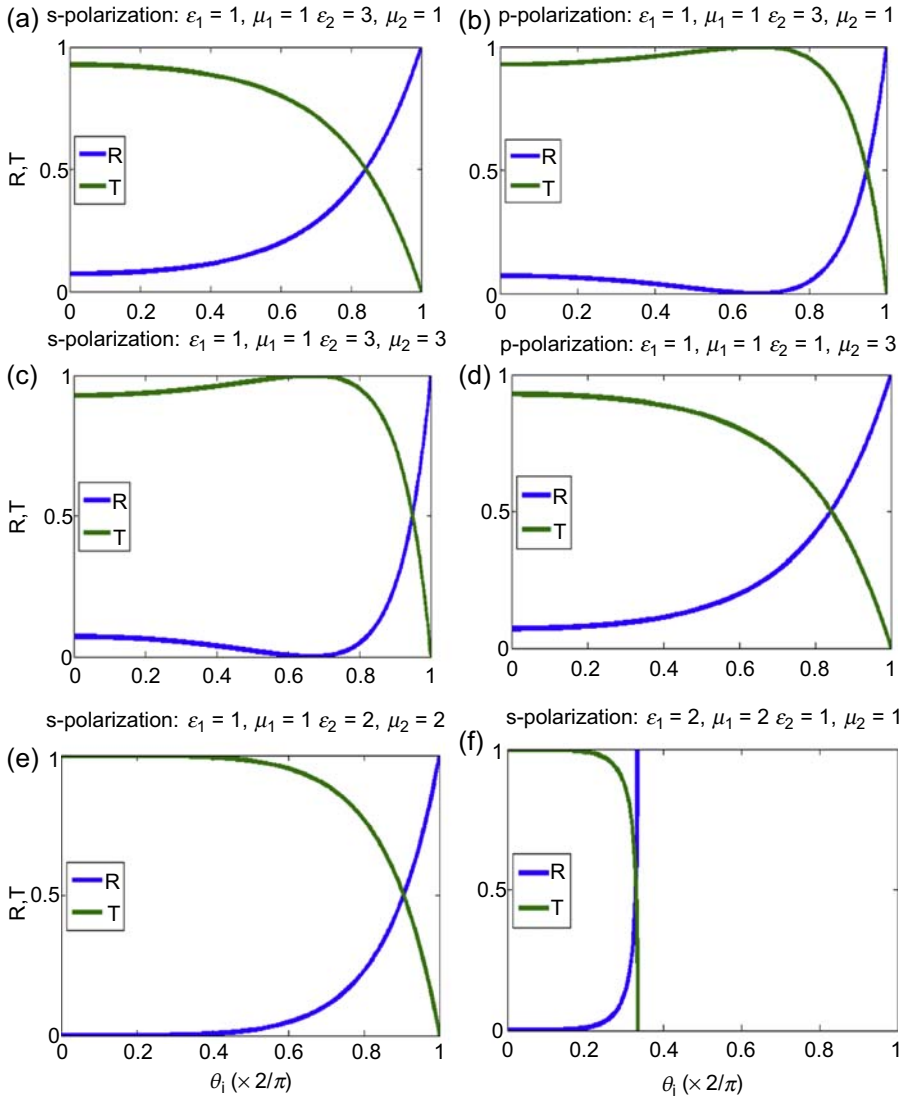
Figure 2.4 Reflection and transmission coefficients that exemplify the characteristics of different media. (a, b) Polarizations for the air: dielectric case with $\varepsilon_{r2} = 3$. (c, d) Polarizations for the air: magnetic case with $\mu_{r2} = 3$. (e) Air: impedance-matched case with $\mu_{r2} = \varepsilon_{r2} = 3$. (f) Impedance matched: air case with $\varepsilon_{r2} = \mu_{r2} = 2$.

materials; the relative magnetic permeability is unity and the dielectric permittivity defines the optical properties. For light incident from the air side, the s- and p-polarization R and T are the same at zero and $\pi/2$ angles of incidence. However, whereas the s-polarization R, T monotonically increases with the angle of incidence, the p-polarization R, T is nonmonotonic and has a zero at the

so-called Brewster angle. The Brewster angle is defined by the zero of the numerator in Eqn (2.64). The Brewster angle for transparent dielectric materials is

$$\tan \theta_i = \frac{n_i}{n_t}, \tag{2.67}$$

where

$$n_i = \sqrt{\varepsilon_{ri}\mu_{ri}} \text{ and } n_t = \sqrt{\varepsilon_{rt}\mu_{rt}}. \tag{2.68}$$

define the indices of refraction of the incident and transmitted media, respectively.

Figure 2.4(c) and (d) explores the Fresnel equations for a pure magnetic material with the dielectric permittivity remaining unity in both materials. This is a situation that does not exist in nature. However, it is interesting to note that the R, T characteristics of the light waves have switched between the two polarization states compared with the previous example.

Figure 2.4(e) and (f) explores two cases in which the impedance is matched between the two media. In the first case, the second medium has the higher index of refraction and in the second case the index of refraction of the first medium is higher. The angle for total internal reflection (TIR) is deduced from Snell's law with the transmission angle $\theta_t = \pi/2$

$$\sin \theta_{\text{TIR}} = \frac{n_t}{n_i}. \tag{2.69}$$

An interesting physical situation arises for $\theta_i > \theta_{\text{TIR}}$ in that the field does not propagate into the medium and $k_{zt} = ik_0\sqrt{n_i^2 \sin^2\theta_i - n_t^2}$ is imaginary; there is an evanescent field present in the transmitted medium and energy is stored in it. A plane-wave evanescent field at the boundary decays as the distance from the boundary; that is,

$$e^{-k_{zt}z}. \tag{2.70}$$

At the TIR angle, the penetration depth of the evanescent wave is infinite; however, for larger angles of incidence, the penetration depth is on the order of the optical wavelength. An interesting and useful consequence of TIR is that the Fresnel equations, Eqns (2.62) and (2.64), are complex, leading to a phase change on reflection of the wave. Because the phase changes are different for each polarization, this implies a change of polarization by reflection at the boundary. A linearly polarized wave shared between the s- and p-polarization states becomes elliptically polarized upon reflection from a boundary for an angle of incidence greater than the TIR angle. For example, this phenomenon has been exploited to transform linear polarized light to circular polarization at the output using a two reflections in a glass material; the device is called a Fresnel rhomb.

2.4.5 Three velocities

The velocity of an electromagnetic wave needs to be carefully examined to avoid pitfalls that lead to confusing and even nonphysical interpretations of experimental results. A cornerstone principle of relativistic physics is that the speed of light in vacuum is a limiting velocity for energy transport. A large body of experimental observations has never violated this principle; indeed, any evidence reporting superluminal wave propagation has not held up under careful scrutiny. There are several velocities that come to mind, including phase, group, energy, and signal velocity. In this section, we briefly define the first three concepts. An optical pulse carries carrier frequency oscillations and an envelope function of the field. An illustration of a scalar plane-wave field is shown in Figure 2.5. The horizontal axis could be either pulse position, representing the field at a moment in time, or temporal, representing the pulse passing a specific position. The sinusoidal oscillations represent the carrier phase of the wave and the smooth envelope function extends over many oscillations of the oscillations.

The mathematical form of an electromagnetic wave is exemplified in the decomposition into plane waves described by Eqn (2.10). To determine the phase velocity, consider the phase in the exponential factor

$$\varphi = \vec{k} \cdot \vec{r} - \omega t \tag{2.71}$$

as constant for the position vector in the direction perpendicular to the phase front (i.e., r is parallel to k). The phase velocity of the propagating phase front is

$$v_{\mathrm{p}} = \frac{\vec{r}}{t} = \widehat{\mathbf{k}} \omega / k, \tag{2.72}$$

where the unit vector $\widehat{\mathbf{k}} = \vec{k} / k$ lies in the direction of the wave vector. In an isotropic material, the phase velocity is perpendicular to the phase fronts.
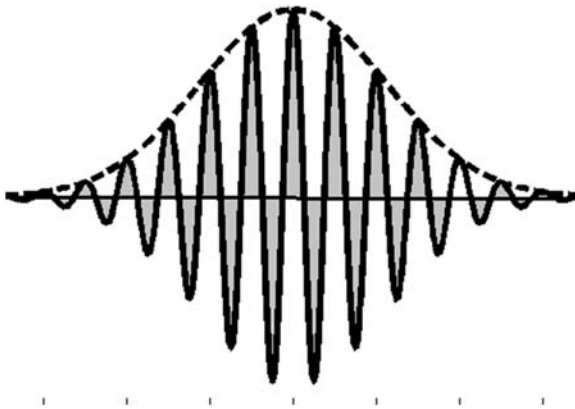
Figure 2.5 Illustration of one field component of a pulse at one instant in time or as it passes one point in space. The spatial period of the field oscillations is the central wavelength λ and the field envelope function highlights the slower variation of the field amplitude.

In dispersive media, the wave number is related to the angular frequency by the dispersion relation Eqn (2.30). Here, we restrict the presentation to isotropic media. A pulse is a superposition of plane waves; we adopt a description of pulses with a carrier angular frequency ω_0 and carrier wave vector $\vec{k}_0$. There is a spread of frequencies ($\Delta\omega$) around the carrier frequency, and we will assume that the pulse has a narrow bandwidth (i.e., $\Delta\omega \ll \omega_0$). Using the dispersion relation $\omega(\vec{k})$, a Taylor series expansion yields

$$\omega(\vec{k}) = \omega(\vec{k}_0) + (\vec{k} - \vec{k}_0) \cdot \vec{\nabla}_{\vec{k}_0} \omega(\vec{k}_0) + \cdots. \tag{2.73}$$

The coefficient of the second term defines the group velocity,

$$\vec{v}_{\mathrm{g}} = \vec{\nabla}_{\vec{k}_0} \omega(\vec{k}_0). \tag{2.74}$$

The group velocity is perpendicular to phase fronts. In an isotropic medium, the phase and group velocities are parallel; this observation does not generalize to anisotropic media in which the velocities point in different directions. In dispersive media, the group and phase velocities are generally different from one another. The group velocity can be negative and even have values that are superluminal. However, a group velocity larger than c does not violate relativity because under these circumstances, it does not propagate information. Keep in mind that it is one term of a Taylor series expansion and higher order terms should be examined when violations of accepted physical principles are discovered.

Energy velocity is defined by the ratio of the Poynting vector and the energy density,

$$\vec{v}_{\mathrm{E}} = \frac{\langle \vec{S} \rangle}{w}, \tag{2.75}$$

where $\langle \vec{S} \rangle$ is defined by Eqn (2.44) and the local energy density w is defined in Eqn (2.58). Under certain circumstances, the energy and group velocity are identical. Again, it is well to keep in mind that the form of the energy density is approximate (also the first contribution of a Taylor series expansion) and physical violations need to be carefully examined.

The mention of signal velocity has been extensively discussed in the literature; especially relevant are the early discussions of Sommerfeld and Brillouin on pulse precursors for which the velocity is bound by the speed of light in vacuum. Over the last century, there have been many studies of pulse propagation in dispersive media and the conundrum of superluminal velocities.

2.5 Quasistatic limits

The electromagnetic interactions in composite materials under the right circumstances can be treated by neglecting propagation effects. This is the lumped circuit model, which can provide useful physical insights in complex systems. These approaches

have a limited usefulness, but they provide physical insight that helps to better understand the physical functionality of the materials. Even extended systems are well described by distributed elements: capacitors, inductors, and resistors.

In a series of papers, Engheta and others promoted the notion that the quasistatic limit could be used to engineer metamaterials. The quasistatic limit is achieved when the size of the lumped circuits is much smaller than the wavelength of the electromagnetic radiation; thus, for metamaterials, this means that for optical or infrared wavelengths the elements must be fabricated on the nanoscale. However, it is not just a matter of scaling the sizes of geometric structures to the nanoscale. Fabrication and materials science technologies have enabled nanometer dimension control. There are additional complications because material parameters, such as the dielectric constant, are not qualitatively the same in the visible and infrared regimes as their values in the RF regime. Furthermore, although lumped circuits are connected by wires and currents are confined to the wires, fabricated nanoscale structures have extended fields that can affect elements that are distant neighbors. These facts mean that use of circuit methods at the nanoscale has limited application. Nevertheless, these techniques are very useful for deriving simple results that can qualitatively and even quantitatively describe the essential underlying physics.

A recent paper formulated the circuit model using a source-free decomposition of the fields. This approach is more physically appealing because it eliminates issues that can plague previous formulations. Before presenting that formulation, we recall salient features of a simple series LRC circuit.

2.5.1   Series LRC circuit

The basic building blocks of circuit theory are resistors, capacitors, and inductors. The rules governing the behavior of circuits made with these lumped parameters are a simplification extracted from Maxwell's equations. There are two simple rules applied to circuits called Kirchhoff's circuit laws. The two laws are illustrated in Figure 2.6.

1. At a node where several current branches combine, the total current flowing into and out of the node is equal. This is a statement that charge is conserved.

$$\sum_{\alpha=1}^{n} I_\alpha = 0. \tag{2.76}$$
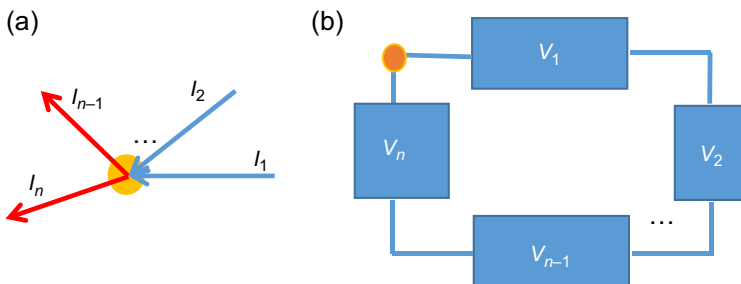
Figure 2.6 Illustrations of Kirchhoff's laws: (a) nodal current law and (b) the voltage loop law.

2. The voltage potential difference around any closed circuit is zero. This is a consequence of conservation of energy. In other words, as a charge moving around the circuit follows the potential changes, the charge gains or loses potential energy in equal amounts.

$$\sum_{\alpha=1}^{n} V_{\alpha} = 0. \tag{2.77}$$

To illustrate Kirchhoff's laws, we adopt a simple-series LRC circuit shown in Figure 2.7. We decompose the voltage into Fourier components and express voltages, current, and charge in complex form as

$$V_{\alpha}(t) = V_{\alpha\omega}e^{-i\omega t}; \quad I(t) = I_{\omega}e^{-i\omega t}, \quad Q(t) = \frac{I_{\omega}}{-i\omega}e^{-i\omega t}. \tag{2.78}$$

α labels the circuit elements $\alpha = L, R, C,$ or a. The nodes between two elements have only two branches; therefore, by Kirchhoff's nodal law, the current through each circuit element is the same.

The voltages across each element are

$$V_L(t) = L\dot{I}(t) = -i\omega L I(t), \quad V_R(t) = RI(t), \quad \text{and} \quad V_C(t) = \frac{I(t)}{-i\omega C}, \tag{2.79}$$

where the dot above the current denotes a time derivative. Application of Kirchhoff's circuit law yields the equation

$$V_a(t) = V_L(t) + V_R(t) + V_C(t). \tag{2.80}$$

The steady-state solution for the current is

$$\dot{I}_{\omega} = \frac{-i\omega V_{a\omega}}{(-L\omega^2 - i\omega R + 1/C)}. \tag{2.81}$$

This result is isomorphic with a mechanical damped harmonic oscillator and can be explored to find a rich set of consequences. We do not explore them here, but only consider the limit of small resistance (i.e., $R^2 \ll L/C$). In this case, the denominator has two frequency zeroes given by

$$\omega_{\pm} = \pm\frac{1}{\sqrt{LC}} - i\frac{R}{2L}. \tag{2.82}$$
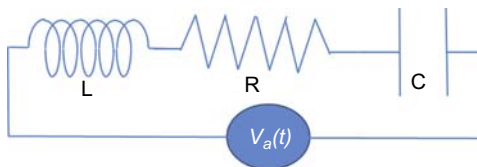
Figure 2.7 A series LRC circuit driven by an applied voltage $V_a(t)$.

The real part represents the resonant frequency of the oscillator, $\omega_0 = \frac{1}{\sqrt{LC}}$, and the imaginary term is one half the damping coefficient, $\gamma = \frac{R}{L}$ defined by the rate of energy loss in the circuit. The quality factor is defined as the maximum energy stored in the circuit at resonance divided by the energy loss per cycle

$$Q_f = 2\pi \frac{E_{stored}}{E_{loss}} = \omega_0 \frac{E_{stored}}{P_{loss}}. \tag{2.83}$$

The last equality equates the energy loss to the power loss in one cycle. In the previously discussed small resistance limit, the power loss is $P_{loss} = \frac{1}{2}R|I|^2$ and the energy stored is $E_{stored} = \left(\frac{|Q|^2}{2C} + L\frac{|I|^2}{2}\right) = L|I|^2$. The quality factor is

$$Q_f = \frac{\omega_0}{\gamma}. \tag{2.84}$$

2.5.2 Source-free formulation of nanocircuits

The treatment of nanostructured materials as consisting of lumped circuit elements (capacitors, resistors, and inductors) has great appeal for bringing a simplification to the analysis of complex systems. Specific results of the nanocircuit treatment depend on the decomposition of the fields. Circuits are described by voltages and currents; therefore, the starting point of the analysis is the treatment of the displacement current. In this case, we have in mind a dielectric material with dielectric permittivity ε surrounded by free space with permittivity ε_0 as illustrated in Figure 2.8. The total displacement current in a region of space is described as

$$\vec{j}_d = -i\omega\vec{D} = -i\omega\varepsilon\vec{E}. \tag{2.85}$$

This can be decomposed in different ways; here, we follow the source-free formulation used by Zhu. The displacement current is decomposed into a free-space
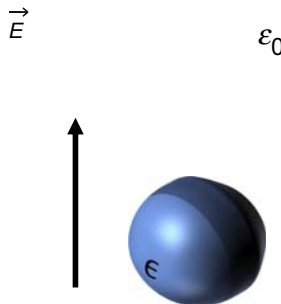
Figure 2.8 Dielectric sphere embedded in material with dielectric constant ε_0.

displacement current contribution $\left(-i\omega\vec{D}_0 = -i\omega\varepsilon_0\vec{E}\right)$ and a current density $\left(\vec{j} = -i\omega(\varepsilon - \varepsilon_0)\vec{E}\right)$ giving the identity

$$\vec{j}_d = \vec{j} - i\omega\vec{D}_0. \tag{2.86}$$

The conductivity is identified from the current density as

$$\sigma = -i\omega(\varepsilon - \varepsilon_0). \tag{2.87}$$

Using this identity, Ampere's law can be expressed as

$$\vec{\nabla} \times \vec{H} = \vec{j} - i\omega\vec{D}_0. \tag{2.88}$$

Taking the divergence of this equation,

$$\vec{\nabla} \cdot \vec{j} - i\omega\vec{\nabla} \cdot \vec{D}_0 = 0. \tag{2.89}$$

Imposing conservation of charge leads to the following identification:

$$\vec{\nabla} \cdot \vec{D}_0 = \rho. \tag{2.90}$$

The other Maxwell equations, Eqns (2.2) and (2.3), remain unchanged. The circuit coefficients are derived by application of Poynting's theorem in Eqn (2.35), incorporating the source-free version of Ampere's law. The revised form of the time-averaged Poynting theorem is

$$-\oint_{\text{in}}\hat{\mathbf{n}} \cdot \vec{S}\mathrm{d}A = \frac{1}{2}\iiint_V \frac{1}{\sigma}|\vec{j}|^2\mathrm{d}V - \frac{1}{2i\omega\varepsilon_0}\iiint_V |\vec{D}_0|^2\mathrm{d}V$$
$$+ \frac{i\omega\mu_0}{2}\iiint_V |\vec{H}|^2\mathrm{d}V. \tag{2.91}$$

The source-free version of Poynting's theorem is separated into real and imaginary contributions. The first integral on the right-hand side is related to the kinetic energy of the electrons. It is separated into its real and imaginary parts, which define the effective resistivity and kinetic inductance:

$$\frac{1}{2}(R - i\omega L_K)|I|^2 = \frac{1}{2}\iiint_V \frac{1}{\sigma}|\vec{j}|^2\mathrm{d}V. \tag{2.92}$$

The real part is called Joule heating and the imaginary part is the electronic contribution to the inductance and is discussed further in the following section. The

capacitance, due to the electric energy density, and Faraday inductance, due to the magnetic energy, are identified in Eqn (2.58) as

$$\frac{1}{2C}|Q|^2 = \frac{1}{2\varepsilon_0} \iiint_V |\vec{D}_0|^2 \mathrm{d}V,$$ (2.93)

$$\frac{1}{2}(L_{\mathrm{F}})|I|^2 = \frac{\mu_0}{2} \iiint_V |\vec{H}|^2 \mathrm{d}V.$$ (2.94)

The total current is determined from the integral

$$I = \iint \vec{j}_{\mathrm{f}} \cdot \mathrm{d}\vec{S}.$$ (2.95)

2.5.3 Kinetic inductance

As mentioned earlier, in the nanoscale limit there is a required fourth circuit element called kinetic inductance related to the kinetic energy of the electrons and that is not included in the magnetic field expression of the (Faraday) inductance. It is invoked in nanocircuits and in superconductor systems in which the kinetic inductance provides additional inertia to imposed current changes and the magnetic contribution is too weak to contribute.

To determine an expression for the kinetic inductance in a bulk medium, a cylinder of displaced charges is used as an example, as shown in Figure 2.9. The displacement of charges creates a restoring force that returns the medium toward neutrality. The equation for the mass acceleration is

$$M\frac{\mathrm{d}v}{\mathrm{d}t} = F = qE.$$ (2.96)

The mass in the thin slice is $M = mnA\mathrm{d}z$ and the total charge is $q = -enA\mathrm{d}z$, where n is the electron density, $\mathrm{d}z$ is the thickness of the slice, and A is the area of the cylinder end cap. The mass and velocity of the charge are related to the current density by

$$\vec{j}_{\mathrm{f}} = -env.$$ (2.97)
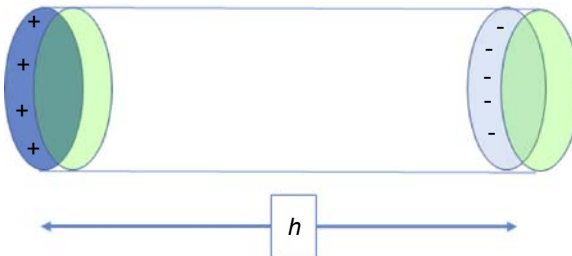
Figure 2.9 A cylinder of length h with electrons displaced from its neutral position.

Neglecting any fringe fields, the capacitor geometry gives a simple relation between the voltage and the field voltage is expressed as $V = \vec{E} \cdot \vec{h}$, the current and charge velocity are related by $I = \vec{j}A = -enAv$. Inserting these relations into Eqn (2.96) we can determine the expression for the kinetic inductance, which is defined by

$$L_k \frac{dI}{dt} = V. \tag{2.98}$$

The kinetic inductance in a bulk medium is given as $L_k = \frac{m}{e^2 n} \frac{h}{A} = \frac{1}{\varepsilon_0 \omega_p^2} \frac{h}{A}$. The frequency ω_p^2 is the plasma frequency defined under Eqn (2.22). Note that for the cylinder geometry considered in Figure 2.9, the capacitance is given by $C = \varepsilon_0 \frac{A}{h}$; then, the plasma frequency is simply given by $\omega_p^2 = 1/L_k C$ and the plasma frequency becomes the natural oscillation frequency for displaced free charges in a simple metal.

The published data on material dielectric functions do not distinguish between the bound and free electron contributions. A more general expression for the kinetic inductance can be obtained by separating the bound and free contributions. The current density in a material is related to the applied field by a generalized form of Ohm's law for isotropic materials,

$$\vec{j}_\omega = \sigma(\omega)\vec{E}_\omega. \tag{2.99}$$

The current density includes contributions from bound and free charges

$$\vec{j}_\omega = \vec{j}_{f\omega} - i\omega\vec{D}_\omega = \sigma(\omega)\vec{E}_\omega. \tag{2.100}$$

The displacement field is related to the material dielectric constant

$$\vec{D}_\omega = \varepsilon_b(\omega)\vec{E}_\omega, \tag{2.101}$$

$$\vec{j}_{f\omega} = \sigma_f(\omega)\vec{E}_\omega. \tag{2.102}$$

The total current is

$$\vec{j}_{T\omega} = \vec{j}_{f\omega} - i\omega\vec{D}_\omega = (\sigma_f(\omega) - i\omega\varepsilon_b(\omega))\vec{E}_\omega = -i\omega\varepsilon(\omega)\vec{E}_\omega \tag{2.103}$$

$$\sigma_f(\omega) = -i\omega\varepsilon_0\chi_f(\omega). \tag{2.104}$$

The free charge susceptibility $\chi_f(\omega)$ defined in Eqn (2.24) is a complex quantity. Solving for the free charge susceptibility, $\chi_f(\omega) = (\varepsilon(\omega) - \varepsilon_b(\omega))/\varepsilon_0$. Applying the inductance definition, a more general expression for the kinetic inductance is found:

$$L_k = -\frac{1}{\omega^2}\text{Re}\left\{\frac{1}{\varepsilon(\omega) - \varepsilon_b(\omega)}\right\}\frac{h}{A}. \tag{2.105}$$

The bound electron dielectric function contribution $\varepsilon_b(\omega)$ may be extracted from the dielectric data by fitting using the Drude and Lorentz models with optimized fitting parameters. The result in Eqn (2.105) is proportional to the plasma frequency and recovers the heuristic results discussed previously as the electron damping vanishes. This result incorporates a modification of the kinetic inductance due to the electronic damping contribution.

2.5.4   Nanocircuit model

We conclude this chapter with a simple application of the aforementioned nanocircuit formulations. Consider a dielectric sphere of radius a is suspended in vacuum with a uniform electric field directed along the z-axis, $\vec{E}_0 = E_0\hat{z}$. The electric field solution of this ubiquitous electrostatic problem is

$$\vec{E}_< = \frac{3\varepsilon_0}{(\varepsilon(\omega) + 2\varepsilon_0)}\vec{E}_0, \tag{2.106}$$

$$\vec{E}_> = \vec{E}_0 + \frac{\left(3n(\vec{p}\cdot\hat{n}) - \vec{p}\right)}{4\pi\varepsilon_0 r^3}. \tag{2.107}$$

The dipole moment for the sphere is $(\hat{n} = \vec{r}/r)$:

$$\vec{p} = 4\pi\varepsilon_0 a^3 \frac{(\varepsilon(\omega) - \varepsilon_0)}{(\varepsilon(\omega) + 2\varepsilon_0)}\vec{E}_0. \tag{2.108}$$

The source-free expressions for the fields (in)side and (out)side the dielectric sphere are

$$\vec{E}_{\text{in}} = \frac{(\varepsilon(\omega) - \varepsilon_0)}{(\varepsilon(\omega) + 2\varepsilon_0)}\vec{E}_0, \tag{2.109}$$

$$E_{\text{out}} = \frac{\left(3n(\vec{p}\cdot\hat{n}) - \vec{p}\right)}{4\pi\varepsilon_0 r^3}. \tag{2.110}$$

Using these fields, the following integrals are evaluated as: sigma2 Ein.Ein.

$$\iiint_{V_<} \vec{j}\cdot\vec{j}\,dV = \frac{4\pi}{3}a^3\sigma^2 \vec{E}_{\text{in}}\cdot\vec{E}_{\text{in}}, \tag{2.111}$$

$$\iiint_V \vec{D}_0\cdot\vec{D}_0\,dV = (\varepsilon_0)^2 \left(\iiint_{V_<} \vec{E}_{\text{in}}\cdot\vec{E}_{\text{in}}\,dV + \iiint_{V_>} \vec{E}_{\text{out}}\cdot\vec{E}_{\text{out}}\,dV \right)$$

$$= (\varepsilon_0)^2 \left(\frac{4\pi a^3}{3}(E_{\text{in}})^2 + \frac{8\pi a^3}{3}(E_{\text{in}})^2 \right). \tag{2.112}$$

The conductivity was previously defined in Eqn (2.86). The LRC coefficients are calculated as

$$R - i\omega L_K = \frac{\iiint\limits_{V_<} \vec{j} \cdot \vec{j} dV}{|I|^2} = \frac{1}{\sigma} \frac{4}{3\pi a}, \tag{2.113}$$

$$C = \varepsilon_0 Q^2 \left/ \iiint\limits_V |\vec{D}_0|^2 dV \right. = \frac{\pi a}{4\varepsilon_0 \omega^2} |\sigma|^2 \tag{2.114}$$

Applying the Drude model, the LRC coefficients are

$$R - i\omega L_K = \frac{4}{3\pi a \varepsilon_0 \omega_p^2}(\gamma - i\omega), \tag{2.115}$$

$$C = \frac{\pi a}{4\varepsilon_0} \frac{\omega_p^4}{\omega^2(\omega^2 + \gamma^2)}. \tag{2.116}$$

From Eqns (2.113) and (2.114), the LC resonance frequency, $\omega = \omega_r(\sigma = \sigma' + i\sigma'')$ is calculated from the product of the capacitance and the inductance:

$$CL_K = \frac{\iiint\limits_{V_<} \vec{j} \cdot \vec{j} dV}{I^2} = \mathrm{Im}\left\{\frac{1}{\sigma}\right\}|\sigma|^2 \frac{-1}{3\varepsilon_0 \omega_r^3} = \frac{\sigma''}{3\varepsilon_0 \omega_r^3} = \frac{1}{3\varepsilon_0 \omega_r^2}(\varepsilon_0 - \varepsilon'). \tag{2.117}$$

The resonant frequency of the electric field in the sphere is found by setting the real part of the denominator in Eqn (2.109) equal to zero:

$$\varepsilon'(\omega_r) = -2\varepsilon_0. \tag{2.118}$$

For this condition, the resonant frequency is identically $CL_K = \frac{1}{\omega_r^2}$. Using the Drude mode results from Eqns (2.115) and (2.116) in Eqn (2.117) and neglecting the damping term, the resonant frequency is $\omega_r = \omega_p/\sqrt{3}$. The same result is also found using the resonance condition in Eqn (2.118).

The results applied here can also be applied to the two-dimensional case of dielectric rods (of infinite length) embedded in vacuum with some care being given to the calculation of the capacitance and inductance. Further treatment of this case is left as a homework problem.

Problems

1. The polarization in Eqn (2.22) can be adopted to describe a simple dielectric material embedded in a static magnetic field $B = B_0\hat{z}$. Consider the dynamical equations for the

polarization as a function of the electric field. Note the equations are best solved using the circular polarization basis defined by $\vec{P} = P_+\hat{e}_+ + P_-\hat{e}_-$ with unit vectors defined as $\hat{e}_\pm = (\hat{x} \pm i\hat{y})/\sqrt{2}$.

   a. Solve for the susceptibility: $\chi_{b,\pm}(\omega, B_0)$.

   b. Calculate the refractive index neglecting the damping contribution and approximate the expression when the magnetic field is a small perturbation.

2. Consider the Drude model of free electrons defined after Eqn (2.23). Write the kinematic equation of motion for the free electron polarization and show that the susceptibility is given by Eqn (2.24).

3. For s- and p-polarizations, verify Eqns (2.62) and (2.64).

4. Derive the transmission and reflection coefficients for s- and p-polarization as a function of angle for a free-standing thin film of thickness d and complex refractive index $n + ik$ surrounded by air. Use the result to numerically explore the R and T for various cases:

   a. $n = 1.5$, $k = 0$, and $d = \lambda/(2n)$, where λ is the wavelength; discuss the Brewster angle results.

   b. $n = 1.5$, $k = 0.1$, and $d = \lambda/(2n)$.

5. Derive the transmission and reflection coefficients for s- and p-polarization as a function of angle for a free-standing thin film of thickness d and complex refractive index $n + ik$ sandwiched between air on the superstrate and material with real refractive index n_1 on the substrate. Use the result to numerically explore the R and T for various cases:

   a. $n = 1.414$, $k = 0$, $n_1 = 2$, and $d = \lambda/(4n)$ where λ is the wavelength; discuss the normal incident results. What is the angular dependence of R and T for this case?

   b. $n = 1.5$, $k = 0.1$, $n_1 = 2$, and $d = \lambda/(2n)$. Plot R and T versus the angle of incidence again.

6. Consider the Drude model (neglecting damping) form of the susceptibility for the dielectric constant and the magnetic permeability:

$$\varepsilon(\omega) = 1 - \frac{\omega_{p,e}^2}{\omega^2} \text{ and } (\omega) = 1 - \frac{\omega_{p,m}^2}{\omega^2}.$$

$\omega_{p,e}$ is the dielectric plasma frequency and $\omega_{p,m}$ is the magnetic plasma frequency. Show that the coefficients of the fields in Eqn (2.58) are positive for all frequencies.

7. Show that for a plane wave for which the energy density is described by Eqn (2.58), the group velocity is defined by $v_g = \hat{z}\partial\omega/\partial k$, where ω is the angular frequency and $k\hat{z}$ is the wave vector of the plane wave. Show that the group velocity is equal to the energy velocity defined by

$$\vec{v}_E = \langle \vec{S} \rangle / w,$$

where $\langle \vec{S} \rangle$ is the average Poynting vector. Assume that the imaginary contributions to the permittivity and permeability are negligible when computing the fields and Poynting vector.

8. An electric field is applied external to a long cylinder of radius a and made with material of dielectric constant $\varepsilon(\omega)$ embedded in a medium of dielectric constant ε_0, as shown in Figure 2.10.

   a. Using electrostatics, calculate the electric field inside and outside of the cylinder.

   b. Calculate the LRC values adapting the relations in Eqns (2.113) and (2.114).

   c. Show that the LC resonance frequency satisfies the condition $\varepsilon'(\omega_r) = -\varepsilon_0$.
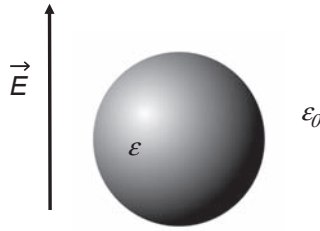
Figure 2.10 Dielectric cylinder embedded in a medium with dielectric constant ε_0.

References for electrodynamics

There are many good books on Electrodynamics. A reference of note is

[1]   L.D. Landau, E.M. Lifshitz, Electrodynamics of Continuous Media second ed., vol. 8, Elsevier, Amsterdam, 2004.

The following books can be consulted for more details.

[2]   M.A. Heald, J.B. Marion, Classical Electrodynamic Radiation, third ed., Saunders College Publishing, Fort Worth, 1995.
[3]   J.D. Jackson, Classical Electrodynamics, third ed., John Wiley and Sons, New York, 1998.
[4]   C.A. Balanis, Advanced Engineering Electromagnetics, second ed., John Wiley and Sons, New York, 2012.
[5]   D.J. Griffiths, Introduction to Electrodynamics, fourth ed., Addison-Wesley, Boston, 2014.

The following book is a useful mathematical primer on the operator calculus.

[6]   H.M. Schey, Div, Grad, Curl, and All that: An Informal Text on Vector Calculus, fourth ed., W.W. Norton and Co, 2004.

A reference covering aspects of electrodynamics in this chapter can be found in

[7]   L. Novotny, B. Hecht, Principles of Nano-optics, Cambridge University Press, Cambridge, 2012.

The nanocircuit concept was advanced in a series of articles by Engheta and collaborators. The following references are an example from the literature.

[8]   N. Engheta, A. Salandrino, A. Alu, Circuit elements at optical frequencies: nano-inductors, nanocapacitors and nanoreistors, Phys. Rev. Lett. 95 (2005) 095504.
[9]   A. Alu, A. Salandrino, N. Engheta, Coupling of lumped optical circuit elements and effect of substrate, Opt. Expr. 15 (2007) 13865.
[10] A. Alu, N. Engheta, Optical 'shorting wires', Opt. Expr. 15 (2007) 13773.
[11] M. Staffaroni, J. Conway, S. Vedantam, J. Tang, E. Yablonovitch, Circuit analysis in metal optics, Photonics Nanostruct. Fundam. Appl. 10 (2012) 166−176.

The nanocircuit presentation discussed in this chapter follows

[12] D. Zhu, M. Bosman, J.K.W. Yang, A circuit model for plasmonic resonators, Opt. Express 22 (2014) 9810.

The Drude model has a long history and has been used with the Lorentz model of a bound electron to fit experimental data. A useful reference with fit parameters for a variety of metals is

[13]   A.D. Rakic, A.B. Djurišic, J.M. Elazar, M.L. Majewski, Optical properties of metallic films for vertical cavity optoelectronic devices, Appl. Opt. 37 (1998) 5271−5283.

A collection of experimental data on dielectric dispersion is found in the collection

[14]   E.D. Palik, Handbook of Optical Constants of Solids (Academic Press, NY) vols. I−V.
[15]   L. Brillouin, Wave Propagation and Group Velocity, Academic, New York, 1960.
[16]   M. Born, E. Wolf, Principles of Optics, fourth ed., Pergamon, Oxford, 1970.

This page intentionally left blank

Quantum mechanics and computation in nanophotonics

3

A. Sarangan
University of Dayton, Dayton, OH, USA

3.1   Introductory concepts

3.1.1   Schrodinger's equation

The motion of electrons at the nanoscale is governed by the Schrodinger's wave equation, given by

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V\right)\psi = i\hbar\frac{\partial\psi}{\partial t}, \tag{3.1}$$

where m is the mass of the electron, $\hbar$ is Plank's constant divided by 2π, V is the potential energy of the system, and ψ is the wave function of the electron. The first term on the left side $-\frac{\hbar^2}{2m}\nabla^2$ is the kinetic energy operator, and the second term V is the potential energy operator. In general, the potential energy will vary with position rather than remain a constant. Therefore, Eqn (3.1) states that the total energy of the electron can be computed from its spatial profile (left side) or from its temporal profile (right side). Assigning ε as the total energy of the electron, we can also write Eqn (3.1) as two separate equations:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V\right)\psi = \varepsilon\psi, \tag{3.2}$$

$$i\hbar\frac{\partial\psi}{\partial t} = \varepsilon\psi, \tag{3.3}$$

Equation (3.2) is often referred to as the time-independent Schrodinger's equation. The time dependence is in Eqn (3.3), from which we can see that ψ has a time harmonic solution of form

$$\psi = \psi_0 e^{-i\left(\frac{\varepsilon}{\hbar}\right)t} = \psi_0 e^{-i\omega t}, \tag{3.4}$$

which states that an electron with an energy of ε will exhibit a temporal oscillation with an angular frequency ω. In most quantum mechanics problems, we often start with a known potential energy profile V, then apply Eqn (3.2) to calculate the total energy ε and the corresponding wave function ψ.
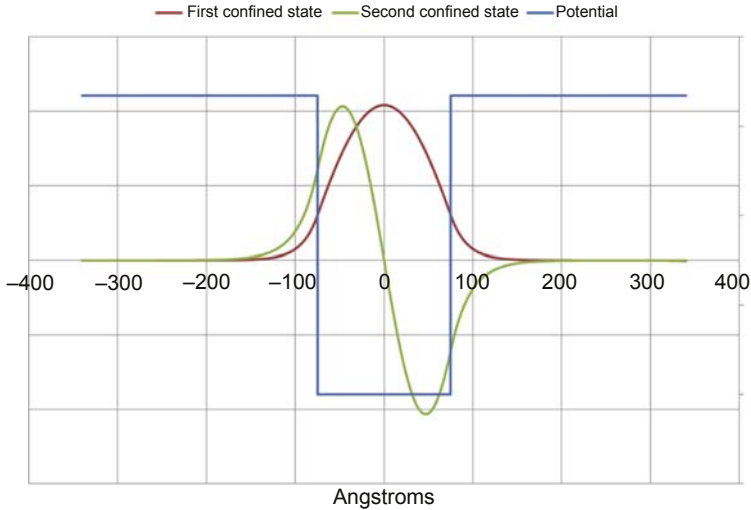
Figure 3.1 Illustration of two electron wave functions in a confining potential.

A simple example is a one-dimensional rectangular quantum well with a potential of zero inside of the well and a higher value outside of the well, as shown in Figure 3.1. Electrons in the system will naturally gravitate toward the center of the well where the potential is lower. However, the wave nature of electrons forces the electrons to exist in discrete states governed by the wavelength of the electron (which is related to its energy) and the width of the well. These solutions can be found by solving the time-independent Schrodinger's wave Eqn (3.2). This results in several different solutions, each having a pair of values (ε, ψ). Figure 3.1 shows an illustration of the first two confined states in a quantum well and their wave functions.

3.1.2 Interpretation of ψ

ψ is a complex number with amplitude and phase and does not have a direct physical meaning. However, the interpretation of $|\psi|^2$ is more meaningful—it is a dimensionless real number interpreted as the probability density function. In the context of a single electron, this interpretation can still lead to the dilemma of an electron's exact location being nondeterministic. However, in practical calculations this is rarely relevant because we deal with systems that have a vast number of electrons. In these cases, $qN|\psi|^2$ can be simply interpreted as the charge density where N is the total number of electrons and q is the unit electronic charge, circumventing the debate of whether an electron is a particle or a wave.

3.1.3 Quantum confinement in one dimension with infinite potentials

Nearly every quantum structure of practical relevance requires a numerical approach because the confining potential V is rarely a simple analytical function. However,

for the sake of completeness, we will first start with an elementary example of a rectangular one-dimensional potential well described by the following function:

$$V(z) = 0 \ \text{for} \ |z| < \frac{L_z}{2}, \tag{3.5}$$

$$V(z) = \infty \ \text{for} \ |z| \geq \frac{L_z}{2}. \tag{3.6}$$

This represents a sandwich structure of three different layered media as shown in Figure 3.2. The center layer is the quantum well and the surrounding layers are the barriers. For the assumed one-dimensional potential, we can recognize that it can be expressed as

$$V = V(z) + V(x) + V(y), \tag{3.7}$$

where incidentally $V(x) = 0$ and $V(y) = 0$. This is a key assumption that allows us to expand Eqn (3.2) in Cartesian coordinates as

$$\left(-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial z^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right] + V(z) + V(x) + V(y)\right)\psi = \varepsilon\psi. \tag{3.8}$$

Recognizing that the operators are independently acting along x, y, and z allows us to write ψ as a product of three independent wave functions and ε as a sum of three energies. This technique is known as the *separation of variables*.
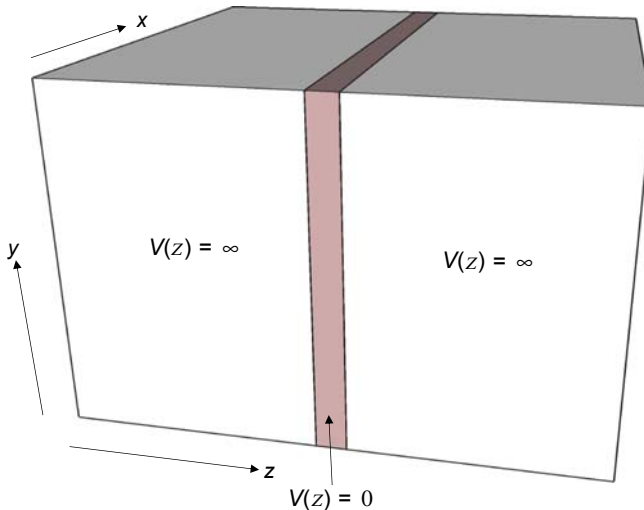
Figure 3.2 Quantum well structure from three-layered films.

$$\psi = \psi(z)\psi(x)\psi(y), \tag{3.9}$$

$$\varepsilon = \varepsilon_x + \varepsilon_y + \varepsilon_z. \tag{3.10}$$

With these substitutions, Eqn (3.8) becomes

$$-\frac{\hbar^2}{2m}\left[\psi(x)\psi(y)\frac{\partial^2\psi(z)}{\partial z^2} + \psi(z)\psi(y)\frac{\partial^2\psi(x)}{\partial x^2} + \psi(z)\psi(x)\frac{\partial^2\psi(y)}{\partial y^2}\right]$$
$$+ (V(z) + V(x) + V(y))\psi(z)\psi(x)\psi(y) = (\varepsilon_x + \varepsilon_y + \varepsilon_z)\psi(z)\psi(x)\psi(y). \tag{3.11}$$

Dividing both sides by $\psi(z)\psi(x)\psi(y)$ results in

$$-\frac{\hbar^2}{2m}\left[\frac{1}{\psi(z)}\frac{\partial^2\psi(z)}{\partial z^2} + \frac{1}{\psi(x)}\frac{\partial^2\psi(x)}{\partial x^2} + \frac{1}{\psi(y)}\frac{\partial^2\psi(y)}{\partial y^2}\right] + (V(z) + V(x) + V(y))$$
$$= (\varepsilon_x + \varepsilon_y + \varepsilon_z). \tag{3.12}$$

From Eqn (3.12), we can decouple the x, y, and z dependent terms into separate equations:

$$-\frac{\hbar^2}{2m}\frac{1}{\psi(z)}\frac{\partial^2\psi(z)}{\partial z^2} + V(z) = \varepsilon_z, \tag{3.13}$$

$$-\frac{\hbar^2}{2m}\frac{1}{\psi(x)}\frac{\partial^2\psi(x)}{\partial x^2} + V(x) = \varepsilon_x, \tag{3.14}$$

$$-\frac{\hbar^2}{2m}\frac{1}{\psi(y)}\frac{\partial^2\psi(y)}{\partial y^2} + V(y) = \varepsilon_y. \tag{3.15}$$

For the one-dimensional potential $V(z)$ that we defined earlier, $V(x) = 0$ and $V(y) = 0$. Therefore,

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2} = \varepsilon_x\,\psi(x), \tag{3.16}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi(y)}{\partial y^2} = \varepsilon_y\psi(y), \tag{3.17}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi(z)}{\partial z^2} + V(z)\psi(z) = \varepsilon_z\psi(z). \tag{3.18}$$

These are the final three equations we need to solve.

The first two equations are straightforward. Their solutions are

$$\psi(x) = A_x e^{\pm i k_x x} \quad \text{and} \quad \psi(y) = A_y e^{\pm i k_y y}, \tag{3.19}$$

where

$$k_x = \frac{\sqrt{2m\varepsilon_x}}{\hbar} \quad \text{and} \quad k_y = \frac{\sqrt{2m\varepsilon_y}}{\hbar}. \tag{3.20}$$

The third equation, which is in the direction of the electron's confinement, requires a few more steps. Because the electron's energy is finite, the probability of its existence outside of the well has to be zero because the potential energy there is infinitely large. Furthermore, because $|\psi|^2$ represents the probability distribution of the electron's location, we can conclude that ψ must have a value of zero for all values $|z| \geq \frac{L_z}{2}$. This becomes the boundary condition for solving the third equation.

Using this assumption, we can obtain

$$\psi(z) = 0 \ \text{ for } \ |z| \geq \frac{L_z}{2}, \tag{3.21}$$

$$\psi(z) = A_z e^{\pm i k_z z} \ \text{ for } \ |z| < \frac{L_z}{2}, \tag{3.22}$$

where

$$k_z = \frac{\sqrt{2m\varepsilon_z}}{\hbar}. \tag{3.23}$$

Furthermore, because $\psi(z)$ has to satisfy both conditions at $|z| = \frac{L_z}{2}$, this results in $k_z L_z$ being an integer number of π, such as

$$k_z L_z = n\pi, \tag{3.24}$$

with

$$\psi(z) = A_{zn} \cos\left(n\frac{\pi}{L_z}z\right) \ \text{ for } \ n = 1, 3, 5, 7...\text{odd multiples}, \tag{3.25}$$

$$\psi(z) = A_{zn} \sin\left(n\frac{\pi}{L_z}z\right) \ \text{ for } \ n = 2, 4, 6, 8...\text{even multiples}. \tag{3.26}$$

The confinement energy becomes

$$\varepsilon_z = \frac{\hbar^2 k_z^2}{2m}, \tag{3.27}$$

$$= \frac{\hbar^2}{2m}\left(\frac{n\pi}{L_z}\right)^2, \tag{3.28}$$

$$= \varepsilon_{z0} n^2, \tag{3.29}$$

where ε_{z0} is the ground state for the confinement given by

$$\varepsilon_{z0} = \frac{\hbar^2}{2m}\left(\frac{\pi}{L_z}\right)^2. \tag{3.30}$$

This result basically says that the energy of the electron along the z axis (direction of confinement) is discretized in steps of $\varepsilon_{z0}, 4\varepsilon_{z0}, 9\varepsilon_{z0}, 16\varepsilon_{z0}$, etc. As a handy reference, $\frac{\hbar^2\pi^2}{2m}$ has a value of 37.6 eV/Å^2, which makes $\varepsilon_{z0} = \frac{37.6}{L_z^2}$ with L_z given in Angstroms.

However, the total energy of the electron is not just ε_z, but the sum of all three energies $\varepsilon_x + \varepsilon_y + \varepsilon_z$. We can see that the equations along x and y directions do not have a boundary condition; therefore, their solutions are essentially a continuum. Therefore, the energies $\varepsilon_x + \varepsilon_y$ can take any value. The discretization only exists along the z axis. However, in practice the electrons are always bounded at some distance L_x and L_y by the material boundaries. However, because these dimensions are typically much larger than the quantum well dimension L_z, the discretization step ε_{x0} and ε_{y0} will be extremely small and can be considered to be zero.

All of the coefficients A_{zn} of the wave functions still remain to be calculated. These can only be determined by assigning the integral $\int_{-\infty}^{+\infty} |\psi|^2 dV$ to a known constant. We choose this to be unity because of the interpretation of $|\psi|^2$ as the probability density function. With this assumption,

$$\int_{-\infty}^{+\infty} |\psi(z)\psi(x)\psi(y)|^2 dz dx dy = 1. \tag{3.31}$$

The separation of variables allows us to normalize each of these functions separately as

$$\int_{-\infty}^{+\infty} |\psi(z)|^2 dz = 1, \tag{3.32}$$

$$\int\limits_{-\infty}^{+\infty} |\psi(x)|^2 dx = 1,$$ (3.33)

$$\int\limits_{-\infty}^{+\infty} |\psi(y)|^2 dy = 1.$$ (3.34)

The normalization integration for $\psi(z)$ (which is left as an exercise for the reader) can be shown to be

$$A_{zn} = \sqrt{\frac{2}{L_z}}.$$ (3.35)

On the other hand, the normalization of $\psi(x)$ and $\psi(y)$ is also identical to Eqn (3.35), but it is usually not necessary unless one needs to know the electronic distributions along the x and y directions. In that case, we can consider A_x and A_y to be the normalization constants in the limiting case of $L_x \rightarrow \infty$ and $L_y \rightarrow \infty$.

In summary, the full solution to the infinite quantum well can be written as

$$\varepsilon = \varepsilon_{z0} n^2 + \varepsilon_{x0} m^2 + \varepsilon_{y0} p^2,$$ (3.36)

and

$$\psi = \sqrt{\frac{2}{L_z}} \left[\cos\left(n\frac{\pi}{L_z} z \right) \quad \text{or} \quad \sin\left(n\frac{\pi}{L_z} z \right) \right] A_x e^{\pm jk_x x} A_y e^{\pm jk_y y}.$$ (3.37)

3.1.3.1   Numerical example

Let us consider an infinite potential well with $L_z = 100$ Å. The energy states for this system are straightforward to calculate. As a reminder, we use the following universal constants:

$$m = 9.1 \times 10^{-31} \text{ kg}$$ (3.38)

$$\hbar = \frac{h}{2\pi} = 1.05 \times 10^{-34} \text{ J}\cdot\text{s}.$$ (3.39)

From this, we can obtain

$$\varepsilon_{z0} = \frac{\hbar^2}{2m} \left(\frac{\pi}{a}\right)^2 = 6.02 \times 10^{-22} \text{ J}$$ (3.40)
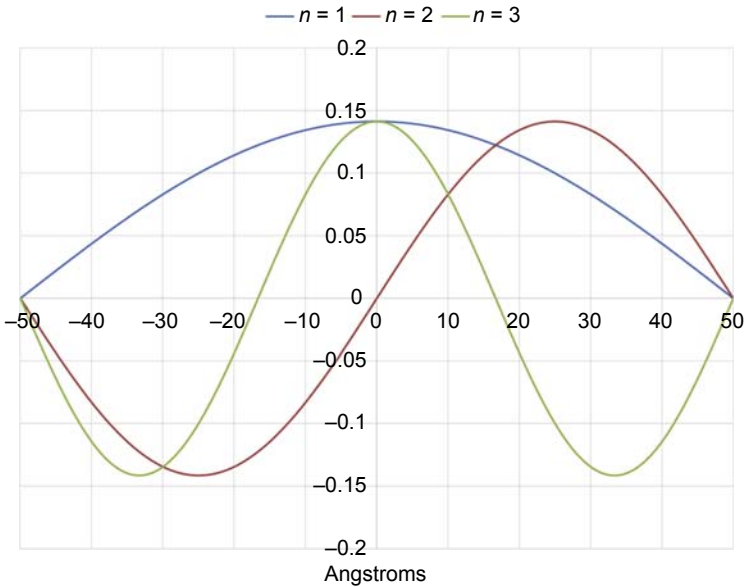
Figure 3.3 First three wave functions in an infinitely deep 100-Å wide potential well.

or, expressed in the more commonly used units of electron-volts (eV) or milli-electron-volts (meV),

$$\varepsilon_{z0} = 3.76 \ \text{meV}. \tag{3.41}$$

Therefore, discrete energy states along the z axis will be 3.76, 15.04, 33.84 meV…, and the wave functions $\psi(z)$ will be $\sqrt{\frac{2}{L_z}}\cos\left(\frac{\pi}{a}z\right)$, $\sqrt{\frac{2}{L_z}}\sin\left(\frac{2\pi}{a}z\right)$, $\sqrt{\frac{2}{L_z}}\cos\left(\frac{3\pi}{a}z\right)$, etc.

Figure 3.3 illustrates these functions and verifies that the functions do indeed fall to zero at the boundaries $z = \pm 50$ Å.

3.1.4 Quantum confinement in one dimension with finite potentials

In the previous section, we examined a quantum well surrounded by infinitely large potential barriers. This is one of only a few special cases for which the energy and wave functions have closed form solutions. When the barriers have finite potentials, the solution requires a semianalytical approach. We write the general solutions for each of the three regions and then apply appropriate boundary conditions at the interfaces to obtain the solution that equally applies to all three regions.
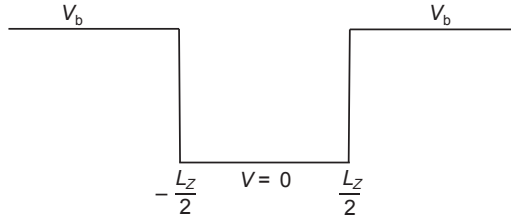
Figure 3.4 Illustration of a finite symmetric potential well structure.

A symmetric finite potential well as shown in Figure 3.4 can be defined mathematically as

$$V(z) = 0 \quad \text{for} \quad |z| < \frac{L_z}{2} \tag{3.42}$$

$$V(z) = V_b \quad \text{for} \quad |z| \geq \frac{L_z}{2}. \tag{3.43}$$

Inside of the well region, we can write the solution as

$$\psi_w(z) = A\cos(k_w z) \quad \text{or} \quad A\sin(k_w z), \tag{3.44}$$

and for $z < -\frac{L_z}{2}$

$$\psi_{bl}(z) = Be^{k_b z}, \tag{3.45}$$

and for $z > \frac{L_z}{2}$

$$\psi_{br}(z) = Ce^{-k_b z}, \tag{3.46}$$

where

$$k_w = \frac{\sqrt{2mE}}{\hbar} \tag{3.47}$$

$$k_b = \frac{\sqrt{2m(V - E)}}{\hbar}. \tag{3.48}$$

First we will consider the symmetric solution of the form $A\cos(k_w z)$. Applying the boundary conditions for $\psi_w(z)$, we can obtain

$$A\cos\left(k_w \frac{L_z}{2}\right) = Be^{-k_b \frac{L_z}{2}}, \tag{3.49}$$

$$A\cos\left(k_w \frac{L_z}{2}\right) = Ce^{-k_b \frac{L_z}{2}}, \tag{3.50}$$

$$Ak_{\mathrm{w}} \sin\left(k_{\mathrm{w}} \frac{L_z}{2}\right) = k_{\mathrm{b}} B e^{-k_{\mathrm{b}}\frac{L_z}{2}}, \tag{3.51}$$

$$-Ak_{\mathrm{w}} \sin\left(k_{\mathrm{w}} \frac{L_z}{2}\right) = -k_{\mathrm{b}} C e^{-k_{\mathrm{b}}\frac{L_z}{2}}, \tag{3.52}$$

From these equations, we can deduce that

$$B = C. \tag{3.53}$$

This results in

$$k_{\mathrm{w}} \tan\left(k_{\mathrm{w}} \frac{L_z}{2}\right) = k_{\mathrm{b}}. \tag{3.54}$$

Manipulating it further,

$$k_{\mathrm{w}} \frac{L_z}{2} \tan\left(k_{\mathrm{w}} \frac{L_z}{2}\right) = k_{\mathrm{b}} \frac{L_z}{2}, \tag{3.55}$$

which then becomes

$$\frac{\sqrt{2mE}}{\hbar} \frac{L_z}{2} \tan\left(\frac{\sqrt{2mE}}{\hbar} \frac{L_z}{2}\right) = \frac{\sqrt{2m(V-E)}}{\hbar} \frac{L_z}{2}. \tag{3.56}$$

Then, we define a few normalized parameters:

$$\frac{\sqrt{2mE}}{\hbar} \frac{L_z}{2} = w, \tag{3.57}$$

and

$$\frac{\sqrt{2mV}}{\hbar} \frac{L_z}{2} = R. \tag{3.58}$$

The final equation then becomes

$$w \tan(w) = \sqrt{R^2 - w^2}. \tag{3.59}$$

This equation can only be solved numerically for w and R, from which we can obtain the energy E.

Next, consider the antisymmetric solution of the form $A \sin(k_w z)$. Applying the boundary conditions results in

$$-A \sin\left(k_w \frac{L_z}{2}\right) = B e^{-k_b \frac{L_z}{2}}, \tag{3.60}$$

$$A \sin\left(k_w \frac{L_z}{2}\right) = C e^{-k_b \frac{L_z}{2}}, \tag{3.61}$$

$$A k_w \cos\left(k_w \frac{L_z}{2}\right) = k_b B e^{-k_b \frac{L_z}{2}}, \tag{3.62}$$

$$A k_w \cos\left(k_w \frac{L_z}{2}\right) = -k_b C e^{-k_b \frac{L_z}{2}}. \tag{3.63}$$

From this, we can see that

$$B = -C, \tag{3.64}$$

which results in

$$-k_w \cot\left(k_w \frac{L_z}{2}\right) = k_b. \tag{3.65}$$

Following the same approach as before, we can obtain

$$-w \cot(w) = \sqrt{R^2 - w^2}, \tag{3.66}$$

which also can only be solved numerically.

If the barrier potentials are different on either side (asymmetric structure), then a similar approach to the one outlined above can still be used, although the algebra does become rather messy.

Nevertheless, these methods are rarely used in practical situations because of their limited utility and applicability. For example, when quantum wells are doped with impurity atoms, the rectangular potential shape will become altered by the distribution of the positive ions and the negative electrons in a manner that is not possible to write with an analytical expression. Likewise, an external applied electric field will also modify the potential profile. Such situations cannot be easily handled by the analytical methods previously outlined. A fully numerical approach is often required. Fortunately, numerical approaches can actually be simpler to perform than the algebraic approaches, at least for the one-dimensional cases. This is the approach we will take in the remainder of this chapter.

A reader who is familiar with optical waveguides may have already noticed the similarity between quantum wells and slab waveguides. The approach to the calculations are nearly identical in both cases except for a few minor differences related to the

polarization of light and the effective mass of the electron. Hence, it is useful for a student who is learning these subject areas for the first time to develop methods and computer codes that will be equally applicable in both situations. The exploration of the similarities between Maxwell's wave equations and Schrodinger's equations is left as an exercise for the reader.

3.2 Computational methods

3.2.1 Numerical shooting method

In this section, we will derive the numerical shooting method, apply it to a few simple cases, and compare them with the analytical results for confirmation. Compared with other numerical techniques, this is an extremely trivial method to understand and implement. The limitation, at least in its current form, is that it can only be applied to one-dimensional problems. However, this is not a serious limitation because the vast majority of quantum engineered devices are planar multilayer structures for which a one-dimensional model is ideally suited.

We start with the one-dimensional decoupled equation (assuming separation of variables):

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(z)}{\partial z^2} + V(z)\psi(z) = \varepsilon_z \psi(z). \tag{3.67}$$

All of the spatial functions will be discretized on a finite-difference grid as

$$\psi(z) = \psi(n\Delta z), \tag{3.68}$$

and

$$V(z) = V(n\Delta z), \tag{3.69}$$

where n is an integer and Δz is the grid size.

The second-order derivative at the grid point n becomes

$$\left.\frac{\partial^2 \psi(z)}{\partial z^2}\right|_n = \frac{\left(\left.\frac{\partial \psi(z)}{\partial z}\right|_{n-\frac{1}{2}} - \left.\frac{\partial \psi(z)}{\partial z}\right|_{n+\frac{1}{2}}\right)}{\Delta z} \tag{3.70}$$

$$= \frac{\left(\frac{\psi((n+1)\Delta z)-\psi(n\Delta z)}{\Delta z} - \frac{\psi(n\Delta z)-\psi((n-1)\Delta z)}{\Delta z}\right)}{\Delta z} \tag{3.71}$$

$$= \frac{\psi((n+1)\Delta z) - 2\psi(n\Delta z) + \psi((n-1)\Delta z)}{\Delta z^2}. \tag{3.72}$$

This allows us to express the one-dimensional Schrodinger's equation in finite-difference notation as

$$-\frac{\hbar^2}{2m}\left[\frac{\psi((n+1)\Delta z) - 2\psi(n\Delta z) + \psi((n-1)\Delta z)}{\Delta z^2}\right] + V(n\Delta z)\psi(n\Delta z)$$
$$= \varepsilon_z\psi(n\Delta z). \tag{3.73}$$

If we know the first two values of the wave function $\psi((n-1)\Delta z)$ and $\psi(n\Delta z)$ as well as the energy ε_z, then we can express the subsequent value $\psi((n+1)\Delta z)$ by rearranging Eqn (3.73) as

$$\psi((n+1)\Delta z) = \left[\frac{2m}{\hbar^2}(V(n\Delta z) - \varepsilon_z)\Delta z^2 + 2\right]\psi(n\Delta z) - \psi((n-1)\Delta z). \tag{3.74}$$

This is the basic equation for the shooting method. First, the numerical procedure for solving this starts by assuming a value for the confinement energy ε_z. This should be somewhat close to the value we are ultimately seeking, so it is predicated on the user having some idea where the solution lies. Second, because we are only seeking solutions for bounded electrons (trapped in the well), we can assume that the wave function falls to zero at large distances outside of the well:

$$\psi(\pm\infty) \rightarrow 0. \tag{3.75}$$

If we set the computational axis such that $m = 1$ is at a large distance left of the quantum well, then this can be stated as

$$\psi(1\Delta z) = 0, \tag{3.76}$$

$$\psi(2\Delta z) = \delta. \tag{3.77}$$

The value of δ at $n = 2$ is entirely arbitrary—it basically states that the wave function has a nonzero gradient a large distance from the quantum well. The exact value of δ only affects the normalization of the wave function, which we will calculate afterward. Because δ is arbitrary, we can go one step further and make $\delta = 1$. Putting all of these together, the first iteration of the numerical process will be

$$\psi(3\Delta z) = \left[\frac{2m}{\hbar^2}(V(2\Delta z) - \varepsilon_z)\Delta z^2 + 2\right]1 - 0. \tag{3.78}$$

The next iteration will be to use these values to calculate $\psi(4\Delta z)$, $\psi(5\Delta z)$ etc. such as

$$\psi(4\Delta z) = \left[\frac{2m}{\hbar^2}(V(3\Delta z) - \varepsilon_z)\Delta z^2 + 2\right]\psi(3\Delta z) - 1, \tag{3.79}$$

$$\psi(5\Delta z) = \left[\frac{2m}{\hbar^2}(V(4\Delta z) - \varepsilon_z)\Delta z^2 + 2 \right]\psi(4\Delta z) - \psi(3\Delta z). \tag{3.80}$$

Of course, this iteration is based on our initial guess for the energy ε_z; therefore, we need to check the validity of the solution and keep iterating until proper convergence is achieved. We perform this check by examining the last point $\psi(N\Delta z)$, where $n = N$ is the last grid point at a large distance to the right of the quantum well. If ε_z is close to the exact solution, then this value of $\psi(N\Delta z)$ should be very close to zero, $\psi(N\Delta z) \to 0$. Several different techniques can be used to solve for $\psi(N\Delta z) = 0$ by driving ε_z toward the exact solution, such as the bisection method and the Newton−Raphson method. Bisection is reliable, but it is slower; Newton−Raphson is faster, but there are some cases in which it may fail because of division by zero. With that precautionary note, we will proceed to use the Newton−Raphson method in this case. We will represent $\varepsilon_z^{(1)}$ to represent the first iteration, $\varepsilon_z^{(2)}$ for the second iteration, etc. and the wave functions as $\psi(N\Delta z)^{\left(\varepsilon_z^{(1)}\right)}$, as $\psi(N\Delta z)^{\left(\varepsilon_z^{(2)}\right)}$, etc.

We can then show that

$$\varepsilon_z^{(q+1)} = \varepsilon_z^{(q)} - \frac{\psi(N\Delta z)^{\left(\varepsilon_z^{(q)}\right)}}{\psi'(N\Delta z)^{\left(\varepsilon_z^{(q)}\right)}} \tag{3.81}$$

where $\psi'(N\Delta z)^{\left(\varepsilon_z^{(q)}\right)}$ is the derivative calculated from the values of $\psi(N\Delta z)$ separated by a small energy value $\Delta\varepsilon_z$ around $\varepsilon_z^{(q)}$:

$$\psi'(N\Delta z)^{\left(\varepsilon_z^q\right)} = \frac{\psi(N\Delta z)^{\left(\varepsilon_z^{(q)} + \Delta\varepsilon_z\right)} - \psi(N\Delta z)^{\left(\varepsilon_z^{(q)}\right)}}{\Delta\varepsilon_z}. \tag{3.82}$$

This iteration process is repeatedly performed and is terminated when the improvement in ε_z falls below a certain threshold ε, such as

$$\left| \frac{\varepsilon_z^{(q+1)} - \varepsilon_z^{(q)}}{\varepsilon_z^{(q)}} \right| < \varepsilon. \tag{3.83}$$

Further details of the numerical shooting method for quantum confinement can be found in the textbook by Harrison [2].

3.2.1.1 Numerical example—finite potential well

In this example, consider a symmetric finite quantum well with $L_z = 50$ Å and with barrier potentials of 50 meV. Using $\Delta z = 0.025$ Å and $N = 10,000$, with a termination condition $\varepsilon = 10^{-9}$ and an initial estimate of 1 meV, we can quickly converge on a solution of $\varepsilon_z = 8.09$ meV. This is the ground state of this quantum well because it is the lowest allowed energy state. This is shown in Figure 3.5(a).

(a)

(b)
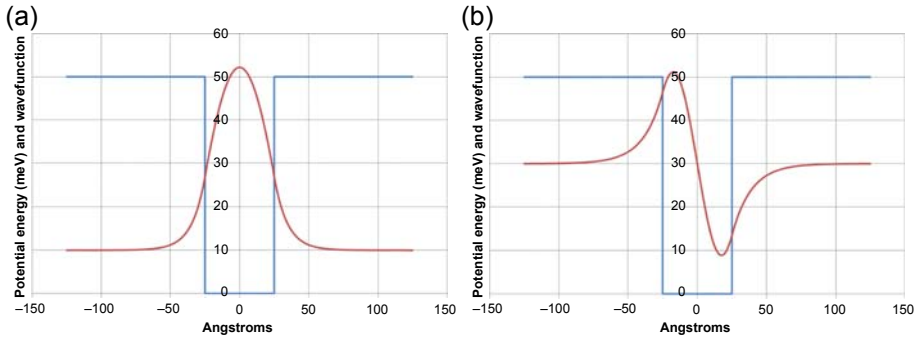
Figure 3.5 (a) First state solution of the 50-meV, 50-Å quantum well: $\varepsilon_z = 8.09$ meV. (b) Second state solution of the 50-meV, 50-Å quantum well: $\varepsilon_z = 30.47$ meV.

We can compare this with the semianalytical equations that we derived earlier. The result is $\varepsilon_z = 8.14$ meV, which is reasonably close to the numerical solution. The agreement can be improved further by refining the mesh size and the termination criteria for the numerical procedure.

We can also find the higher order solutions by using a larger initial estimate for ε_z. Starting from an estimate of 15 meV, we can converge on a solution of $\varepsilon_z = 30.47$ meV, whereas the analytical solution gives 30.65 meV. This is shown in Figure 3.5(b).

However, the numerical process cannot distinguish between symmetric and antisymmetric solutions or the order of the solutions. If we are seeking a specific solution, then this can only be determined by examining the wave function. In a large problem, this can become cumbersome, but the process can be automated to examine the wave function for certain symmetries and the number of peaks and valleys. This is left as an exercise for the interested reader to pursue.

3.2.1.2   Numerical example—two coupled wells

In this example, consider two 50 Å quantum wells with 50 meV barriers separated by 50 Å. Using $\Delta z = 0.03$ Å with $N = 10,000$, with a termination condition $\varepsilon = 10^{-9}$ and an initial estimate of 1 meV, we can converge on a solution of $\varepsilon_z = 8.078$ meV. The reader may notice that this solution is very close to the solution we obtained in the previous example for the single quantum well. In fact, we can find a second solution that has a value of 8.117 meV. This is the effect of energy splitting due to the interaction between two quantum wells. The wave function of the first solution turns out to be a symmetric combination of the single quantum well solution, and the second solution is an antisymmetric combination, as shown in Figure 3.6(a) and (b). The energy of the symmetric combination is slightly lower than the single well, and the energy of the antisymmetric combination is slightly higher. These states are also sometimes referred to as the bonding and antibonding states in the context of the interaction between two neighboring atoms.

Coupled quantum wells are extremely useful structures in quantum engineering. They are used in devices such as quantum cascade lasers.

(a)

(b)
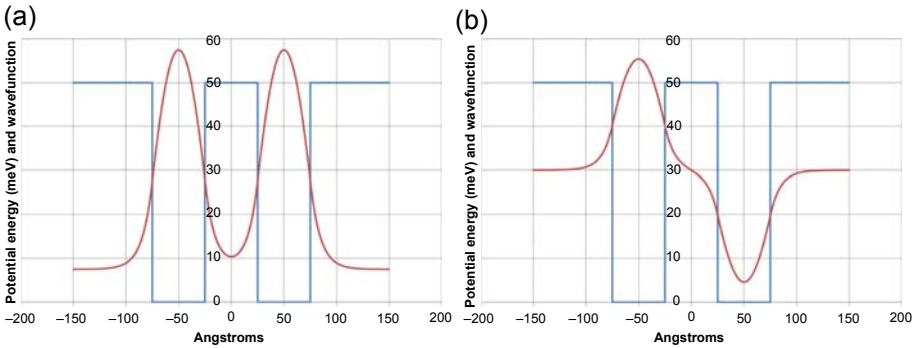
Figure 3.6 (a) First state solution of the two 50-meV, 50-Å coupled wells: $\varepsilon_z = 8.078$ meV. (b) Second state solution of the two 50-meV, 50-Å coupled wells: $\varepsilon_z = 8.117$ meV.

3.2.1.3 Numerical example—coupled wells with an applied electric field

In this example, consider the same two coupled quantum wells but with an applied static electric field across the quantum well structure. The field modifies the potential profile as follows:

$$V(z) = V_{QW}(z) + qF \tag{3.84}$$

where $V_{QW}(z)$ is the potential profile without the electric field and F is the field intensity.

As expected, the degeneracy of the coupled states breaks when an asymmetric potential is superimposed by the applied electric field. The electrons migrate into the quantum well with the lower potential instead of being equally split between the two wells. As the field intensity is increased, almost all of the electrons move into the lower potential well. These situations are shown in Figure 3.7(a) and (b).

(a)

(b)
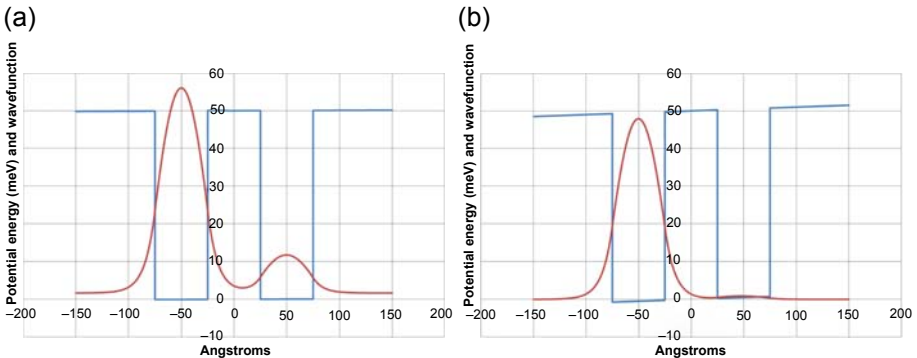
Figure 3.7 (a) First state solution of the two 50-meV, 50-Å coupled wells with an applied field of $F = 0.1 \frac{kV}{cm}$: $\varepsilon_z = 8.044$ meV. (b) First state solution of the same two wells with a larger applied field of $F = 1.0 \frac{kV}{cm}$: $\varepsilon_z = 7.596$ meV.

3.2.1.4 Numerical example—10-coupled wells

In this example, consider ten 50 Å quantum wells with 50 Å wide 50 meV barriers, with $\Delta z = 0.12$ Å and $N = 10{,}000$. We can find 10 different solutions that correspond to the first state of the single quantum well. The first four of these are shown in Figure 3.8.

We can see that the energies of all of these states are extremely close together, which is what we would expect from our earlier example of two coupled quantum wells. This example can be extended to a large number of coupled quantum wells, which then becomes known as the superlattice. If we had 50 coupled quantum wells, the ground-state energy will split 50 ways, which can then be considered as a band of energies. This is how artificial band structures are synthesized in quantum engineered structures. The larger the number of quantum wells in a coupled system, the smaller the energy spacing within the energy band.

3.2.2 Additional notes on the numerical shooting method

The numerical shooting method has some quirks that the user must be aware of to avoid getting caught with unexpected results.
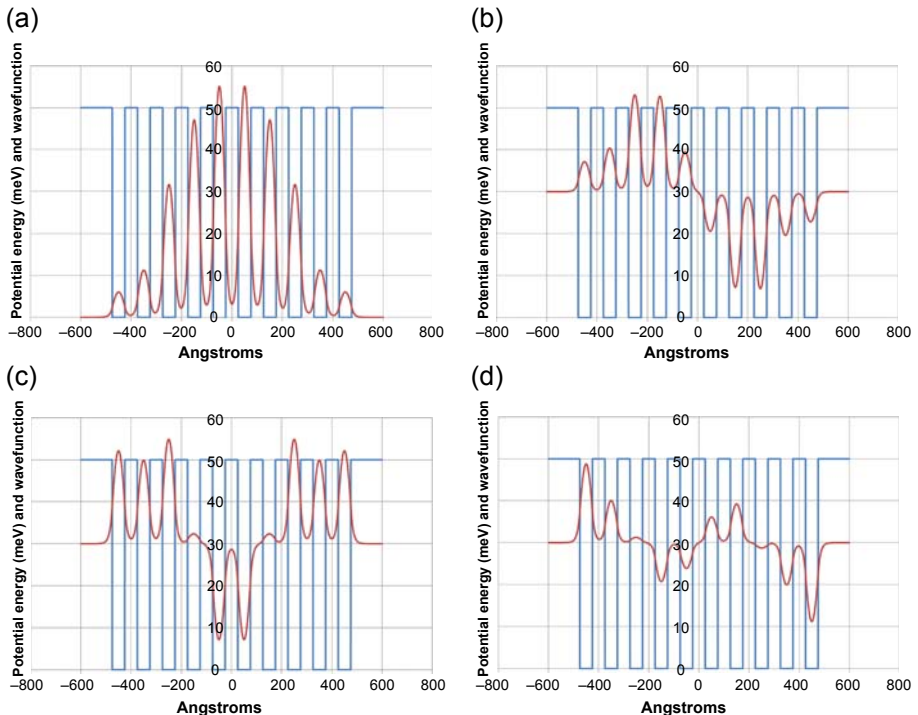
Figure 3.8 (a) First solution of the 10-coupled wells: $\varepsilon_z = 8.047$ meV. (b) Second solution of the 10-coupled wells: $\varepsilon_z = 8.055$ meV. (c) Third solution of the 10-coupled wells: $\varepsilon_z = 8.066$ meV. (d) Fourth solution of the 10-coupled wells: $\varepsilon_z = 8.073$ meV.

(a)

(b)
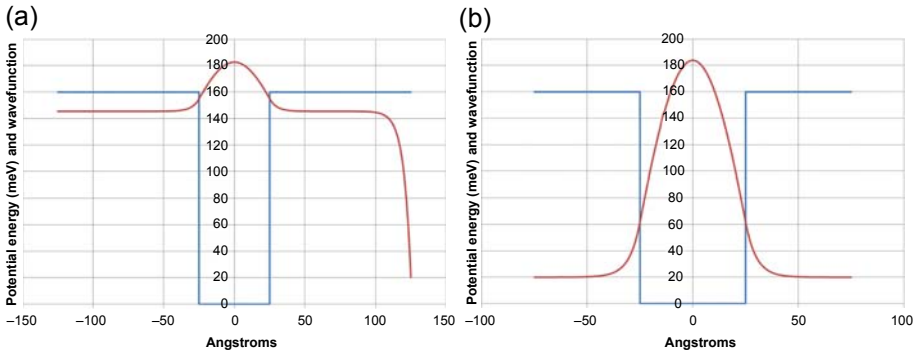
Figure 3.9 (a) Illustration of a typical convergence problem on a 50-Å quantum well, 160 meV barriers, $\Delta z = 0.025$ Å, $N = 10,000$ (window size is 250 Å): $\varepsilon_z = 10.41$ meV. (b) The same example with $\Delta z = 0.015$ Å, $N = 10,000$ (window size is 150 Å): $\varepsilon_z = 10.42$ meV.

The distance of the first point ($n = 1$) and the last point ($n = N$) from the quantum well cannot be too large. Because the functions are exponentially decaying outside of the quantum well, if these distances are made too large, then the numerical values can get extremely small beyond the resolution of the computer or exceedingly large. For example, in a strongly confined quantum well, it is not uncommon for the shooting method to produce a wave function that initially settles down but then grows to an indefinitely large positive or negative value. This growth is induced by numerical discretization errors and can significantly affect the results of normalization and subsequent calculations. It is important to watch for this occurrence and correct it by moving the computational boundaries closer to the quantum well. The plots in Figure 3.9 illustrate this scenario.

3.3   Quantum tunneling across barriers

As a result of their wave nature, electrons can penetrate a potential barrier to a certain depth even when the total energy of the electron is smaller than the barrier height energy. This can produce the unusual effect of electrons traversing through a barrier rather than over it. The result is often expressed as a tunneling probability, from which we can calculate a tunneling current when multiplied by the rate of incident electrons at the barrier.

3.3.1   Tunneling across a single barrier

Consider a rectangular one-dimensional potential barrier, as shown in Figure 3.10, with electrons incident from the left side. Similar to before, we will assume the separation of variables still apply and only consider the z axis for the calculations.

Because the incident electrons are traveling waves along the z axis, they can be written as

$$\psi_i(z) = A_{zi}e^{+ik_z z}, \tag{3.85}$$

Barrier

Backscattered electrons

Incident electrons   Tunneling electrons   Transmitted electrons

$z = 0$ $z = w$
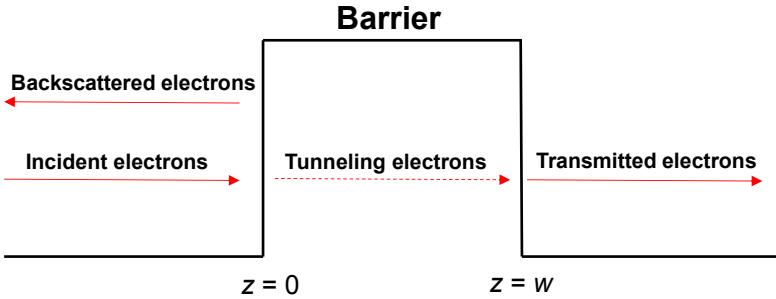
Figure 3.10 Single-barrier tunneling.

where we have used the positive sign to denote forward-traveling waves, and

$$k_z = \frac{\sqrt{2m\varepsilon_z}}{\hbar}. \tag{3.86}$$

Likewise, the backscattered (reflected) electrons are

$$\psi_r(z) = A_{zr}e^{-ik_z z}, \tag{3.87}$$

and the transmitted electrons on the other side of the barrier become

$$\psi_t(z) = A_{zt}e^{+ik_z z}. \tag{3.88}$$

The tunneling electrons inside of the barrier have an energy lower than the potential; therefore, they will exhibit an exponential behavior. Because the barrier has a finite dimension, both exponential factors need to be included:

$$\psi_b(z) = A_{zb1}e^{-k_b z} + A_{zb2}e^{+k_b z}, \tag{3.89}$$

where

$$k_b = \frac{\sqrt{2m(V_b - \varepsilon_z)}}{\hbar}. \tag{3.90}$$

Applying the boundary conditions at the left interface, we can obtain

$$A_{zi} + A_{zr} = A_{zb1} + A_{zb2}, \tag{3.91}$$

$$ik_z A_{zi} - ik_z A_{zr} = -A_{zb1}k_b + A_{zb2}k_b, \tag{3.92}$$

and at the right interface we obtain

$$A_{zb1}e^{-k_b w} + A_{zb2}e^{+k_b w} = A_{zt}e^{+ik_z w}, \tag{3.93}$$

$$-k_b A_{zb1}e^{-k_b w} + k_b A_{zb2}e^{+k_b w} = ik_z A_{zt}e^{+ik_z w}. \tag{3.94}$$

The amplitude of the incident electron can be arbitrarily set to $A_{zi} = 1$, which allows us to interpret A_{zr} as the reflection amplitude and A_{zt} as the transmission amplitude.

The above four equations can be solved for the four unknowns. Despite the straight-forward nature of the problem, the algebra for this solution can run into several pages (which is left as an exercise for the reader). The solutions are

$$A_{zr} = \frac{\left(k_z^2 + k_b^2\right)\left(1 - e^{-2k_b w}\right)}{(k_z + ik_b)^2 - (k_z - ik_b)^2 e^{-2k_b w}}, \tag{3.95}$$

$$A_{zb1} = \frac{2k_z(k_z + ik_b)}{(k_z + ik_b)^2 - (k_z - ik_b)^2 e^{-2k_b w}}, \tag{3.96}$$

$$A_{zb2} = \frac{-2k_z(k_z - ik_b)e^{-2k_b w}}{(k_z + ik_b)^2 - (k_z - ik_b)^2 e^{-2k_b w}}, \tag{3.97}$$

$$A_{zt} = \frac{i4k_z k_b e^{-(ik_z + k_b)w}}{\left(k_z^2 - k_b^2\right)\left(1 - e^{-2k_b w}\right) + 2ik_b k_z\left(1 + e^{-2k_b w}\right)}. \tag{3.98}$$

The tunneling and reflection probabilities can be calculated by taking the magnitude of the complex amplitudes and further simplified to produce the following two expressions:

$$T = |A_{zt}|^2 = \frac{1}{\frac{1}{4}\left(\frac{k_z^2 + k_b^2}{k_z k_b}\right)^2 \left(\frac{e^{+k_b w} - e^{-k_b w}}{2}\right)^2 + 1}, \tag{3.99}$$

$$R = |A_{zr}|^2 = \frac{1}{4\left(\frac{k_z k_b}{k_z^2 + k_b^2}\right)^2 \left(\frac{2}{e^{+k_b w} - e^{-k_b w}}\right)^2 + 1}. \tag{3.100}$$

We can also verify that $T + R = 1$, which is required for the conservation of charge and mass.

If the incident current is I_i, the tunneling and reflected currents become

$$I_t = I_i T, \tag{3.101}$$

$$I_r = I_i R. \tag{3.102}$$

The above derivation only considers the case of $\varepsilon_z < V_b$, which is what we generally consider as tunneling. However, even when $\varepsilon_z > V_b$, the approach is still valid as long as we replace the real term k_b with the imaginary terms $\pm i k_b$. This case would represent a reflection off a potential step rather than tunneling through a barrier, much like ordinary light reflecting off a refractive index step.

3.3.1.1 Numerical example—tunneling across a single barrier

Consider a barrier height of 100 meV and a barrier width of 50 Å. We can calculate the tunneling probability of an electron through this barrier as a function of its energy using the above expression. The resulting plot is shown in Figure 3.11.

We can see that the tunneling probability is a very small number when the electron's energy is significantly smaller than the barrier height. As the energy approaches the barrier height, the tunneling exponentially increases. It does not level off at a value of 1.0 as the energy approaches the barrier height as one might expect. This is because even when the electron energy exceeds the barrier height, it still experiences reflection off of the barrier. It tends toward 1.0 only when the electron energy is significantly greater than the barrier height.
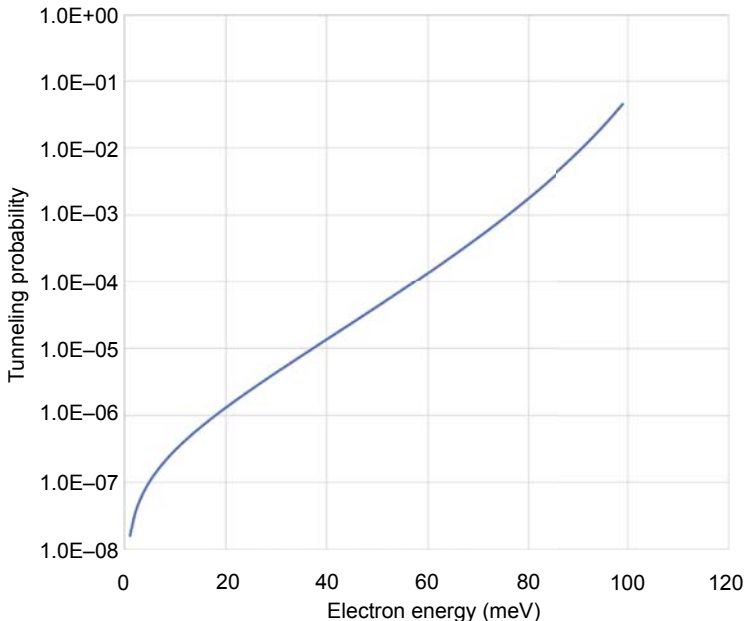
Figure 3.11 Tunneling probability for 100-meV, 50-Å barrier calculated from the analytical expression in Eqn. 3.99.

3.3.2   Numerical shooting method for tunneling problems

Instead of the analytical expressions, which apply only to rectangular-shaped barriers, the tunneling probability can also be easily calculated using the same foundations laid out in the earlier section on the numerical shooting method. However, in the earlier case, we had assumed a confining potential, but the tunneling potential is an inverse of the quantum confining potential. Whereas the wave function in a confining potential decays exponentially to zero outside of the wells, in a tunneling barrier the wave functions will contain traveling waves on either side of the barrier. To account for these differences, we will modify the shooting method as follows.

Although it may seem odd at first, it is easier if we start the numerical shooting method from the transmission side and proceed towards the incident side. The reason for this is because the initial condition is easier to define on the transmission side as it contains just one traveling wave $A_{zt}e^{+ik_z z}$, whereas the incident side has a combination of two waves. This is shown in Figure 3.12. Furthermore, for the ease of calculations, we will assume that the amplitude of the electron on the transmitted side is $A_{zt} = 1$. This is an arbitrary choice of course, which can be easily scaled once we find the amplitudes of the waves on the other side. Therefore, on the transmission side, we will have

$$\psi_t(z) = A_{zt}e^{+ik_z z}, \tag{3.103}$$

and on the incident side we will have

$$\psi_i(z) = A_{zi}e^{+ik_z z}, \tag{3.104}$$

$$\psi_r(z) = A_{zr}e^{-ik_z z}. \tag{3.105}$$

Therefore, the initial conditions will be

$$\psi(\Delta z) = 1, \tag{3.106}$$

$$\psi(2\Delta z) = e^{-ik_z \Delta z}. \tag{3.107}$$
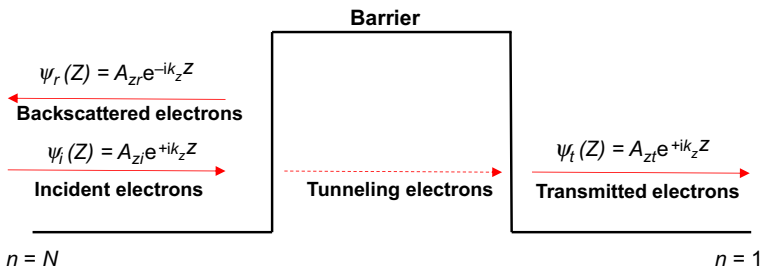
Figure 3.12 Numerical shooting method for the tunneling problem.

This is all we need to proceed with the calculation of $\psi(3\Delta z), \psi(4\Delta z), \ldots$ up to the last point $\psi(N\Delta z)$, which is on the left side, using the shooting algorithm defined in Eqn (3.74). However, unlike the quantum confinement calculations, we do not have to iterate the calculation until convergence is achieved—it is a one-time sweep from $n = 1$ to $n = N$. Therefore, it is much easier than the quantum well calculations.

The next step is to calculate the amplitudes A_{zi} and A_{zr}. Because the incident side contains a mixture of two waves, a few extra steps are required to separate these wave functions and obtain their amplitudes. If the numerical solutions at the last two points are $\psi(N\Delta z)$ and $\psi((N-1)\Delta z)$, then these can be expressed as

$$\psi(N\Delta z) = A_{zr}e^{-ik_z N\Delta z} + A_{zi}e^{+ik_z N\Delta z}, \tag{3.108}$$

$$\psi((N-1)\Delta z) = A_{zr}e^{-ik_z (N-1)\Delta z} + A_{zi}e^{+ik_z (N-1)\Delta z}. \tag{3.109}$$

The value of k_z is known because we know the potential function, which we assume to be flat at the edge of the computation window. That is,

$$V(N\Delta z) = V((N-1)\Delta z). \tag{3.110}$$

This results in

$$k_z = \frac{\sqrt{2m(\varepsilon_z - V(N\Delta z))}}{\hbar}. \tag{3.111}$$

Therefore, Eqns (3.108) and (3.109) contain only two unknowns, A_{zr} and A_{zi}, which can be calculated by simple substitution. This results in

$$A_{zr} = \frac{\psi((N-1)\Delta z)e^{+ik_z \Delta z} - \psi(N\Delta z)}{e^{-ik_z N\Delta z}(e^{+2ik_z \Delta z} - 1)}, \tag{3.112}$$

and

$$A_{zi} = e^{-ik_z N\Delta z}\left(\psi(N\Delta z) - A_{zr}e^{-ik_z N\Delta z}\right). \tag{3.113}$$

From this, we can obtain the tunneling probability

$$T = \left|\frac{A_{zt}}{A_{zi}}\right|^2 = \left|\frac{1}{A_{zi}}\right|^2. \tag{3.114}$$

We can also obtain the reflection probability

$$R = \left|\frac{A_{zr}}{A_{zi}}\right|^2. \tag{3.115}$$

3.3.2.1 Numerical example—tunneling across a single barrier with shooting method

We will consider the same earlier example with a barrier height of 100 meV and a barrier width of 50 Å. We can sweep the energy ε_z, plot the tunneling probability as a function of ε_z, and compare against the analytical results for confirmation. For this, we have used $\Delta z = 0.015$ Å with $N = 10{,}000$. As we can see in Figure 3.13, both curves are identical.

Shown next in Figure 3.14 is the plot of the wave functions at $\varepsilon_z = 50$ meV. Remembering that the input wave is on the right side of the barrier, we can see the exponential decay of the wave function inside of the barrier as well as the result of counterpropagating incident and reflected waves resulting in an oscillation pattern.

3.3.2.2 Numerical example—resonant tunneling across a double barrier

Compared with a single barrier, a structure that contains multiple barriers exhibits a unique feature known as resonant tunneling. In this example, consider two 50-Å wide 100-meV barriers separated by 50 Å. Using the numerical shooting method, we can calculate the tunneling probability as a function of energy as shown in Figure 3.15.
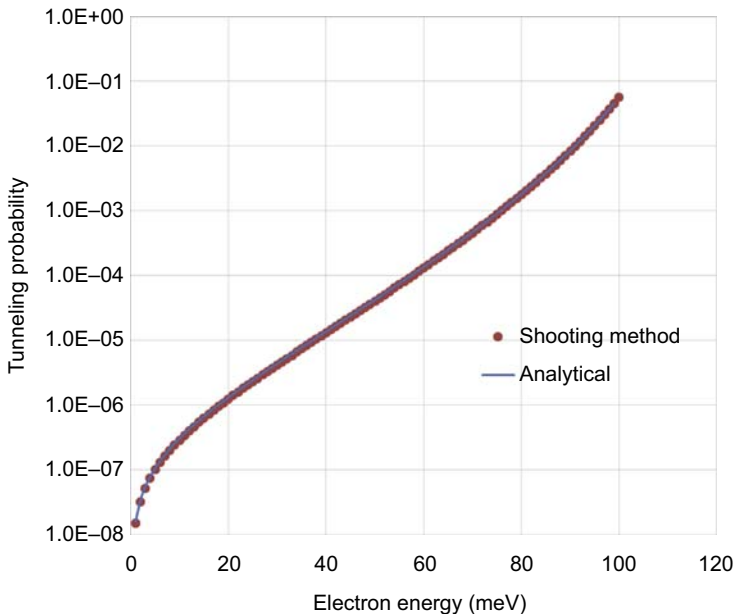
Figure 3.13 Comparison of the tunneling results for the 100-meV, 50-Å barrier from the analytical expression and the numerical shooting method.
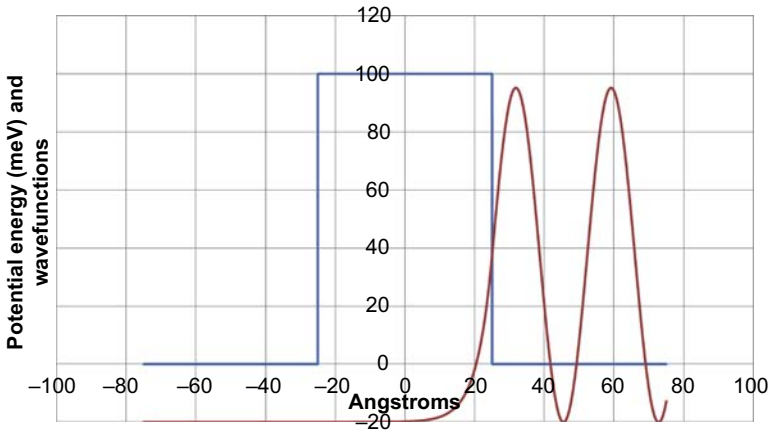
Figure 3.14 Wave functions for an incident energy of 50 meV.
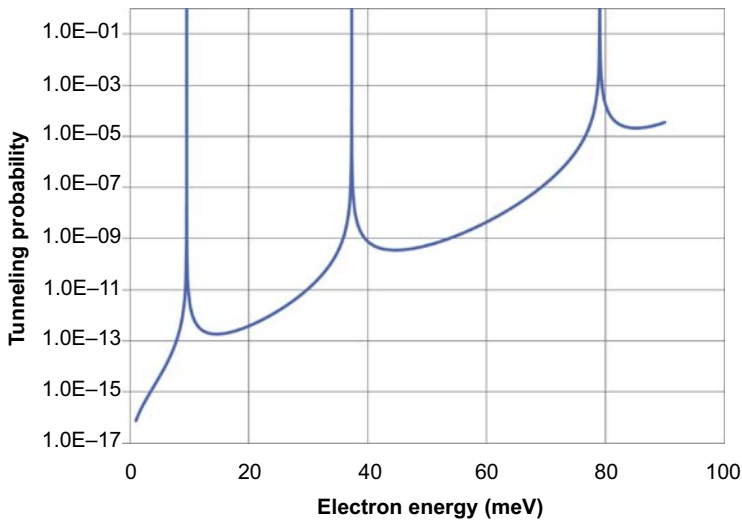
Figure 3.15 Resonant tunneling across two 50-Å wide 100-meV barriers.

The most important features are the sharp peaks that increase to 1.0, the energies of which are all substantially below the barrier energy. These energies actually correspond to the resonant states of the electrons because of the confinement between the two barriers, very similar to that of quantum wells. When the incident energy is exactly equal to the resonant energy, electrons are able to penetrate the barrier with very large probabilities. This is a very useful feature in quantum engineered structures. It allows specific energies to pass through a barrier unimpeded.

It should be noted that the widths of these peaks are extremely narrow. This makes it easy to miss them in numerical calculations, especially if the step size (in energy) is

(a)

(b)
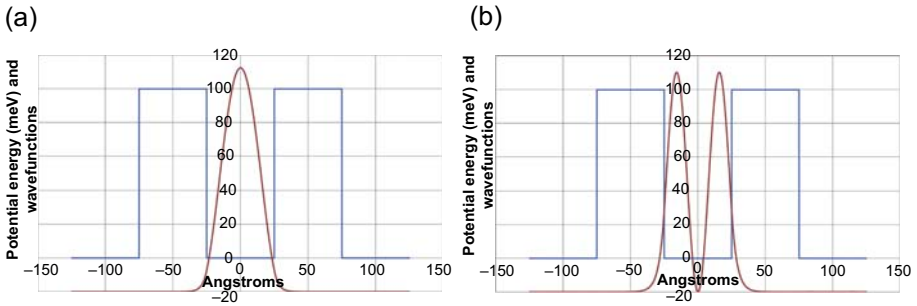
Figure 3.16 (a) Wave function in the double barrier structure at the first resonant energy of $\varepsilon_z = 9.54$ meV. (b) Wave function in the double-barrier structure at the second resonant energy of $\varepsilon_z = 37.30$ meV.

coarse. Ideally, an adaptive step size should be used so that the step size is fine in the vicinity of the peaks and coarse elsewhere.

For illustration, the plots in Figure 3.16(a) and (b) show the wave functions when the electron energy is equal to the first two resonant energies. We can see that the electrons are trapped between the barriers as in a quantum well. This results in a large accumulation of electrons in the cavity that produces a high transmission, very similar to the Fabry−Perot effect in optical cavities.

3.4   Quantum well structures in semiconductors

Thus far, we have defined the potential energy functions $V(z)$ without specifying how they are created in practice. By far the most commonly used method is by sandwiching semiconductor materials of different band gaps and utilizing the discontinuities in their conduction and valence bands to produce the step-like functions. GaAs and $Al_xGa_{1-x}As$ are the most widely used material systems primarily because $Al_xGa_{1-x}As$ is perfectly lattice matched to GaAs. Many other material systems such as InP, GaSb, InAs, and their alloys are also used, but lattice matching requirements often limit the range of compositions that can be used. The material aspects of this subject are beyond the scope of this chapter. The detailed calculation of electronic band structures of solids is also beyond our scope. Nevertheless, it is useful to understand some basic concepts and properties of common material systems used in quantum devices, which we will briefly explore next.

3.4.1   Origin of band structures

We can illustrate the origin of band structures by going back to Example 3.2.1.4, in which we had 10-coupled quantum wells. This resulted in 10 different energy levels all clustered around 8 meV, which also happens to be the confinement energy of a solitary quantum well of the same size. Furthermore, examining the wave functions,

we can see that they all contain a single peak inside of each quantum well with an over-all envelope that spans across the 10 quantum wells. We will designate the wave function inside of each well as the core function $\psi_c(z)$, and designate the envelope function that links all of these core functions together as $\psi_e(z)$. We can also see that the difference between one solution and another lies primarily in the envelope functions, whereas the core functions all remain pretty much the same. This can be mathematically approximated as

$$\psi(z) = \psi_e(z) \sum_q^Q \psi_c(z - qa) \qquad (3.116)$$

where a is the distance between each quantum well. Furthermore, the envelope function $\psi_e(z)$ is a purely sinusoidal function bounded by the 10 quantum wells. This allows it to be written as

$$\psi_e(z) = Ae^{+ik_e z} + Be^{-ik_e z}, \qquad (3.117)$$

where

$$k_e = \frac{n\pi}{Qa}, \qquad (3.118)$$

and $Q = 10$ in this case and $n = 1\ldots Q$. The first solution will have $k_e = \frac{\pi}{10a}$, the second solution will have $k_e = \frac{2\pi}{10a}$ etc. and the last solution will have $k_e = \frac{10\pi}{10a}$. If we repeat Example 3.2.1.4 with the next higher confined state for the core function, there will be 10 additional energy states clustered around 30 meV. These energy values as a function of k_e will produce a curve that looks like Figure 3.17.

   We can identify two lines of energies, each containing 10 states, with each energy clustered around the confinement energy of the single quantum well. Each line represents an energy band. The lower band spans 8.05–8.13 meV and the upper band spans 30.04–30.94 meV. If there are more states within the individual quantum wells, then more bands can be found. In this case, we only had 10 wells; therefore, the energies are still discrete. However, if the number of quantum wells is increased to a very large number, we can see that the energy states will start to form a continuous line rather than discrete points. This is the basis of how energy bands are formed in a solid, although this example only illustrates a one-dimensional case. In a three-dimensional crystal, many atoms couple together to form bands similar to the one-dimensional coupled quantum wells.

3.4.2 Effective mass

Calculations involving many coupled cells would be greatly simplified if we represent the electrons by $\psi_e(z)$ alone instead of as a product of the envelope and core functions $\psi_e(z)\sum_q^Q \psi_c(z - qa)$. The envelope function $\psi_e(z)$ is a simple plane wave; therefore, it
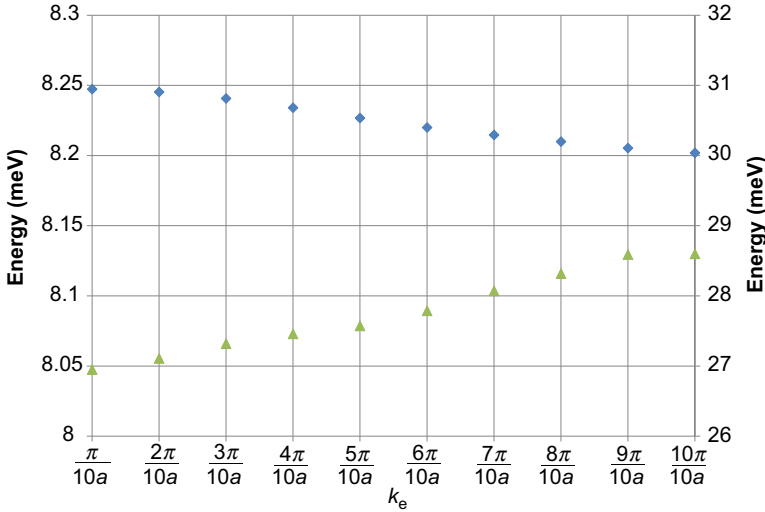
Figure 3.17 Calculated energy versus envelope k_e for the 10 quantum wells example.

is much easier to handle than the full wave function consisting of a plane wave multiplied by many confined state wave functions. However, a correction needs to be applied in the Schrodinger's equation to account for the omission of the core functions. This is where the concept of effective mass comes in. The mass of the electron is multiplied by a dimensionless parameter m^*(effective mass) so that the envelope functions yield the same energy values as the full wave functions. The modified Schrodinger's equation then becomes

$$-\frac{\hbar^2}{2m_o m^*}\frac{\partial^2 \psi_e(z)}{\partial z^2} + V(z)\psi_e(z) = \varepsilon_z \psi_e(z), \tag{3.119}$$

where m_o is the free electron mass. Furthermore, m^* is not a single constant, but a function of energy because the magnitude of the correction required at each energy will be different. In practice, m^* is computed and specified over a small energy range near the top and bottom of an energy band because electrons or holes accumulate primarily in these areas. Using $\psi_e(z) = Ae^{+ik_e z} + Be^{-ik_e z}$, we can show that

$$\frac{\hbar^2 k_e^2}{2m_o m^*} + V(z) = \varepsilon_z \tag{3.120}$$

and

$$m^* = \frac{\hbar^2}{m_o}\left(\frac{\partial^2 \varepsilon_z}{\partial k_e^2}\right)^{-1}, \tag{3.121}$$

where the term $\frac{\partial^2 \varepsilon_z}{\partial k_e^2}$ is computed from the previously calculated band structure of ε_z versus k_e.

Returning to the 10-coupled well example, we can calculate the effective masses to be -5.67 at the top of the lower band and 3.15 at the bottom of the upper band. Notice the negative effective mass for the lower band, which simply means that the curvature of the ε_z versus k_e is concaved. This is often found in the valence band of semiconductors. In addition, the effective mass can have a value smaller than 1.0, which is also frequently encountered in semiconductors.

3.4.3 GaAs/Al$_x$Ga$_{1-x}$As quantum wells

A simple quantum well can be constructed by sandwiching a thin layer of GaAs between two layers of Al$_x$Ga$_{1-x}$As, such as Al$_x$Ga$_{1-x}$As|GaAs|Al$_x$Ga$_{1-x}$As. The bandgap of GaAs is 1.42 eV and that of Al$_x$Ga$_{1-x}$As is empirically given as $1.42 + 1.247x$ for $x < 0.45$. More importantly, the relative offset in the conduction band minimum between GaAs and Al$_x$Ga$_{1-x}$As is $\Delta E_c = 0.836x$. The electron effective mass is also a function of the alloy composition, which is empirically determined to be $m^* = 0.063 + 0.083x$ for $x < 0.45$. Therefore, by changing the composition of the Al$_x$Ga$_{1-x}$As alloy, we can construct different quantum wells, multiquantum wells, and barrier structures. A GaAs/Al$_x$Ga$_{1-x}$As quantum well structure is illustrated in Figure 3.18. These are the type of structures one often encounters in devices such as quantum well infrared photodetectors (QWIPs) and quantum cascade lasers (QCLs).

Quantum wells can also be constructed using the valence band offsets. The valence band offset between GaAs and Al$_x$Ga$_{1-x}$As is $\Delta E_v = -0.412x$ and the effective mass is $m^* = -(0.51 + 0.25x)$. This actually results in an inverted quantum well, but because the masses are negative, it still results in a confinement inside of the well. This is illustrated in the next two examples.

3.4.3.1 Numerical example of a GaAs/Al$_x$Ga$_{1-x}$As single quantum well

Consider a three-layer quantum-well structure consisting of Al$_{0.2}$Ga$_{0.8}$As/GaAs/Al$_{0.2}$Ga$_{0.8}$As in which the thickness of the central GaAs layer is 100 Å. This results in a confining potential of 167.1 meV because of the conduction band offset. The electron effective mass in GaAs is 0.063 and in Al$_{0.2}$Ga$_{0.8}$As it is 0.0796. Using this, we
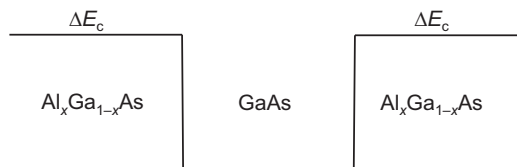
Figure 3.18 Illustration of a conduction band quantum well in the GaAs/Al$_x$Ga$_{1-x}$As material system.

(a)

(b)
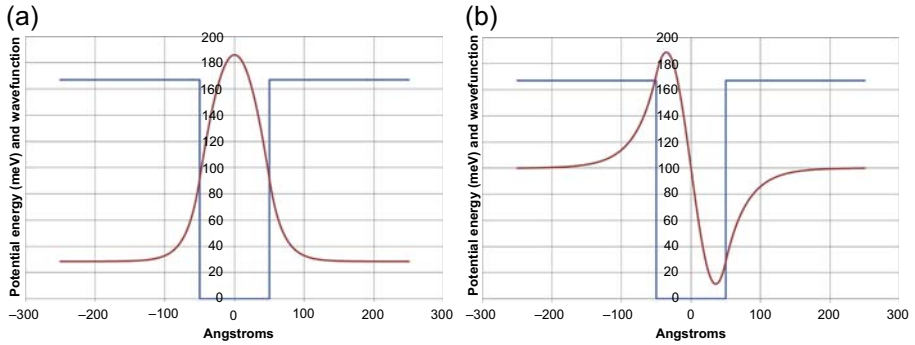
Figure 3.19 (a) First state solution of an $Al_{0.2}Ga_{0.8}As/GaAs/Al_{0.2}Ga_{0.8}As$ 100 Å conduction band quantum well: $\varepsilon_z = 32.28$ meV. (b) Second state solution of a $Al_{0.2}Ga_{0.8}As/GaAs/$ $Al_{0.2}Ga_{0.8}As$ 100 Å conduction band quantum well: $\varepsilon_z = 116.75$ meV.

(a)

(b)
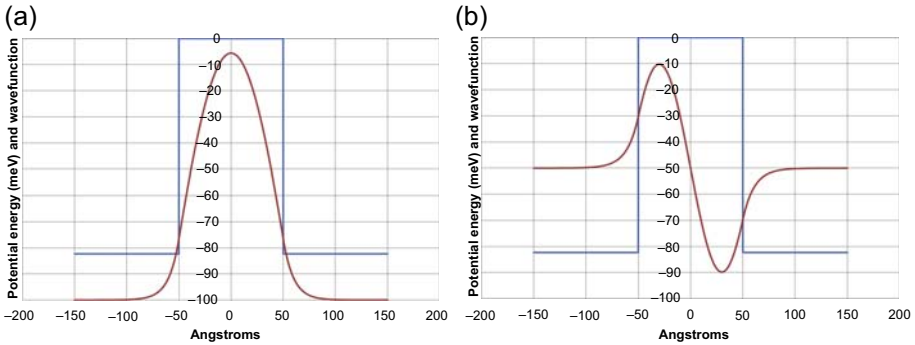
Figure 3.20 (a) First state solution of an $Al_{0.2}Ga_{0.8}As/GaAs/Al_{0.2}Ga_{0.8}As$ 100 Å valence band quantum well: $\varepsilon_z = -5.16$ meV. (b) Second state solution of a $Al_{0.2}Ga_{0.8}As/GaAs/Al_{0.2}Ga_{0.8}As$ 100 Å valence band quantum well: $\varepsilon_z = -20.40$ meV.

can calculate the first and second confinement energies to be 32.28 and 116.75 meV, respectively. The two wave functions are shown in Figure 3.19.

The same structure also exhibits an inverted confining potential in the valence band, with an offset of 82.4 meV. The effective masses are -0.51 and -0.56 in GaAs and $Al_{0.2}Ga_{0.8}As$, respectively. The calculated first two energies become -5.15 meV and -20.39 meV, respectively, using the valence band edge of the GaAs layer as the zero reference. These wave functions are shown in Figure 3.20.

3.5 Two- and three-dimensional quantum confined structures

One-dimensional quantum confined structures naturally arise from thin film technology and have been widely used to build devices. However, it is also possible to confine

electrons in more dimensions. A two-dimensional confinement would produce a quantum wire, and confinement in all three directions would produce a quantum dot or box. All of these structures result in different confinement states.

3.5.1 Quantum wire

When electrons are confined along two directions and unconfined in the third direction, the structure can be described as a quantum wire, or as a nanowire. This can be thought of as similar to an optical fiber in which light is trapped by two-dimensional confinement and propagates along the third direction. In Cartesian coordinates, the simplest structure would be a rectangular channel as shown in Figure 3.21. This can be solved by separating the two-dimensional (y, z) plane from the x axis as

$$V = V(y,z) + V(x),\tag{3.122}$$

$$\psi = \psi(y,z)\psi(x)\tag{3.123}$$

$$\varepsilon = \varepsilon_{yz} + \varepsilon_z.\tag{3.124}$$

The one-dimensional shooting method we developed earlier will not work for this solution because it requires a two-dimensional boundary value solution. Numerical methods for implementing two-dimensional solutions can be found elsewhere, but here we will focus on making certain approximations so that we can still use the one-dimensional approach. For this, we will attempt to write the potential function as a sum of three one-dimensional functions, such as

$$V = V(x) + V(y) + V(z).\tag{3.125}$$
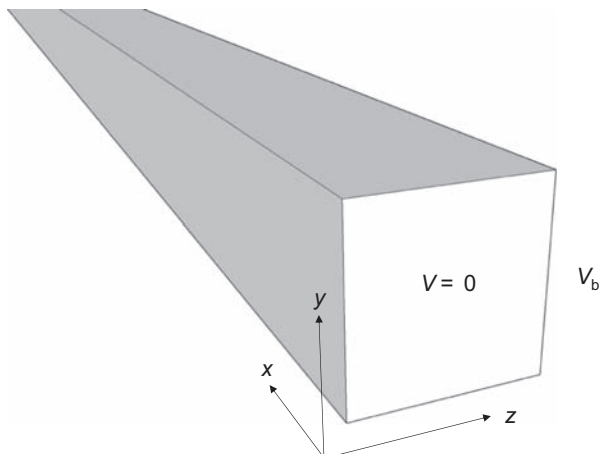
Figure 3.21 Illustration of a quantum wire.

If $V(z)$ and $V(y)$ are both quantum well functions with $V = 0$ inside of the well and $V = V_b$ outside of the well, then the overlap between the two functions can produce a two-dimensional rectangular quantum wire, but with one important difference—the corner regions will contain a potential of $2V_b$ instead of V_b. This is illustrated in Figure 3.22. However, depending on the specific problem, this may be an acceptable approximation; therefore, we will proceed with that assumption.

Following the same separation of variables approach as we did before, we can obtain

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(x)}{\partial x^2} = \varepsilon_x \psi(x), \tag{3.126}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(y)}{\partial y^2} + V(y)\psi(y) = \varepsilon_y \psi(y), \tag{3.127}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(z)}{\partial z^2} + V(z)\psi(z) = \varepsilon_z \psi(z). \tag{3.128}$$
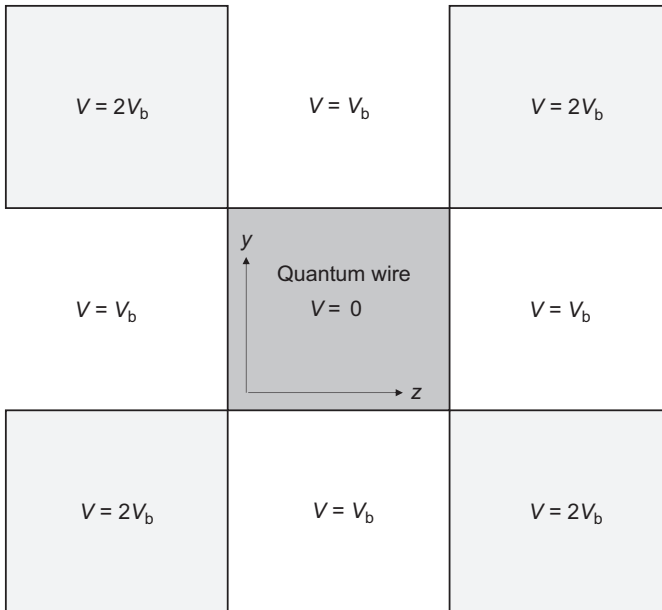
Figure 3.22 Confining potential distribution in a quantum wire under the separation of variables approximation.

These three equations are all one-dimensional and can be solved using the approaches we described earlier. The wave function and energy become

$$\psi = \psi(z)\psi(y)A_x e^{\pm ik_x x}, \tag{3.129}$$

$$\varepsilon = \varepsilon_x + \varepsilon_y + \varepsilon_z. \tag{3.130}$$

Until recently, quantum wires remained mostly a laboratory curiosity because there were no compelling device applications. They were grown by self-assembly with randomly nucleated droplets as catalysts. With the emergence of multigate transistors such as FinFETs and Gate-All-Around (GAA) field-effect transistors (FETs) in recent years, the interest in silicon-based quantum wires has gained significant interest. Although they are primarily being sought to improve the switching characteristics of transistors and not necessarily for their quantum confinement properties, it nevertheless becomes necessary to include the quantum properties in the modeling and simulation of these devices. This is an area that is likely to gain significant attention in the coming years.

3.5.2 Quantum box

A quantum box results when electrons are confined along all three directions, as illustrated in Figure 3.23. Just as with the quantum wire, we can use separation of variables to solve this problem. However, this results in a potential of $3V_b$ at all eight corners of the box and V_b along the sides. Assuming this to be an acceptable approximation of the structure, we can solve the three directions separately as we did before:

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(x)}{\partial x^2} + V(x)\psi(x) = \varepsilon_x \psi(x), \tag{3.131}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(y)}{\partial y^2} + V(y)\psi(y) = \varepsilon_y \psi(y), \tag{3.132}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi(z)}{\partial z^2} + V(z)\psi(z) = \varepsilon_z \psi(z). \tag{3.133}$$

The solution becomes

$$\psi = \psi(z)\psi(y)\psi(x), \tag{3.134}$$

$$\varepsilon = \varepsilon_x + \varepsilon_y + \varepsilon_z. \tag{3.135}$$

3.5.2.1 Numerical example of a quantum box

Consider a 100-Å cubic structure of GaAs surrounded by $Al_{0.2}Ga_{0.8}As$ on all sides. This produces a conduction band offset of 167.1 meV. The fundamental confinement
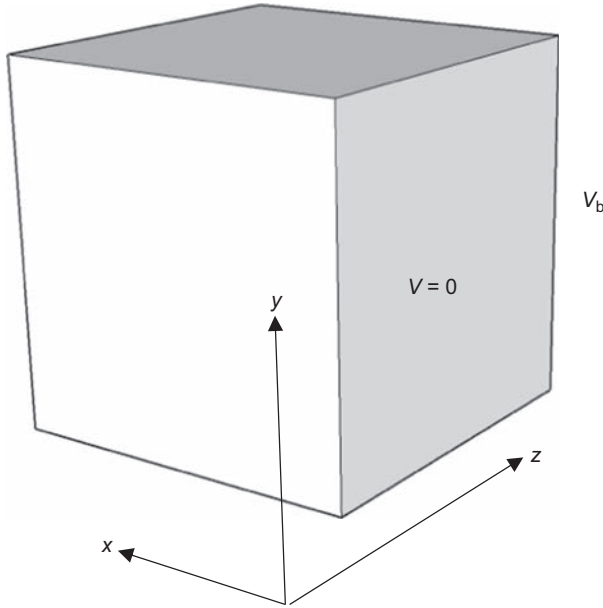
Figure 3.23 Illustration of a quantum box.

energy along any one of the axes can be calculated using the numerical shooting method. This results in a solution of 32.27 meV. The second confinement energy is 116.8 meV. If all three directions have the same fundamental solution, then we can designate this solution as (1, 1, 1) to indicate their quantum states and the energy will become 96.8 meV. If one direction is at the second confinement, then this can be represented as (2, 1, 1), (1, 2, 1), or (1, 1, 2) depending on which direction contains the second state. Because we considered a cubic structure, they will all result in a so-lution of 181.3 meV. If it seems curious that the confinement energy is larger than the barrier height of 167.1 meV, it should be remembered that the confinement energies are kinetic energies and have directions; therefore, along any one direction the energy is still smaller than the barrier height.

3.6 Quantum device structures

In this section, we will examine a few examples of quantum engineered devices that have been successfully developed for commercial applications. Although their basic concepts are based on the quantum confinement principles that we have discussed so far, electron scattering is another important concept that we have not discussed in this chapter. Scattering rates from photons, phonons, and Coulombic effects play a profound role in determining the electron population densities in various quantum confined states. Readers interested in these topics should refer the articles and books listed at the end of this chapter.

3.6.1 Quantum well lasers

The ability to create offsets in the conduction and valence bands was the key capability that ushered the invention of the room temperature semiconductor laser. A step discontinuity in the conduction band is designed such that it blocks the flow of electrons, and a similar discontinuity in the valence band blocks the flow of holes. This way the electrons and holes become trapped within the central region of the device region (known as the active region), making it possible to reach population inversion at low current injections. This is the double-heterostructure semiconductor laser, and it is still the most widely used device structure in semiconductor lasers (Figure 3.24).

The quantum well laser is an extension of the double heterostructure, in which thin quantum wells are grown inside of the active region. Because the width of the quantum well is very small compared with the optical wavelength, in most practical devices, several quantum wells are stacked to increase the extent of the overlap between the optical waveguide mode and the quantum wells. These are referred to as multi-quantum well (MQW) lasers. These quantum wells are usually not coupled together; they are simply a multiple number of single quantum wells. Additional band discontinuities are also used outside of the heterostructure to induce a refractive index step and create an optical waveguide. These are referred to as separate-confinement heterostructure multi-quantum well (SCH-MQW) lasers (Figure 3.25).

The quantum confinement in these lasers makes it possible to realize several key objectives: (1) the emission wavelength can be adjusted by changing the confinement energies; (2) the two-dimensional staircase-like density of states pins the gain peak at a constant energy value, resulting in a stable lasing wavelength; (3) differential gain becomes higher, which increases the high-speed modulation bandwidth of the laser.

3.6.1.1 Numerical example of InP/In$_{0.53}$Ga$_{0.47}$As/InP quantum well laser

Consider an MQW laser consisting of InP|In$_{0.53}$Ga$_{0.47}$As|InP with 50-Å wells. Unlike the GaAs/Al$_x$Ga$_{1-x}$As system, the only composition of In$_x$Ga$_{1-x}$As that is lattice
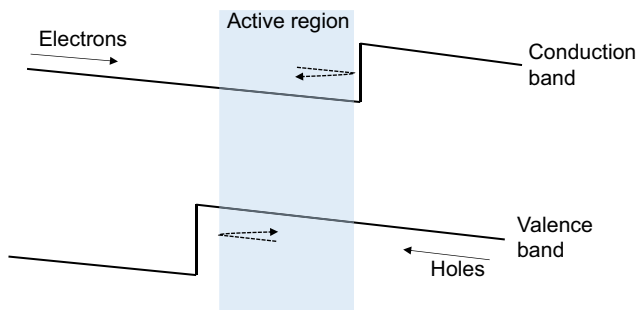
Figure 3.24 Band alignment in a double-heterostructure semiconductor laser.
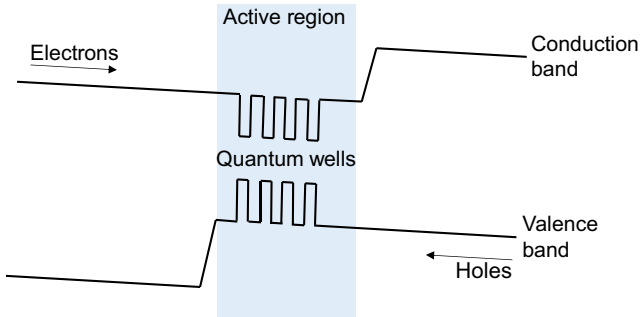
Figure 3.25 Band alignment in a quantum well laser enclosed within a double heterostructure.

matched to InP is $In_{0.53}Ga_{0.47}As$. The bandgap of InP is 1.35 eV, and the bandgap of $In_{0.53}Ga_{0.47}As$ is 0.75 eV. The conduction band and valence band offsets are 0.25 and 0.35 eV, respectively. The InP and $In_{0.53}Ga_{0.47}As$ electron effective masses in the conduction band are 0.08 and 0.041, respectively, and for the valence band they are -0.6 and -0.45. Using these, we can calculate the confinement energy of the conduction band quantum well to be 54.04 meV (measured from the bottom of the $In_{0.53}Ga_{0.47}As$ conduction band) and the valence band confinement to be -7.04 meV (measured from the top of the $In_{0.53}Ga_{0.47}As$ valence band). Therefore, the effective energy gap will be $(54.04 + 7.04) \times 10^{-3} = 61.1$ meV higher than the native bandgap of $In_{0.53}Ga_{0.47}As$. The emission wavelength from this quantum well will be 1529 nm. In comparison, the emission wavelength from bulk $In_{0.53}Ga_{0.47}As$ is 1653 nm. Therefore, the quantum confinement results in a wavelength shift of approximately -124 nm, which is a significant shift to be of practical value in communication applications. What is even more significant is that this wavelength shift is completely within our control, and it can be varied by adjusting the quantum well width (Figure 3.26).
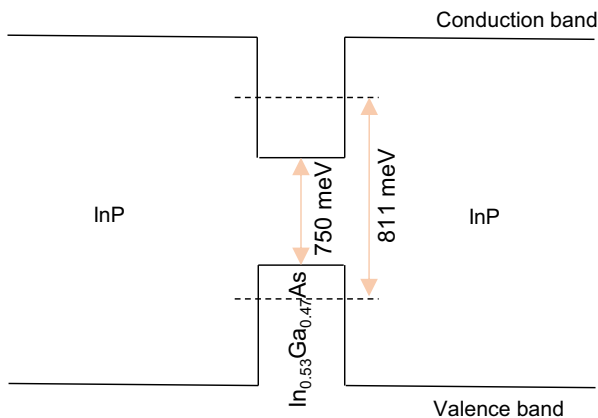
Figure 3.26 InP/$In_{0.53}Ga_{0.47}As$/InP quantum well laser example.

3.6.2   Quantum well infrared photodetectors

QWIPs are quantum-engineered devices for the absorption of long-wavelength infrared radiation whose energy is much smaller than the native material bandgap. As we saw in the previous example with the quantum well lasers, the optical emission (and absorption lines) of a quantum well can be engineered by adjusting the well width. This absorption can be from a confinement state in the valence band to a confinement state in the conduction band (interband transition) or between two adjacent confinement states within the same band (inter-subband transition). These are illustrated in Figure 3.27. Whereas the emission in quantum well lasers is due to interband transitions, the absorption process in QWIPs is due to inter-subband transitions within the conduction band. In addition, because thermal excitation can significantly interfere with the photogenerated current, these devices are generally operated at cryogenic temperatures. Because the transition process takes place entirely within one band, QWIPs are referred to as unipolar devices. They typically operate as photoconductive devices and do not produce energy from the incident radiation, although some level of photovoltaic activity can be realized by introducing a structural asymmetry in the QWIP structure.

A photodetector has to not only absorb a photon, it has to also release an electron into the external circuit in response to the photon. The quantum well structure is designed carefully to accomplish both of these objectives. There are several mechanisms that have been used:

- *Bound-to-bound transition*: The quantum well is designed such that the upper state has a larger tunneling probability through the barrier than the lower state. Upon absorption of a photon, the electrons are elevated to the upper state and then tunnel through the thin triangular barrier into the unconfined conduction band continuum, as illustrated in Figure 3.28. Some of them can also escape to the continuum through thermionic emission. However, the higher electron density in the lower state combined with the finite tunneling probability generally results in a large dark current in these devices.
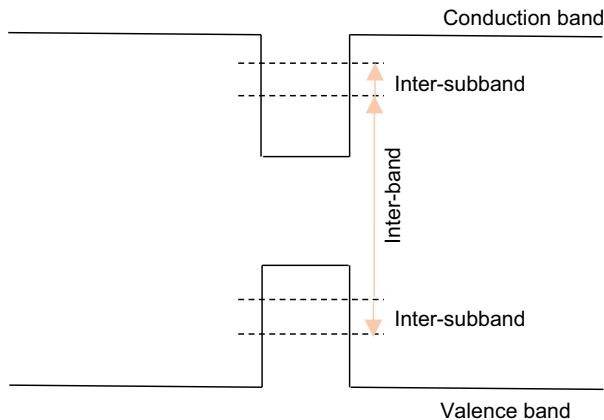
Figure 3.27 Interband and inter-subband transitions in a quantum well.
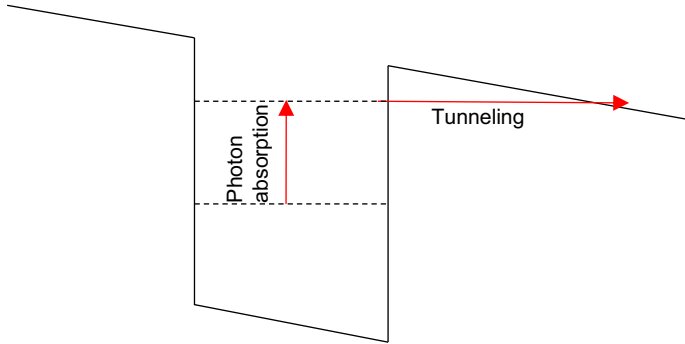
Figure 3.28 Bound-to-bound transition followed by tunneling in a QWIP.

- *Bound-to-continuum transition*: In this design, the quantum well is designed to support only one state. Upon absorption of a photon, the electrons are elevated out of the quantum well into the unconfined conduction band continuum to produce a current flow. This is illustrated in Figure 3.29. Because the carriers do not have to tunnel, the barriers in these devices can be made wide, which can lead to low dark currents. However, at higher operating temperatures, thermionic emission from the confined state to the conduction band continuum can still contribute to an increase in dark current. This dark current can be reduced significantly if the upper state can be dropped from the continuum to just barely at the top of the well, known as the quasibound state.
- *Bound-to-miniband transition*: In this design, each quantum well is surrounded by several barriers and wells coupled together to produce a band of closely spaced energy levels to coincide with the upper state of the quantum well. An electron elevated from the lower state to the upper state will couple with the miniband and will be transported across the band, just like the conduction band continuum, as illustrated in Figure 3.30.
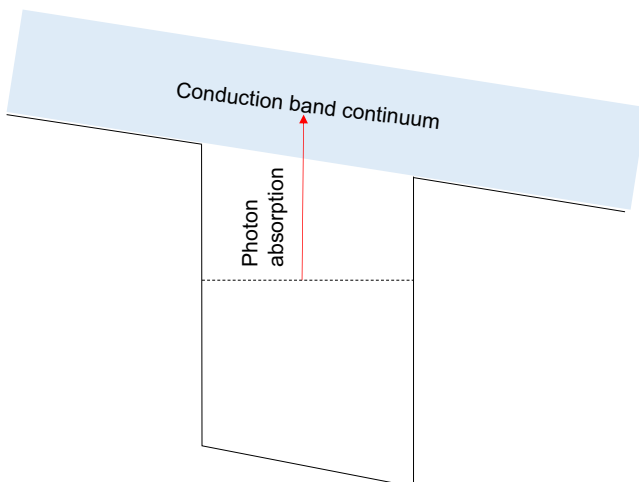
Figure 3.29 Bound-to-continuum transition in a QWIP.
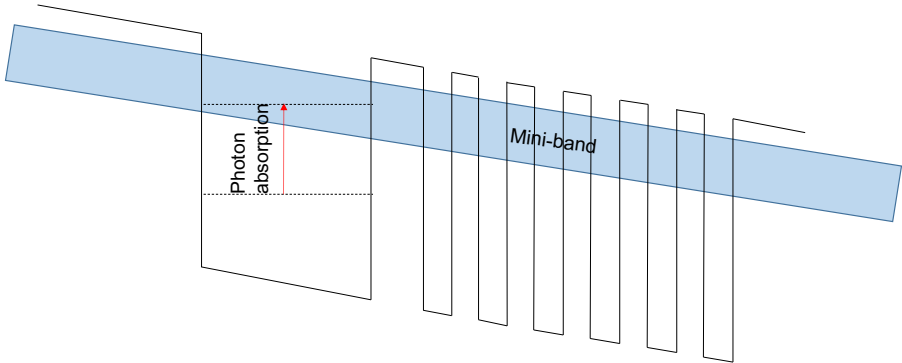
Figure 3.30 Bound-to-miniband transition in a QWIP.
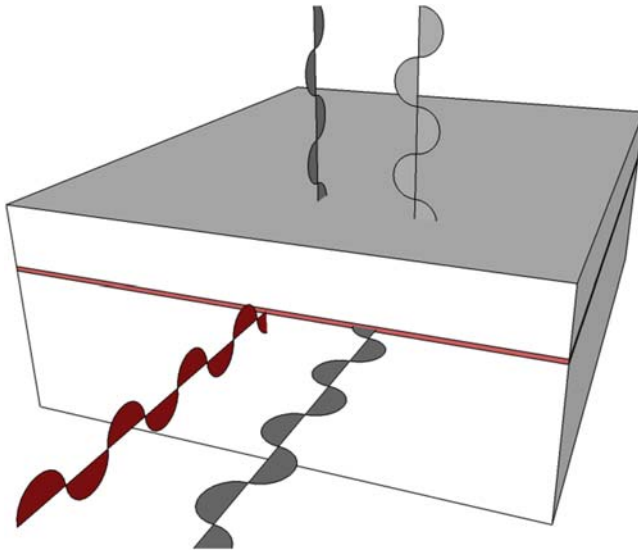
Figure 3.31 Different incident electric field polarizations on a QWIP structure. Only the wave shown in red will be absorbed.

Most of the QWIP devices that have been demonstrated so far have been fabricated using the GaAs/Al$_x$Ga$_{1-x}$As material system. This is due to the lattice-matching feature of this material system and the relatively mature processing technology compared with other III−V semiconductors. QWIPs have been successfully used as the sensor elements in thermal imaging cameras in the mid-infrared and long-wavelength regimes.

A major consideration in QWIPs is their sensitivity to the incident polarization. The electric field vector of the dipole moment between two confined states in a quantum well turns out to be directed perpendicular to the quantum well layers. This means that any electromagnetic radiation that is normally incident to the layer structure will not be absorbed. This is illustrated in Figure 3.31. To be absorbed, the radiation has to be tilted, or otherwise turned with gratings or reflectors. This is one of the biggest challenges with QWIPs, especially in pixel-level implementations.

3.6.3   Quantum cascade lasers

Similar to QWIPs, QCLs are also devices that exploit subband transitions whose energies are much smaller than the material bandgap, but for the purpose of photon emission rather than absorption. Because these are lasing devices, an additional requirement that has to be met is population inversion. However, because population inversion is not possible in a two-level system, QCL structures are designed to be three-level systems. In addition, the structures are also designed so that the electrons in the lowest state of one unit cell tunnel into the upper state of the adjacent cell, thereby recycling the electrons in a cascading fashion. This results in a significant improvement in the quantum efficiency. This is schematically represented in Figure 3.32.

If we represent the wave functions in each unit cell as ψ_1, ψ_2 and ψ_3, with ψ_3 as the upper state and ψ_1 as the lowest state, then the $\psi_3 \rightarrow \psi_2$ is designed to be the photon emission transition and the $\psi_2 \rightarrow \psi_1$ is designed to be a fast, nonradiative transition. If we represent the lifetimes of these transitions as τ_{32} and τ_{21}, then to reach population inversion between ψ_3 and ψ_2 we would need to have

$$\tau_{32} \gg \tau_{21}. \tag{3.136}$$

The scattering rates between the various states are designed to achieve the largest ratio of $\frac{\tau_{32}}{\tau_{21}}$ at the lowest current through the device.

In the most common implementation of the QCL, three coupled quantum wells are used in each cell. This allows a greater flexibility to manipulate the individual wave functions instead of a single well structure with three states. An example is shown in Figure 3.33. The quantum wells are designed such that ψ_3 (shown in magenta) is dominantly localized in the left quantum well, ψ_2 (green) is localized in the right
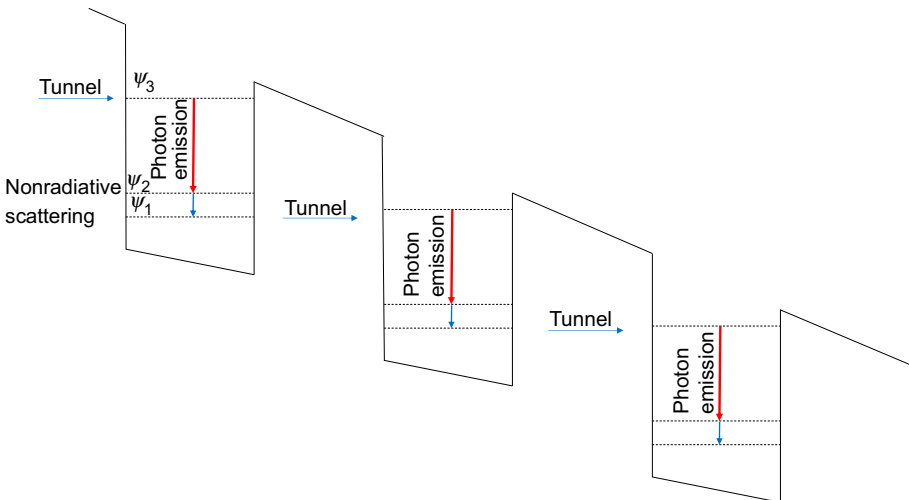
Figure 3.32 Schematic representation of the quantum cascade process.
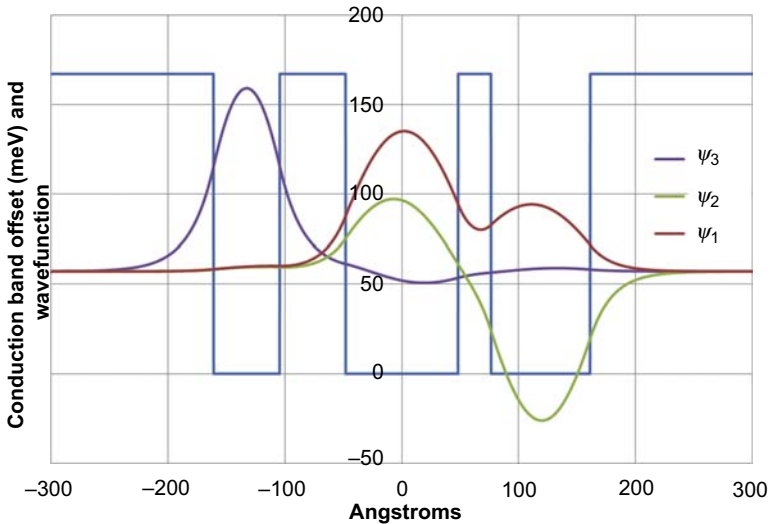
Figure 3.33 Three-coupled quantum well structure of QCL.

quantum well, and the lowest energy state ψ_1 (red) is localized in the central quantum well, although the latter two states are somewhat delocalized between the two quantum wells because they are designed to be closely spaced in energy. The lifetime of the photon scattering process τ_{32} is generally in the nanosecond range, whereas the non-radiative τ_{21} process (which includes phonon and electron–electron scattering) is much faster in the picosecond range. To further reduce the τ_{21} scattering lifetime, the energy separation between ψ_2 and ψ_1 is often matched to certain resonant energy levels of the semiconductor material. For instance, in GaAs, the longitudinal optical (LO) phonon energy is approximately 36 meV and QCLs with $\psi_2 \rightarrow \psi_1$ transition matched to this energy will experience a fast scattering rate.

Another effect is electron–electron scattering, which reaches a maximum value when ψ_2 and ψ_1 go through what is known as the anticrossing state. This occurs when an electric field is applied to this structure. As the field is increased, the confined energy states of the central quantum well and the right quantum well will converge toward each other and then diverge as the field is increased further. When the energies are closest together, the electron–electron scattering rate between the two states will reach a maximum value. Figure 3.34 illustrates ψ_2 and ψ_1 when their energies are closest to each other.

Many cascading units are typically used in QCL devices. Therefore, in principle, each electron can produce many photons. In addition, unlike other semiconductor lasers, the QCL is a unipolar device because the entire process takes place in the conduction band. The GaAs/Al$_x$Ga$_{1-x}$As material system is the most commonly used for QCL. Al$_x$Ga$_{1-x}$As being perfectly lattice matched to GaAs allows a greater flexibility to design the quantum wells without introducing strain and dislocations. QCLs are currently commercially available for various wavelengths ranging from mid-infrared to terahertz emission.
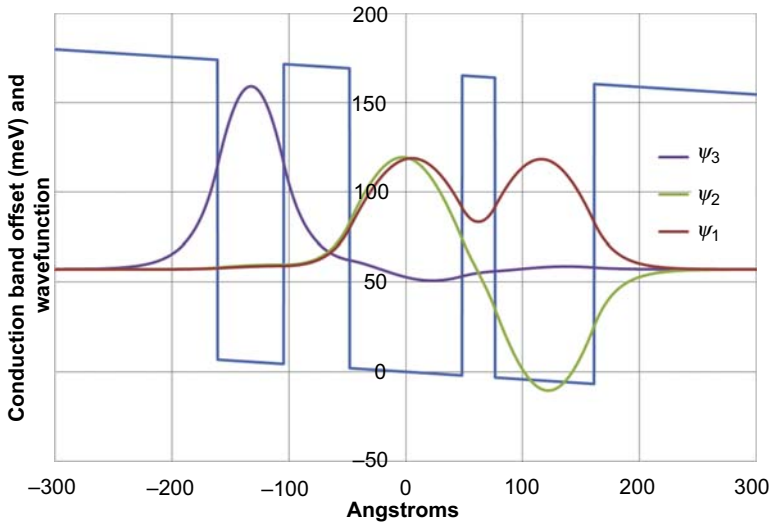
Figure 3.34 Anticrossing state between the first and second solution.

Problems

1. Consider a 100-Å quantum well with infinite barriers. If an electron has a total energy of 50 meV, then determine the number of possible confined states of the electron. Assume that the electron effective mass is 1.0.
   (Three possible states with energies of 3.76, 15.04, 33.84 meV)

2. Use the numerical shooting method to solve for the first confined state of a 30-Å quantum well with 25-meV barriers. Compare this result against the semianalytical solution. Assume that the electron effective mass is 1.0.
   (11.45 meV)

3. Consider the $Al_{0.2}Ga_{0.8}As/GaAs/Al_{0.2}Ga_{0.8}As$ quantum well structure. Calculate the maximum quantum well thickness that supports just one confined state in the conduction band and the valence band.
   (65 nm, 33 nm)

4. Calculate the tunneling probability through a 150-Å, 200-meV barrier as a function of electron energy normal to the barrier from 0 to 500 meV. At approximately what energy does the tunneling probability exceed 99%?
   (350 meV)

5. A double tunnel barrier consists of two 50-Å, 100-meV barriers separated by 100 Å. Calculate the tunneling probability through this structure as a function of electron energy and identify the resonant tunneling energies.
   (27.2 meV, 96 meV)

6. For an asymmetric double tunneling barrier consisting of one 50-Å, 100-meV barrier and another 50-Å, 50-meV barrier separated by 100 Å, calculate the tunneling probability through this structure as a function of electron energy and identify the resonant tunneling energies.
   (23.3 meV)

7. For the quantum cascade laser structure shown in Figure 3.33, the well widths (from left to right) are 56.5, 96.1, and 84.8 Å and the barrier widths (also from left to right) are 56.5 and 28.25 Å. The wells are composed of GaAs and the barriers are composed of $Al_{0.2}Ga_{0.8}As$. Calculate and plot all three energy levels of this structure as a function of applied electric field from 0 to 10 kV/cm.

Further reading

[1]   D.M. Kim, Y.-H. Jeong, Nanowire Field Effect Transistors: Principles and Applications, Publisher: Springer, New York, 2014.
[2]   P. Harrison, Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures, third ed., Wiley Publications, 2010.
[3]   S. Adachi, Physical Properties of III−V Semiconductor Compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP, Wiley Publications, 2005.
[4]   P.S. Zory, Quantum Well Lasers, Academic Press, 1993.
[5]   K.K. Choi, The Physics of Quantum Well Infrared Photodetectors, World Scientific Pub Co Inc., 1997.
[6]   A. Rogalski, Infrared Detectors, second ed., Taylor and Francis Group, 2011.
[7]   J. Faist, Quantum Cascade Lasers, Oxford University Press, 2013.

This page intentionally left blank

Materials

4

J.W. Haus
University of Dayton, Dayton, OH, USA

> *Only within the moment of time represented by the present century has one species –*
> *man – acquired significant power to alter the nature of the world.*
>
> **Rachel Carson, Silent Spring**

4.1 Introduction

An important ingredient of many technological advances has been the discovery of new materials. In this regard, the Bronze Age and Iron Age are important historical markers of materials that drove technology and propelled civilization forward. The pace of new materials development has accelerated in the past century mainly because of the advancement in understanding the underlying quantum nature of atomic and molecular constituents. There has been a transformation from a materials-centric view to an information-based approach to developing technology. In other words, extensive numerical simulations are performed to design the material composition and structure of a potential useful device. Experiments are performed to validate the findings and help lead to new photonic devices.

The periodic table of the elements is a profound descriptive picture of the chemical properties of the elements. Each row is descriptive of filling another electronic shell. Remember that this table was devised by Dmitri Mendeleev before the existence of electrons was discovered by J.J. Thomson. Mendeleev's organization of the elements was based on the similarity of their chemical properties and their atomic mass. He observed that the elements had properties that were recurring or periodic in relation to their atomic masses. Of course, there were holes in Mendeleev's table, but over time the missing elements were found. The table was also refined by the eventual recognition that atomic number (i.e., the number of protons in the nucleus) should replace the atomic mass for ordering the elements in the periodic table. The open mystery at the time, of course, was why are the properties periodic? This question was definitively answered by the advent of quantum mechanics, which reveals physical phenomena that have a profound effect on our understanding of nature.

Surprisingly, the quantum mechanical solution of the hydrogen atom provides deep insight into the chemical properties of the multielectron elements. The details of solving Schrödinger's equation for the hydrogen atom are given in Supplement A. The atomic wave functions of the hydrogen atom and their symmetry provide a reliable guide to properties of many electron atoms. The atomic wave function has a principal

quantum number, n, that labels the electron shells; an orbital quantum number, l; and a magnetic quantum number m that determines its angular symmetry and the number of states for each shell ($n > l \geq |m|$). There are two additional physical properties that are invoked to understand the periodic table. One is the electron has an intrinsic spin; it is denoted by a spin quantum number with two values $m_s = \pm 1/2$. The electron is a spin $^1/_2$ Fermion. The second property is the Pauli exclusion principle, which states that no two Fermions can occupy the same quantum state.

The first six rows of elements in the periodic table are shown below with the atomic number at the top and the outer orbital at the bottom. As a notational fresher, recall that the occupied electronic states are written in the following notation using the previously defined quantum numbers $n[1]^{[m,m_s]}$ for each of the elements, where the superscript counts the number of electrons occupying states summing the magnetic quantum numbers and intrinsic spin states. Of course, because of Pauli's exclusion principle for Fermions, no two electrons occupy states having the same values of all of the quantum numbers.

The electrons occupy states starting from the ground state and fill up the lower energy quantum states first. The levels are built up according to the *Aufbau* principle, which states that the electrons fill up orbital states by increasing order of $n + l$ quantum numbers, starting with the smallest n value and in accordance with the Pauli exclusion principle for electrons. For the s shell ($n = 1,2,3...$, $l = 0$, $m = 0$), two electrons fill the shell ($m_s = \pm 1/2$). For the p shell ($n = 2,3...$, $l = 1$, $m = -1,0,1$) six electrons fill the shell. For the d and f shells (i.e., $l = 2$ and 3, respectively), a simple enumeration of the states shows that 10 and 14 electrons will fill those shells. For example, the inert gas argon with 18 electrons has the filled electronic states $1s^2 2s^2 2p^6 3s^2 3p^6$ (i.e., $n + l = 1,2,3,4$) and germanium with 32 electrons has the states $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^2 = [Ar]3d^{10}4s^2 4p^2$ (i.e., $n + l = 1,2,3,4,5$). The symbol [Ar] is shorthand for the electronic states of argon (Table 4.1).

Table 4.1 Periodic table of the elements to element 86, but excluding the Lanthanide series of rare-earth elements

IA																		VIIIA
1 H 1s¹	IIA												IIIA	IVA	VA	VIA	VIIA	2 He 1s²
3 Li 2s¹	4 Be 2s²												5 B 2p¹	6 C 2p²	7 N 2p³	8 O 2p⁴	9 F 2p⁵	10 Ne 2p⁶
11 Na 3s¹	12 Mg 3s²	IIIB	IVB	VB	VIB	VIIB	←—	VIIIB	—→	IB	IIB		13 Al 3p¹	14 Si 3p²	15 P 3p³	16 S 3p⁴	17 Cl 3p⁵	18 Ar 3p⁶
19 K 4s¹	20 Ca 4s²	21 Sc 3d¹	22 Ti 3d²	23 V 3d³	24 Cr 3d⁴	25 Mn 3d⁵	26 Fe 3d⁶	27 Co 3d⁷	28 Ni 3d⁸	29 Cu 3d¹⁰4s¹	30 Zn 3d¹⁰4s²	31 Ga 4p¹	32 Ge 4p²	33 As 4p³	34 Se 4p⁴	35 Br 4p⁵	36 Kr 4p⁶	
37 Rb 5s¹	38 Sr 5s²	39 Y 4d¹	40 Zr 4d²	41 Nb 4d³	42 Mo 4d⁴	43 Tc 4d⁵	44 Ru 4d⁶	45 Rh 4d⁶5s¹	46 Pd 4d¹⁰	47 Ag 4d¹⁰5s¹	48 Cd 4d¹⁰5s²	49 In 5p¹	50 Sn 5p²	51 Sb 5p³	52 Te 5p⁴	53 I 5p⁵	54 Xe 5p⁶	
55 Cs 6s¹	56 Ba 6s²	57 La 5d¹	72 Hf 5d²	73 Ta 5d³	74 W 5d⁴	75 Re 5d⁵	76 Os 5d⁶	77 Ir 5d⁷	78 Pt 5d⁹6s¹	79 Au 5d¹⁰6s¹	80 Hg 5d¹⁰6s²	81 Tl 6p¹	82 Pb 6p²	83 Bi 6p³	84 Po 6p⁴	85 At 6p⁵	86 Rn 6p⁶	

The occupied or partially occupied electron orbital is indicated on the bottom of each box. The orbitals generally follow Hund's rule, but there are exceptions. The transition metals are found under the B-labeled columns. Notable for photonic applications are the noble metals in the IB column.

Figure 4.1 Electron configurations for three elements. The occupation of the state sublevels follows Hund's rule.

The Aufbau principle describes the order that shells are occupied, but it does not explain how the suborbitals are filled. Hund's rule addresses this issue. It states that every suborbital (i.e., s, p, d,…) is filled by a single electron before being doubly occupied and the electron spins in each suborbital align to maximize the total spin. Three electron configuration examples applying Hund's rule are shown in Figure 4.1.

4.2 Crystal structure

4.2.1 Periodic lattices

Many phenomena explored in this and later chapters exploit properties of periodic structures called crystals. Advanced fabrication tools are available to build thin structures to confine electrons, and a later chapter will examine the electromagnetic properties of materials, so-called *photonic crystals*, built from a periodic arrangement of dielectric structures. Understanding of the energy and dispersion properties of periodic materials is based on a solid mathematical foundation of periodic functions.

A theory of crystalline solids is simplified by the existence of translational symmetry where a set of basis vectors are defined to translate to equivalent positions in the lattice when multiplied by a set of integers. The crystal structure is defined by the lattice and the atomic basis. The unit cell, defined by a set of primitive basis vectors, is repeated throughout the solid, as illustrated in Figure 4.2. The placement of the unit cell can be defined as a matter of convenience. In Figure 4.2, two unit cells are depicted: one with its corners aligned with the vertices of the lattice and a second centered about a vertex. In Figure 4.3, an important third unit cell choice that preserves the symmetry of the lattice around the lattice point is shown for the example of a triangular lattice. This construction is called the Wigner−Seitz (WS) unit cell. The WS unit cell for the triangular lattice is invariant, for example, under rotations around its center by 60°, which is also an invariant rotation for the lattice. The construction of the WS cell is illustrated by the unit cell labeled 2 in Figure 4.2 and is especially useful for the reciprocal lattice.

For Bravais lattices, the entire space is filled by translations of the unit cell. In two dimensions there are 5, and in three dimensions there are 14 Bravais lattices. They are
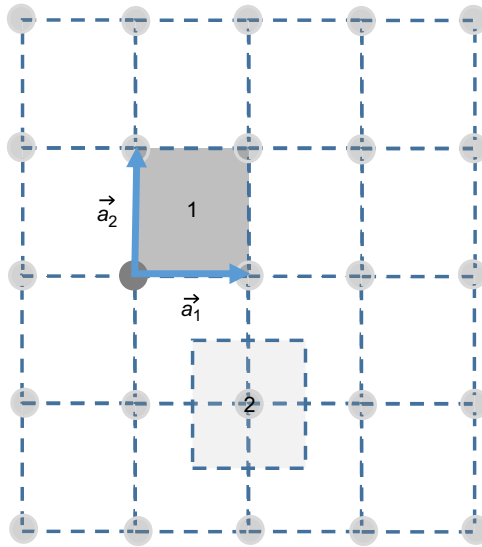
Figure 4.2 A two-dimensional rectangular lattice is used to illustrate the translational properties of a crystalline solid. The lattice vertices are occupied, and the shaded region labeled 1 is a unit cell with two primitive basis vectors $(\vec{a}_1, \vec{a}_2)$, which are translations to nearest-neighbor lattice vertices. The choice of unit cells is arbitrary, a second unit cell, centered on a lattice vertex is labeled 2.
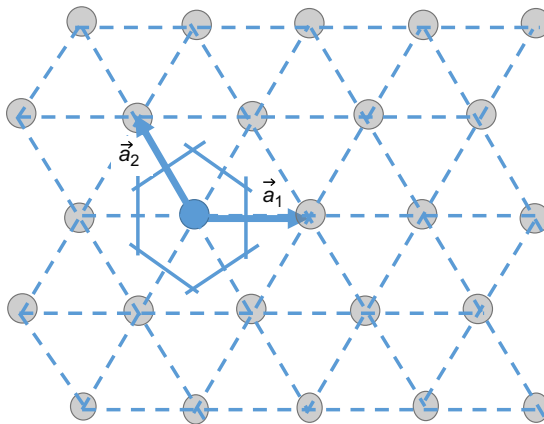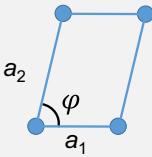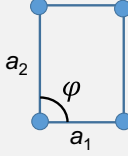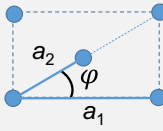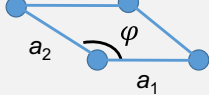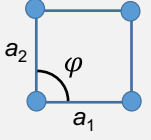
Figure 4.3 Construction of the Wigner−Seitz unit cell for the triangular lattice.

listed in Tables 4.2 and 4.3. These are lattices where all of the lattice positions can be reached by linear combinations of two $(\vec{a}_1, \vec{a}_2)$ or three $(\vec{a}_1, \vec{a}_2, \vec{a}_3)$ primitive basis vectors, respectively. The lattice positions are determined by multiplying the basis vectors by a set of integers (n_1, n_2) and (n_1, n_2, n_3).

$$\vec{R}(n_1, n_2) = n_1\vec{a}_1 + n_2\vec{a}_2, \tag{4.1}$$

Table 4.2 Bravais lattices in two dimensions defined by relations between two lattice constants and the angles between the two sides

Oblique	$a_1 \neq a_2, \varphi \neq \frac{\pi}{2}$	
Rectangular	$a_1 \neq a_2, \varphi = \frac{\pi}{2}$	
Rhombic	$a_1 \neq a_2, \varphi \neq \frac{\pi}{2}$ $\vec{u}_1 \cdot \vec{u}_1 = 2\vec{u}_1 \cdot \vec{u}_2$	
Hexagonal	$a_1 \neq a_2, \varphi \frac{2\pi}{3}$	
Square	$a_1 = a_2, \varphi = \frac{\pi}{2}$	
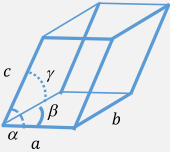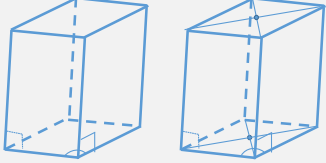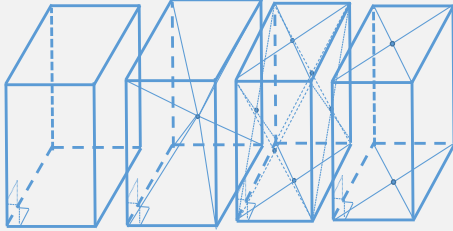
$$\vec{R}(n_1, n_2, n_3) = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3. \qquad (4.2)$$
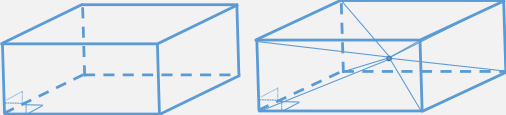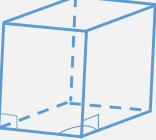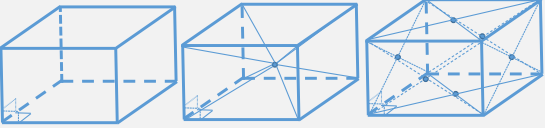
The classification of lattices by their symmetries (translations, rotations, mirror symmetries, inversion, etc.) is an important tool that can be applied to work out the number of independent parameters that are needed to fully describe the properties of a material within a physical tensor. The set of lattice transformations that leave the lattice unchanged or invariant is a space group. An understanding of group theory is beneficial in working out the connections between the lattice symmetry and the physical properties. However, the study of group theory is too complex to delve into and would divert attention from the main topics.

4.2.2 The reciprocal lattice

Periodic functions of one variable are defined by a coordinate translation $f(x + a) = f(x)$. The function can be represented by an infinite one-dimensional lattice with lattice constant a. The treatment of periodic, one-dimensional functions is familiar in most

Table 4.3 **Bravais lattices in three dimensions**

Triclinic (1)	$a \neq b \neq c \neq a$, $\alpha, \beta, \gamma \neq \frac{\pi}{2}$	
Monoclinic (2)	$a \neq b \neq c \neq a$, $\alpha = \gamma = \frac{\pi}{2}, \ \beta \neq \frac{\pi}{2}$	
Orthorhombic (4)	$a \neq b \neq c \neq a$, $\alpha = \beta = \gamma = \frac{\pi}{2}$	

Tetragonal (2)	$a = b \neq c$ $\alpha = \beta = \gamma = \frac{\pi}{2}$	 S B
Trigonal (1)	$a = b = c,$ $\alpha = \beta = \gamma \neq \frac{\pi}{2}, < \frac{2\pi}{3},$	
Hexagonal (1)	$a = b \neq c,$ $\beta = \frac{2\pi}{3},\ \alpha = \gamma = \frac{\pi}{2}$	
Cubic (3)	$a = b = c,$ $\alpha = \beta = \gamma = \frac{\pi}{2}$	 S B F

S, simple; B, body-centered; F, face centered; Ba, base-centered. The lattice lattices are determined by the relations among three lattice constants and three angles defined in the first row.

scientific disciplines, where it is well known that the function can be decomposed into a familiar Fourier series

$$f(x) = \sum_{n=-\infty}^{\infty} A_n e^{iG_n x}, \quad G_n = n\frac{2\pi}{a}. \tag{4.3}$$

This expression manifestly satisfies the periodicity condition because $G_n a = 2\pi n$; the set of amplitudes $\{A_n, n \in \text{Integers}\}$ are determined from the original function by the integral

$$A_n = \frac{1}{a} \int_0^a f(x) e^{-iG_n x} dx. \tag{4.4}$$

The extension of the Fourier decomposition to higher dimensions is straightforward when the function is periodic on a rectangular or simple orthorhombic lattice,

$$f(x+a, y+b) = f(x,y), \tag{4.5}$$

$$f(x+a, y+b, z+c) = f(x,y,z). \tag{4.6}$$

The decomposition is a product of the expressions given Eqn (4.3) for each dimension. Using vector notation, the Fourier series for rectangular and orthorhombic lattices are

$$f(x,y) = \sum_{n_1,n_2=-\infty}^{\infty} A_{n_1,n_2} e^{i\vec{G}_{n_1,n_2} \cdot \vec{r}}, \quad \vec{G}_{n_1,n_2} = n_1\frac{2\pi}{a}\hat{x} + n_2\frac{2\pi}{b}\hat{y}, \tag{4.7}$$

$$f(x,y,z) = \sum_{n_1,n_2,n_3=-\infty}^{\infty} A_{n_1,n_2,n_3} e^{i\vec{G}_{n_1,n_2,n_3} \cdot \vec{r}},$$

$$\vec{G}_{n_1,n_2,n_3} = n_1\frac{2\pi}{a}\hat{x} + n_2\frac{2\pi}{b}\hat{y} + n_3\frac{2\pi}{c}\hat{z}. \tag{4.8}$$

The functions $\vec{G}_{n_1,n_2}$ and $\vec{G}_{n_1,n_2,n_3}$ are reciprocal lattice vectors.

Constructing the Fourier series for a multivariable function on a lattice that has an orthogonal basis vectors (e.g., the triangular lattice in Figure 4.3) is managed using a straightforward formulation. In general symmetry cases, the basis vectors $(\vec{a}_1, \vec{a}_2, \vec{a}_3)$ are used to form the reciprocal lattice basis vectors and are defined by

$$\vec{g}_1 = 2\pi\frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3}, \quad \vec{g}_2 = 2\pi\frac{\vec{a}_3 \times \vec{a}_1}{\vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3},$$

$$\vec{g}_3 = 2\pi\frac{\vec{a}_1 \times \vec{a}_2}{\vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3}. \tag{4.9}$$

For a nonorthogonal set of basis vectors, the volume of the unit cell is $v_c = \vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3$ (i.e., the volume of a parallelepiped with side lengths $|\vec{a}_1| \cdot |\vec{a}_2|$, $|\vec{a}_3|$). For the reciprocal lattice vector in two dimensions, set $\vec{a}_3 = c\hat{z}$ (the constant, c, is entirely arbitrary) and use two basis vectors restricted to the (x, y) plane. The reciprocal lattice basis vectors satisfy the following relation:

$$\vec{g}_i \cdot \vec{a}_j = 2\pi\delta_{ij}. \tag{4.10}$$

The general Fourier series for the function is

$$f(\vec{r}) = \sum_{n_1,n_2,n_3=-\infty}^{\infty} A_n e^{i\vec{G}_{n_1,n_2,n_3} \cdot \vec{r}}, \quad \vec{G}_{n_1,n_2,n_3} = n_1\vec{g}_1 + n_2\vec{g}_2 + n_3\vec{g}_3. \tag{4.11}$$

The set of variables $\{\vec{G}_{n_1,n_2,n_3}, n_1, n_2, n_3 \in \text{Integers}\}$ form a periodic reciprocal lattice. The volume of the unit cell in the reciprocal lattice is inversely proportional to the volume of the unit cell,

$$\vec{g}_1 \cdot \vec{g}_2 \times \vec{g}_3 = \frac{(2\pi)^3}{\vec{u}_1 \cdot \vec{u}_2 \times \vec{u}_3} = \frac{(2\pi)^3}{v_c}. \tag{4.12}$$

The reciprocal lattice is a momentum space representation of the functions; therefore, the variables $\vec{G}_{n_1,n_2,n_3}$ represent wave vectors. On the reciprocal lattice, the WS unit cell is invoked to examine the momentum space properties of physical functions (wave functions, electromagnetic fields, etc.). The WS unit cell is the first Brillouin zone; it is simply called the Brillouin zone (BZ) and is used to quantify the electronic or optical properties in periodic materials in a momentum space representation. Band structure plots are energy or dispersion bands for wavenumbers along specific symmetry directions in the BZ.

For example, consider the triangular lattice with the basis vectors defined as

$$a_1 = a\hat{x}, \quad \vec{a}_2 = a\left(-\frac{1}{2}\hat{x} + \frac{\sqrt{3}}{2}\hat{y}\right), \tag{4.13}$$

and shown in Figure 4.3. The corresponding reciprocal lattice wave vectors are

$$\vec{g}_1 = \frac{2\pi}{a}\hat{x} + \frac{2\pi}{\sqrt{3}a}\hat{y}, \quad \vec{g}_2 = \frac{4\pi}{\sqrt{3}a}\hat{y}. \tag{4.14}$$

They also form a triangular lattice, and the BZ has the shape of a regular hexagon. The reciprocal lattice is illustrated in Figure 4.4(a). The BZ is highlighted in Figure 4.4(b) with a triangle inserted. The BZ in Figure 4.4(b) has three labeled points of high symmetry.
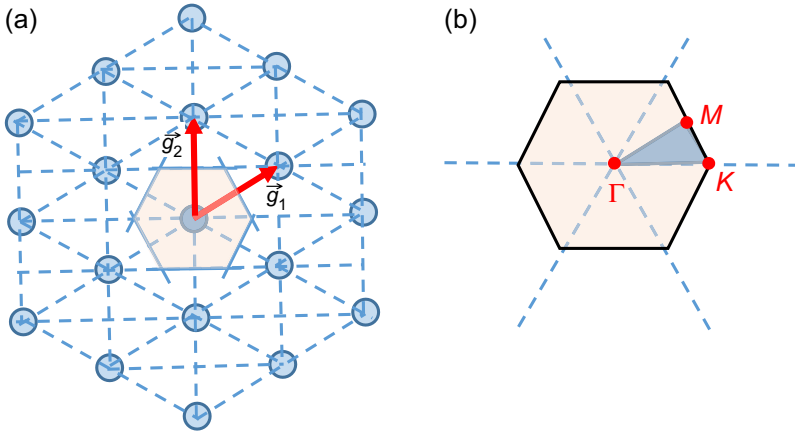
Figure 4.4 (a) A portion of the reciprocal lattice for the triangular lattice shown in Figure 4.3 with the basis vectors $(\vec{g}_1, \vec{g}_2)$ and the shaded region is the BZ. (b) The BZ with high symmetry points indicated.

$$\Gamma = 0\widehat{x} + 0\widehat{y}, \quad M = \frac{\pi}{a}\left(\widehat{x} + \frac{1}{\sqrt{3}}\widehat{y}\right) \quad \text{and} \quad K = \frac{4\pi}{3a}\widehat{x} + 0\widehat{y}. \tag{4.15}$$

The determination of the energy or dispersion bands may be simplified by using the crystal symmetry. In addition to the lattice symmetry, the structures at the lattice sites also have symmetry properties. For example, a triangular lattice with circular structures on the lattice sites (Figure 4.4) satisfies the same symmetry operations as the lattice point group. In this case, the band structure properties within the shaded triangular region of the BZ in Figure 4.4(b) capture the same characteristics over the entire zone. The band structure can be extended over the entire zone area by using the group transformations. For high-symmetry systems, there are 6 equivalent M and K points in the BZ and only 1/12th of its area needs to be examined. An itinerary for presenting the band structure is along lines connecting the high-symmetry points (e.g., Γ to K to M to Γ). The band structure images represent points along the periphery of the triangle.

The three-dimensional lattices of interest are the simple and face-centered cubic lattice and the hexagonal lattice. Unit cells of these lattices are drawn in Figure 4.5.
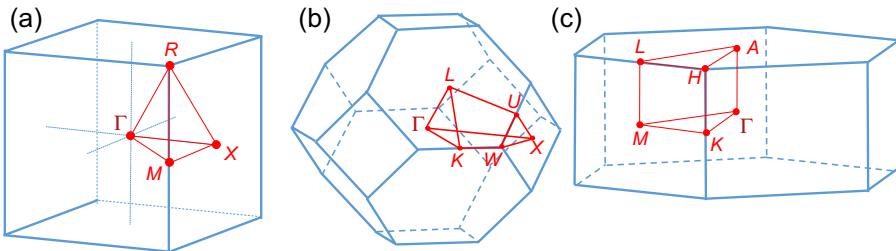
Figure 4.5 BZs constructed for three lattices: (a) SC, (b) FCC, and (c) hexagonal. The high-symmetry points on the surface are labeled.

A set of basis vectors together with their corresponding reciprocal lattice basis vectors are as follows:

Simple cubic (SC; lattice constant a)

$$\vec{a}_1 = a\widehat{x}, \quad \vec{a}_2 = a\widehat{y}, \quad \vec{a}_3 = a\widehat{z}. \tag{4.16}$$

$$\vec{g}_1 = \frac{2\pi}{a}\widehat{x}, \quad \vec{g}_2 = \frac{2\pi}{a}\widehat{y}, \quad \vec{g}_3 = \frac{2\pi}{a}\widehat{z}. \tag{4.17}$$

Face-centered cubic (FCC; lattice constant a)

$$\vec{a}_1 = \frac{a}{\sqrt{2}}\left(\widehat{x}+\widehat{y}\right), \quad \vec{a}_2 = \frac{a}{\sqrt{2}}\left(\widehat{x}+\widehat{z}\right), \quad \vec{a}_3 = \frac{a}{\sqrt{2}}\left(\widehat{y}+\widehat{z}\right). \tag{4.18}$$

$$\vec{g}_1 = \frac{\sqrt{2}\pi}{a}\left(\widehat{x}+\widehat{y}-\widehat{z}\right), \quad \vec{g}_2 = \frac{\sqrt{2}\pi}{a}\left(\widehat{x}-\widehat{y}+\widehat{z}\right),$$
$$\vec{g}_3 = \frac{\sqrt{2}\pi}{a}\left(-\widehat{x}+\widehat{y}+\widehat{z}\right). \tag{4.19}$$

Hexagonal (lattice constants: (x, y) plane a, z plane c)

$$\vec{a}_1 = a\widehat{x}, \quad \vec{a}_2 = a\left(-\frac{1}{2}\widehat{x}+\frac{\sqrt{3}}{2}\widehat{y}\right), \quad \vec{a}_3 = c\widehat{z}. \tag{4.20}$$

$$\vec{g}_1 = \frac{2\pi}{a}\widehat{x}+\frac{2\pi}{\sqrt{3}a}\widehat{y}, \quad \vec{g}_2 = \frac{4\pi}{\sqrt{3}a}\widehat{y}, \quad \vec{g}_3 = \frac{2\pi}{c}\widehat{z}. \tag{4.21}$$

The SC lattice and its reciprocal lattice are identical in form. The BZ is also a cube and is illustrated in Figure 4.5(a). The gamma point always denotes the point in the center of the zone. A subset of the designated high-symmetry points on the surface of the BZ is

$$X = \frac{\pi}{a}\widehat{x}, \quad M = \frac{\pi}{a}\widehat{x}+\frac{\pi}{a}\widehat{y}, \quad R = \frac{\pi}{a}\widehat{x}+\frac{\pi}{a}\widehat{y}+\frac{\pi}{a}\widehat{z}. \tag{4.22}$$

The FCC lattice in Figure 4.5(b) displays a 14-sided BZ (6 sides are diamond shaped and 8 sides are hexagonal shaped). Labeled high-symmetry points at representative positions are

$$X = \frac{\sqrt{2}\pi}{a}\widehat{y}, \quad W = \frac{\pi}{\sqrt{2}a}\widehat{x}+\frac{\sqrt{2}\pi}{a}\widehat{y}, \quad K = \frac{3\pi}{2\sqrt{2}a}\widehat{x}+\frac{3\pi}{2\sqrt{2}a}\widehat{y},$$
$$U = \frac{\pi}{2\sqrt{2}a}\widehat{x}+\frac{\pi}{\sqrt{2}a}\widehat{y}-\frac{\pi}{2\sqrt{2}a}\widehat{z}, \quad L = \frac{\pi}{\sqrt{2}a}\widehat{x}+\frac{\pi}{\sqrt{2}a}\widehat{y}-\frac{\pi}{\sqrt{2}a}\widehat{z}. \tag{4.23}$$

Finally, the hexagonal reciprocal lattice BZ in Figure 4.5(c) is composed of a base with the triangular lattice BZ and a basis vector in the z direction. A subset of high-symmetry points on the surface of the BZ is

$$M = \frac{\pi}{a}\left(\widehat{x} + \frac{1}{\sqrt{3}}\widehat{y}\right), \quad K = \frac{2\pi}{\sqrt{3}a}\widehat{x}, \quad A = \frac{\pi}{c}\widehat{z},$$

$$L = \frac{\pi}{a}\left(\widehat{x} + \frac{1}{\sqrt{3}}\widehat{y}\right) + \frac{\pi}{c}\widehat{z}, \quad H = \frac{2\pi}{\sqrt{3}a}\widehat{x} + \frac{\pi}{c}\widehat{z}.$$

For high-symmetry physical systems, the reduced volume of the BZ sufficient to plot the entire band structure is reduced by 32, 48, and 24 times for each lattice.

4.2.2.1 Hybridization

Examining atomic orbitals of the electrons for the elements is the key to a deeper understanding of the electronic properties of solid materials. The atomic orbitals were previously introduced and electron configurations are tabulated in Table 4.1. Electrons in the filled shells are strongly bound to the nucleus and interact very weakly with neighboring atoms. The inner electrons partially shield the nucleus from the open shell electrons. The outer electrons are called *valence electrons*.

The valence electrons in the unfilled electronic shells extend their electronic wave functions over a wider radius that overlaps with neighboring atomic lattice sites. The coupling between all of the lattice sites contributes to the formation of bands that extend over the entire lattice. The valence electrons are shared among all of the atoms in the solid and they fill available states in the bands.

The mathematical treatment of bands was previously presented in the chapter on quantum mechanics. The reader may examine the tight binding model (TBM) treating the quantum well wave functions as atomic wave functions and using the overlap be-tween them to form bands. The TBM illustrates how the overlap of wave functions among neighboring atoms leads to the formation of electronic bands. The bands are a consequence of bonding atoms together, and they contribute to the energy of cohesion, which is discussed in Section 4.3.

In the treatment of covalent bonds, a pair of electrons with a linear combination of atomic orbitals. For instance, a hydrogen atom with a single electron in the 1s state can complete the shell by binding two hydrogen atoms together to form a hydrogen molecule. The formation of the molecule is energetically favored over the atomic state. To understand more complex atomic systems, bond formation is understood by combining or hybridizing multiple orbitals. There are three hybridized orbitals that will be highlighted here that are simply designated as sp, sp^2, and sp^3. There are addi-tional cases that are relevant involving d orbitals, but they are not considered further here.

The carbon atom containing six electrons is an excellent example of sp orbital hybridization. Its electronic ground state has four valence electrons in the $n = 2$ orbitals; the 2s state has two electrons and two additional electrons are shared between the three 2p states, in accordance with Hund's rule, as shown on the left in
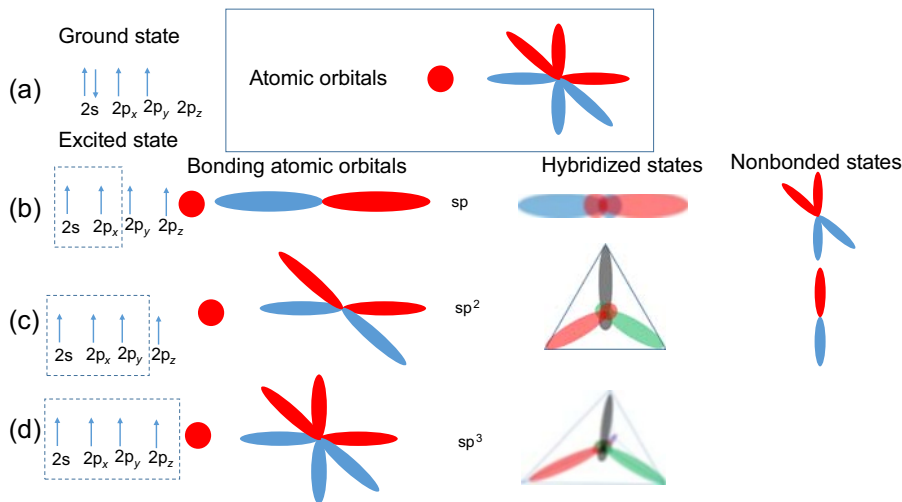
Figure 4.6 Hybridized atomic orbitals.

Figure 4.6(a). The arrows indicate the electron spin state, either up or down. The s state is spherically symmetric and the p state is cylindrically symmetric with equal probability to be on either side of the node. The atomic orbitals are depicted in Figure 4.6(a) with the s state as a circle and the p states as three orthogonal double ellipses. The π phase difference between the two lobes in the p states is indicated by different shading.

For sp hybrid states, an s and p atomic orbital are combined to form the new orbitals. The two hybrid states indicated by the box in Figure 4.6(b) are 2s and $2p_x$. There are two sp hybrid orbitals each with a probability distribution that is asymmetric around the node. This characteristic is shown as a large lobe on one side of the node. The other two atomic p orbitals (p_y, p_z) do not participate in the sp hybrid bond. The sp hybrid state bonding is along the line of symmetry and can participate in the formation of a linear molecule.

The sp^2 hybrid states are formed by combining the s state with two p states, as shown in Figure 4.6(c). The atomic p states lie orthogonal to one another in the (x, y) plane and their mixing forms three hybrid states that have asymmetric lobes and lie in a plane. The hybrid orbitals minimize their overlap by distributing the lobes at a 120° angle from one another (i.e., their lobes point along different vertices of an equilateral triangle). The sp^2 hybrid bond can be identified when molecular bond angles display the same symmetry.

The four sp^3 hybrid states are illustrated in Figure 4.6(d). All four states are mixed to form the hybrid wave functions. The nucleus is found at the center of a triangular pyramid and the lobes of the four hybrid orbitals point to the vertices. The angle that each hybrid orbital makes with the others is the characteristic 109.5°.

The hybrid orbitals are applied to understand the covalent bonding of atoms to form molecules. Hybrid orbitals from two atoms overlap and pair electrons to

complete the occupation of the state, consistent with the Pauli exclusion principle. In particular, they impose a symmetry that is different from the original atomic orbitals and the number of states for a particular symmetry is equal to the number of electrons occupying the atomic states. The symmetry of the hybrid orbitals conforms to expectations based on the repulsion between the numbers of electrons that occupy the bonding orbitals.

Understanding electron occupation of states when bonding atoms to form molecules or solids is based on simple, previously discussed physical principles. Electrons will occupy the lowest energy orbitals first; as the lower energy states are filled, electrons will occupy the next higher energy state.

Diatomic molecules are formed by sharing the electrons. This results in a modification of the electron density between the two nuclei. Consider two hydrogen atoms in their ground (1s) state. The spherical distributions of the electrons in the 1s state around each nucleus are distorted by linear combinations of the s states. When the phase of the two wave functions is the same, the linear combination results in an electron density that is concentrated between the nuclei. When the two atomic wave functions are combined with the opposite phase, the electron density has a node between the two nuclei and the electron density is displaced closer to the nuclei. The two states are illustrated in Figure 4.7. The energy is higher when there is a node in the electron density and is called an antibonding state. The lower energy state is called the bonding state; it has a lower energy because the electron density is concentrated at a greater distance from the nuclei.
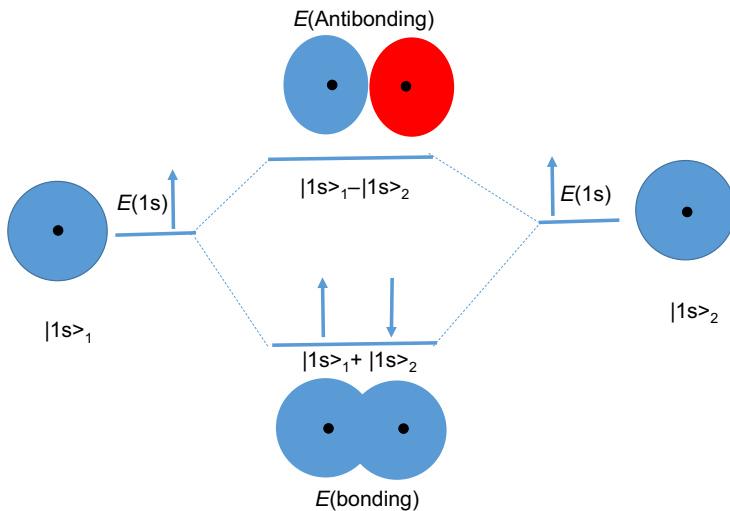
Figure 4.7 Two separate hydrogen atoms have $|1s>$ ground states; the arrows (up or down) depict the two intrinsic electron spin states. As the atoms are placed closer together, their wave functions overlap and mix to form two states. The bonding state is the lower energy molecular state. The higher energy molecular state is called the antibonding state.
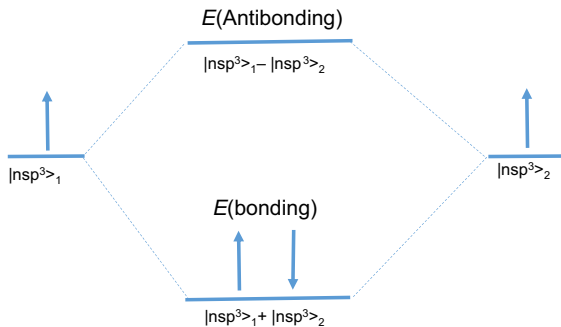
Figure 4.8 Two atoms possessing sp^3 states that overlap with one another will form bonding and antibonding states when bound together.

The same bonding principle applies to hybrid states and is illustrated in Figure 4.8. Hybrid orbitals from neighboring atoms combine to form a set of bonding and antibonding states. The bonding state is occupied by two electrons in the hybrid orbitals. The shift of the energy levels can be determined by the overlap of the hybrid wave functions.

The principles of hybrid orbitals will be invoked when we examine the properties of specific materials in the next section. The symmetry of the solid-state materials and our knowledge about the atomic constituents will be a guide to a deeper understanding of their physical properties. Specifically, the theoretical considerations start with the identification of the relevant atomic valence electrons as drawn from the periodic table. The physical properties are then calculated using quantum mechanics based on the linear combination of atomic orbitals of the valence electrons centered on the nuclei.

4.3 Metals

The optical properties of metals are complicated by the free electrons that act like a fluid inside of the solid. For instance, the observation of collective electronic motion called plasmons has opened up many new applications of metals, and they are of special interest in the field of nanophotonics. Moreover, metallic elements dominate the periodic table and they have been widely studied using density functional theory (DFT), which is a method that has been refined, expanded, and applied to understanding materials properties. The literature on this topic is extensive. The origins of DFT are based on the original contributions of Fermi and Thomas almost a century ago. They recognized the central role of electron density in computing the properties multiatomic systems. Hohenberg and Kohn's 1964 paper proving that the ground-state energy of an electronic system is determined by the density alone is an important marker that propelled the future development of DFT. Since that time, over 100,000 DFT-related papers have been published.

Simple metals from Groups IA, IIA, and IIIA of the periodic table have valence electrons in s and p orbitals with closed-shell inner orbitals that are tightly bound to the nuclei. The group of transition metals has valence electrons in the d orbitals, and in rare-earth metals (not shown) the valence electrons occupy f orbitals.

For the free electron model, the electron density is homogeneous throughout the solid (volume V). At zero temperature, the N (factor of two for spin $1/2$ degree of freedom) free electrons fill up the volume to the Fermi wave number k_F,

$$N = \frac{2V}{(2\pi)^3} \iiint d^3k = \frac{V}{3\pi^2} k_F^3. \tag{4.24}$$

From this result, the Fermi wave number is related to the electron density by $k_F = (3\pi^2 n)^{1/3}$; the electron density is $n = N/V$. In monovalent metals (i.e., one free electron per atom) the number n is the number density of atoms in the volume. In multivalent metals, the number of electrons per metal atom is denoted as z and $N = zN_c$, where N_c is the number of metal atoms in the volume V.

The wave function for a multielectron is represented by the Slater determinant in Eqn 4.25, which incorporates the Pauli exclusion principle, into the determinant of an $N \times N$ matrix composed of single-electron atomic wave functions. Thus, the multi-electron wave function is antisymmetric in the interchange of any pair of atomic wave functions. The functions $\psi_{\vec{k}}(\vec{r})$ are single-electron wave functions, where the subscript also implicitly includes the electron spin to designate the quantum state of the electron. The normalized N particle wave function can be defined as

$$\Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots \vec{r}_N) =$$

$$\frac{1}{\sqrt{N!}} \det \begin{vmatrix} \psi_{\vec{k}_1,\sigma_1}(\vec{r}_1) & \psi_{\vec{k}_2,\sigma_2}(\vec{r}_1) & \cdots & \psi_{\vec{k}_N,\sigma_N}(\vec{r}_1) \\ \psi_{\vec{k}_1,\sigma_1}(\vec{r}_2) & \psi_{\vec{k}_2,\sigma_2}(\vec{r}_2) & \cdots & \psi_{\vec{k}_N,\sigma_N}(\vec{r}_2) \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{\vec{k}_1,\sigma_1}(\vec{r}_N) & \psi_{\vec{k}_2,\sigma_2}(\vec{r}_N) & \cdots & \psi_{\vec{k}_N,\sigma_N}(\vec{r}_N) \end{vmatrix}. \tag{4.25}$$

For any two states taking the same quantum number values (e.g., $\vec{k}_1, \sigma_1 = \vec{k}_2, \sigma_2$) the determinant vanishes, thus enforcing Pauli's principle.

The quantum mechanical kinetic energy operator for the N electron system is defined as

$$\widehat{T}_{KE}(N) = \sum_{i=1}^{N} -\frac{\hbar^2}{2m} \nabla_i^2. \tag{4.26}$$

Each term is a single-electron contribution. The average value is expressed as

$$
\begin{aligned}
T_{\mathrm{KE}}(N) &= \int \Psi^*(\vec{r}_1, \vec{r}_2, \vec{r}_3, \dots \vec{r}_N)\widehat{T}_{\mathrm{KE}}(N) \\
&\quad \times \Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \dots \vec{r}_N)\mathrm{d}^3\vec{r}_1\mathrm{d}^3\vec{r}_2\dots\mathrm{d}^3\vec{r}_N \\
&= N \sum_{\vec{k},\sigma} \iiint -\frac{\hbar^2}{2m}\psi^*_{\vec{k},\sigma}(\vec{r})\nabla^2\psi_{\vec{k}\sigma}(\vec{r})\mathrm{d}^3\vec{r}.
\end{aligned}
\tag{4.27}
$$

The last single-electron integral is a consequence of the identical contribution for each electron. The sum over quantum numbers $(\vec{k}, \sigma)$ reflects the properties of the Slater determinant. The single particle properties assume orthogonality of the set of wave functions.

To be specific, consider the free electron wave function as a normalized plane-wave function:

$$
\psi_{\vec{k},\sigma}(\vec{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{i\vec{k}\vec{r}\cdot}|\sigma>, \quad \psi^*_{\vec{k},\sigma}(\vec{r}) = \frac{1}{\sqrt{V}}\mathrm{e}^{i\vec{k}\vec{r}\cdot}<\sigma|.
\tag{4.28}
$$

The bra ($|\sigma>$) and ket ($<\sigma|$) notation is adopted for the spin coordinate; the inner product, a bra and ket, has the following orthogonality property:

$$
<\sigma|\sigma'> = \delta_{\sigma,\sigma'}.
\tag{4.29}
$$

The sum over wave vectors is transformed into an integral using the relation

$$
\sum_{\vec{k},\sigma} \Rightarrow \frac{2V}{(2\pi)^3}\iiint \dots \mathrm{d}^3\vec{k}.
\tag{4.30}
$$

The factor of two assumes that the integrand is independent of the spin coordinate. This restriction is true of the kinetic energy discussed above. For a single particle wave function, the zero temperature kinetic energy is given by

$$
T_{\mathrm{KE}}(N) = \frac{2V}{(2\pi)^3}\frac{\hbar^2}{2m}\iiint k^2\mathrm{d}^3\vec{k} = \frac{1}{5}\frac{V}{(\pi)^2}\frac{\hbar^2}{2m}k_F^5 = \frac{3}{5}NE_F.
\tag{4.31}
$$

In the last equality, the Fermi wavenumber is used to define the Fermi level $E_F = \frac{\hbar^2}{2m}k_F^2$. Recall that the Fermi wavenumber, defined in Eqn (4.24) is a function of the free electron density.

In a noninteracting electron system, the internal energy of a system is determined from the kinetic energy. The kinetic energy vanishes as V goes to infinity; therefore,

there is no energy of cohesion in the model. The inclusion of interactions among the constituents in a material leads to additional attractive contributions to the internal energy that can have a minimum value at a finite volume, and this stabilizes the material. The Jellium model incorporates Coulomb interactions among the constituents with the proviso that the ion density is uniform throughout (i.e., n_{ion} is constant). The internal energy for Jellium is composed of three additional terms

$$U = T_{\text{KE}} + V_{\text{ion}-e} + V_{\text{ion}-\text{ion}} + V_{e-e};$$

The Coulomb interaction energy among the smeared ions is

$$V_{\text{ion}-\text{ion}} = \frac{1}{2}\iiint n_{\text{ion}}\frac{\text{e}^2}{4\pi\varepsilon_0|\vec{r}-\vec{r}'|}n_{\text{ion}}\text{d}^3\vec{r}\text{d}^3\vec{r}'. \tag{4.32}$$

It is not affected by the electronic wave function and appears as a constant shift of the electron internal energy.

The interaction energy between electrons and ions is

$$V_{\text{ion}-e} = -\sum_{\vec{k},\sigma}\iiint n_{\text{ion}}\frac{\text{e}^2}{4\pi\varepsilon_0|\vec{r}-\vec{r}'|}\left|\psi_{\vec{k},\sigma}(\vec{r})\right|^2\text{d}^3\vec{r}\text{d}^3\vec{r}'. \tag{4.33}$$

In this expression, the nuclei play the role of applying a smeared Coulomb field acting on the electrons. The integrand contains the electron density defined as $n_{\text{e}} = \sum_{\vec{k},\sigma}\left|\psi_{\vec{k},\sigma}(\vec{r})\right|^2$, which, analogous to the kinetic energy contribution, is composed of a sum of single-electron wave functions. The electron−electron interactions have a richer form based on the Slater determinant. There are two electron contributions appearing in the integrals:

$$\begin{aligned}V_{e-e} =&\frac{1}{2}\sum_{\vec{k},\sigma}\sum_{\vec{k}',\sigma'}\iiint\left|\psi_{\vec{k},\sigma}(\vec{r})\right|^2\frac{\text{e}^2}{4\pi\varepsilon_0|\vec{r}-\vec{r}'|}\left|\psi_{\vec{k}',\sigma'}(\vec{r}')\right|^2\text{d}^3\vec{r}\text{d}^3\vec{r}'\\[6pt]&-\frac{1}{2}\sum_{\vec{k},\sigma}\sum_{\vec{k}',\sigma'}\iiint\frac{\text{e}^2}{4\pi\varepsilon_0|\vec{r}-\vec{r}'|}\psi_{\vec{k}',\sigma'}(\vec{r})\psi_{\vec{k},\sigma}^*(\vec{r})\psi_{\vec{k},\sigma}(\vec{r}')\\[6pt]&\times\psi_{\vec{k}',\sigma'}^*(\vec{r}')\delta_{\sigma,\sigma'}\text{d}^3\vec{r}\text{d}^3\vec{r}'.\end{aligned} \tag{4.34}$$

The first term corresponds to the classical repulsion that two electron densities in the same locality would experience. The last term is a consequence of the Pauli exclusion principle that no two Fermions can occupy the same state (including the spin degree of freedom). This is called the exchange term and it has no classical analog. It is

present because of the antisymmetrized Slater determinant. Note that this term is only nonzero when the spins on the two wave functions are the same.

Invoking the plane-wave electron wave functions to calculate the interaction contributions, the Coulomb terms cancel one another ($n = n_{\text{ion}}$) and the internal energy is

$$
U = \frac{3}{5}NE_{\text{F}} - \frac{1}{(2\pi)^6}
$$
$$
\times \iiint \left(\iiint \frac{e^2}{4\pi\varepsilon_0|\vec{r} - \vec{r}'|} e^{\mathrm{i}\left(\vec{k}-\vec{k}'\right)\cdot\left(\vec{r}-\vec{r}'\right)} \mathrm{d}^3\vec{r}\mathrm{d}^3\vec{r}' \right) \mathrm{d}^3\vec{k}\mathrm{d}^3\vec{k}'. \tag{4.35}
$$

The second term is the exchange term and it is evaluated by transforming the six-dimensional spatial integral in parentheses to "center of mass" and relative coordinates. The integral over the center of mass yields the volume V and the integral over the relative coordinates is the Fourier transform of the Coulomb potential of a point charge, which is

$$
\frac{e^2}{\varepsilon_0\left|\vec{k} - \vec{k}'\right|^2}. \tag{4.36}
$$

The six integrals over the wave vectors are likewise evaluated by decomposing the variables into center of mass and relative coordinates. The center of mass coordinate integral is $\frac{4\pi}{3}k_{\text{F}}^3$ and the relative integral has the value $4\pi k_{\text{F}}$. After some reorganization of the last expression, the internal energy is

$$
U = N\left(\frac{3}{5}E_{\text{F}} - \frac{3e^2k_{\text{F}}}{16\pi^2\varepsilon_0}\right). \tag{4.37}
$$

This result is elegant and simple and a useful physical quantity for deriving additional results. The cohesion energy of a metal is primarily based on electrons that do not share a bond with a specific atom but instead are delocalized throughout the volume. The lack of bond-specific interactions in metals is what endows them with their ductility and malleability. The internal energy is a thermodynamic quantity. The electron pressure is defined by

$$
P = -\left.\frac{\partial U}{\partial V}\right|_N = N\left(\frac{2}{5}E_{\text{F}} - \frac{e^2k_{\text{F}}}{16\pi^2\varepsilon_0}\right). \tag{4.38}
$$

The first term is the repulsive free electron contribution and the exchange term lowers the pressure. In equilibrium, the pressure is zero in this expression. The electron

pressure is used to derive the bulk modulus, which measures the compressibility or rigidity of the metal. Equation (4.38) has serious drawbacks; for instance, it does not differentiate between elements. Additional effects were included to remedy this deficiency, such as correlated energy and core repulsion effects.

It is instructive to minimize the internal energy with respect to the Fermi wavenumber. The value of the wavenumber at the minimum is

$$k_{F,\min} = \frac{5}{4\pi a_0},$$ (4.39)

where the denominator is expressed in terms of the Bohr radius $a_0 = \frac{4\pi\varepsilon_0 h^2}{me^2} = 0.0529$ nm. The Fermi wavenumber is a function of the electron density. For multivalent metal the electron density is $n = z\, n_{\text{atom}}$ with n_{atom} the atomic density of the material and z is the electron valence ($z = 1$ in monovalent metals, $z = 2$ in divalent metals etc.). The atomic density is related to the WS radius, r_{ws} which is defined as the radius of a sphere around a single atom:

$$V_{ws} = \frac{4\pi}{3}(R_{ws})^3.$$

The atomic density is $n_{\text{atom}} = 1/V_{ws}$. Using these relations, the Fermi wavenumber is related to the WS radius as $k_F = \left(\frac{9\pi z}{4}\right)^{1/3} \frac{1}{R_{ws}}$. Using Eqn (4.39), the WS radius ($z = 1$) is 0.255 nm and the internal energy minimum is $\frac{U_{\min}}{N} = -\frac{15}{16\pi^2}E_1 = -1.29$ eV, the symbol E_1 is the ground-state energy of the hydrogen atom (13.6 eV). An estimate of lattice constant is determined by doubling the WS radius. The lattice constant depends weakly on the electron valency in the unit cell, but it does not depend on specific properties of the elements.

The severe shortcomings of the model have resulted in a reexamination of the underlying assumptions, and corrections have been added. For instance, using the many-body theory, Low and Pines derived an additional contribution called the correlation energy. It is a small correction to the energy and the lattice constant, but it has the troublesome logarithmic divergence at large radii. The contribution of the Coulomb energy between the charges in a WS volume is another physical effect; the mutual repulsive effect between electrons in the core is represented as a repulsive constant potential core of radius R_c. The effective ion-electron potential is represented in Figure 4.9. The potential has the form

$$V(r) = \begin{cases} \dfrac{-1}{4\pi\varepsilon_0 r}, & r > R_c \\[2mm] 0, & r < R_c \end{cases}$$ (4.40)
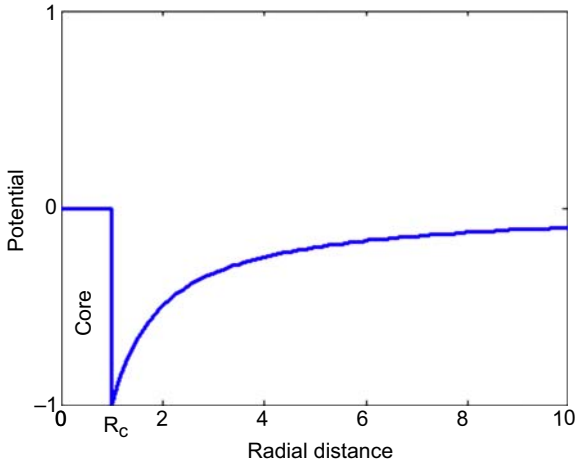
Figure 4.9 Core repulsion potential for the valence electrons.

The potential introduced by Ashcroft is called a pseudopotential to repel the electrons from the core region. By pursuing these corrections, more faithful correspondence can be constructed with experimental results. We do not pursue those here; the literature and books can be consulted for further results.

The tight binding approximation is described in Supplement B. Here a simple example is considered by adopting only a single s wave electron orbital basis. The wave function is a summation over the single electron atomic wave function $\psi_s(\vec{r})$ has the Bloch wave function has the form $(M = 1)$

$$\Phi_{\vec{k}}(\vec{r}) = a_{\vec{k}}\psi_s(\vec{r}). \tag{4.41}$$

The TBM, Eqn (4.B13), reduces to a single equation of the form

$$\left(E_{\vec{k}} - E_s\right)a_{\vec{k}} = \left(\sum_{n=1}^{NN} \xi_{ss}(\vec{R}_n)e^{i\vec{k}\cdot\vec{R}_n}\right)a_{\vec{k}}. \tag{4.42}$$

The wave function overlap integral is restricted to nearest neighbor (NN) sites as shown by a sum over nearest neighbors; the positions $\vec{R}_n$ are atom displacement positions relative to the central atom; the structure function defines the electron wave vector dependence of the band structure, defined by

$$f\left(\vec{k}\right) = \sum_{n=1}^{NN} e^{i\vec{k}\cdot\vec{R}_n}. \tag{4.43}$$

Its form is specific to the symmetry of the lattice. Consider two cases: an SC lattice and an FCC lattice. The SC lattice has six nearest neighbors, $(\pm a, \pm a, \pm a)$, and the

FCC lattice has 12 nearest neighbors, $(\pm a, \pm a, 0)$, $(\pm a, 0, \pm a)$, $(0, \pm a, \pm a)$. The single energy band function is

$$E_{\vec{k}}(\mathrm{SC}) = E_\mathrm{s} - 2\xi_\mathrm{ss}(\cos{(k_x a)} + \cos{(k_y a)} + \cos{(k_z a)}), \qquad (4.44)$$

$$E_{\vec{k}}(\mathrm{FCC}) = E_\mathrm{s} - 2\xi_\mathrm{ss}(\cos{(k_x a + k_y a)} + \cos{(k_x a - k_y a)} + \cos{(k_y a + k_z a)}$$
$$+ \cos{(k_y a - k_z a)} + \cos{(k_x a + k_z a)} + \cos{(k_x a - k_z a)}).$$
$$(4.45)$$

The band structure for the two geometries is plotted in Figure 4.10. The height of the band is adjustable by modifying the overlap coefficient ξ_ss. The itinerary chosen depends on the lattice symmetry. The wave vector follows the line connecting the two endpoints. The SC lattice itinerary ΓXMRΓ shows the shape of the energy eigenvalues along one path. The width of each segment between two symmetry points is the length of the wave vector's path for that segment. The itinerary WKΓLUXΓ of the FCC lattice has more segments as it captures all of the high-symmetry points.

The noble metals copper, silver, and gold are commonly a material component in nanophotonic systems because of their useful optical properties. The noble metals appear in the IB column of the periodic table and they are grouped together with so-called transition metals, which have d shell, s shell, and p shell valence electrons. Note that the noble metals form an exception to Hund's rule: the d state is not closed and one electron instead occupies an s state in the $n + 1$ shell. The d state wave functions are more tightly bound to the nucleus than the open shell s state/p state wave functions, which means that in the tight binding calculations, d state bands are narrower than sp state bands. The electronic density of states (DOS) illustrated in Figure 4.11 depicts contributions from sp and d states. The Fermi level separates the filled from the unfilled states at zero temperature. As discussed earlier, the Pauli exclusion principle prohibits the occupation of any state by more than two electrons. In the parlance of chemistry, the states having lower energies are *bonding* orbitals for
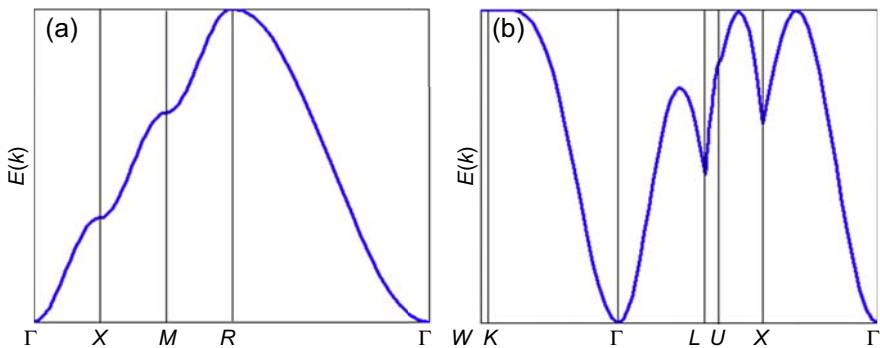
Figure 4.10 Band structure for a single band approximation: (a) SC lattice and (b) FCC lattice. The width of the band is determined by the value of the overlap coefficient ξ_ss.
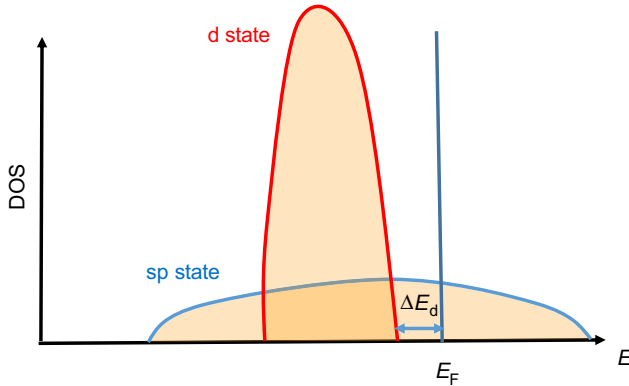
Figure 4.11 Schematic of the electronic sp and d band DOS for transition metals. The Fermi level lies within the s band and the top of the d band DOS is indicated by the energy ΔE_{d}.

the sp and d states, and the *antibonding* orbitals are found at the higher energies. In other words, the states below the Fermi level are bonding states.

Copper, silver, and gold are monovalent metals forming an FCC lattice. The individual atoms have filled d shells and one electron in the s shell (Cu(3d^{10}4s^1), Ag(4d^{10}5s^1), Au(5d^{10}6s^1)). The d band and sp bands overlap in the metals, as indicated in Figure 4.11. The d states are more tightly bound to the nucleus; hence, their electronic DOS is narrower. Furthermore, because there are more electronic states for the d orbitals, the DOS is higher than the sp state band DOS.

For the noble metals, *intraband* transitions (sp state to sp state) are responsible for the free electron-like behavior using an electron mass close to the free electron value. At long wavelengths, the intraband transitions near the Fermi level dominate the optical properties. The *interband* d state to sp state transitions contribute to the optical properties when the photon energy can promote d band electrons to an unoccupied state above the Fermi level, $\hbar\omega > \Delta E_{\mathrm{d}}$ (indicated in Figure 4.11). As a consequence, there will be nonfree electron contributions to the dielectric function. The interband transitions contribute to an increase in the absorption losses; the free electron (i.e., Drude) model predicts that the imaginary contribution to the dielectric function decreases as the inverse cube of the photon energy. The free electron behavior expected of intraband transitions is apparent for the noble metals for photon energies below 2 eV. The value of the imaginary part of the dielectric function, Im$\{\varepsilon(\hbar\omega)\}$, plotted for the noble metals in Figure 4.12, increases as the interband transitions contribute to electron dynamics, which is approximately 2 eV for Cu and Au and 3.7 eV for Ag.

4.4   Semiconductors

In the context of electronic band structure, a semiconductor is characterized by a filled valence band and empty conduction band; the two bands are separated from one
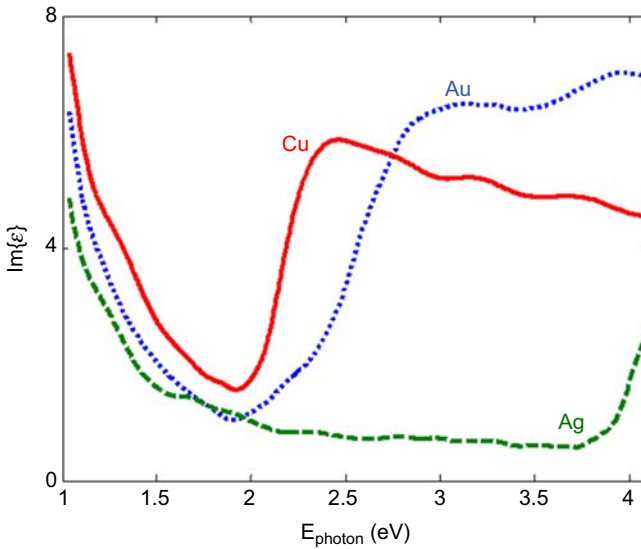
Figure 4.12 The imaginary portion of the dielectric for the noble metals (Cu, Ag, Au) as a function of photon energy.

another by an energy bandgap (E_g). There is no typical bandgap energy for a semiconductor material. As the bandgap approaches zero, the material exhibits properties of semimetals. As the bandgap becomes larger, the material approaches the properties of an insulator. The room temperature conductivity of a semiconductor is typically in the range of 10^{-8}–10^3 $(\Omega\,m)^{-1}$. This range distinguishes a semiconductor from an insulator, such as diamond from a semiconductor such as Si or Ge or a semimetal, such as Bi or graphite. Below the lower end of the range, the material is an insulator and above it the material is classified as a semimetal (conductivity range 10^3–10^6 $(\Omega\,m)^{-1}$) or a metal.

Electrical conductivity is one measure that can be used to classify a material. For a semiconductor, the electrical conductivity of a semiconductor is determined by the carrier density in the conduction band, which is proportional to the Boltzmann distribution function of the bandgap energy and the temperature T: $n \propto \exp(-E_g/k_B T)$, where k_B is the Boltzmann constant. Typical electron concentrations for semiconductors are 10^4–10^{14}/cm^3, for metals it is 10^{22}/cm^3, for semimetals it is 10^{18}/cm^3, and for insulators it is less than 10^4/cm^3.

Table 4.4, containing semiconductor bandgap properties, illustrates the range that is available from a selection of elemental and binary compound materials. The elemental solids Si and Ge are Group IVA semiconductors, simply referred to as Group IV; the Group IV carbon allotrope diamond is also included for comparison. Diamond with a large bandgap is classified as an insulator, but its other allotropes are technologically interesting in nanophotonic's devices. The binary semiconductors are drawn from Groups III−V and II−VI of the periodic table. Table 4.4 also lists the crystal structure, the classification of the material as a direct (D) or indirect (I) bandgap material, and the bandgap.

Table 4.4 **Selected material bandgaps**

Element groups	Material	Crystal structure	Bandgap type	Bandgap (eV)
IV	C	D	I	5.4
IV	Si	D	I	1.12
IV	Ge	D	I	0.66
III–V	GaAs	ZB	D	1.42
III–V	InAs	ZB	D	0.35
III–V	InP	ZB	D	1.34
III–V	InSb	ZB	D	0.17
III–V	AlAs	ZB	I	2.17
III–V	GaN	ZB/W	D	3.2/3.39
III–V	GaP	ZB	I	2.26
II–VI	ZnS	ZB/W	D	3.534/3.91
II–VI	ZnSe	ZB	D	2.82
II–VI	ZnTe	ZB	D	2.24

Crystal structure: D, diamond; ZB, zincblende; W, wurtzite. Bandgap type: I, indirect; D, direct.

There are three crystal structures listed in Table 4.4. In the diamond lattice structure, the atoms have four nearest neighbors; the diamond lattice is a non-Bravais lattice type that is related to the FCC lattice with two atoms per unit cell. One atom pair occupies the FCC lattice site and the second is displaced by a quarter of the edge length a in the (111) direction (i.e., $\frac{1}{4}a(\widehat{x} + \widehat{y} + \widehat{z})$). The zincblende structure denotes the periodicity of binary compound materials; it is equivalent to the diamond structure with the replacement of the atom pair by different elements (e.g., Ga and As). The diamond and zincblende structures are shown in Figure 4.13(a). The wurtzite structure is a non-Bravais lattice that has the symmetry of the hexagonal lattice with two atoms per lattice site. The wurtzite structure is illustrated in Figure 4.13(b).

An empirical TBM for calculating semiconductor band structures with sp^3 covalent bonding and an additional excited s state added (the so-called sp^3s* model) is discussed in Supplement C. The method restricts overlap integral contributions to nearest-neighbor atoms and uses on-site energy eigenvalues obtained from free atom orbital energies (i.e., only weakly dependent on the local environment and experimentally derived overlap parameters). Selected electronic band structures of materials with a diamond or with a zincblende lattice are shown in Figure 4.14. The "simple" TBM is able to capture the main features of indirect and direct bandgap semiconductors; a discussion of the TBM is available in Supplement B with examples presented in Supplement C. It has been applied to Group IV elements and III–V and II–VI direct and indirect gap
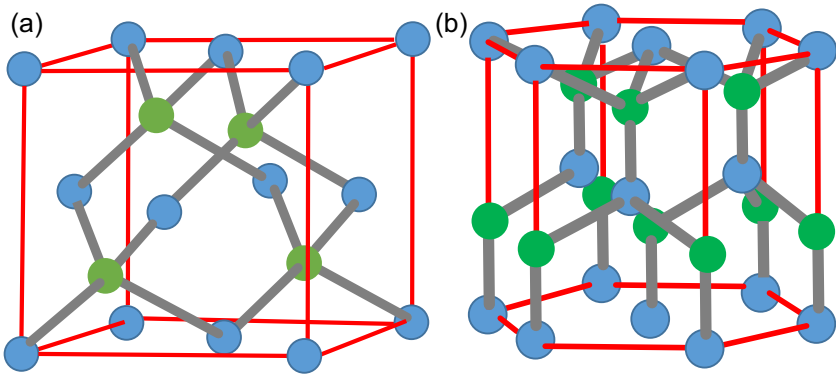
Figure 4.13 Zincblende/diamond structure and the wurtzite structure.
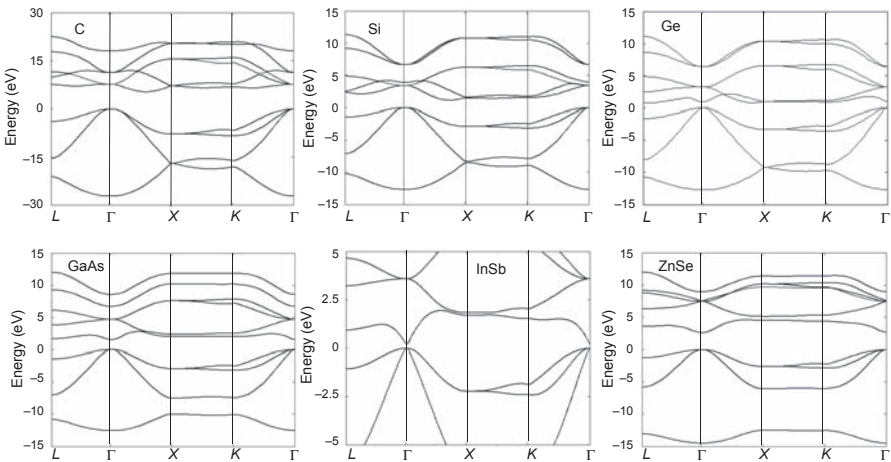
Figure 4.14 TBM based on five states per atom called the sp^{3}s*. See Supplement C for details. After Vogl et al. [1].

compound semiconductors. The model input is based on the available empirical data at selected symmetry points.

In each example, the top of the valence is at 0 eV. The bandgap is emphasized by a rectangular band. Silicon and germanium are indirect bandgap semiconductors. The featured band structures miss several details that can be captured by extensions. One feature of multielectron interactions is spin−orbit coupling, which is ignored in the images in Figure 4.14. The placement of the bands with spin−orbit coupling alters the valence band characteristics.

4.4.1 Doping

Semiconductors are noted for their ability to incorporate doping impurities into their lattice. The electrical and optical properties of semiconductors can be modified by

doping with atoms. Impurities can either donate an electron to the conduction band, so called donors create n-doped materials, or accept an electron from the valence band, which are thereby dubbed acceptors. Acceptors create p-doped materials with holes in the valence band left by the captured electrons. Modern electronics nanotechnology places p- and n-doped silicon to form a pn junction interface to fabricate diodes or transistors. However, pn junctions are also important for photonic devices, such as solar cells, semiconductor lasers, and photodetectors. Junctions made from the same material (e.g., silicon) are called *homojunctions*. Junctions made from different materials (e.g., alloys of AlGaAs) are called *heterojunctions*.

4.4.2 Group IV

Group IV elements have electronic orbitals that can be expressed in the form $[X]$ ns^2np^2, $n = 2, 3, 4$. The X in square brackets contains the closed electronic shells for the element. The diamond lattice has tetrahedral coordinated atoms, which suggests an obvious sp^3 hybridization of the valence electrons to form covalent bonds.

4.4.3 Carbon [He]$2s^2 2p^2$
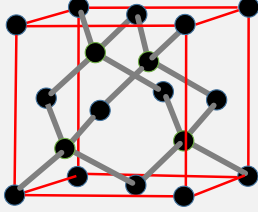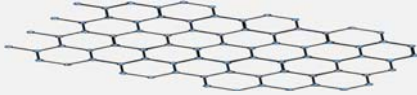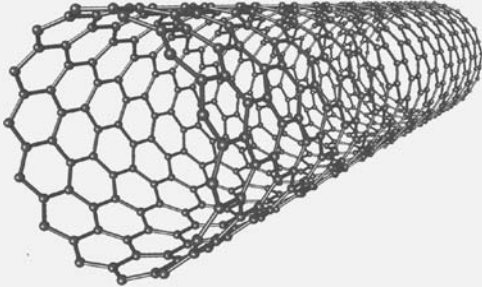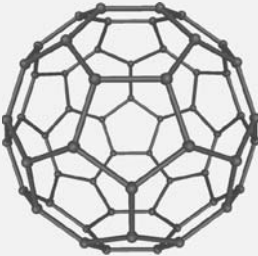
The building blocks of life on Earth owe their existence to the abundance of carbon. Carbon is a versatile and abundant material with four carbon *allotropes* widely mentioned in the nanotechnology literature: diamond, graphene or graphite, Buckminster fullerenes or simply Buckyballs, and carbon nanotubes. The list in Table 4.5 displays prominent carbon allotropes and their dimensionality. The table demonstrates some of the complex structures that can be formed that span all of the dimensions from extended three-dimensional crystals to zero-dimensional molecules.

4.4.4 Diamond

Diamond is the hardest natural material on earth. Synthetic diamonds are made on an industrial scale for use as abrasives. In nature, the diamond phase is formed at high temperatures and high pressures. The diamond synthesis process involves using specially designed anvils to achieve the growth conditions required by the carbon phase diagram. Carbon atoms are covalently bonded with four other carbon atoms in a network that forms a crystalline network. The diamond lattice with two atoms per unit cell is related to the Bravais FCC lattice. The lattice is named the diamond lattice.

The electronic bonds between the carbon atoms are formed from sp^3 hybrid orbitals. The combination of four atomic orbitals have the tetrahedral symmetry of the bond directions for the diamond lattice. The diamond band structure in Figure 4.14 has a large bandgap (5.5 eV) separating the fourth and fifth bands. Surprisingly, although diamond is an electrical insulator, it also has excellent thermal conductivity. This property has promoted diamond thin film applications as a heat sink layer to help cool electronic and photonic devices. Diamond films can be synthesized using chemical vapor deposition (CVD). Despite its large bandgap, diamond can be doped with

Table 4.5 **Four carbon allotropes**

Dimensionality	Allotrope name	Image
3	Diamond, graphite	
2	Graphene	
1	Carbon nanotubes	
0	Buckminster fullerenes	

Images created by Michael Ströck.

impurities to endow it with semiconductor electrical properties. It can be p doped using boron impurities that are introduced during the deposition process. On the other hand, n doping using nitrogen is problematic because the donor doping level is deep compared with thermal energy, and few electrons are released from the trap state and excited into the conduction band.

4.4.5   Graphene

Graphite is a completely different structure in comparison with diamond. The carbon atoms have three nearest neighbors and align themselves in a two-dimensional sheet forming a honeycomb lattice. In this material, the electrons hybridize into the planar symmetry of sp^2 orbitals. The sheets stack up to form a three-dimensional structure called a hexagonal lattice. As the image in Table 4.2 suggests, the stacked layers are weakly bonded and can be slid over one another by applying a shear stress. The stacked layers form a very soft material that has been used for millennia for writing. The word graphite is derived from the Greek meaning "writing stone." Graphite powders also form excellent dry lubricants and have a property called superlubricity that is a manifestation of the sheets sliding effortlessly over one another after local bonds between them have been broken. Moreover, graphite has a high electrical conductivity because of the unbonded valence electron; this makes it useful as an electrode material in arc lamps.

The isolation of the two-dimensional allotrope of carbon graphene was first reported in 2004 by Andre Geim and Konstantin Novoselov. The discoverers isolated a single layer from graphite using sticky tape and transferred the layer to a substrate in their extensive studies of its physical properties. They received the Nobel Prize in 2010 for the discovery. In earlier attempts at single-layer graphene synthesis, it was electronically bonded to a substrate, which altered its electrical properties. Single-layer graphene has several unusual properties including an electrical conductivity 10 times higher than silver and mechanical strength 100 times higher than steel.

Electrically graphene is a semimetal with a zero bandgap between the valence and conduction bands. The effective mass of the electrons inferred from the band structure is zero, and they are described by a two-dimensional analogue of the Dirac equation, rather than the usual Schrödinger equation. In other words, the energy-momentum dispersion relation is linear for graphene converging to so-called Dirac points at zero momentum. This is in contrast to the quadratic energy-momentum dispersion relationship expected from Schrödinger's equation.

Graphene has three nearest neighbors, $120°$ bond angles, and a lattice constant $a = 0.246$ nm. This makes graphene a prototypical example of a material with sp^2 covalent bonding of the valence atomic orbitals. These covalent bonds make single sheets mechanically strong. However, stacking graphene sheets builds the graphite solid that is familiarly used in pencils because of its very soft nature. This property is an indication that the bonding between graphene layers is a weak bond using the unfilled valence p_z orbital and van der Waals forces. The unfilled p_z orbital is also responsible for the unusual electronic properties of a single sheet of graphene.

Graphene has the geometry of a honeycomb lattice structure as shown in Figure 4.15. The honeycomb lattice is related to the previously discussed triangular lattice in the following way. There are two sets of circles in Figure 4.15 distinguished by different shading. Each set of circles form a triangular lattice with primitive lattice unit vectors $\vec{a}_1$ and $\vec{a}_2$. The second triangular lattice is displaced from the first by a displacement vector $\vec{d}_1$. The length of the primitive unit vectors is a and
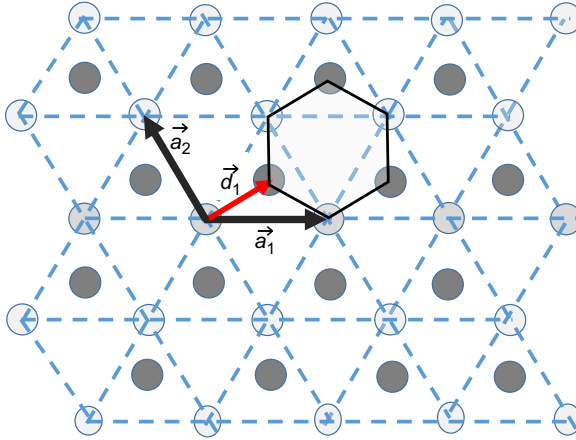
Figure 4.15 Graphene honeycomb lattice structure. Primitive cell lattice vectors $\{\vec{a}_1, \vec{a}_2\}$ of the triangular lattice are drawn and the displacement vector, $\vec{d}_1$, for the second lattice is also drawn.

$\left|\vec{d}_1\right| = a/\sqrt{3}$. A hexagon is drawn in the figure to emphasize that the lattice sites are connected to form a set of contiguous hexagons; this is called a honeycomb lattice. Each atom has three nearest-neighbor atoms and bond angles are $120°$.

The band structure of graphene due to the p_z orbital has interesting and unusual properties. Using the TBM, the energy band structure derived in Supplement C has the following form (ΔV is the overlap integral contribution):

$$E_{\vec{k}\pm} = E_0 \pm \Delta V \sqrt{1 + 4\cos^2\left(\frac{k_x a}{2}\right) + 4\cos\left(\frac{k_x a}{2}\right)\cos\left(\frac{k_y \sqrt{3} a}{2}\right)}. \qquad (4.46)$$

Surface plots of the energy bands in Figure 4.16 demonstrate that graphene is a perfect semimetal with touching valence and conduction bands. The two bands touch at the six K points at the edge of the BZ and have intersecting dispersion curves at the Dirac point. Near a Dirac point, the energy is a linear function of the wave vector yielding back-to-back cones that touch at the point. This functional behavior is the same as expected for a free, massless Fermion particle solution of Dirac's equation; thus, the name *Dirac point* has been coined for the touching points. In contrast, the shape of the band edges of many three-dimensional materials, as seen for all cases shown in Figure 4.13, is parabolic.

At the Dirac points, the electronic DOS vanishes and the p_z states share one electron, to fill the band up to the Dirac point, which is the Fermi level. The (antibonding) states above the Dirac point are empty. These properties make graphene an ideal example of a semimetal.
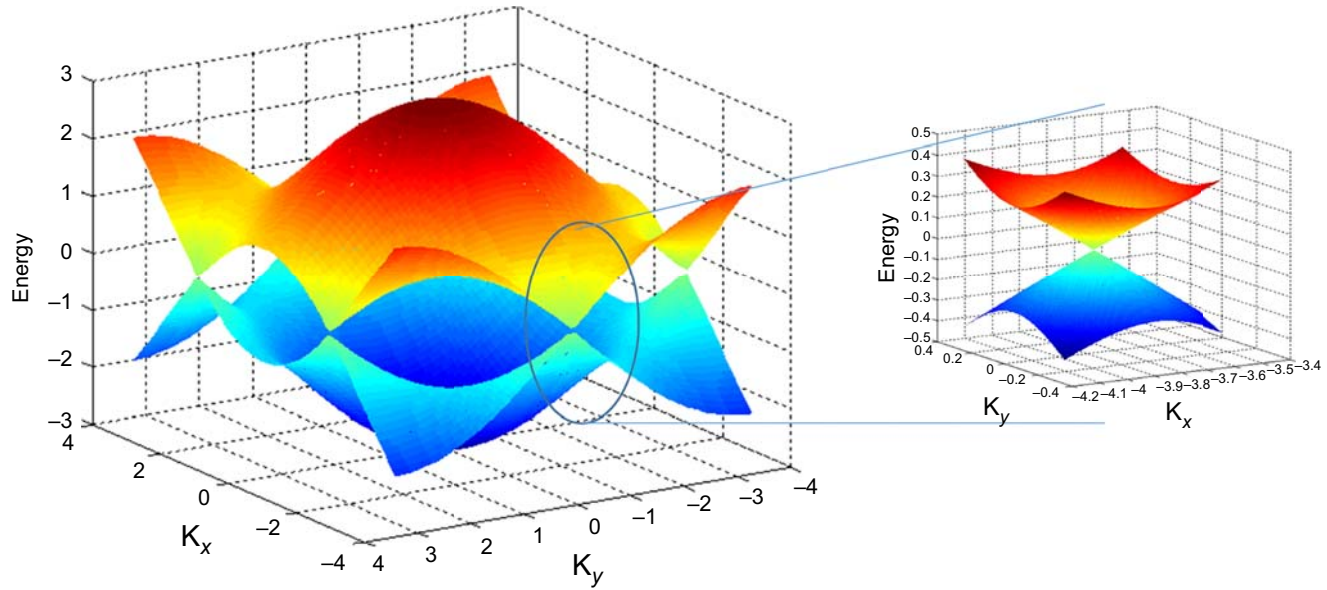
Figure 4.16 Graphene band structure. There are six Dirac points at K symmetry points on the edge of the BZ. A region around one of the points is enlarged to accentuate the cone-shaped band structure around the Dirac point.

4.4.6 Carbon nanotubes

Carbon nanotubes (CNTs) are structurally related to graphene by rolling the sheet into a cylindrical form. The strength and electrical properties of CNTs are also determined by the sp^2 bonds; thus, CNT properties are similar to those of graphene. Two important designations of this carbon allotrope are the single-walled carbon nanotube (SWCNT) and the multiwalled carbon nanotube (MWCNT). SWCNTs are closed tubes for which the length is defined by a displacement vector from one side of the cut to the other, $\vec{C}_{n,m} = n\vec{a}_1 + m\vec{a}_2$, as illustrated in Figure 4.17. The variables (n,m) are integers and $\vec{C}_{n,m}$ is called the chiral vector. The symmetry axis of the cylinder is perpendicular to the chiral vector. Two special cases are commonly defined: the zigzag SWCNTs defined by $(n,0)$ and the armchair SWCNT (n,n). The circumference of the CNTs is $C_{n,m} = a\sqrt{n^2 + nm + m^2}$. The length of the tubes can extend to more than 100 million lattice constants. The diameter of nanotubes are indeed on the order of nanometers in size.

The electronic properties of SWCNTs are sensitively dependent on the (n,m) variables. SWCNT band structure can be classified as semiconducting to metallic. When $|n - m| = 3p$, where p is an integer, the electronic properties are metallic; therefore, all armchair SWCNTs are metallic. For other values of $|n - m|$, the SWCNTs are semiconducting. Of course, because the aspect ratio of length to diameter is much greater than unity, the electronic and optical properties are highly anisotropic.

Two varieties of MWCNTs are widely reported. One variety is a nesting of SWCNTs around one another and may be called a Matryoshka MWCNT. Of course, the Matryoshka is the hollow Russian wooden doll that contains other smaller hollow wooden dolls inside of them. The second variety is the jellyroll MWCNT, in which a single graphene sheet is rolled in a spiral shape several times around a central axis.

Optical studies of CNTs reveal their electronic structure to characterize the samples. Measurements of optical absorption, photoluminescence, and Raman scattering spectra reveal the specific characteristics of a sample. The electronic DOS with a high narrow peak at discrete energies has the properties of other one-dimensional
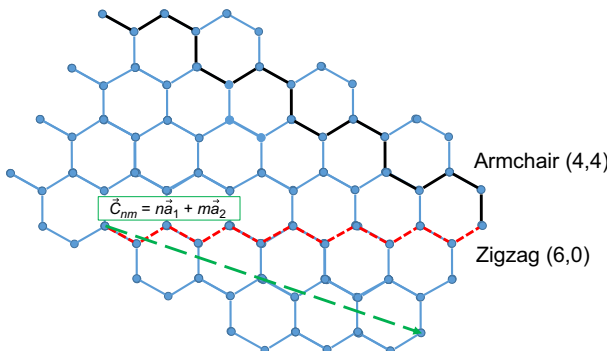
Figure 4.17 Cut lines that join to form two types of CNTs.

electronic systems. The electronic DOS of single CNTs are probed using scanning tunneling microscope measurements (see Chapter 6 for further details), which confirm their expected peaked features. CNTs have potential applications in electronics and photonics that make it an active field of study. Applications have been proposed to design and make photodetection, energy harvesting, and light-emitting devices. However, they have shown interesting nonlinear optical properties that can make them useful elements to mode-lock or Q-switch lasers.

CNTs are now commercially available. They are synthesized using different technologies: arc discharge, laser ablation, and CVD. The CVD method with its variants has been widely used to synthesize CNTs; the substrate is covered with metal nanopatches, which serve as a catalyst that seeds CNT growth. The diameter of the CNT is controlled by the nanopatch size. CNT growth is often randomly oriented from the substrate patches; however, plasma-enhanced CVD creates a forest of aligned CNTs that cover the substrate. These methods produce a distribution of diameters with either or both SWCNT and MWCNT types. Many applications of CNTs depend on purifying the samples to a single diameter and chirality. Several techniques have been proposed for sorting SWCNTs, including ultracentrifuge, gel filtration, and chromatography techniques. Individual nanotubes can align themselves into chains bound through van der Waals forces.

4.4.7 Buckminster fullerenes

Buckyballs, a shorter name for Buckminster fullerenes, are a cluster of covalently bonded carbon atoms with atoms placed at the vertices of pentagons and hexagons; the shape of a buckyball resembles a soccer ball. The name applies to the structure with 60 carbon atoms, but there are stable carbon-based structures, notably with 70 atoms. There are two bond lengths for C_{60}: the pentagons with single electron bonds have a bond length of 0.144 nm and there are three double electron bonds on the hexagons with a bond length of 0.14 nm. They were first identified in 1985 by using laser ablation of a graphite target and mass spectroscopy to identify the products. In 1996, Richard Smalley, James R. Heath, and Harold Kroto were awarded a Nobel Prize for their discovery. Buckyball applications have been impeded by the availability of large and inexpensive quantities of the product. The Krätschmer−Huffman method is an arc discharge heated using carbon rods in a helium atmosphere, which produces gram quantities of C_{60} in the carbon soot formed during the process.

In solid form, buckyballs bind together by van der Waals forces to form an FCC lattice. The weak binding makes a C_{60} solid a soft material at atmospheric pressures. The solid forms a direct bandgap semiconductor with a 1.9-eV bandgap. When metal atoms are placed at interstitial positions in a C_{60} solid, it becomes a conductor, and at low temperatures ($\sim 20-30$ K) the compounds A_3C_{60} (A = K, Rb, Cs as single elements or in combination) transition to a superconducting state.

4.4.8 Silicon [Ne]$3s^2 3p^2$

The rise of silicon is one of the most fascinating success stories chronicled in the history of technology. It has become the quintessential semiconductor material for

electronic applications. Because of silicon's technological preeminence, a worldwide scientific effort was supported over many decades to measure and catalog its fundamental properties. Table 4.6 lists a few electronic properties of silicon; however, the data does not really reveal silicon's exceptional technological importance. For instance, although it has an indirect bandgap, silicon has found commercial uses in photonic devices, such as photodetectors and solar cells.

Silicon is a major element available in the Earth's crust (see Figure 1.2). Its abundance and electronic properties make it an ideal material that launched the microelectronics revolution. The main raw material for growing silicon crystals is silica or amorphous silicon dioxide. It is a tetravalent chemical with four valence electrons. It has an atomic number of 14 with the inner shells (1s, 2s, 2p) filled. The 3s shell is also filled and the 3p shell has two electrons out of a possible six electrons to be filled. It crystallizes to a solid with diamond lattice symmetry. The simplified band structure shown in Figure 4.14 is based on the sp^3s^* model. Some essential features of silicon's band structure are illustrated in Figure 4.18.

Silicon is commercially available in three grades. Silicon is separated from oxygen at high temperatures using carbon electrodes in an electric arc furnace. The yield is *metallurgical-grade* silicon, which is approximately 95% pure. The electronics industry requires higher purity silicon for its products. *Solar-grade* (purity >99.9999%) and *electronic-grade* (purity 99.9999999% or nine 9s) silicon can be directly made from silica applying molten salt electrolysis.

Higher grade silicon can be refined from metallurgical-grade materials. The zone melting technique exploits the properties of the phase diagram where equilibrium coexists between a liquid and solid to collect impurities in the liquid region. The
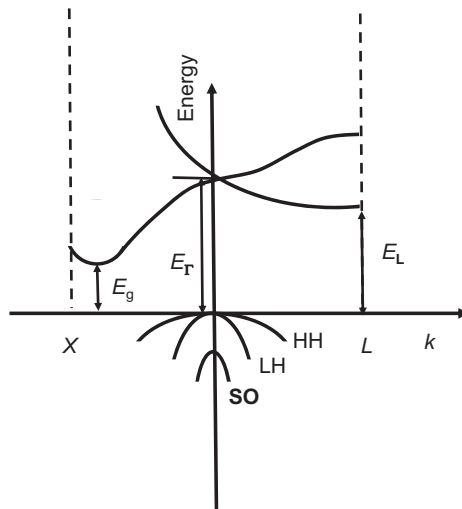
Figure 4.18 Selected features of silicon's band structure along two wave vector directions from the center (Γ) of the BZ: $X = <100>$ and $L = <111>$. The heavy-hole (HH), light-hole (LH), and split-off (SO) bands are the main features to the valence states. $E_\Gamma = 3.4$ eV, $E_L = 2.0$ eV, $E_{SO} = 0.044$ eV. Illustration after http://www.ioffe.rssi.ru/SVA/NSM/Semicond/.

process called zone melting refines metallurgical-grade silicon by locally melting a small region or *zone* of the silicon ingot and slowly moving the liquid section to one end of the solid. As the liquid part collects impurities, the melting temperature of that section is raised. The impurities in the liquid plug are transported to one end of the rod. The process is repeated to improve the purity. A localized zone can be heated by several methods, including direct flame heating and magnetic induction.

High-purity silicon is also formed using gaseous reactions from a silicon compound. Perhaps the best known uses silane (SiH_4) or silicon tetrachloride ($SiCl_4$) reactions to grow silicon films. High-concentration silane is a highly reactive material in air demanding safe handling procedures. In fluidized beds, silicon tetrachloride forms silicon films in a continuous process.

Silicon wafers are cut from silicon ingots or booles. Single crystalline ingots can be made in large diameters using the Czochralski process. In this process, a crystal is grown from a silicon melt by using a seed crystal that is slowly rotated and lifted out of the melt. The rotation reduces the defect density in the crystalline material, producing ingots of excellent quality. Dopants can also be directly added to the melt to grow p- or n-doped crystals.

Silicon can be p-doped using boron or gallium impurities or n-doped using phosphorus or arsenic. The uses of silicon span many technologies from electronics to photonics. This makes silicon and other semiconductors useful for designing and fabricating electronic and optoelectronic devices that use, say, diode action of a pn junction (Table 4.6).

4.4.9 Compound semiconductors

Binary compound semiconductors are made from many elements; for instance, SiC binds two group IV elements, but many can be recognized as elemental pairs taken from either side of the Group IV elements. Examples of widely used compound semiconductors are GaAs, InSb, and GaN from the III−V element groups and CdS, ZnSe, and HgTe from the II−VI groups. Many bind together in the zincblende (FCC diamond symmetry) or wurtzite (hexagonal symmetry) structures, as illustrated in Figure 4.13. The crystals have the tetrahedral nearest-neighbor coordination that is indicative of sp^3 hybrid orbital covalent binding. Atoms from different groups of the periodic table share electrons, which also endows the bonds with an ionic contribution. Figure 4.14 shows exemplar band structures for binary compounds using the sp^3s* TBM discussion in Supplement C.

Compound semiconductors, such as GaAs and alloys have found commercial applications as optoelectronic devices. The direct bandgap of many makes them more efficient light emitters. Fabrication methods can be adopted to alloy the elements to design a specific bandgap and confine the carriers and modify their DOS. In this section, we will confine our discussion to selected III−V compounds, which have widespread applications throughout modern technology.

Simplified electronic band structure diagrams for GaAs and AlAs are illustrated in Figure 4.19. GaAs has a direct bandgap and in AlAs the bandgap is indirect. These two materials are important for technological reasons because their lattices are only slightly different from one another, which is a requirement for growing epitaxial films, and

Table 4.6 **Electronic properties of selected semiconductors**

	Si	GaAs	AlAs	InSb	Units
Bandgap type	I	D	I	D	
Bandgap energy, E_g	1.12	1.42	2.17	0.17	eV
Electron effective mass	0.98	0.063	0.146	0.014	m_0
Hole effective mass	0.49	0.51	0.76	0.43	m_0
Static dielectric constant	11.7	12.9	10.06	16.8	
Refractive index $(E \sim E_g)$	3.42	3.3	2.86	4.0	
Lattice constant	0.5431	0.56533	0.56611	0.6479	nm

I and D denote the indirect and direct bandgap, respectively. m_0 is the mass of the electron in free space.
Data taken from http://www.ioffe.rssi.ru/SVA/NSM/Semicond/.

their bandgaps are different. Epitaxy refers to crystalline film growth over a crystalline substrate that is registered with the substrate lattice. An AlAs crystalline film deposited on GaAs exhibits very little strain as the film gets thicker because of the matched lattice constants. A diagram of the bandgap energy and lattice constants can be found in Figure 4.20. Binary semiconductor compounds are denoted as points in the diagram. The lines between two points show the change of the two parameters as the one atom is substituted for another.

A larger lattice mismatch when one film is grown over another can lead to defects at the interface in the bulk material that reduce optical and electronic device performance. Figure 4.21 is a simple depiction of film/substrate interfacial characteristics. Figure 4.21(a) is the ideal case of a coherent interface with minuscule lattice mismatch. Placing two materials with different lattice constants will result in lattice strain. The strain can be large enough that the lattices are incoherent and do not register their atoms and the lattice is incoherently grown over the substrate, as shown in Figure 4.21(b); the film lattice constant a_F is different from the substrate lattice constant a_S. The thin film lattice may choose to accommodate the substrate lattice by forming strong ionic/covalent bonds between the atoms on the two surfaces; this distorts the natural lattice constant of the film lattice, as in Figure 4.21(c). The lattice, so distorted, can be compressive ($a_F < a_S$) or under tension ($a_S < a_F$). However, when the film
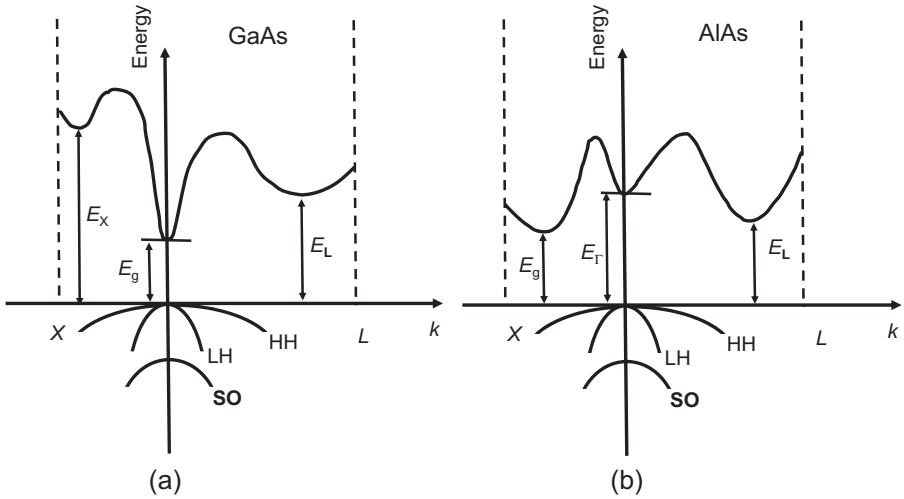
Figure 4.19 Illustrated band structures for GaAs and AlAs along two wave vector directions from the center (Γ) of the BZ: $X = <100>$ and $L = <111>$. The heavy-hole (HH), light-hole (LH), and split-off (SO) bands are main features of the valence states. $E_\Gamma = 3.4$ eV, $E_L = 2.0$ eV, $E_{SO} = 0.044$ eV.
Illustration after http://www.ioffe.rssi.ru/SVA/NSM/Semicond/.
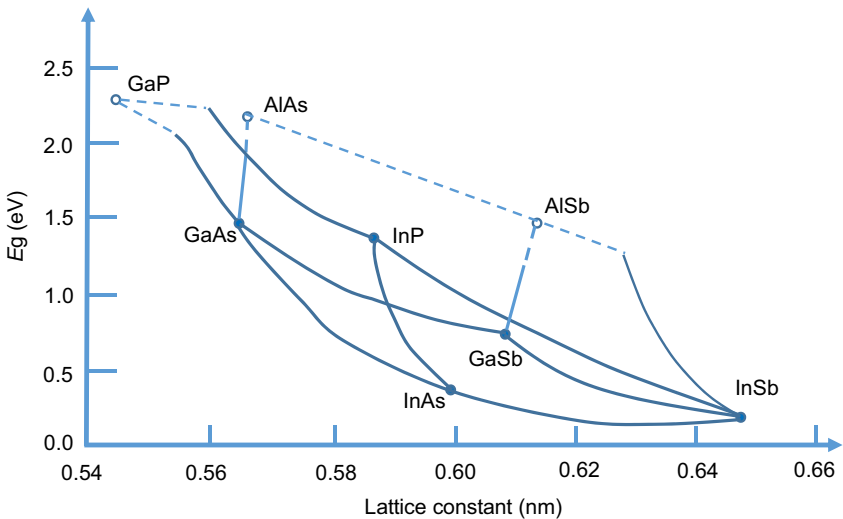
Figure 4.20 Bandgap energy (eV) vs. lattice constant (nm) for selected III−V compounds and their alloys. The line connecting two points represents the change of the variables as one atom is substituted for another. The dashed lines represent indirect bandgap regions. Open circles are indirect bandgap binary compounds.
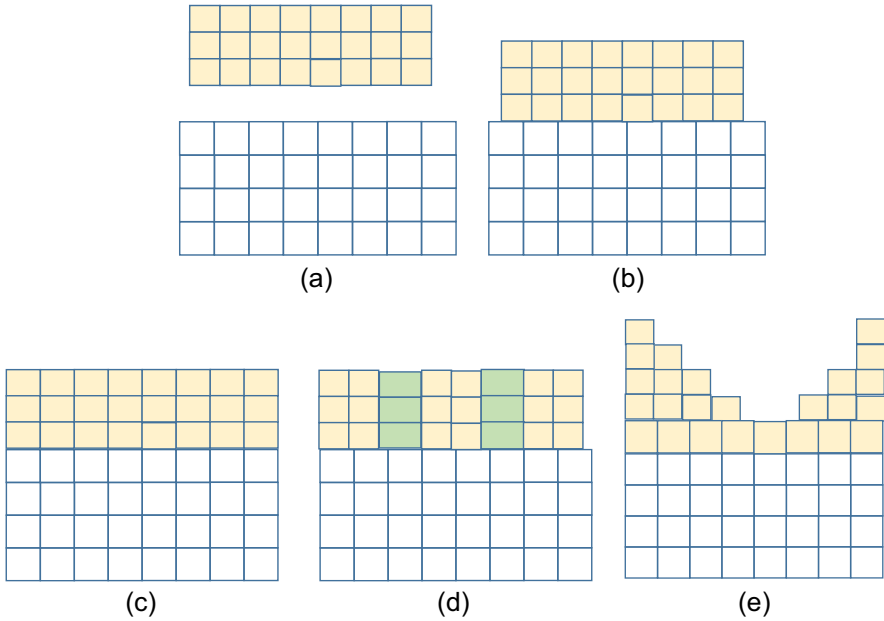
Figure 4.21 Lattice drawings showing different interfacial atomic arrangements. (a) The crystalline film lattice is coherently registered with the substrate lattice. (b) A nonregistered film with different lattice constants. (c) The film lattice is distorted to fit the substrate crystalline lattice constant. (d) The lattice remains misregistered but forms dislocations to relieve the strain. (e) Island formation by the Kransky–Krastanov high lattice mismatch mechanism.

thickness reaches a critical thickness, the film relieves the strain by forming disloca-tions. In some cases, the dislocations can extend through the entire film, as shown in Figure 4.21(d). Dislocations extending through the lattice are undesirable because they affect electron dynamics throughout the bulk. When dislocations are formed, it is desirable to confine them near the interface.

On the other hand, the effects of lattice strains can be harnessed to design high-performance devices in which built-in strain layers modify the electronic properties of heterostructures for very thin films (i.e., less than the critical thickness). Another interesting growth mode of films with large mismatched lattice constants is the forma-tion of islands; after a few monolayers are formed, the strain is relieved by atoms that migrate to self-assemble into islands distributed over the wafer surface. This is called the Kransky–Krastanov growth mechanism and is illustrated in Figure 4.21(e). The islands are small enough to confine the carriers in all three dimensions (i.e., quantum dots). For very high lattice mismatch, the islands can form directly without first form-ing a few monolayers, a mode called the Volmer–Weber growth mechanism.

Returning now to the bandgap diagram, the GaAs and AlAs are alloyed in the form of a ternary compound $Al_xGa_{1-x}As$; as Al content is increased, the bandgap of the material increases. The substitution concentration of a few properties are given in Table 4.7. The bandgap crosses from direct to indirect at $x = 0.45$. Bandgap engineer-ing and quantum confinement applications of AlGaAs require films that are narrower than the critical thickness.

Table 4.7 $Al_xGa_{1-x}As$ parameters

Bandgap energy, Γ point (eV)	$E_{g,\Gamma} = 1.424 + 1.247x$, $x < 0.45$ $E_{g,\Gamma} = 1.424 + 1.247x + 1.147$ $\quad (x - 0.45)^2$, $x > 0.45$
Bandgap energy, L point (eV)	$E_{g,L} = 1.708 + 0.642x$
Bandgap energy, X point (eV)	$E_{g,X} = 1.900 + 0.125x + 0.143x^2$
Lattice constant (nm)	$a = 0.56533 + 0.00078x$
Effective electron mass (m_0)	$0.063 + 0.085x$
Heavy hole mass (m_0)	$0.62 + 0.085x$
Light hole mass (m_0)	$0.087 + 0.063x$
Effective hole mass (m_0)	$0.51 + 0.25x$
Static dielectric permittivity	$12.90 - 2.84x$

The linear relationship between composition and bandgap energy is called Vegard's law.

Quaternary compound semiconductors provide another degree of freedom in designing new devices. InP substrates are used to fabricate photonic devices, such as light-emitting diodes, laser diodes, and photodetectors. The quaternary compound InGaAsP materials are used to design and fabricate photonic devices around 1.5 μm wavelength which is a dominant wavelength regime for fiber-optic communications systems. InP has a bandgap of 1.35 eV and a lattice constant of 0.58687 nm at room temperature. The ternary compound $Ga_{0.47}In_{0.53}As$ is lattice matched to InP with a bandgap of 0.75 eV and thin films can be grown over InP wafers. Furthermore, lattice-matched thin film growth of the quaternary compound $Ga_{0.47}In_{0.53}As_yP_{1-y}$ remains lattice matched to InP for all values of y allowing another degree of freedom in bandgap engineering photonic devices. The energy bandgap change ($E = 1.344 - 0.738\,y + 0.138y^2$ eV) can be used to refine designs and functionality of multiple thin film devices.

Problems

1. The TBM for a single atomic state of a diatom molecule (e.g., H_2) has a Hamiltonian with the form

$$H = \varepsilon \sum_{\alpha=1,2} |\alpha><\alpha| + V(|1><2| + |2><1|).$$

Using the wave function,

$$\Psi = \sum_{\alpha=1,2} c_\alpha |\alpha>.$$

    a. Write the coupled equations for the wave function amplitudes $\{c_1, c_2\}$.

    b. Solve the coupled equations to find the two energy eigenvalues.

    c. Identify the eigenvectors $\{c_1, c_2\}$ for each eigenvalue. The lowest eigenvalue corresponds to a bonding ground state (i.e., no node) for $V < 0$ and the wave function has no nodes.

2. For the diatom in problem 1, include a contribution of the wave function overlap γ.

    a. Write the coupled equations using the TBM with the new terms added.

    b. Solve the equations for the two eigenvalues.

    c. Find the corresponding eigenvectors.

3. A triatom cluster is assumed to be either in a ring or linear geometry.

    a. For both triatom geometries, extend the Hamiltonian with only one nearest-neighbor overlap parameter V.

    b. Find the energy eigenvalues for both cases.

    c. Find the eigenvectors of the wave function amplitudes and comment on the bonding and antibonding state energies.

4. An electron and a hole in GaAs combine to form a hydrogen-like atom are called an exciton.

    a. Calculate the ground-state binding energy and the ground-state Bohr radius of the three-dimensional exciton. Use the following parameters: electron and hole effective mass: $m_h = 0.47\ m_0$, and $m_e = 0.067\ m_0$, relative dielectric constant: $\varepsilon_r = 13$.

    b. Find the same quantities as in part (a) or the two-dimensional exciton using the same material parameters as in part (a).

    c. From the results of parts (a) and (b), comment on the effects of quantum well confinement on the exciton stability and size.

5. Hydrogen atom solutions can help us to understand the size of the wave functions. We explore that aspect in this problem.

    a. For the hydrogen atom, calculate the averages $<r>$ and $<r^2>$ defined by

$$< r^q > = \int_0^\infty r^q |R_{nl}(r)|^2 r^2 \mathrm{d}r, \quad q = 1, 2.$$

    Answers: $<r \geq a_B(3n^2 - l(l+1))$ and $<r^2> \frac{a_B^2 n^2}{2}(5n^2 + 1 - 3l(l+1))$. Note that the size is reduced for higher angular momentum.

    b. Plot the square of the radial wave functions ($n = 1$, 2, and 3 with available l values) given in Tables 4.A1 and 4.A2 to verify the trend in the average and variance of the radius from part (a).

6. An infinite linear chain consists of two types of atoms (A, B) separated by the lattice constant a; a segment is shown in the figure below.

Using TBM with lattice site energies and nearest-neighbor overlap parameters $\{E_A, E_B V_{AB}, \gamma_{AB}\}$,

    a. Deduce the wave function amplitude equations for each sublattice (A, B).

    b. Calculate the energy eigenvalues using Fourier methods and diagonalization of a 2×2 matrix.

    c. Plot the two bands and restrict your wave vector to the first BZ.

7. The phase velocity of a free electron in a solid is given by $\vec{v}_p = \hat{k}\omega/k$, where ω is the frequency that appears in the time-dependent wave function. Calculate the electron group velocity $(\vec{v}_g = \vec{\nabla}_k\omega)$ and compare the result with the phase velocity.

8. Using the coefficients in Table 4.C2 calculate the band structure for C, Si, and GaAs by numerically evaluating the 10×10 matrix given in Table 4.C1. Compare with the results found in Figure 4.13.

Appendices

Supplement A: Quantum mechanical hydrogen atom

The hydrogen atom in its simple form is a nontrivial, solvable model of a system that can be applied to many different problems. The quantum numbers were introduced in the Materials chapter to demonstrate how they explain the properties of the elements. It is surprising that such a simple model provides a deep understanding of the chemical elements and the formation of molecules and solid-state symmetries. Moreover, the hydrogen atom provides a model to understand the binding of electron and hole carriers in semiconductors, called excitons, and the binding of carriers to impurities in a semiconductor. With that short exposition on the wide applicability of the quantum model, let us proceed to examine the hydrogen atom in greater detail.

Hydrogen atom in three dimensions

The hydrogen atom is a prototypical model that has found applications throughout science and technology. In materials, its solutions form the basis of our understanding of chemical properties organized in the periodic table and the hydrogen analogy gives insight into the binding of electrons and holes to form excitons and the binding of electrons to impurities in semiconductors.

We start by considering two charged particles with masses m_1 and m_2 and charges Ze and $-e$, where Z is an integer and e is the charge of the electron. The positions of the charges are $\vec{r}_1 = (x_1, y_1, z_1)$ and $\vec{r}_2 = (x_2, y_2, z_2)$. The Schrodinger equation for the two-particle system is

$$\left(-\frac{\hbar^2}{2m_1}\nabla_1^2 - \frac{\hbar^2}{2m_2}\nabla_2^2 + V(|\vec{r}_2 - \vec{r}_1|)\right)\psi = E\psi, \tag{4.A1}$$

where the Laplacian for each particle in Cartesian coordinates is

$$\nabla_\alpha^2 = \frac{\partial^2}{\partial x_\alpha^2} + \frac{\partial^2}{\partial y_\alpha^2} + \frac{\partial^2}{\partial z_\alpha^2} \quad \alpha = 1, 2$$

and the potential is given by

$$V(|\vec{r}_2 - \vec{r}_1|) = -\frac{Ze^2}{4\pi\varepsilon}\frac{1}{|\vec{r}_2 - \vec{r}_1|}. \tag{4.A2}$$

The parameter ε is the dielectric constant of the surrounding medium. Because the potential only depends on the relative coordinates, the equation can be separated into relative and center of mass coordinates

$$\vec{r} = \vec{r}_2 - \vec{r}_1 \quad \text{and} \quad \vec{R} = \frac{(m_1\vec{r}_1 + m_2\vec{r}_1)}{(m_1 + m_2)}. \tag{4.A3}$$

Using this transformation in Eqn (4.A1), we find using a subscript to denote the Laplacian in each coordinate system

$$\left(-\frac{\hbar^2}{2M}\nabla_R^2 - \frac{\hbar^2}{2m_r}\nabla_r^2 + V(r)\right)\psi = E\psi, \tag{4.A4}$$

where $M = m_1 + m_2$ is the total mass and $m_r = m_1 m_2/(m_1 + m_2)$ is the reduced mass. The wave function is separated into functions of the center of mass and relative coordinates,

$$\psi(R, r) = \psi_R(\vec{R})\psi_r(\vec{r}). \tag{4.A5}$$

The Schrodinger equation is separated into two equations

$$-\frac{\hbar^2}{2M}\nabla_R^2\psi_R(\vec{R}) = E_R\psi_R(\vec{R}), \tag{4.A6}$$

$$\left(-\frac{\hbar^2}{2m_r}\nabla_r^2 + V(r)\right)\psi_r(\vec{r}) = E_r\psi_r(\vec{r}). \tag{4.A7}$$

The first equation has the plane-wave solutions of a free particle. The wave functions are written as

$$\psi_R(\vec{R}) = e^{i\vec{K}\cdot\vec{R}}, \tag{4.A8}$$

where $\vec{K}$ is the center of mass wave vector and the energy of the plane-wave state is

$$E_R = \frac{\hbar^2}{2M}K^2. \tag{4.A9}$$

Equation (4.A7) for the center of mass dynamics requires further detailed analysis. Because of the symmetry of the potential, the natural choice for coordinates is the spherical coordinate system $\vec{r} = (r, \theta, \phi)$. When the Laplacian is expressed in spherical coordinates, Schrodinger's equation is

$$\left(-\frac{\hbar^2}{2m_r}\left(\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right) + V(r)\right)\psi_r(\vec{r})$$

$$= E_r\psi_r(\vec{r}). \tag{4.A10}$$

The angular coordinate contributions to the Laplacian are related to the angular momentum contribution to the particle dynamics. The angular momentum operator is defined as

$$L^2 = -\hbar^2 \left(\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial\phi^2} \right). \tag{4.A11}$$

The wave function is separated into a product of functions that depend on the radial and angular coordinates:

$$\psi_r(\vec{r}) = R(r)Y(\theta, \phi). \tag{4.A12}$$

The two parts of Schrodinger's equation are

$$\frac{1}{R(r)} \left(\left(\frac{d}{dr} r^2 \frac{d}{dr} \right) + \frac{2m_r}{\hbar^2} r^2 (E_r - V(r)) \right) R(r)$$

$$= -\frac{1}{Y(\theta, \phi)} \left(\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial\phi^2} \right) Y(\theta, \phi) \tag{4.A13}$$

There is a separation constant for the two sides, and we write this as two equations

$$\frac{1}{R(r)} \left(\left(\frac{d}{dr} r^2 \frac{d}{dr} \right) + \frac{2m_r}{\hbar^2} r^2 (E_r - V(r)) \right) R(r) = C_Y, \tag{4.A14}$$

$$\frac{1}{Y(\theta, \phi)} \left(\frac{1}{\sin\theta} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial\phi^2} \right) Y(\theta, \phi) = -C_Y. \tag{4.A15}$$

The function $Y(\theta, \phi)$ further separates the equations into a set of two ordinary differential equations using the definition $Y(\theta, \phi) = P(\theta)H(\phi)$,

$$\frac{1}{P(\theta)} \left(\frac{1}{\sin\theta} \frac{d}{d\theta} \sin\theta \frac{d}{d\theta} \right) P(\theta) + C_Y = \frac{1}{H(\phi)} \frac{d^2 H(\phi)}{d\phi^2}. \tag{4.A16}$$

First, consider the right-hand side, which is the equation for the azimuthal angle ϕ, and the function $H(\phi)$ is periodic in 2π (i.e., $H(\phi) = H(\phi + 2\pi)$). Hence, the separation constant is taken as $-m^2$,

$$\frac{d^2 H(\phi)}{d\phi^2} = -m^2 H(\phi). \tag{4.A17}$$

The solutions of Eqn (4.A17) are

$$H(\phi) = e^{im\phi}, \quad m = 0, \pm 1, \pm 2, \ldots \tag{4.A18}$$

The equation for the polar angle θ is

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\sin\theta\frac{dG(\theta)}{d\theta} - \frac{m^2}{\sin^2\theta}G(\theta) = -C_Y G(\theta). \tag{4.A19}$$

This equation is the associated Legendre differential equation in canonical form when the new variable $s = \cos\theta$ is defined. The solutions are finite and normalizable when the coefficient on the right-hand side is $C_Y = l(l+1)$. Equation (4.A19) is written as

$$\frac{d}{ds}\left(1-s^2\right)\frac{dG(s)}{ds} - \frac{m^2}{1-s^2}G(\theta) = -l(l+1)G(\theta). \tag{4.A20}$$

This is the differential equation for associated Legendre polynomials. It has solutions for $l = 0,\ 1,\ 2,\dots$ and the integer m is also restricted to the values $m = -l,\ -l+1,\dots,\ l$. The solutions are associated Legendre polynomials,

$$G(\theta) = P_l^m(\cos\theta). \tag{4.A21}$$

The solutions of Eqn (4.A20) form a set of orthogonal functions for different quantum numbers called spherical harmonics, which are written as

$$Y_l^m(\cos\theta,\phi) = (-1)^m\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos\theta)e^{im\phi}. \tag{4.A22}$$

The numerical factors are chosen so that the integral is normalized

$$\iint Y_l^m(\cos\theta,\phi)Y_{l'}^{m'*}(\cos\theta,\phi)d\Omega = \delta_{ll'}\delta_{mm'}. \tag{4.A23}$$

The spherical harmonics are closely connected with the angular momentum operator defined by Eqn (4.A11). The z component of the angular momentum operator is

$$L_z = \frac{\hbar}{i}\left(\frac{\partial}{\partial\phi}\right). \tag{4.A24}$$

The angular momentum operators satisfy the following relations when operating on the spherical harmonics:

$$L^2 Y_l^m(\cos\theta,\phi) = \hbar^2 l(l+1)Y_l^m(\cos\theta,\phi),\quad L_z Y_l^m(\cos\theta,\phi) = m\hbar Y_l^m(\cos\theta,\phi). \tag{4.A25}$$

In other words, the values (l, m) partially label the angular momentum properties of the electronic state wave function.

The differential equation for the radial contribution to the wave function is

$$\left(\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr} + \frac{2m_r}{\hbar^2}\left(E_r + \frac{e^2}{4\pi\varepsilon_0 r}\right) - \frac{l(l+1)}{r^2}\right)R(r) = 0. \tag{4.A26}$$

Using the definition of the length $a_0 = 4\pi\varepsilon_0\hbar^2/[e^2 m_r]$ and an energy constant $E_0 = e^4 m_r/[2(4\pi\varepsilon_0)^2\hbar^2]$, the radius and energy can be scaled $r = a_0 q$ and $E_r = E_0 E_s$ with the result

$$\left(\frac{d^2}{dq^2} + \frac{2}{q}\frac{d}{dq} + E_s + \frac{2}{q} - \frac{l(l+1)}{q^2}\right)R(q) = 0. \tag{4.A27}$$

This is a form of Laguerre's differential equation and can be cast into a new form by the transformation $F(q) = u(q)/q$,

$$\left(\frac{d^2}{dq^2} + E_s + \frac{2}{q} - \frac{l(l+1)}{q^2}\right)u(q) = 0. \tag{4.A28}$$

The convergent solutions of the differential equation are found for negative and discrete values of the energy given by

$$E_s = -\frac{1}{n^2} , \quad n = 1, 2, \ldots \tag{4.A29}$$

The energy is the familiar Rydberg series of energy states and it is related to the bound-state eigenvalues of the differential equation. The symbol n is called the principal quantum number; the angular momentum integer eigenvalue l is restricted to $l < n$. The energy eigenvalues are degenerate with a degeneracy $2(2l + 1)$. The factor of 2 accounts for the intrinsic pin $1/2$ $(\sigma = \pm\frac{1}{2})$ of the electrons; the electrons are part of a family of particles called Fermions with half-integer intrinsic spin. The Pauli exclusion principle prohibits the occupation of a quantum state (n, l, m, σ) by more than one electron.

The eigenfunctions with subscripts added to indicate the quantum state are

$$R_{nl}(r) = \mathcal{N}_{nl}\left(\frac{2r}{na_0}\right)^l L_{n-l-1}^{2l+1}\left(\frac{2r}{na_0}\right)e^{-r/na_0}. \tag{4.A30}$$

The functions $L_{n-l-1}^{2l+1}(x)$ are Laguerre polynomials and the normalization constant

$$\mathcal{N}_{nl} = \sqrt{\left(\frac{2}{na_0}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]^3}}, \tag{4.A31}$$

is defined so that

$$\int_0^\infty |R_{nl}(r)|^2 r^2 \mathrm{d}r = 1. \tag{4.A32}$$

Wave functions and atomic orbitals

The treatment of the hydrogen atom has applications across the fields of physics, chemistry, and materials science. The organization of the elements in the periodic table is a testament to the application of atomic orbitals beyond the hydrogen atom.

Selected radial and angular wave functions are provided in Tables 4.A1 and 4.A2. The definition of the spherical harmonics in Eqn (4.A22) can be applied to deduce the angular wave functions for negative values of the quantum number m.

Table 4.A1 **Selected radial wave functions**

n	L	Radial wave function
1	0	$R_{10} = \left(\dfrac{2}{a_0^{\frac{3}{2}}}\right) \mathrm{e}^{-r/a_0}$
2	0	$R_{20} = \left(\dfrac{1}{\sqrt{2}a_0^{\frac{3}{2}}}\right)\left(1 - \dfrac{r}{2a_0}\right)\mathrm{e}^{-r/2a_0}$
2	1	$R_{21} = \left(\dfrac{1}{\sqrt{6}a_0^{\frac{3}{2}}}\right)\left(\dfrac{r}{2a_0}\right)\mathrm{e}^{-r/2a_0}$
3	0	$R_{30} = \left(\dfrac{2}{\sqrt{27}a_0^{\frac{3}{2}}}\right)\left(1 - \dfrac{2r}{3a_0} + \dfrac{2r^2}{27a_0^2}\right)\mathrm{e}^{-r/3a_0}$
3	1	$R_{31} = \left(\dfrac{8}{9\sqrt{6}a_0^{\frac{3}{2}}}\right)\left(1 - \dfrac{r}{6a_0}\right)\left(\dfrac{r}{3a_0}\right)\mathrm{e}^{-r/3a_0}$
3	2	$R_{32} = \left(\dfrac{4}{9\sqrt{30}a_0^{\frac{3}{2}}}\right)\left(\dfrac{r}{3a_0}\right)^2 \mathrm{e}^{-r/3a_0}$

Table 4.A2 Selected angular wave functions

l	M	Angular wave function
0	0	$Y_0^0(\theta, \phi) = \frac{1}{\sqrt{4\pi}}$
1	0	$Y_1^0(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos\theta$
1	1	$Y_1^1(\theta, \phi) = -\sqrt{\frac{3}{4\pi}} \sin\theta \mathrm{e}^{\mathrm{i}\phi}$
2	0	$Y_2^0(\theta, \phi) = \sqrt{\frac{5}{16\pi}}(3\cos^2\theta - 1)$
2	1	$Y_2^1(\theta, \phi) = -\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta \mathrm{e}^{\mathrm{i}\phi}$
2	2	$Y_2^2(\theta, \phi) = \sqrt{\frac{15}{32\pi}} \sin^2\theta \mathrm{e}^{\mathrm{i}2\phi}$

In atomic physics, a letter symbol is used to designate the orbital angular momentum state. The s state is $l = 0$, the p state is $l = 1$, the d state is $l = 2$, and additional states f, g, h,..., skipping the letter "j", correspond to l values in ascending order.

Hydrogen atom in two dimensions

Under conditions where the charges are confined to two dimensions, the hydrogen atom solutions are modified in important ways. For instance, this situation is relevant, in quantum-confined structures called quantum wells, where confinement of an electron or hole, which forms an exciton, freezes out one degree of freedom in the kinetic energy. The reduction of kinetic energy without affecting the Coulomb binding energy results in an increased binding energy of the exciton in two dimensions.

In two dimensions, the Hamiltonian in the center of mass coordinates is

$$\left(-\frac{\hbar^2}{2m_r} \left(\frac{1}{\rho} \frac{\partial}{\partial\rho} \rho \frac{\partial}{\partial\rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial\phi^2} \right) + V(\rho) \right) \psi_r(\vec{r}) = E_r \psi_r(\vec{r}). \tag{4.A33}$$

The wave function is separated into the radial and angular coordinates,

$$\psi_r(\vec{r}) = F(\rho)G(\phi). \tag{4.A34}$$

The function $G(\phi)$ satisfies the differential equation with the separation variable m^2,

$$\frac{\partial^2}{\partial\phi^2} G(\phi) = -m^2 G(\phi). \tag{4.A35}$$

The solutions are already familiar from the previous analysis,

$$G(\phi) = \mathrm{e}^{\mathrm{i}m\phi}. \tag{4.A36}$$

Defining the parameters $\alpha^2 = -2m_r E_r/\hbar^2$ and $\beta = -2m_r e^2/(4\pi\varepsilon\hbar^2)$, the differential equation for the radial coordinate is

$$\left(\frac{1}{\rho}\frac{\partial}{\partial\rho}\rho\frac{\partial}{\partial\rho} - \frac{m^2}{\rho^2} - \frac{\beta}{\rho} - \alpha^2\right)F(\rho) = 0. \tag{4.A37}$$

Using the coordinate transformation $u = 2\alpha\rho$ and introducing the new function $L(u)$,

$$F(\rho) = u^{|m|}L(u)\bar{e}^{-\alpha\rho}. \tag{4.A38}$$

This equation is cast into the form of the associated Laguerre differential equation

$$uL'' + \left(2|m| - 1 + \frac{1}{u}\right)L' + \left(-|m| - \frac{1}{2} - \frac{\beta}{2\alpha}\right)L = 0. \tag{4.A39}$$

Solutions that are finite at $u = 0$ are associated Laguerre polynomials $L_{n+|m|}^{2|m|+1}(u)$. The subscripted index is a non-negative integer and it satisfies the equation

$$n + |m| = |m| - \frac{1}{2} - \frac{\beta}{2\alpha}. \tag{4.A40}$$

The values of the principal quantum number are $n = 0, 1, 2,\ldots$ and the azimuthal angle quantum number is $|m| \leq n$. Rewriting Eqn (4.A40), the energy eigenvalues are determined as

$$E_r = -\frac{\hbar^2}{2m_r}\frac{\beta^2}{\left(n+\frac{1}{2}\right)^2} = -\frac{E_0}{(2n+1)^2}, \tag{4.A41}$$

where the ground-state energy is $E_0 = \frac{2m_r}{\hbar^2}\frac{e^4}{(4\pi\varepsilon)^2}$. In free space for the reduced mass equal to the free electron mass, the ground-state energy is $E_0 = 54.4$ eV. This is 4 times the magnitude of the hydrogen atom ground-state energy in three dimensions.

As mentioned in the introduction to this section, the higher binding energy of an electron to a hole when they are confined to two dimensions is to be expected because one degree of freedom of the kinetic energy has been frozen out while the same Coulomb potential attraction still acts on the electronic system.

Supplement B: Tight binding method

Lattice wave function

The TBM is one of the simplest models for calculating the band structure of solids and is based on the atomic structure of its elements. The Bravais lattice is defined by a set

of ionic position vectors, $\{\vec{R}\}$, which are defined by all linear combinations of the primitive basis vectors with integer coefficients as shown in Eqn (4.1) or (4.2). To streamline the notation, we drop the explicit dependence on the integers. When the Hamiltonian is a periodic operator (i.e., $H(\vec{r} + \vec{R}) = H(\vec{r})$), the wave function satisfies the Bloch theorem, which is expressed in the form

$$\Psi_{\vec{k}}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u_{\vec{k}}(\vec{r}) = \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}}\Phi_{\vec{k}}(\vec{r} - \vec{R}). \tag{4.B1}$$

The function $u_{\vec{k}}(\vec{r})$ is periodic for translations by a lattice vector (i.e., $u_{\vec{k}}(\vec{r} + \vec{R}) = u_{\vec{k}}(\vec{r})$). The wave functions under the summation $\Phi_{\vec{k}}(\vec{r} - \vec{R})$ consist of a linear combination of atomic orbital wave functions that are localized near the lattice position, $\vec{R}$. Bloch's theorem is based on the requirement that

$$\left|\Psi_{\vec{k}}(\vec{r} + \vec{R})\right|^2 = \left|\Psi_{\vec{k}}(\vec{r})\right|^2. \tag{4.B2}$$

In other words, the individual sites in a periodic lattice are indistinguishable. As mentioned previously, each site is represented as a linear combination of atomic orbital wave functions. The choice of wave functions is taken from the set of valence electron states, such as corresponding to s and p orbitals. For a representation of M atomic orbitals on a Bravais lattice, the local wave function is a linear combination of the basis wave functions, $\{\psi_\alpha(\vec{r}),\ \alpha = 1,...M\}$

$$\Phi_{\vec{k}}(\vec{r}) = \sum_{\alpha=1}^{M} a_{\alpha,\vec{k}}\psi_\alpha(\vec{r}). \tag{4.B3}$$

The set of amplitude coefficients $\{a_{\alpha,\vec{k}},\ \alpha = 1,...M\}$ use atomic wave functions that are orthonormal to one another; in other words, they satisfy the integral condition

$$\iiint \psi_\beta^*(\vec{r})\psi_\alpha(\vec{r})d^3\vec{r} = \delta_{\alpha\beta}, \quad \alpha,\beta = 1,...M. \tag{4.B4}$$

The set of unknown amplitude coefficients are determined by solving Schrödinger's equation in the form

$$E_{\vec{k}}\Psi_{\vec{k}}(\vec{r}) = H\Psi_{\vec{k}}(\vec{r}). \tag{4.B5}$$

$E_{\vec{k}}$ is the energy eigenvalue of the electron in the lattice.

The Hamiltonian operator consists of two contributions. One is composed of a sum over single atom contributions; each single atom Hamiltonian has a basis set of atomic

wave functions $\{\psi_\alpha(\vec{r})\}$. The single-atom Schrodinger equation for the atom centered at lattice position $\vec{R}$ is

$$H_{\text{atom},\vec{R}}\psi_\alpha(\vec{r}-\vec{R}) = \boldsymbol{E}_\alpha\psi_\alpha(\vec{r}-\vec{R}), \quad \alpha = 1,...M. \tag{4.B6}$$

The energy eigenvalues (for which the single-atom energy eigenvalue is emphasized by using a bold letter) $\boldsymbol{E}_\alpha$ are independent of the lattice site. The M atomic wave functions chosen as basis wave functions are chosen after making an assessment of the contributing valence electron states; they are determined from an examination of the open shell states binding the atoms together.

The second contribution to the Hamiltonian operator includes the deformation of the potential due to the neighboring atoms. The importance of the overlap can be gauged by looking at the potential difference, $\Delta V(\vec{r})$ between the single lattice site potential and the superposition of neighboring atomic potentials. The potential $\Delta V(\vec{r})$ is a periodic function on the lattice as illustrated by the shaded region in Figure 4.B1. The total Hamiltonian is explicitly written in terms of these two contributions,

$$H(\vec{r}) = H_{\text{atom}} + \Delta V(\vec{r}). \tag{4.B7}$$

where H_{atom} is a sum over all lattice sites, and

$$H_{\text{atom}} = \sum_{\vec{R}} H_{\text{atom},\vec{R}}. \tag{4.B8}$$
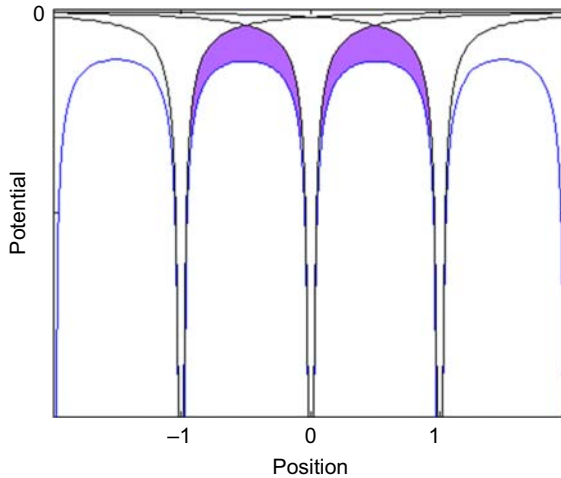
Figure 4.B1 Periodic atomic potential illustrated on a background of single atomic potentials. The difference between the two curves, shown as the shaded region, is the difference potential $\Delta V(\vec{r})$.

The band structure is calculated by calculating the eigenvalues from Schrodinger's equation (Eqn (4.B5)). Multiply both sides by the wave function of the atom labeled $\vec{R} = \vec{0}$: $\psi_\beta^*(\vec{r})\mathrm{d}^3\vec{r}$ and integrate over the crystal volume. The M coupled equations are written as

$$E_{\vec{k}} a_{\alpha,\vec{k}} \delta_{\alpha\beta} = \sum_{\alpha=1}^{M} \sum_{\vec{R}} \iiint \psi_\beta^*(\vec{r}) H_{\mathrm{atom}} \psi_\alpha(\vec{r} - \vec{R}) \mathrm{d}^3\vec{r} \mathrm{e}^{i\vec{k}\cdot\vec{R}} a_{\alpha,\vec{k}} + \sum_{\alpha=1}^{M}$$

$$\times \sum_{\vec{R}} \iiint \psi_\beta^*(\vec{r}) \Delta V(\vec{r}) \psi_\alpha(\vec{r} - \vec{R}) \mathrm{d}^3\vec{r} \mathrm{e}^{i\vec{k}\cdot\vec{R}} a_{\alpha,\vec{k}}. \tag{4.B9}$$

The sum over $\vec{R}$ is over all lattice sites. The energy on the left-hand side has a Kronecker delta as a result of assumed orthonormality among the set of atomic states. The integrals appearing on the right-hand side can be defined as coefficients

$$\iiint \psi_\beta^*(\vec{r}) H_{\mathrm{atom}} \psi_\alpha(\vec{r} - \vec{R}) \mathrm{d}^3\vec{r} = \gamma_{\beta\alpha}(\vec{R}). \tag{4.B10}$$

$$V_{\beta\alpha}(\vec{R}) = \iiint \psi_\beta^*(\vec{r}) \Delta V(\vec{r}) \psi_\alpha(\vec{r} - \vec{R}) \mathrm{d}^3\vec{r}. \tag{4.B11}$$

Note that the same site wave function overlap satisfies $\gamma_{\beta\alpha}(\vec{0}) = E_\beta \delta_{\beta\alpha}$, but in general, $\gamma_{\beta\alpha}(\vec{R}) \neq 0$ because of the overlap of neighboring wave functions. Note also that E_β is the energy deduced from the atomic Hamilitonian including neighbors; in general, it will be different from the single atom eigenvalue $\boldsymbol{E_\beta}$, which was introduced above. The set of coefficients $\{\gamma_{\beta\alpha}(\vec{R}), \; V_{\beta\alpha}(\vec{R})\}$ are assumed to be short ranged, often extending only to the nearest-neighbor sites. Equation (4.B9) is reduced to a set of M linear algebraic equations,

$$\left(E_{\vec{k}} - E_\beta\right) a_{\alpha,\vec{k}} \delta_{\alpha\beta} = \sum_{\alpha=1}^{M} \sum_{\vec{R} \neq \vec{0}} \gamma_{\beta\alpha}(\vec{R}) \mathrm{e}^{i\vec{k}\cdot\vec{R}} a_{\alpha,\vec{k}}$$

$$+ \sum_{\alpha=1}^{M} \sum_{\vec{R}} V_{\beta\alpha}(\vec{R}) \mathrm{e}^{i\vec{k}\cdot\vec{R}} a_{\alpha,\vec{k}}. \tag{4.B12}$$

This secular equation is the main TBM result. The set of coefficients $\{E_\alpha, \; V_{\beta\alpha}(\vec{R}), \; \gamma_{\beta\alpha}(\vec{R})\}$ are either calculated from assumed potentials or deduced using an empirical fit to available data. The energy eigenvalues, $E_{\vec{k}}$, are calculated using the coefficients as input parameters. For each value of $\vec{k}$, the energy eigenvalues are found from the determinant of the $M \times M$ matrix representation of Eqn (4.B12). The values of $\vec{k}$ are restricted to lie within the first BZ.

The TBM has wide application in electronic systems, including metal, semiconductor, and molecular systems. Because the potential of a many-electron system is difficult to capture, the TBM can be empirically applied using experimental data as input to determine the coefficients. The number of different coefficients is reduced by invoking point group operations, rotations, and mirror symmetries that leave the lattice invariant. So, for instance, restricting the coefficients to nearest neighbors, a lattice with NN nearest neighbors could have $M \times M \times NN$ coefficients. When the lattice possesses high symmetry, the same value is found for all nearest neighbors, reducing the number of contributions coefficients to NN, i.e., $V_{\beta\alpha}(\vec{R}) = V_{\beta\alpha}(NN)$ for $\vec{R}$ a nearest neighbor to the origin, 0, and on the same site they are denoted as $V_{\beta\alpha}(\vec{0})$, $\gamma_{\beta\alpha}(\vec{0}) = E_\beta \delta_{\alpha\beta}$. The last equality expresses the assumed orthogonality of the atomic wave functions.

For this case, the matrix can be simplified by defining the variables

$$D_{\alpha\beta} = -\left(E_\beta \delta_{\beta\alpha} + V_{\alpha\beta}(NN)\right)f\left(\vec{k}\right) + V_{\alpha\beta}(0),$$

$$f\left(\vec{k}\right) = \sum_{n,NN} e^{i\vec{k}\cdot\vec{R}},$$

$$\Delta E_\alpha = E_{\vec{k}} - E_\alpha - V_{\alpha\alpha}(0) - D_{\alpha\alpha}. \tag{4.B13}$$

Note that $f(\vec{k})$ is a sum over nearest neighbors. The last coefficient of $D_{\alpha\beta}$ is the on-site overlap integral including the perturbation potential. The $M \times M$ equation has the form

$$D = \begin{pmatrix} \Delta E_{11} & D_{12} & \cdots & D_{1M} \\ D_{21} & \Delta E_{22} & \cdots & D_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ D_{M1} & D_{M2} & \cdots & \Delta E_{MM} \end{pmatrix} \tag{4.B14}$$

Solving specific examples and extensions using TBM provides valuable experience in applying this method. The following two supplements present two specific cases that are relevant to semiconductors with non-Bravais lattice types and binary compounds. One case is graphene, which is a two-dimensional application of TBM with two sublattices. The second case is a three-dimensional model that may be applied to elemental and compound semiconductors. The overlap integral parameters are empirically extracted from available data.

Non-Bravais lattices

The central TBM result in Eqn (4.B12) can be adopted to treat many other important and useful semiconductor materials that have an additional complication; namely, they form a non-Bravais lattice. However, the lattice can be decomposed into sublattices

that are Bravais lattices. This means that all of the sites of the non-Bravais lattice can be reached by introducing an additional set of position vectors. For instance, consider two atoms per unit cell. One atom (labeled A) is positioned at the lattice sites of a Bravais lattice, which means all atoms are reached by integer coefficients of linear combinations of the primitive basis vectors. The second atom (labeled B) can be reached from any site on sublattice A by a displacement vector $\vec{d}$ (i.e., $\vec{R}_B = \vec{R}_A + \vec{d}$). The Hamiltonian incorporating both atom types is written as

$$H = \sum_{\vec{R}_A} H_{\text{atom},\vec{R}_A} + \sum_{\vec{R}_B} H_{\text{atom},\vec{R}_B} + \Delta V(\vec{r}). \tag{4.B15}$$

The potential $\Delta V(\vec{r})$ contains all of the interactions between A and B atoms. There are two sets of wave functions usually consisting of valence electron states: one for the A atoms, $\{\psi_{\alpha,A}(\vec{r}), \alpha = 1, \ldots M_A\}$, and one for the B atoms $\{\psi_{\alpha,B}(\vec{r}), \alpha = 1, \ldots M_B\}$ that describe the specific chemical properties of those atoms. The notation in Eqn (4.B12) is extended to incorporate the new different atom or lattice positions. The set of lattice positions covers all sublattices; for instance, for two sublattices the set includes $\{\vec{R}\} = \{\vec{R}_A, \vec{R}_B\}$. Using the atomic wave functions, the TBM can be separated into two expressions: one for the A atom sites,

$$\left(E_{\vec{k}} - E_{\beta A}\right) a_{A\alpha,\vec{k}} \delta_{\alpha\beta} = \sum_{\alpha=1}^{M_A} \sum_{\vec{R}_A \neq \vec{0}} E_{\alpha A} \gamma_{\beta\alpha}(\vec{R}_A) e^{i\vec{k}\cdot\vec{R}_A} a_{A\alpha,\vec{k}}$$
$$+ \sum_{\alpha=1}^{M_X} \sum_{\vec{R}} V_{\beta\alpha}(\vec{R}) e^{i\vec{k}\cdot\vec{R}} a_{X\alpha,\vec{k}}, \tag{4.B16}$$

and one for the B atom sites,

$$\left(E_{\vec{k}} - E_{\beta B}\right) a_{B\alpha,\vec{k}} \delta_{\alpha\beta} = \sum_{\alpha=1}^{M_B} \sum_{\vec{R}_B \neq \vec{0}} E_{\alpha B} \gamma_{\beta\alpha}(\vec{R}_B) e^{i\vec{k}\cdot\vec{R}_B} a_{B\alpha,\vec{k}}$$
$$+ \sum_{\alpha=1}^{M_X} \sum_{\vec{R}} V_{\beta\alpha}(\vec{R}) e^{i\vec{k}\cdot\vec{R}} a_{X\alpha,\vec{k}}, \tag{4.B16}$$

where $X = $ A or B depending on the type of atom at the lattice site $\vec{R}$.

In the cases in which the A and B atoms are the same element, the TBM coefficients at each site are identical; treatment of this situation is covered in examples found in Supplement C. The extension of this result to more than two sublattices of a Bravais lattice introduces additional displacement vectors to the atomic positions in each unit cell.

Supplement C: TBM examples

Two instructional examples of TBM are presented here. These examples are interesting because of their relevance to materials being studied today. The model for graphene is a two-dimensional non-Bravais lattice type and the model for the diamond/zincblende lattice, which is also a non-Bravais lattice, is applied to describe applicable elemental and compound semiconductors. The TBM of the previous supplement is used to numerically calculate the band structure for these two models.

Example 1: Graphene

A single layer of graphene is a two-dimensional hexagonal structure with carbon atoms at each site; it has the symmetry of the triangular lattice with two atoms per unit cell. The two atomic

positions in the unit cell will be called A ($\vec{d}_A = 0\hat{x} + 0\hat{y}$) and B $\left(\vec{d}_B = \frac{a}{2}\hat{x} + \frac{\sqrt{3}a}{4}\hat{y}\right)$. This

pair of atoms is periodically repeated throughout all of the unit cells in the lattice. To construct the wave function for grapheme, the valence band atomic orbitals are composed of the sp^2 hybrid states and the atomic p_z state. The sp^2 states produce strong covalent C—C bonds; they are not the subject of our treatment here (Figure 4.C1).
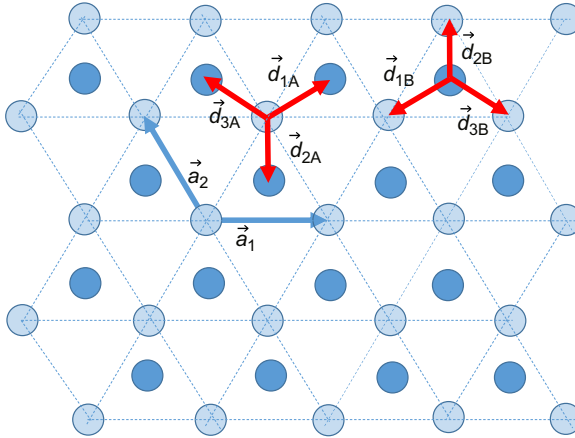
Figure 4.C1 Honeycomb lattice geometry with two interwoven triangular sublattices. The triangular lattice primitive lattice vectors are $\vec{a}_1$ and $\vec{a}_2$. The three nearest-neighbor displacements for each sublattice have the relationship $\vec{d}_{\alpha A} = -\vec{d}_{\alpha B}, \alpha = 1, 2, 3$.

The bonds of interest are the bonds formed from the p_z states, the wave functions of which lie along a symmetry line that is perpendicular to the graphene plane. The primitive lattice basis vectors and reciprocal lattice basis vectors are

$$\vec{a}_1 = a\hat{x}, \quad \vec{a}_2 = a\left(-\frac{1}{2}\hat{x} + \frac{\sqrt{3}}{2}\hat{y}\right), \tag{4.C1}$$

$$\vec{g}_1 = \frac{2\pi}{a}\hat{x} + \frac{2\pi}{\sqrt{3}a}\hat{y}, \quad \vec{g}_2 = \frac{4\pi}{\sqrt{3}a}\hat{y}. \tag{4.C2}$$

The wave function on the A sites $\phi_A(\vec{r}) = c_A \psi_{p_z}(\vec{r})$ and on the B sites $\phi_B(\vec{r}) = c_B \psi_{p_z}(\vec{r})$ have an identical functional form. The Bloch wave function is

$$\Psi_{\vec{k}}(\vec{r}) = \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}} \Phi_{\vec{k}}(\vec{r}-\vec{R}). \tag{4.C3}$$

The wave function at each lattice site is ($M_A = M_B = 1$)

$$\Phi_{\vec{k}}(\vec{r}-\vec{R}) = \begin{cases} c_{A\vec{k}} \psi_{p_z}(\vec{r}-\vec{R}), & \vec{R} \in \{A \text{ sites}\} \\ c_{B\vec{k}} \psi_{p_z}(\vec{r}-\vec{R}), & \vec{R} \in \{B \text{ sites}\} \end{cases}. \tag{4.C4}$$

This information is inserted into Eqn (4.B13). For only nearest-neighbor overlap integral contributions, the following two equations are

$$\left(E_{\vec{k}} - E_{pz} - V_{AA}(0)\right)c_{A,\vec{k}} = V_{AB}(1)f_A\left(\vec{k}\right)c_{B,\vec{k}},$$
$$\left(E_{\vec{k}} - E_{pz} - V_{AA}(0)\right)c_{B,\vec{k}} = V_{BA}(1)f_B\left(\vec{k}\right)c_{A,\vec{k}}. \tag{4.C5}$$

The structure factors are

$$f_B\left(\vec{k}\right)^* = f_A\left(\vec{k}\right) = e^{i\vec{k}\cdot\vec{d}_{1A}} + e^{i\vec{k}\cdot\vec{d}_{2A}} + e^{i\vec{k}\cdot\vec{d}_{3A}}.$$

The nearest-neighbor position vectors are

$$\vec{d}_{1A} = \frac{a}{2}\left(\hat{x} + \frac{1}{\sqrt{3}}\hat{y}\right), \quad \vec{d}_{2A} = -\frac{a}{\sqrt{3}}\hat{y}, \quad \vec{d}_{3A} = \frac{a}{2}\left(-\hat{x} + \frac{1}{\sqrt{3}}\hat{y}\right).$$

The wave function overlap contributions are assumed to vanish. In matrix form, Eqn (4.C5) is

$$D = \begin{pmatrix} E_{\vec{k}} - E_{pz} - V_{AA}(0) & -V_{AB}(1)f_A\left(\vec{k}\right) \\ -V_{BA}(1)f_B\left(\vec{k}\right) & E_{\vec{k}} - E_{pz} - V_{AA}(0) \end{pmatrix} = 0. \tag{4.C6}$$

The two eigenvalues of Eqn (4.C6) are determined from the determinant of the matrix D,

$$E_{\vec{k}\pm} = E_{pz} + V_{AA}(0) \pm \sqrt{(V_{AB}(1))^2 \left|f_A\left(\vec{k}\right)\right|^2}. \tag{4.C7}$$

$$f_A\left(\vec{k}\right) = e^{-ik_y\frac{a}{\sqrt{3}}}\left(1 + 2\cos\left(\frac{k_x a}{2}\right)e^{ik_y\sqrt{3}a/2}\right). \tag{4.C8}$$

The eigenvalues are equal in the BZ at points

$$f_A\left(\vec{k}\right) = 0.$$

Not surprisingly, there are six points that satisfy this condition by analyzing the expression in parentheses. For $k_y = 0$, the cosine function satisfies

$$f_A\left(\vec{k}\right) = e^{-ik_y\frac{a}{\sqrt{3}}}\left(1 + 2\cos\left(\frac{k_x a}{2}\right)e^{ik_y\sqrt{3}a/2}\right). \tag{4.C9}$$

For $k_y \sqrt{3}a = \pm\pi$, the cosine function for either case satisfies

$$\cos\left(\frac{k_x a}{2}\right) = \frac{1}{2} \Rightarrow \frac{k_x a}{2} = \pm\frac{\pi}{3}. \tag{4.C10}$$

These correspond to the six K points at the edge of the BZ.

$$K_1 = \frac{4\pi}{3a}\hat{x} + 0\hat{y}, \quad K_2 = \frac{2\pi}{3a}\hat{x} + \frac{2\pi}{\sqrt{3}a}\hat{y}, \quad K_3 = -\frac{2\pi}{3a}\hat{x} + \frac{2\pi}{\sqrt{3}a}\hat{y},$$

$$K_4 = -\frac{4\pi}{3a}\hat{x} + 0\hat{y}, \quad K_5 = -\frac{2\pi}{3a}\hat{x} - \frac{2\pi}{\sqrt{3}a}\hat{y}, \quad K_6 = -\frac{2\pi}{3a}\hat{x} - \frac{2\pi}{\sqrt{3}a}\hat{y}. \tag{4.C11}$$

Example 2: Zincblende semiconductor model

Semiconductors formed from sp^3 bonding states can be modeled by adapting the TBM from Supplement B. The treatment here is restricted to the diamond or zincblende crystal structures, which applies to Si and Ge as well as materials such as GaAs. An illustration of the sp^3 orbitals appears in Figure 4.C2(a). The s orbital is illustrated at the center site and three p orbitals are drawn on the nearest-neighbor site in the (111) direction. The p orbitals are oriented with the positive phase in the positive direction of their symmetry axis. In the diagram in Figure 4.C2(b), four nearest neighbors are illustrated: two lie above the $z = 0$ plane (solid boxes) and two lie below it (dotted boxes). The overlap integrals will have positive or negative contributions depending on the position of the nearest-neighbor site. The shading of the orbital label in each box indicates whether it has a positive or negative contribution.
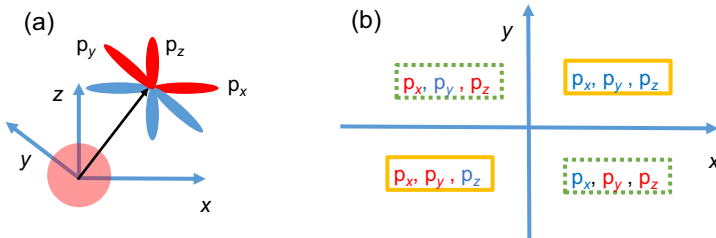
Figure 4.C2 (a) The s orbital at the origin has an overlap with the three p orbitals shown at the nearest-neighbor sites. (b) The two-dimensional diagram illustrates the relative positions of the orbitals relative to the center. The solid box represents sites at $+z$ positions and the dotted box represents sites at $-z$ positions.

The usual set of atomic wave functions (sp^3) is expanded to incorporate an additional state called s*, which is treated as s state symmetry; the model with the additional atomic wave function is dubbed the sp^3s^* model. We restrict the examples here to the zincblende lattice illustrated in Figure 4.13, which is the structure found for many of the compound semiconductors. For the binary semiconductors, the atoms pair up and share their bonds with four nearest neighbors. The sp^3s^* model is represented by five atomic wave functions ($M_A = M_B = 5$), and there are two sublattices (e.g., GaAs): one for the cations (Ga ions, labeled c) and one for the anions (As ions, labeled a). The TBM is a 10×10 matrix with structure functions that are defined by the lattice structure and potential coefficients representing the overlap integrals.

The cation and anion lattices form interlocking FCC lattices. In this section, the TBM is restricted to nearest-neighbor overlap integrals. The four nearest-neighbor positions from the cation lattice are

$$\vec{d}_{1c} = \frac{a}{4}\hat{x} + \frac{a}{4}\hat{y} + \frac{a}{4}\hat{z}, \ \vec{d}_{2c} = -\frac{a}{4}\hat{x} - \frac{a}{4}\hat{y} + \frac{a}{4}\hat{z},$$

$$\vec{d}_{3c} = -\frac{a}{4}\hat{x} + \frac{a}{4}\hat{y} - \frac{a}{4}\hat{z} \ \text{and} \ \vec{d}_{4c} = \frac{a}{4}\hat{x} - \frac{a}{4}\hat{y} - \frac{a}{4}\hat{z}.$$

The anion sites have cation nearest neighbors, which are displaced from the anion lattice positions. The anion nearest neighbors are defined by the displacement vectors $\vec{d}_{ia} = -\vec{d}_{ic}$ and those neighbors are cations. The treatment here is restricted to only nearest-neighbor overlap integrals.

The overlap between neighboring s wave functions has the same phase for all nearest neighbors; for instance, an s wave on a cation lattice side overlapping with an s wave on an anion lattice site $\left(\left\langle sc|V|sa \right\rangle = \frac{1}{4}V_{ss} \right)$ yields a contribution $V_{ss}f_0(\vec{k})$, where the function is related to the structure of the lattice. For the four nearest neighbors of a cation-central atom,

$$f_0(\vec{k}) = \sum_{\alpha=1}^{4} e^{i\vec{k}\cdot\vec{d}_{\alpha c}} = \cos\left(\frac{k_x a}{4}\right)\cos\left(\frac{k_y a}{4}\right)\cos\left(\frac{k_z a}{4}\right) - i\sin\left(\frac{k_x a}{4}\right)\sin\left(\frac{k_y a}{4}\right)\sin\left(\frac{k_z a}{4}\right).$$

The last equality follows by applying the Euler formulas from complex analysis. For the case when the central site is an anion, the result is $V_{ss}f_0^*(\vec{k})$, where the complex conjugate of the function $f_0(\vec{k})$ appears.

To derive the matrix elements, the phase and orientation of the p-type wave functions need special attention. For the p-type wave functions, we assume that the lobe directed in the positive direction of the symmetry axis has a positive phase and the lobe on the opposite side is shifted by 180°, as illustrated in Figure 4.C2(a). The phase difference results in a change of sign for the overlap integrals for nearest-neighbor atoms. As an example, consider the s-wave function overlap with the p_x wave function. The overlap integral for either the cation or the anion has identical matrix elements:

$$< sc|V|p_x a > = < sa|V|p_x c > = \frac{1}{4}V_{sapc}$$

The sum over nearest neighbors yields

$$V_{sapc}f_1\left(\vec{k}\right) = \sum_{\alpha=1}^{4}(-\operatorname{sign}(d_{x\alpha}))e^{i\vec{k}\cdot\vec{d}_{xc}}.$$

The function $(-\operatorname{sign}(d_{x\alpha}))$ is -1 for $d_{x\alpha} > 0$ and $+1$ otherwise. The function on the left-hand side is

$$f_1\left(\vec{k}\right) = \cos\left(\frac{k_x a}{4}\right)\sin\left(\frac{k_y a}{4}\right)\sin\left(\frac{k_z a}{4}\right) - i\sin\left(\frac{k_x a}{4}\right)\cos\left(\frac{k_y a}{4}\right)\cos\left(\frac{k_z a}{4}\right).$$

Similarly treating the remaining cases, we define two more structure functions:

$$f_2\left(\vec{k}\right) = \sin\left(\frac{k_x a}{4}\right)\cos\left(\frac{k_y a}{4}\right)\sin\left(\frac{k_z a}{4}\right) - i\cos\left(\frac{k_x a}{4}\right)\sin\left(\frac{k_y a}{4}\right)\cos\left(\frac{k_z a}{4}\right)$$

and

$$f_3\left(\vec{k}\right) = \sin\left(\frac{k_x a}{4}\right)\sin\left(\frac{k_y a}{4}\right)\cos\left(\frac{k_z a}{4}\right) - i\cos\left(\frac{k_x a}{4}\right)\cos\left(\frac{k_y a}{4}\right)\sin\left(\frac{k_z a}{4}\right).$$

Table 4.C1 10×10 matrix to calculate the band structure of elemental and compound semiconductors for the zincblende lattice

E_{sa}	$V_{ss}\cdot f_0$	0	0	0	$V_{sapc}\cdot f_1$	$V_{sapc}\cdot f_2$	$V_{sapc}\cdot f_3$	0	0
$V_{sa}\cdot f_0^*$	E_{sc}	$-V_{sapc}\cdot f_1$	$-V_{sapc}\cdot f_2$	$-V_{sapc}\cdot f_3$	0	0	0	0	0
0	$-V_{sapc}\cdot f_1^*$	E_{pa}	0	0	$V_{xx}\cdot f_0$	$V_{xy}\cdot f_3$	$V_{xy}\cdot f_2$	0	$-V_{pas*c}\cdot f_1$
0	$-V_{sapc}\cdot f_2^*$	0	E_{pa}	0	$V_{xy}\cdot f_3$	$V_{xx}\cdot f_0$	$V_{xy}\cdot f_1$	0	$-V_{pas*c}\cdot f_2$
0	$-V_{sapc}\cdot f_3^*$	0	0	E_{pa}	$V_{xy}\cdot f_2$	$V_{xy}\cdot f_1$	$V_{xx}\cdot f_0$	0	$-V_{pas*c}\cdot f_3$
$V_{sapc}\cdot f_1^*$	0	$V_{xx}\cdot f_0^*$	$V_{xy}\cdot f_3^*$	$V_{xy}\cdot f_2^*$	E_{pc}	0	0	$V_{s*apc}\cdot f_1^*$	0
$V_{sapc}\cdot f_2^*$	0	$V_{xy}\cdot f_3^*$	$V_{xx}\cdot f_0^*$	$V_{xy}\cdot f_1^*$	0	E_{pc}	0	$V_{s*apc}\cdot f_2^*$	0
$V_{sapc}\cdot f_3^*$	0	$V_{xy}\cdot f_2^*$	$V_{xy}\cdot f_1^*$	$V_{xx}\cdot f_0^*$	0	0	E_{pc}	$V_{s*apc}\cdot f_3^*$	0
0	0	0	0	0	$V_{s*apc}\cdot f_1$	$V_{s*apc}\cdot f_2$	$V_{s*apc}\cdot f_3$	E_{s*a}	$V_{s*s*}\cdot f_0$
0	0	$-V_{pas*c}\cdot f_1^*$	$-V_{pas*c}\cdot f_2^*$	$-V_{pas*c}\cdot f_3^*$	0	0	0	$V_{s*s*}\cdot f_0^*$	E_{s*c}

With permission from Vogl et al. [1].

Table 4.C2 Empirically determined parameters for the sp3s* model

Variable	C	Si	Ge	AlAs	GaAs	GaP	InAs
a (nm)	0.3557	0.5431	0.5680	0.56611	0.56533	0.5450	0.6051
E_{sa}	−4.545	−4.200	−5.880	−7.527	−8.343	−8.112	−9.538
E_{pa}	3.840	1.715	1.610	0.983	1.041	1.125	0.910
E_{sc}	−4.545	−4.200	−5.880	−1.163	−2.657	−2.198	−2.722
E_{pc}	3.840	1.715	1.610	3.587	3.669	4.115	3.720
E_{s*a}	11.370	6.685	6.390	7.483	8.591	8.515	7.410
E_{s*c}	11.370	6.685	6.390	6.727	6.739	7.185	6.740
V_{ss}	−22.725	−8.300	−6.780	−6.664	−6.451	−7.471	−5.605
V_{xx}	3.840	1.715	1.610	1.878	1.955	2.152	1.840
V_{xy}	11.670	4.575	4.900	4.292	5.078	5.137	4.469
V_{sapc}	15.221	5.729	5.4649	5.111	4.480	4.277	3.035
V_{scpa}	15.221	5.729	5.4649	5.497	5.784	6.319	5.439
V_{s*cpa}	8.211	5.375	5.2191	4.522	4.842	4.654	3.374
V_{pas*c}	8.211	5.375	5.2191	4.995	4.808	5.095	3.910
V_{s*s*}	0	0	0	0	0	0	0

The band structure is calculated by finding the energy eigenvalues of the matrix in Table 4.C1. The empirically determined coefficients for a select group of materials is found in Table 4.C2. The band structure results for six different materials are plotted in Figure 4.14.
With permission from Vogl et al. [1].

The matrix elements are generated in a similar fashion. The order of the vector components used to generate the matrix in Table 4.C1 is

$$\vec{V} = \left\{ |s_a>, \ |s_c>, \ |p_x a>, \ |p_y a>, \ |p_z a>, \ |p_x c>, \ |p_y c>, \ |p_z c>, \right.$$

$$\left. |s_a^*>, \ |s_c^*> \right\}$$

Reference

[1] P. Vogl, H.P. Hjalmarson, J.D. Dow, A semi-empirical tight-binding theory of the electronics structure of semiconductors, J. Phys. Chem. Solids 44 (1983) 365.

Further reading

[1] J.M. Ziman, Principles of the Theory of Solids, Cambridge University, London, 1969.
[2] W.A. Harrison, Solid State Theory, McGraw-Hill, NY, 1970.
[3] A.P. Sutton, Electronic Structure of Materials, Oxford University Press, Oxford, 2004.
[4] J. Singh, Electronic and Optoelectronic Properties of Semiconductor Structures, Cambridge University Press, Cambridge, 2003.
[5] S. Dutta, Quantum Transport Atom to Transistor, Cambridge University Press, Cambridge, 2006.
[6] V.V. Mitin, V.A. Kochelap, M.A. Strossio, Introduction to Nanoelectronics, Cambridge University Press, Cambridge, 2008.
[7] B.E.A. Saleh, M.C. Teich, Fundamentals of Photonics, second ed., John Wiley and Sons, Hoboken, NJ, 2007.

Nanofabrication

5

A. Sarangan
University of Dayton, Dayton, OH, USA

5.1 Nanofabrication

There is no single accepted definition of nanofabrication, nor a definition of what separates nanofabrication from microfabrication. To meet the continuing challenge of shrinking component size in microelectronics, new tools and techniques are continuously being developed. Component sizes that were in tens of micrometers became single-digit micrometers, and then hundreds of nanometers, and then went down to a few tens of nanometers where they stand today. As a result, what used to be called microfabrication was rebranded as nanofabrication, although the governing principles have remained essentially the same. The main driver of this technology has been the manufacture of integrated circuits, but there have been tremendous fallout benefits to other areas, including photonics.

Nanofabrication can be loosely divided into three major areas: thin films, lithography, and etching. Each of these are vast subject areas in and of themselves, but in this chapter we attempt to cover their essential concepts in a concise fashion for someone new to these areas. The goal is not to provide a working experience that allows one to walk into a laboratory and perform these tasks, but to provide an overall understanding of what these areas are as well as the pros and cons of the most commonly used techniques. Hopefully it will build a foundation for more specific training for anyone who wants to venture further into these areas.

5.2 Thin films

Thin film science is a vast and mature area that permeates nearly all disciplines. Thin films are found in almost every manufactured product, from eyeglasses to display screens to automobiles and aerospace components. They are used to achieve various purposes, such as optical modification (antireflection), chemical modification (corrosion inhibition), mechanical modification (scratch resistance), electrical modifications and thermal modifications.

In this section, we focus primarily on films on the order of a few tens of nanometers thick for electronic and photonic applications. One example is the gate dielectric in metal-oxide-semiconductor (MOS) transistors. This dielectric is often less than 10 nm thick, and in many cases it approaches 2 nm. Its properties such as dielectric constant, thickness, and defect density are critical parameters that affect the transconductance and switching characteristics of the devices. Another area in which nanoscale films are widely used is in optical filters. In the visible and infrared range, film

thicknesses are in the range of 10−1000 nm. In the extreme-ultraviolet (EUV) range, the multilayer films used to make photomasks are in the range of 1−2 nm. Hence, these films span a wide range of thicknesses depending on their applications.

In the following sections, we will discuss some of the commonly used methods for producing thin films of interest in the field of nanophotonics. We only cover the most salient features of each technique and their pros and cons. The reference list at the end of the chapter provides the interested reader more sources to learn about these topics in greater depth [1−8].

The properties of thin films vary greatly depending on the method used to make them. The typical properties of interest are

- *Conformal nature of the film*: Conformal films grow on all topographic features, whereas nonconformal films are line-of-sign and grow only on certain planes.
- *Film density*: Although materials have standard densities, the films made from different techniques can have a lower or higher density.
- *Dielectric constant*: Higher densities generally result in a higher refractive index, although other factors such as contamination and compound formations can alter the refractive index from their nominal bulk values.
- *Stress*: Films under high stress can introduce curvatures to the substrate, or in extreme cases they can separate from the substrate by peeling or cracking. Stress will also affect the refractive index of the film.
- *Chemical composition*: In single elements, this is not usually a concern, but in compounds, the film stoichiometry can vary greatly depending on the growth conditions.
- *Electrical conductivity*: Grain size, packing density, and impurities can significantly alter the conductivity of thin films.

The methods used to make thin films can be broadly classified as physical methods and chemical methods. Physical methods involve transferring the material from the source to the substrate without changing its chemical state. In chemical methods, the film is created as a by-product of a chemical reaction.

5.2.1   Physical methods

By far the most commonly used physical methods are evaporation and sputtering. Both are vapor coating methods in which the source material is transferred atom-by-atom to the substrate. These are generally referred to as physical vapor deposition (PVD). PVD is usually done in a high-vacuum chamber to reduce the interaction with gas species in the environment.

5.2.1.1   Evaporation

Evaporation is easy to understand—the source material is heated to a high temperature until it starts to evaporate. The important parameter here is vapor pressure. Every material has a characteristic vapor pressure as a function of its temperature, as shown in Figure 5.1. In other words, every material is always evaporating. For most solids at room temperature, this vapor pressure is extremely low. For example, consider gold. Its vapor pressure at room temperature is well below 10^{-15} torr and cannot be detected.
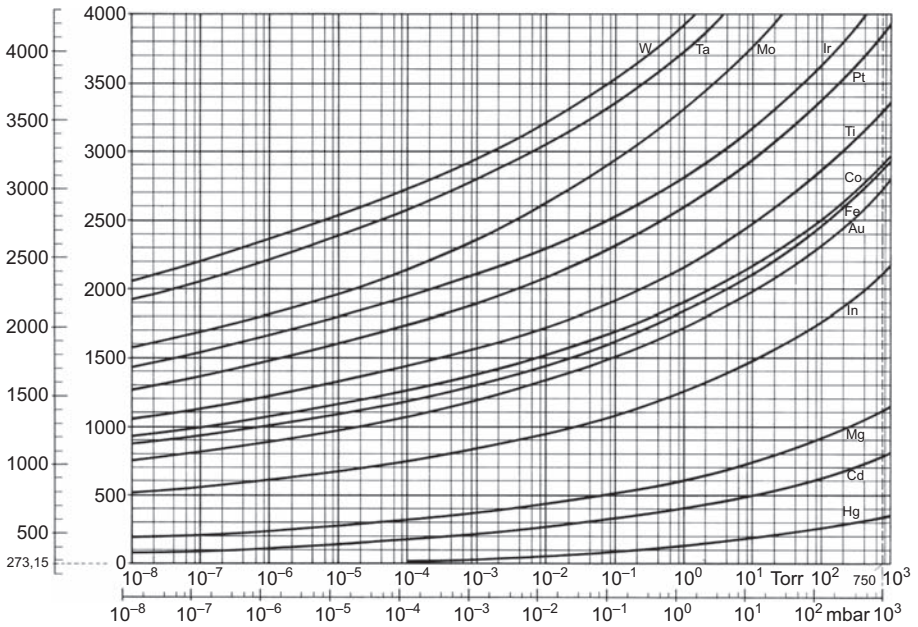
Figure 5.1 Vapor pressure versus temperature curve for metals.
Source: Fundamentals of Vacuum Technology, Oerlikon Leybold Vacuum, Figure 9.13, 2007.
Reproduced with permission.

At 1000 °C, its vapor pressure rises to 20 µTorr, and at 1500 °C it becomes 100 mTorr. In addition, the vapor pressure will also be a function of position inside of the vapor flux. The pressure will drop as the vapor diffuses away from the evaporating source. The vapor pressure at the substrate surface can be estimated from simple geometrical considerations of the vapor flux. From ideal gas laws, we can then relate pressure to the impingement rate of atoms. If we further assume that all impinging atoms condense on the substrate surface, then we can calculate the growth rate of the film on the substrate. Although this is the physical description of the evaporation—condensation phenomena, such calculations are rarely done in practice because the source temperature and vapor pressures are typically not measured. Instead, one monitors the deposition rate on the substrate by controlling the power delivered to the evaporating source, leaving source temperature and vapor pressure as intermediate parameters.

The unit of measure in thin film science is Angstroms rather than nanometers, but we will adopt the units of nanometers here for consistency with other chapters. Typical deposition rates are in the range of 0.1—5 nm/s. To reach these levels, most materials have to be heated to near or above their melting temperature. Some materials evaporate by sublimation well below their melting temperature. Examples of sublimating materials are chromium and silicon dioxide. Boiling temperature is not a relevant concept here—this is the temperature at which the vapor pressure is equal to the ambient pressure. Because the ambient pressure in a vacuum chamber is extremely low, every material can be considered to be boiling all of the time in vacuum.

There are two major types of evaporation. They differ by the method of power delivered to the source. In one method, the source in the form of pellets is placed in a metallic boat and is heated by passing a very high current through it, in the range of 100 A. This is known as resistively heated evaporation. An advantage of this technique is its simplicity because only a high current (low-voltage) DC source is needed to power the source. The disadvantage is the lack of heating efficiency and contamination. Because all parts of the heating circuit will be hotter than the source pellet, they will also evaporate to some extent, which can lead to contamination and outgassing.

The second method is electron-beam heating. In this technique, electrons emitted from a heated filament are accelerated by a high voltage in the range of 10 kV and focused to a small spot on the source pellet, as shown in Figure 5.2. Because the process is done in a vacuum, electrons can be easily accelerated and manipulated by magnetic fields without colliding with gas molecules. This method is more efficient because energy is delivered accurately to the source pellet with minimal heating of the other fixtures. As a result, contamination is kept to a minimum. A disadvantage of this method is the complexity of the power source. A high-voltage source is more complex in its design than a low-voltage, high-current source. Safety features are also a major consideration because a 10-kV DC can be lethal. Nevertheless, electron-beam evaporation continues to be one of the most commonly used evaporation methods in thin film research and development.

Because of the physical trajectories of the evaporating species and the high-vacuum environment, the films produced by evaporation tend to be directional and nonconformal. This can be an advantage in many applications, such as lift-off lithography. The highly directional nature of the evaporating flux is also utilized in a special class of thin film known as nanostructured thin films. Instead of containing a randomly packed isotropic structure, these films contain nanocolumns and nanochiral structures, which can be used to engineer the electrical and optical properties of these films. A class of patterning known as shadow-masked lithography also relies on a line-of-sight deposition for which evaporation is ideally suited.
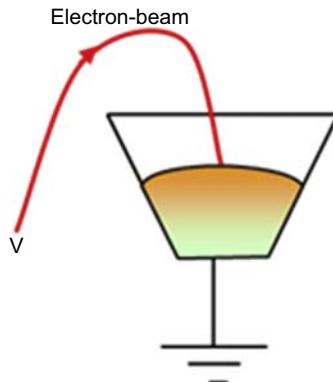

Figure 5.2 Electron-beam evaporation.

5.2.1.2   Sputtering

Sputter deposition utilizes an electrically excited gas plasma in a vacuum system. The ions in the plasma are accelerated toward the cathode, which upon bombardment eject neutral atoms from the cathode surface. The ejected atoms collect on all surfaces including the substrate surface. This is illustrated in Figure 5.3. Therefore, the cathode has to be constructed from the same source as the material being deposited. Unlike evaporation, the source material here is kept at a low temperature by flowing cooling water behind it. Atoms are ejected from the source (which in this context is referred to as the target) by momentum transfer rather than by heat. This is a fundamental difference between sputtering and evaporation. Because of the incident momentum, the resulting films will be more compacted and denser than in evaporation. The low target temperature also enables the deposition of certain compounds such as oxides and nitrides, which may otherwise decompose at elevated temperatures encountered in evaporation. Nevertheless, for complex compounds, pulsed laser deposition (see Section 5.2.1.3) is the preferred technique over sputtering.

The atomic weight of the plasma gas species and the gas pressure play a significant role in the sputter yield, which is defined as the number of target atoms that are ejected for each incident ion. Ordinarily, the gas species is selected to be chemically inert so that it does not interact with the target material and form compounds. For this reason, argon gas is typically used in most sputter systems. It has a reasonably high atomic mass ($\sim$40 u, where u is the unified atomic mass unit of 1 g/mol), is inert, and is relatively inexpensive. Sputter yield data for various gas species, target materials, and ion energies can be found in the literature. In general, sputter
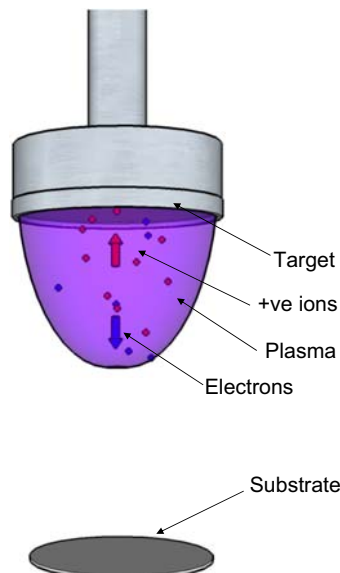
Figure 5.3 Sputter deposition configuration.

yield is related to the atomic number and masses of the plasma ion and the target material. Sputter yields also exhibit a threshold energy. These can be determined through empirical models as well as numerical models using Monte Carlo methods. For example, the sputter yield of Ag from Ar^+ ions is shown in Figure 5.4. This was calculated using the empirical model presented in Ref. [3]. We can see that it shows a threshold energy of 10 eV and rises from 0.7 at 100 eV to 2.7 at 500 eV. When He^+ ions are used, the sputter yield is about an order of magnitude smaller. At very high energies, the yield declines again because of a process known as ion implantation, in which the incident ions bury themselves deeper in the target instead of ejecting surface atoms.

Sputtering is more versatile than evaporation and is more widely used in industrial processes because the targets and plasma sources can be constructed in various shapes to accommodate different coating configurations. They can be circular, rectangular, cylindrical, or other exotic shapes to fit specific needs. By far the most common type of targets used in research laboratories are circular, 2 or 3 inch in diameter. Sputter deposition can take place upward or downward or even sideways, whereas evaporation is typically limited to only an upward configuration.

Figure 5.4 Calculated sputter yield versus energy [3].

Power is delivered to the plasma by a DC source or radiofrequency (RF) source. For metallic target materials, a DC source can be used. Insulating targets will require an RF source, where the target acts as a coupling capacitor to the plasma. RF excitatio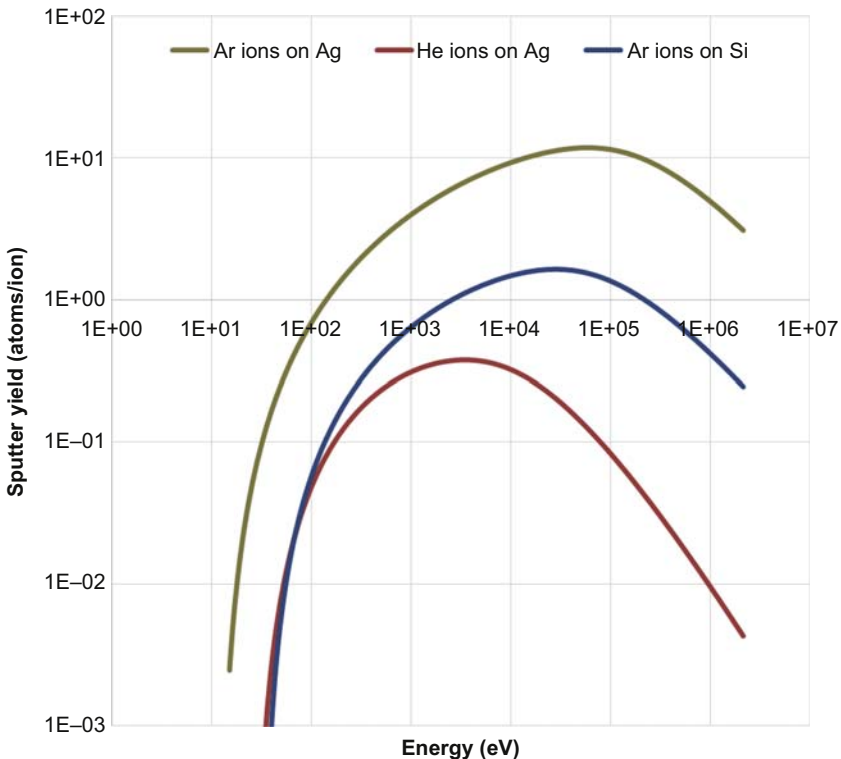n also requires an automatic impedance matching network because the plasma will not have a constant impedance but will vary with its operating conditions such as pressure, material type, target thickness, etc. Many sputter systems used in research tend to utilize RF excitation because it can be used on conductive and nonconductive targets, whereas a DC excitation can be used only on conductive targets.

Because of the presence of the plasma gas, depending on the operating pressure, the sputtered atoms may undergo a few collisions before reaching the substrate. Therefore, sputter deposition can be somewhat less directional than evaporation.

In most cathodes designed for sputtering, magnets are placed behind the target to alter the trajectories of the electrons in the plasma. This has the effect of focusing the plasma close to the target, resulting in a higher ion density and a higher sputter removal rate. These are referred to as magnetron cathodes.

Another variant of sputter deposition is reactive sputtering. In this technique, a small amount of reactive gas such as oxygen or nitrogen is mixed with the argon gas to cause the ejected target species to form compounds. For example, it is possible to create TiO_2 from a metallic Ti by flowing a small amount of oxygen with the argon. However, this is not as simple as it may first appear. Excessive oxygen can cause the target surface to become oxidized and will result in an extremely low sputter yield. This condition is known as target poisoning. The ideal condition is achieved when the oxidation rate of the target surface is exactly balanced by the removal rate of the oxide, and the ejected species are dominantly metallic, which are subsequently oxidized on the substrate surface to achieve the correct stoichiometry. Furthermore, these oxides and nitrides may have several different stable stoichiometric states; therefore, simply mixing a reactive gas will not yield the desired result. For example, vanadium oxide (VO_2) has interesting thermal properties and is used in MEMS actuators. However, during reactive sputtering of a metallic vanadium target with an oxygen gas, the resulting film will contain a mixture of VO_2, V_2O_3, V_3O_5, etc. Deposition conditions such as oxygen partial pressure, substrate temperature, and plasma discharge power are all used in combination to direct the resulting species toward a dominantly VO_2 composition. The same is true with TiO_2, which is used as the high-index film in multilayer optical filters because it has the highest refractive index in the visible spectrum.

5.2.1.3   Pulsed laser deposition

Pulsed laser deposition (PLD) uses high-energy laser pulses on the order of a few nanoseconds to ablate the target material. The laser radiation is focused on the target surface, is absorbed, and rapidly evaporates the material, resulting in the ablation of atoms, which are subsequently collected on the substrate. Because the energy source (laser) is outside of the vacuum chamber, this method has the advantage of being done in an ultra-high vacuum or under a wide range of ambient pressures and gas species.

The target is generally rotated so that the same spot is not repeatedly ablated. A schematic of a PLD setup is illustrated in Figure 5.5.

The biggest advantage of PLD is the stoichiometric removal of the target material (i.e., the atoms are removed without discrimination). This is due to the fast, transient nature of the ablation process and the high laser fluence, which creates a surface temperature of approximately 5000 K within a few nanoseconds. Short pulses of fluence levels much higher than the ablation threshold allow all components of the target to be ablated equally, irrespective of their binding energies. In contrast, in evaporation and sputtering the volatility of the species affects the removal rate. In sputtering, the material with the higher sputter yield will be removed faster, and in evaporation the material with the higher vapor pressure will be removed faster. As a result, PLD is most often used for the deposition of complex ceramic films such as yttrium barium copper oxide (YBCO); lead zirconium titanate (PZT); and many other carbides, oxides, and nitrides that are difficult or impossible to deposit with other methods. However, stoichiometric removal does not always imply stoichiometric deposition because some of the more volatile species can sputter off of the substrates or simply evaporate. Some compensation techniques are necessary, such as background reactive gases and the use of multiple targets.

Most of the currently used PLD systems utilize excimer lasers emitting in the ultraviolet (UV) range, such as KrF (248 nm), ArF (193 nm), and F_2 (157 nm). UV light is desirable because of its short penetration depth in materials, resulting in the removal of atoms closest to the surface. Longer wavelengths would result in a deeper penetration, which would cause subsurface evaporation and eruption that can lead to larger clusters being ejected. However, in practice, the availability of UV lasers, vacuum windows, and other optical elements limit the lower wavelength to approximately 200 nm.

PLD also allows the user to tune the energy of the species arriving at the substrate by introducing an inert gas atmosphere such as argon. The gas atmosphere increases the scattering rate of the ejected species and slows them down. This technique can be used to produce changes in the film morphology and stress. It is also possible to create nanoparticle deposition by increasing the gas pressure until the ejected species collide with each other to form clusters of atoms. The size of the nanoparticles can also be controlled this way.
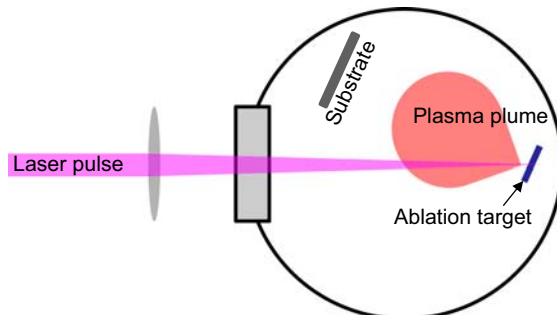
Figure 5.5 Pulsed laser deposition.

Although PLD has advanced significantly in recent years, there are still many challenges. Film uniformity is one of the biggest challenges. Because the ablation occurs at a single point and the plasma plume has a narrow angle, only a small substrate area will get coated during each pulse. The source-to-substrate distance also plays a role. Larger distances will produce a more uniform film, albeit at the expense of a reduced deposition rate. Although the laser beam and the substrate can be continuously translated to improve the film uniformity, it is still inferior to evaporation or sputtering. Another challenge with PLD is the deposition of larger particulates, sometimes up to 10 μm in size, which occur because of molten droplets being ejected during the ablation. These can also be mitigated to some extent by using physical filters.

5.2.2   Chemical methods

Chemical methods can produce films with excellent uniformity, coverage, and stoichiometry, but they require different gases and sometimes different chambers for each film type. The most common chemical method is chemical vapor deposition (CVD). In this technique, gas precursors are introduced into a chamber and the substrate is heated to a sufficiently high temperature to cause a reaction and produce the film of interest. There are different types of CVD, such as low-pressure CVD (LPCVD), atmospheric pressure CVD, plasma-enhanced CVD (PECVD), and atomic layer deposition (ALD).

In CVD, the pressure has to be low enough and the substrate temperature high enough that the reaction only takes place on the substrate surface and not in the gas phase. Gas-phase reactions would result in particles forming and depositing on the surface.

5.2.2.1   Low-pressure CVD

In LPCVD, the substrate is placed in a quartz tube and heated while flowing the precursor gases to maintain a constant pressure, typically on the order of a few Torr. In this regime, the deposition rate is primarily governed by the pressure and temperature and not by the gas flow characteristics. Because maintaining pressure and temperature is relatively easy, LPCVD films tend to be very uniform and very conformal. LPCVD is used for creating films such as silicon nitride, silicon dioxide, silicon carbide, and several germanium compounds. The required substrate temperature is governed by the reaction chemistry and is typically above 700 °C.

5.2.2.2   Plasma-enhanced CVD

PECVD is a variant of LPCVD in which a plasma is used to reduce the substrate temperature to less than 300 °C. This was developed to meet the needs of the complementary MOS (CMOS) manufacturing process in which high-quality dielectrics were required as insulation layers between the metal interconnect traces, but the LPCVD temperature was too high for integrated circuits in their later stages of manufacture. In PECVD reactors, the plasma is in close proximity to the substrate and is typically

at very low discharge power levels such that gas-phase reactions do not occur. The chemistries are very similar to LPCVD, except for the lower substrate temperatures.

5.2.2.3 Atomic layer deposition

ALD is a CVD technique that has recently become popular because of its highly conformal pinhole-free coverage and its ability to be run under a wide range of pressures and temperatures. An example of a conformal deposition is shown in Figure 5.6. The highly conformal nature allows one to achieve a uniform coverage around high-aspect-ratio structures and in deep holes. Unlike other CVD techniques, the gas precursors in ALD are flowed one at a time, not simultaneously. The gases are switched on and off with inert purge gas flows in between. Each gas flow duration is referred to as a pulse. During the first precursor pulse, the reactant molecules are allowed to adsorb onto the substrate surface. The extent of adsorption will be a function of temperature, pressure, and time. The chamber is then purged and the second precursor pulse is applied. The second precursor will react with the adsorbed gases from the first precursor to produce the desired film. Because the gases mix only at the surface, it is inherently a surface reaction. It is also highly conformal and can be used to deposit a film over high-aspect-ratio structures. Furthermore, the reaction is self-limiting because once all of the adsorbed molecules from the first precursor have been consumed, the reaction comes to a stop. The film growth proceeds by repeating the above cycle many times. During each cycle, the growth is typically on the order of a monolayer or even smaller. By counting the number of cycles, the layer thickness can be very accurately controlled. The average growth rate in ALD is only a few Angstroms per cycle, so the overall growth rate is slow. This is one of its biggest disadvantages.
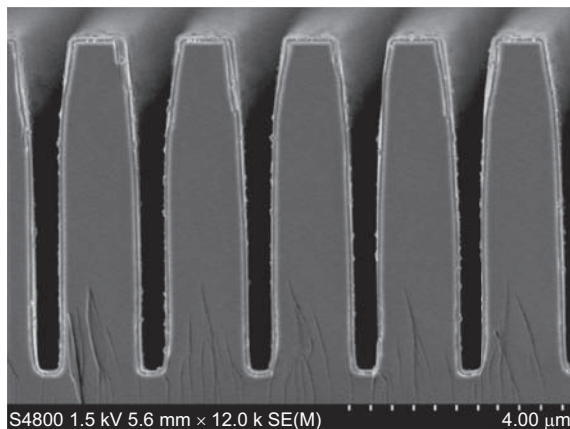

Figure 5.6 Conformal film by ALD deposition.
Reprinted with permission from J. Am. Chem. Soc., 2009, 131 (10), pp. 3478−3480. Copyright 2009 American Chemical Society.

ALD chemistries have been developed for several different oxides, nitrides, and metals, and the list of chemistries and applications continues to grow. Because both precursors are never flowed simultaneously, gas-phase reaction will not occur regardless of pressure or temperature. However, the pressure and temperature will affect the adsorption rate and hence the film growth rate during each cycle. This flexibility allows one to develop a deposition process for virtually any substrate, including polymer substrates. ALD is currently most actively being developed for CMOS gate dielectrics, such as aluminum oxide (Al_2O_3) and hafnium oxide (HfO_2), because of ALD's ability to produce defect-free ultrathin films.

5.2.3   Epitaxy

All of the aforementioned thin film growth techniques produce amorphous films. Amorphous means that the atoms in the film are randomly packed with no specific order. These are the simplest kind of films to produce. On the other hand, a crystalline film has a long range order and can be grown only under very specific conditions. Most materials exist in amorphous, crystalline, or polycrystalline states. Their mechanical, optical, thermal, and electrical properties will also change depending on their state. For example, carbon in its amorphous state is a black powder, whereas in its crystalline state it is diamond—an optically clear solid with a high refractive index. Silica in its amorphous form is a white powder and in its crystalline form is quartz.

Because crystalline films are more difficult to produce, amorphous films are used whenever their properties are sufficient for the application. Such is the case with optical coatings—the refractive index of most films can be made with excellent reliability and repeatability even in their random amorphous state. Metal films are also used in amorphous form because their electrical conductivity and optical reflectivity can be reliably reproduced. The major difference between amorphous and crystalline materials appears in their electronic band structure. Amorphous materials do not have a distinct band structure because of their random orientations; crystal structures do. Certain applications require well-defined electronic band structures for their operation. Examples are semiconductor electronics, optoelectronics, and components that require birefringence or piezoelectric properties. In some limited cases, it is possible to crystallize an amorphous film by annealing it at high temperatures. However, epitaxy is the formal process of creating high-quality crystalline films. Epitaxy can be chemical or physical, just like their amorphous counterparts, but we discuss epitaxy as a separate topic because of its distinct differences compared with other types of film growth.

A crystalline film can only be grown on a substrate whose lattice structure is closely matched to the film being grown. If the substrate is identical to the film, then it is known as homoepitaxy. If the substrate is slightly different, but is still compatible, then it is called heteroepitaxy.

Let us examine how silicon crystalline wafers are grown, which is the cornerstone of all of today's electronics. One starts with amorphous or polysilicon (which can be a powder or chunks) reduced from silicon dioxide mined from sand. This silicon is melted into a liquid, and a seed crystal is inserted into the melt. This starts a layer-by-layer solidification process on the seed crystal that eventually grows in size.

This is a homoepitaxial liquid-phase growth known as the Czochralski process. There is another method called the float zone process that can produce higher purity silicon crystals. In this method, the polysilicon is made into a rod and is heated from the outside with a moving coil without contacting the silicon. Certain II−V compounds use another technique known as the Bridgeman method. In this method, the melt is slowly cooled within the container from one end with a seed crystal. All of these methods produce a cylindrical-shaped crystalline rod known as the boule. The crystal is then shaped, sliced into thin wafers, and individually polished.

For some applications, the purity and defect density of silicon grown from melts is not adequate because of the contamination from the crucibles and the mechanical finishing process. Therefore, a thin layer of crystalline silicon film is grown on top of these wafers using a vapor-phase epitaxy. This is a homoepitaxial vapor-phase growth. Vapor-phase epitaxy can produce higher purity films with a lower defect density. This requires appropriate precursor gases that contain silicon, such as silane. These silicon wafers are known as epitaxial silicon, which is widely used in high-performance electronics.

In addition to silicon, epitaxy is most extensively used in III−V semiconductors such as GaAs, InP, InAs, etc. Many III−V semiconductors exhibit an interesting property in which their band structures can be adjusted by including other elements, such as $Ga_xAl_{1-x}As$, without significantly changing their original crystal structure. This allows one to stack epitaxial layers with differing electronic band structures on top of each other. This is heteroepitaxy and has become commonplace in optoelectronics devices such as laser diodes, light-emitting diodes (LEDs), and quantum well devices.

5.2.3.1   Metal organic CVD

Metal organic CVD (MOCVD) is a CVD process for growing epitaxial films, very similar to LPCVD, and is done by flowing precursor gases over the substrate. In III−V semiconductors, the metallic element is carried by an organic gas such as trimethylgallium ($Ga(CH_3)_3$) and trimethylindium ($In(CH_3)_3$) along with arsine (AsH_3) or phosphine (PH_3). The gases are allowed to decompose due to pyrolysis on the heated substrate surfaces to produce the desired film. The process pressures are typically in the range of 10−100 torr, resulting in relatively fast growth rates. One drawback of MOCVD is the toxic and explosive nature of the precursor gases, which makes them difficult to use in small research laboratories. Nevertheless, MOCVD is a scalable process amenable to volume manufacturing because many substrates can be simultaneously placed in the chamber. As a result, it is widely used in the manufacture of quantum well lasers, LEDs, and other components.

5.2.3.2   Molecular beam epitaxy

Whereas MOCVD is similar to LPCVD, molecular beam epitaxy (MBE) can be considered similar to PVD evaporation and is performed in ultra-high vacuum. This makes MBE better suited for applications that require very high purity levels. Solid sources such as gallium or indium from different effusion cells are typically allowed

to sublime and condense on the substrate. The cells are shuttered to allow rapid and precise transition from one material to another. The high-vacuum environment also allows one to use a wide range of diagnostic tools during growth. Many MBE systems use reflection high-energy electron diffraction (RHEED) to monitor the progression of growth, with the ability to count monolayers as they grow. The source configurations also make this system a lot less hazardous than MOCVD. Chemical beam epitaxy is a variant of MBE in which gas sources are used instead of solid sources but the principles are very similar. The biggest drawback of MBE compared with MOCVD is the slow growth rate and the inability to grow many wafers at once. Nevertheless, it is much more widely used than MOCVD in research facilities to study the fundamental properties of epitaxial film growth and in some limited production environments.

Table 5.1 is a brief summary of each thin film growth technique discussed in this chapter.

5.3   Lithography

Lithography literally means printing artwork or text on a surface. In the context of nano- and microfabrication technology, lithography is used to apply a pattern on a substrate surface so that it can be subsequently transferred to the underlying substrate. The lithography material itself is used as a sacrificial film that is discarded after the pattern is transferred.

Optical lithography uses light to replicate a pattern from a master (photomask) onto the substrate. The process is very similar to traditional photographic reproduction from negative films. A photosensitive polymer (photoresist) is applied to the substrate and then exposed to light through the photomask. The exposed photoresist undergoes a chemical reaction that results in a change in solubility. This is subsequently used to dissolve parts of the photoresist, leaving a patterned photoresist. There are several different photolithography techniques, such as contact, projection, immersion, and interference as well as UV, deep-UV, EUV, and X-ray based methods.

Nonoptical lithography involves electron beams, ion beams, or mechanical forces to create the pattern on the resist film. These are e-beam lithography (EBL), focused-ion beam (FIB) lithography, and nanoimprint lithography (NIL). A brief review of these techniques and their limitations are described in the following sections.

5.3.1   Photolithography

5.3.1.1   Light sources

UV light sources are commonly used in photolithography. This is not only because short wavelengths lead to better image resolution, but also because of the widespread availability of photochemicals sensitive to UV light. The mercury vapor lamp still continues to be the dominant UV light source in photolithography with emission lines at 405 nm (h-line), 365 nm (i-line), and 254 nm. Of these, 365 nm (i-line) is used the most, and many photoresists have been developed for this spectral range. The demand

Table 5.1 Summary of thin film deposition techniques

	Substrate temperature	Deposition energy	Pressure	Step coverage	Defect density	Uniformity	Deposition rate	Commonly used materials	Common application
Evaporation	Wide range	Low	Vacuum or reactive gas	Highly directional	High	High	Fast	Most metals, single elements, and stable dielectrics, such as Au, Ag, Cu, Si, SiO_2, MgF_2, etc.	Optical and electrical films, other generic applications
Sputtering	Wide range	High	Moderate. Mostly argon, but it can also include reactive gases	Directional	Moderate	High	Fast	Same materials as evaporation, plus additional metals and dielectrics such as W, VO_2, etc.	Optical and electrical films, other generic applications
PLD	Wide range	High	Wide range	Directional	Moderate	Poor	Slow	Complex compounds such as YBCO, PZT, and ferroelectric materials	Currently mostly used for exploration
LPCVD	High	Surface reaction	Moderate	Conformal	Very low	High	Fast	Si_3N_4, SiO_2	Masking and MEMS
PECVD	Moderate	Surface reaction	Moderate	Somewhat directional	Low	High	Fast	Si_3N_4, SiO_2, polySi	Electrical insulation, passivation, masking
ALD	Wide range	Surface reaction	Wide range	Highly conformal	Very low	High	Slow	Al_2O_3, HfO_2, SiO_2, and certain metals	Gate dielectrics, passivation
MOCVD	High	Surface reaction	Moderate	Epitaxial	Low	High	Moderate	Compound semiconductors— GaAs, InP, AlGaAs	Manufacture of optoelectronic devices
MBE	Wide range	Surface reaction	Vacuum	Epitaxial	Very low	High	Slow	Compound semiconductors— GaAs, InP, AlGaAs	Research and development in epitaxy and optoelectronics

for higher and higher resolution has led to the utilization of deep-UV light sources, such as excimer lasers at 248 and 193 nm. EUV sources are currently being developed at 13.5 nm. In lithography applications, the illumination intensity has to be uniform across the entire substrate surface. Gaussian beams are not acceptable and speckle patterns from laser sources need to be eliminated. Therefore, significant effort is spent in shaping the beam into a flat, uniform profile.

5.3.1.2 Photoresists

Photoresists are light-sensitive, organic polymers in a solvent and are most commonly applied to the substrate by spin coating. Photoresists consist of a photoactive compound, a resin, and a solvent. The purpose of the solvent is simply to allow the photoresist to be spin coated. After the spin coating step, the solvents are removed by heating the photoresist. The resin is the structural component of the photoresist that is used for subsequent pattern transfer. In i-line (365 nm) photoresists, the most commonly used resin is the novolac resin (belongs to the phenolic family). The photoactive compound is diazonaphthoquinone (DNQ). Upon exposure to UV light, the DNQ releases a photoacid that increases the solubility of the resin. This type of photoresist is also referred to as a positive-tone photoresist because the exposed areas will eventually be removed and the unexposed areas will remain. The dynamic range of the solubility can be greater than three orders of magnitude and highly nonlinear, which is what gives photolithography its excellent contrast.

The exposure is commonly measured in the units of millijoules per square centimeter (mJ/cm^2). This is the illumination intensity in mW/cm^2 multiplied by the exposure time. Typical dose values range from 50 to 500 mJ/cm^2.

Although the above description is for the most common type of photoresists, there are other photoresists based on different chemistries and mechanisms. One common variant is the negative-acting photoresist. These photoresists have the opposite behavior, in which the solubility of the photoresist is reduced after exposure to light. This occurs because of polymerization reactions that increase the molecular weight of the resin.

In the deep-UV wavelengths, the photosensitivity becomes greatly reduced because of absorption by the resin. As a result, a mechanism known as chemical amplification is utilized, in which a single photoacid molecule can act as a catalyst to create many reactions. Photoresists in the EUV range are far less common and are still under development.

5.3.1.3 Photomasks

Photomasks are transparent glass substrates on which metal patterns are created to block the transmission of light. This metal is typically chromium because it has excellent adhesion to glass and is very opaque to UV wavelengths. Standard photolithography processes are used to create the photomask, including the application of photoresists, exposure, and pattern transfer to the underlying chromium layer. However, for the exposure, instead of a photomask, scanning tools such as a laser

scanner or an electron-beam scanner are used. A UV laser source such as HeCd or ArF is used to raster scan the laser spot across the entire surface while turning the laser beam on or off with a shutter driven by software that contains the photomask design. Assuming a HeCd laser at a wavelength of 325 nm, the smallest spot size that can be achieved is on the order of 300 nm. A binary feature such as a line may require multiple widths of this spot. As a result, the generally advertised limit for laser-written photomasks is approximately 1 μm. Electron-beam writing is used when smaller features are required.

5.3.1.4   Contact photolithography

Contact photolithography is the simplest to describe and the most widely used method in research laboratories. The photomask is placed in physical contact with the photoresist-coated substrate and exposed to UV light, typically with a 365-nm mercury lamp, as illustrated in Figure 5.7. The extent of the contact is important because it determines the resolution of the features that can be imaged. The term critical dimension (CD) is often used in photolithography to refer to the width of the smallest line that can be printed. An unintentional small gap can cause the exposed patterns to diffract and become enlarged. Even in the ideal case of a perfect contact with a zero gap, there will still be diffraction through the thickness of the photoresist. The aerial image can be calculated, but because this is a near-field phenomenon it would require a numerical evaluation of the Fresnel integrals. However, if we can approximate the light exiting a photomask opening as a Gaussian shape, it becomes possible to derive a rough analytical formula. Using this approximation, the following formula shows the width of the Gaussian aerial image as a function of the photoresist thickness z, air gap s, and UV wavelength λ:

$$W_{\min} \approx \frac{3}{2} \sqrt{\lambda \left(s + \frac{z}{2} \right)}. \tag{5.1}$$

At a wavelength of 365 nm, with a zero gap and a photoresist thickness of 500 nm, the smallest feature we can print with contact lithography is approximately 450 nm. This is the best-case scenario, and any contamination on the substrate or photoresist will introduce additional gaps between the mask and photoresist and will degrade
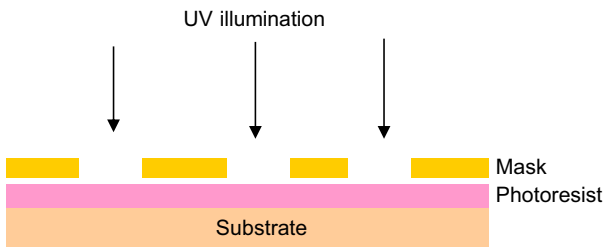

Figure 5.7 Contact photolithography.

the resolution. A dust particle 1 μm in size will create a 1-μm air gap. Using the above formula, we can verify that W_{min} will enlarge to 1 μm. This is one reason why contact photolithography must be performed in an ultra-clean particle-free environment.

One can reduce the photoresist thickness to improve resolution, but this results in a compromise. A photoresist that is too thin will not withstand the subsequent pattern transfer steps. In addition, the scaling of resolution with photoresist thickness is not linear—a 50% reduction in thickness will only produce a 30% improvement in resolution. The same is also true with the source wavelength. Therefore, the often claimed resolution limit of contact photolithography is about 500 nm.

The biggest advantage of contact photolithography is the simplicity of the equipment. The optical components are fairly simple and are used primarily for creating a uniform illumination profile. As a result, most research and development facilities rely primarily on contact photolithography. If the required resolution is beyond what can be achieved with this method, then EBL or FIB lithography can be used.

5.3.1.5 Projection photolithography

In projection systems, the photomask is held at a distance from the substrate and an optical imaging system is used to project the image of the photomask onto the substrate surface. The system has many similarities to an optical microscope, and even the resolution conditions are the same. A simplified projection system is shown in Figure 5.8.

The resolution limit using Abbe's criteria is

$$R = \frac{\lambda}{2\mathrm{NA}},\tag{5.2}$$

where λ is the UV wavelength and NA is the numerical aperture. This equation can be derived by considering a cone of incident waves up to a maximum angle of $\sin \theta$. All of these waves combine together to produce the aerial image. The features with the highest resolution will arise from the waves that have the largest wave vectors in the plane of the image. If k is the free space wave vector, then the largest wave vector parallel to the image plane will be $nk \sin \theta$, as illustrated in Figure 5.9.

Two of these counterpropagating waves will produce a standing wave with a wave vector of $2nk \sin \theta$. This will contain the largest spatial frequency of the image. If we convert this to spatial period, then we can get $R = \frac{\lambda}{2n \sin \theta}$. If we substitute $n \sin \theta = \mathrm{NA}$, we can easily obtain Abbe's equation. R is the smallest distance between two bright fringes in the image and is defined as the resolution of the image. In lithography, this quantity is called the pitch. If we assume the lines and spaces have equal widths, then the width of a single line will be half of this value. The half-pitch is typically defined as the CD of the lithography system.

Therefore, the half-pitch, HP, can be written as

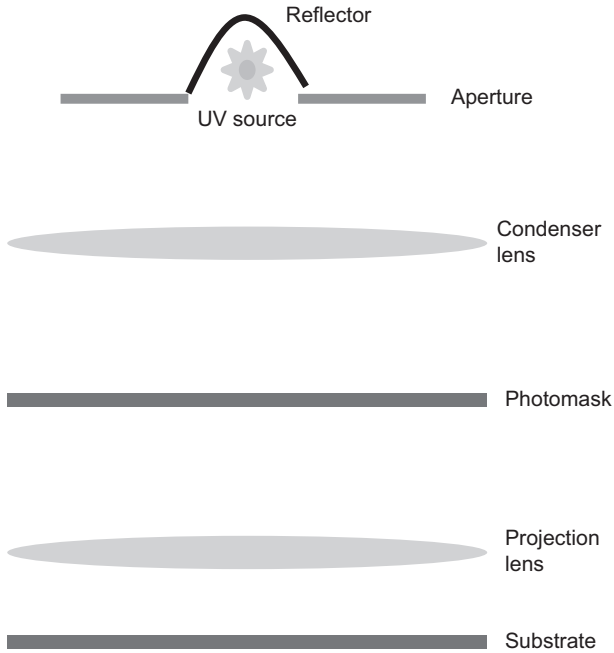$$\mathrm{HP} = \frac{\lambda}{4\mathrm{NA}}.\tag{5.3}$$

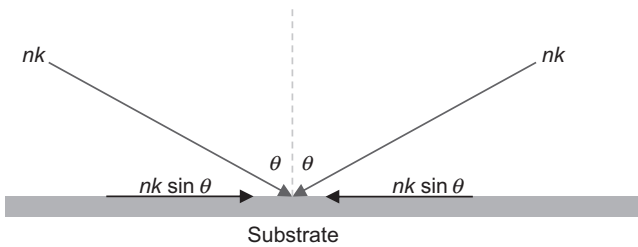**Figure 5.8** Elements of a projection photolithography system.

Figure 5.9 Numerical aperture and maximum spatial resolution.

In general, we lump the $^1/_4$ and several other optical system and photoresist factors into a single parameter k_1 and write the equation as

$$\mathrm{HP} = k_1 \, \frac{\lambda}{\mathrm{NA}}. \tag{5.4}$$

In older projection lithography systems, NA was lower than 1.0 and k_1 was higher than 0.25. For example, using a mercury lamp at 365 nm, if NA $= 0.5$ and $k_1 = 0.5$, we get a half-pitch of 365 nm. This was a fairly typical result, and the wavelength was often loosely treated as being equal to the half-pitch. Although this was not significantly better than contact lithography, most manufacturing systems used projection lithography because of their increased throughput and reliability. Because the mask

is never brought in contact with the substrate, the mask can be kept clean and the substrates can be easily switched. Furthermore, projection systems also allow one to demagnify the image. For example, using 5:1 reduction optics, 500-nm features can be 2.5 µm on the mask. This relaxes the feature sizes required on the photomask and the cost of producing photomasks.

The depth of focus is the vertical distance above and below the image plane where the smallest features remain in focus. This is

$$\text{DOF} = k_2 \frac{\lambda/n}{\sin^2 \theta}. \tag{5.5}$$

For a 365-nm illumination with NA $= 0.5$ (i.e., $\sin \theta = 0.5$) and $k_2 = 0.5$, we can get the DOF $= 730$ nm. This means the substrate and photomask have to be held perfectly parallel within 730 nm, and the photoresist film has to be perfectly flat with no topographic features exceeding this DOF value. However, even in the best wafers, minor substrate curvatures can easily exceed this DOF. This is one reason why an entire wafer is not exposed with a single image in a projection system. It becomes nearly impossible to satisfy the DOF when the image size is larger than approximately 1 in. Instead, projection lithography systems use a step-and-repeat configuration. The reduced image of the photomask is projected and exposed, then the wafer is translated and the same pattern is exposed over and over again to fill the entire wafer. Hence, these systems are also known as steppers. This is illustrated in Figure 5.10. The maximum size of a single exposure is defined as the field size and is determined by the substrate and mask flatness as well as the quality of the optics and spherical aberrations in the system compared with its DOF. For example, if a system can only maintain the DOF specifications within a 1- by 1-in area, and if the reduction optics are 5:1, then a 5- by 5-in photomask will be printed as a 1- by 1-in field on the substrate, which can be repeated several times to cover the entire substrate.

Although device dimensions have been shrinking, the demand for larger and larger chips have been driving the field sizes upward. This requires larger imaging
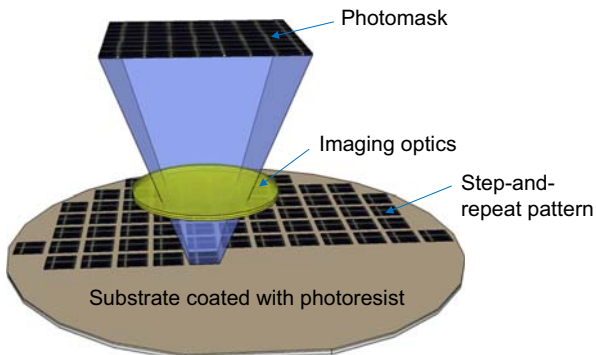

Figure 5.10 Step-and-repeat configuration.

lenses; however, this leads to increased aberrations. One way to overcome this problem is by projecting only a small portion of the image through a slit aperture using a smaller lens. The mask and the wafer are then linearly scanned synchronously to cover the entire aerial image. This is known as the step-and-scan system, and it is currently the most widely used commercial system.

To improve resolution, projection systems have moved toward shorter and shorter wavelengths, from 254 nm (KrF) and then to 193 nm (ArF) excimer lasers in the deep-UV range. In addition, the k_1 parameter has also significantly improved because of better optics with lower aberrations and photoresist performance and is currently very close to the theoretical limit of 0.25. The numerical aperture has also increased over the years and is now at 1.35. If it seems odd that the numerical aperture is higher than 1.0, its because this is the effective numerical aperture by replacing air with water as the imaging medium. This is known as immersion lithography. In microscopy, this is a commonly used method in which high-refractive-index oils are used to increase the numerical aperture. However, many oils are incompatible with photoresists, and many are not transparent in the deep-UV wavelengths; therefore, water has become the preferred medium for immersion lithography. The refractive index of water at 193 nm is 1.44. Putting all of these together, the HP becomes 35 nm and the DOF becomes 38 nm. This is a very small number, and it is difficult to maintain the parallelism between the substrate and photomask over more than 1 in. Even then, elaborate methods such as laser interferometric methods are used to measure the substrate-to-photomask distance and make fine corrections in real time.

Thirty-five nanometers was once considered the limit of the 193-nm immersion lithographic systems (abbreviated as "193i"). In recent years, techniques such as phase-shifted masks, optical proximity corrections, off-axis illumination, double patterning, dual tone resists, self-aligned double-patterning, etc. have been successfully developed to push the limit of this resolution even further. Currently, features as small as 20 nm are being produced by major integrated circuit manufacturers using the 193-nm laser source. This is an impressive $\frac{\lambda}{10}$ resolution that is being mass produced in consumer electronics without circumventing the effects of diffraction. Some of these resolution enhancement techniques are briefly described here.

Double patterning

Double patterning is based on the concept that exposed lines and spaces do not have be of equal widths. Assuming a positive-tone photoresist, a higher dose will result in narrower photoresist lines and wider spaces. Because the aerial image formed on the photoresist surface will always contain diffused edges due to diffraction, increasing the dose will enlarge the area that will receive the threshold dose required for dissolution. As a result, it is possible to print lines that are smaller than half-pitch. If the line width is 25% of the pitch, then that would leave 75% of the empty space to be used for printing another line, effectively doubling the density of lines. The half-pitch then becomes 17.5 nm.

In practice, this is done with two layers of hard masks (typically dielectrics) under the photoresist. The first set of lines are exposed, developed, and etched into the top hard mask layer. A new coating of photoresist is then applied, and the second mask

is carefully aligned to the previously created features and exposed. The second photoresist pattern combined with the previously patterned hard mask are then used to etch into the bottom hard mask layer. The top hard mask can then be removed leaving the bottom hard mask with the desired high-density pattern. This process is referred to as litho-etch-litho-etch. To simplify this process and to increase the reliability, another process known as litho-freeze-litho-etch was developed. In this process, only one hard mask layer is used. After the first photoresist pattern is developed, it is chemically treated to prevent it from dissolving during the second application of the photoresist. The second photoresist is then applied, aligned, and patterned. Finally, a single etch step is performed to transfer the high-density pattern into the underlying hard mask.

Dual tone photoresists

Positive-tone photoresists have low solubility when unexposed and become highly soluble when exposed beyond their threshold dose. Dual tone photoresists are designed to have low solubility when the dose is too low and also when the dose is too high. This allows the photoresist to dissolve only when the dose falls within a certain band in the middle. When exposed with an array of diffused lines and spaces, the resulting photoresist pattern will produce two lines for each line in the aerial image, effectively doubling the density of lines.

Self-aligned double-patterning

Self-aligned double-patterning (SADP) is widely used in today's manufacturing systems. The process uses the edges of each line to create a new line. Because each line contains two edges, it effectively doubles the line density. Furthermore, this process does not require a second lithography step or alignment. Hence, it is referred to as a self-alignment process.

First, the photoresist is exposed and developed with the initial features. These features are then isotropically coated with a film. The film is then directionally etched so that it is removed from all horizontal surfaces. This leaves the films on the sidewalls of each feature. The density of these features will be twice that of the original line density. This is illustrated in Figure 5.11. The isotropic coating and directional etch can be repeated for a second time, doubling the line density again. This will result in four times the original density. This is referred to as self-aligned quadruple patterning. Yet another iteration will result in self-aligned octuplet patterning.

Because of its simplicity and self-alignment capability, SADP has allowed the manufacturers to use their existing 193 immersion systems to push the limits further without significant retooling.

5.3.1.6   EUV lithography

F_2 lasers emitting at 157 nm were expected to supersede the ArF 193-nm lasers in projection lithography. That did not happen because of several technical hurdles, one of which was the requirement for CaF_2 optics because silica is not transparent at this wavelength. In addition to the high cost of CaF_2 optics, special design considerations were required to overcome its birefringence. As a result, the 157-nm method
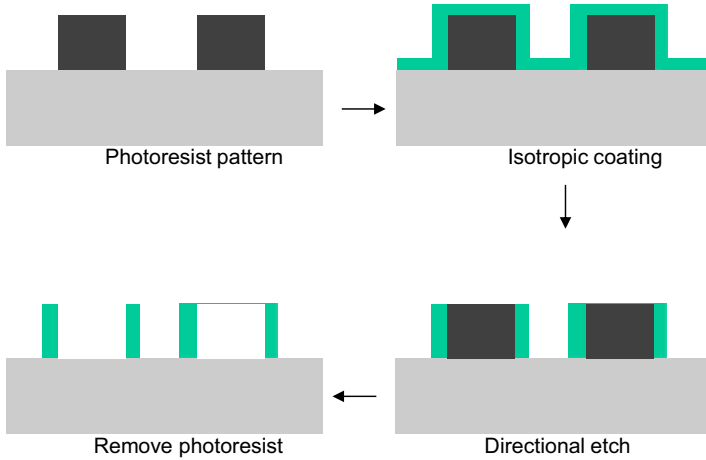
Figure 5.11 Illustration of the SADP process.

never became a mainstream lithography method, and the industry continued to look for sources in the EUV range. One of these is the 13.5-nm EUV method. This is not just a shorter wavelength, but it is also a dramatic shift in all aspects of lithography because the wavelength is almost in the regime of X rays. The whole optical train has to be in vacuum. High-power light sources with efficiency and uniformity are required. New photoresists and new approaches for making photomasks had to be developed. Because of the difficulty in making refractory optics in EUV, an all-reflective approach was developed using multilayer Bragg mirrors. The photomasks are also reflective instead of transparent. This is currently an active area of research, and many tool manufacturers are heavily investing in the EUV method. For more information on EUV lithography, the reader is referred to the special sections on EUV in the SPIE JM3 journal [9,10].

5.3.1.7 *Laser interference lithography*

Laser interference lithography is a maskless projection method to create periodic lines of exposure with line widths as small as $\frac{\lambda}{4n}$. A laser beam is split in two and recombined at the substrate surface to create a periodic array of bright and dark lines (see Figure 5.12). If two coherent beams of wavelength λ are incident at an angle θ from opposite sides of the surface normal, then Figure 5.9 can be used to show that

$$\Lambda = \frac{\lambda}{2n \sin \theta}, \tag{5.6}$$

where Λ is the pitch and n is the refractive index of the incident medium. The smallest pitch will occur when the incident angle is 90°. Using a 266-nm laser, the smallest pitch will be 133 nm with a line width of 66 nm (assuming a 50% duty cycle). In practice, it is not possible to illuminate the substrate at 90° incidence. The maximum
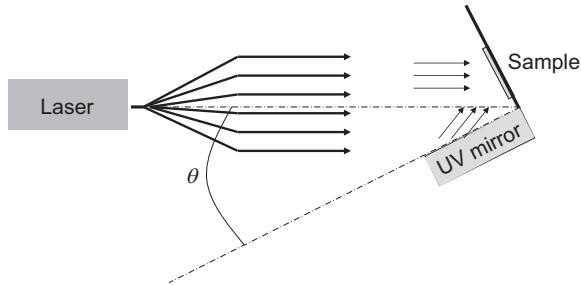
Figure 5.12 Laser interference lithography using a Lloyd's mirror setup.

angle is likely to be close to 60°. This would result in a pitch of 153 nm and a line width of 76 nm. It is also possible to perform two consecutive exposures after rotating the substrate in plane by 90° to achieve a two-dimensional grid pattern.

One commonly used configuration for interference lithography is the Lloyd's mirror setup, which contains two planes at right angles to each other. A mirror is installed on one plane, the substrate is placed on the other plane, and the laser beam is expanded and collimated to cover both planes. From simple geometry, we can show that the angles of incidence of each beam on the substrate will be equal, which creates the interference pattern. Furthermore, the entire fixture can be rotated to change the angle of incidence, which gives it the flexibility to easily change the pitch of the periodic fringes.

Interference lithography is used to make diffraction gratings and photonic crystals. Its advantage is that it is a maskless method and requires a fairly simple setup to achieve line widths that rival even the most advanced projection lithography system. It can cover a large substrate area, much larger than what is currently possible with mask projection systems. It also does not suffer from any depth-of-field issues. Compared with the tens of millions of dollars it costs to acquire and set up a 193-nm projection lithography tool, an interference lithography setup can be built for far less. Its biggest disadvantage is that it can only make periodic structures. Nevertheless, interference lithography is still a powerful tool that is used anytime large-area periodic nanostructures are required.

Figure 5.13 shows two typical examples of one- and two-dimensional periodically patterned photoresist films made by interference lithography.

5.3.2 Nonoptical lithography

Although optical lithography is the dominant technique, there are other techniques that do not use light to print a pattern on a resist layer. We will use the term resist rather than photoresist to indicate the lack of photosensitivity in this application.

5.3.2.1 Electron-beam lithography

Of all of the nonoptical lithographic methods, this is the most commonly used method. It uses a beam of electrons rather than photons to expose the resist to induce a chemical
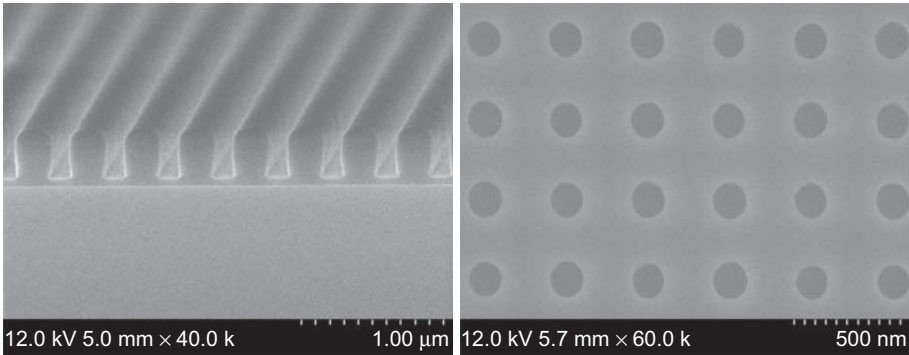
Figure 5.13 Scanning electron microscope images of photoresist features made using laser interference lithography.

change, which subsequently results in a change in solubility. However, unlike photo-lithography, there is no mask and the beam does not illuminate the entire substrate. In that sense, EBL does not have any similarities with contact lithography or projection lithography. Instead, the electron beam is generated, accelerated, and focused down to a small spot and scanned across the substrate to create the image. The scanning mechanism is done by a combination of mechanical translation of the substrate stage as well as tilts in the electron beam. The pattern is generated by modulating the beam current on and off as the beam is scanned. The equipment is similar to a scanning electron microscope, where the e-beam, steering coils, and the substrate are all housed in a high-vacuum chamber. The resolution limit follows the same relation as with photons:

$$\text{HP} = k_1 \frac{\lambda}{\text{NA}}. \tag{5.7}$$

The wavelength of electrons can be computed from de Broglie's principle, which states $\lambda = \frac{h}{p}$, where h is Plank's constant and p is the momentum. If the kinetic energy of the electron is E, then we can use $E = \frac{1}{2}mv^2$ and $p = mv$ to obtain $p = \sqrt{2mE}$ where m is the mass of the electron. This allows us to express the wavelength in terms of kinetic energy as

$$\lambda = \frac{h}{\sqrt{2mE}}. \tag{5.8}$$

If the electron is accelerated by 10 kV, then it will acquire a total energy of 10 keV and the wavelength can be calculated to be approximately 12 pm. This is many orders of magnitude smaller than UV wavelengths and is the main attraction of electron beams. It has the potential for an extremely small resolution limit, although much of it is still unrealized as of today.

The numerical apertures of electron-beam systems are typically very low, on the order of 0.01. This has the benefit of an extremely large depth of focus. Despite the

low numerical aperture, the expected resolution should still be on the order of 1 nm. However, in practice, this is not the case. The k_1 parameter in electron-beam systems is much larger than in optical systems. As we saw earlier, the smallest theoretically possible value for k_1 is 0.25, which with 10-keV electrons would lead to a subnanometer resolution. However, in current EBL systems, the value of k_1 is on the order of 5. This is due to the poor spherical and chromatic aberrations in the magnetic focusing system as well as the interaction of the electron beam with the substrate that emits secondary electrons that cause blurring. As a result, the practical resolution of EBL is only in the range of 5 nm.

The biggest advantage of EBL is its high resolution. The depth of focus is also very large because of the small numerical aperture. This makes it a very forgiving system for performing lithography on substrates with topographic features and small curvatures. Currently EBL is widely used in the manufacture of photolithography masks, especially when the required resolution on the mask is beyond what can be achieved with a laser scanner. This is by far the largest commercial application of EBL today. In research and development, EBL is used whenever the required features are smaller than approximately 500 nm, by directly writing the pattern on the resist layer, bypassing a mask. Although deep-UV projection with immersion can easily reach below 100 nm, such tools are generally not available outside of large production environments. Hence, for research and development, EBL becomes the tool of choice where contact photolithography leaves off. The biggest disadvantage of EBL is its slow speed. Because it is a scanning system, it is inherently slow. This speed is inversely related to the resolution—as the beam spot size is reduced to achieve a higher resolution, its speed will also decline. To write a 1- by 1-in area could take several hours depending on the density of the pattern and the required dose. Another drawback is the requirement for the substrate to be conductive. As in the scanning electron microscope, the beam current needs a path to ground through the substrate to maintain charge neutrality. Any localized charging effects can significantly diminish the resolution. For photomask writing, although the substrate is glass, a metal film is deposited first before coating the resist film, and this metal film is electrically grounded. Writing on a resist film on a silicon substrate is also possible because silicon is partially conductive. Other cases of purely insulating systems will require careful consideration of how to dissipate the charge accumulation.

The dose in EBL is the charge per unit area, commonly expressed in microcoulombs per square centimeter ($\mu C/cm^2$). The beam current in combination with the scan speed affects the dose that is deposited on the substrate. Because long write times are the main disadvantage of EBL, optimization algorithms are used to minimize the write times. The current and beam size are dynamically adjusted based on the resolution of the local feature being written. They are increased when writing large features and reduced for finer features.

The resist mechanism is different from UV photoresists. In positive-tone resists, a process known as chain scission decreases the molecular weight of the exposed areas and increases the solubility. In negative-tone resists, a cross-linking process increases the molecular weight and reduces the solubility.

5.3.2.2 FIB milling

This is a maskless, resistless scanning lithography technique. It is very similar to EBL except it uses an ion beam instead of an electron beam to write the pattern. In addition, the beam is not used for exposing a resist; instead, the beam directly sputters and removes the substrate (or thin film), hence bypassing the need for a sacrificial resist layer.

Liquid metals are used for the ion source because of their heavier atomic weights, which will result in a higher sputter yield. Of these, gallium is the most common because of its low melting temperature. The gallium atoms are heated, ionized, accelerated, and focused to a small spot on the order of 5 nm on the substrate stage. The acceleration energy of the incident ions and substrate type will determine the sputter yield and hence the removal rate of atoms.

Redeposition of the sputtered atoms is a major consideration in FIB. This is when the sputtered atoms land back in the vicinity of the milled site, causing surface roughness. For this reason, a gas injector is used to volatilize the sputtered species and accelerate the removal rate of atoms. With suitable precursor gases, it is also possible to induce a CVD at the ion bombardment site. These different operational modes—sputter removal, reactive etch, and deposition—make FIB a very versatile tool. However, similar to EBL, its main disadvantage is its low speed. FIB is currently used to make repairs to photomasks by milling away unwanted metal traces and depositing missing metal traces as well as surgical repairs to integrated circuit chips.

Another very useful aspect of FIB is that it can be easily combined with a scanning electron microscope in the same vacuum chamber in a dual beam configuration. An electron gun and an ion gun can both be directed at the substrate, which allows one to direct the ion beam and monitor the milling process in real time. As in EBL, the substrate must be electrically grounded.

5.3.2.3 Nanoimprint lithography

NIL is very simple to describe. A mold (stamp) is first constructed with surface relief features on it and is then pressed against a polymer material to transfer the pattern. This is illustrated in Figure 5.14. The polymer material is typically a sacrificial resist film that is used as a mask to etch the underlying substrate or film. Interest in NIL has grown significantly because of its potential for nanoscale manufacturing.

There are several different NIL approaches. Thermocompression NIL uses elevated temperatures to soften the polymer film to allow it to flow around the surface reliefs of
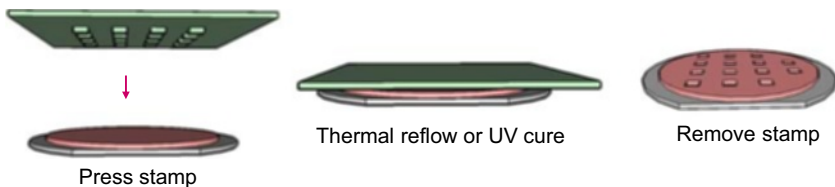

Press stamp   Thermal reflow or UV cure   Remove stamp

Figure 5.14 Nanoimprint lithography.

the stamp. The stamp is removed after allowing it to cool back to room temperature. This is also known as hot embossing lithography. Another variant uses UV light to cross-link and harden the polymer film. A soft liquid film is used, and during the imprint the film is exposed to UV light. This can also be used in combination with elevated temperature. UV-NIL requires the substrate and/or stamp to be transparent to UV wavelengths such as quartz.

The main advantage of NIL is the extreme simplicity of the process. There are no expensive optics, ultra-high-vacuum chambers, electron beams, or excimer lasers. Feature sizes also do not have a fundamental resolution limit. However, in practice, there are several challenges that have prevented NIL from being adopted on a larger scale. NIL requires physical contact between the stamp and the substrate. After several cycles, the stamp will have to be cleaned. In that sense, this is similar to contact photo-lithography, except NIL can transfer much finer features than contact photolithography. The flatness of the substrate and stamp are also critical factors. Any protrusions or topographic features will compromise uniform contact between the stamp and the film. However, this is not a significant disadvantage, at least when compared with deep-UV projection systems, which also have limited depth-of-focus problems.

NIL is currently used in many research and development laboratories, but it is still under development for large-scale manufacturing applications. The proposed NIL manufacturing systems use a step-and-repeat configuration just like in the UV steppers. Some of the manufacturing challenges are air entrapment between the stamp and the substrate, particulates, stamp contamination, and throughput.

Table 5.2 summarizes some of the main advantages and disadvantages of the different lithography methods discussed in this section.

5.4 Pattern transfer

Once the resist film is patterned by lithography, chemical processes are used to transfer that pattern into the underlying substrate or film. This is the pattern transfer process. There are exceptions, such as FIB lithography, in which the pattern is directly etched without a resist, and the lift-off process, in which a patterned resist is used to lift an overlying film simply by dissolving the resist and not by etching the film. The etching processes are generally divided into wet-chemical etching and plasma etching.

5.4.1 Wet-chemical etching

This process requires only liquid chemicals; therefore, it is very simple and inexpensive to implement. For example, for etching a copper film, as is routinely done in the manufacture of printed circuit boards, the resist is applied and patterned over the copper film and then a chemical such as ammonium persulfate or ferric chloride is used to selectively etch the copper, leaving only the areas of copper that are protected under the resist film. The resist is then stripped off with a solvent.

Table 5.2 Summary of lithographic methods

Method	Wavelength	Approximate half-pitch[a]	Depth of focus	Common applications
Contact lithography	365 nm (Hg)	500 nm		R&D
Projection lithography	365 nm (Hg)	350 nm		R&D and small production
Projection lithography	193 nm (ArF)	75 nm		Production systems
Projection immersion lithography	193 nm (ArF)	35 nm		Production systems
Projection lithography with immersion and resolution enhancement	193 nm (ArF)	20 nm		Production systems
EUV lithography	13.5 nm	5 nm		Still under development for production systems
Laser interference lithography	325 nm (HeCd), 266 nm (YAG), 248 nm (KrF), 193 nm (ArF), etc.	100−500 nm	Infinite	Periodic structures such as gratings
EBL	0.01 nm (e-beam)	5 nm	Large	R&D laboratories and for making masks for production systems
FIB lithography	Gallium ions	10 nm	Large	Mask repair, chip repair, R&D
NIL	N/A	N/A	N/A	R&D with potential for commercial use

R&D, research and development.
[a]Half-pitch values given here are realistic values for these systems and not their theoretical limit.

Although simple, there are several limitations to this type of etch. First, the etch chemistry is directionally isotropic. This means the film is etched vertically and horizontally at nearly equal rates. The consequence of this is a curved sidewall that results in a larger opening at the top than the original opening in the resist. The excess lateral etch under the resist is known as the undercut.

The etch process can be described as a sequence of three steps: (1) the etchant diffuses from the bulk liquid to the surface being etched, (2) the surface reaction takes place, and (3) the by-products diffuse out to the bulk liquid. The rate of each step determines the overall etch rate. This is known as the rate-limiting step. The actual model is complex because it involves the diffusion of different species, all of which depend on their respective concentration gradients, but we can consider a simplified model as follows:

$$\frac{1}{R} = \frac{1}{R_i} + \frac{1}{R_r} + \frac{1}{R_o}, \tag{5.9}$$

where R_i is the diffusion rate of the fresh reactants to the reaction site, R_r is the etching reaction rate, and R_o is the diffusion rate of the by-products away from the site. The smallest of these factors determines the overall etch rate R. The surface reaction rate R_r is a function of the concentration of the reacting species and is exponentially dependent on temperature. This can be written as

$$R_r = k_0 [A]^n [B]^m e^{-\frac{E_a}{kT}}, \tag{5.10}$$

where $[A]$ and $[B]$ are the concentrations of the reacting species (assuming there are only two species), E_a is the activation energy for the reaction, k is the Boltzmann constant, T is the temperature, n and m are constants, and k_o is the rate constant. The diffusion rates R_i and R_o are functions of the geometry, such as mask opening and aspect ratio of the etch profile. The latter will be a dynamically varying parameter as the etch progresses.

Wet-chemical etching is mostly used for etching large geometries where the diffusion rates are large; therefore, the rate-limiting step is primarily the surface reaction rate. Therefore, the overall rate depends primarily on the concentration of the fluid and the temperature. In this regime, the etch profile will be ideally isotropic. The diagram in Figure 5.15 illustrates this, where a thin film of thickness t is wet-etched through a mask opening of width w.

Using geometry, when the width of the etched trace is w at the bottom of the trench, the width at the top of the trench will be:

$$w' = w + 2t, \tag{5.11}$$

or expressed as a fraction of the original space width,
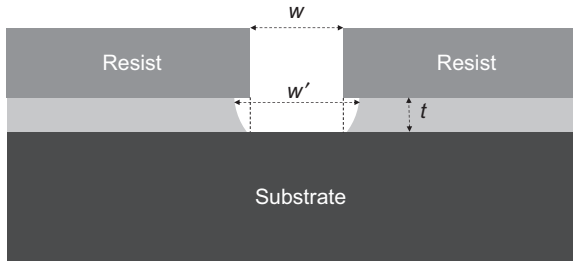
$$\frac{w'}{w} = 1 + 2\frac{t}{w}. \tag{5.12}$$

Figure 5.15 Illustration of wet etching characteristics.

Here are some numerical results to illustrate the line width:

- $t = 100$ nm, $w = 500$ nm $\rightarrow \frac{w'}{w} = 1.4$
- $t = 100$ nm, $w = 250$ nm $\rightarrow \frac{w'}{w} = 1.8$
- $t = 100$ nm, $w = 100$ nm $\rightarrow \frac{w'}{w} = 3.0$

Clearly, as the mask dimensions shrink toward and below the thickness of the film, we will not be able to etch the film without a significant loss of resolution. The film thickness is often limited by other design criteria such as the optical dielectric constant or electrical resistivity; therefore, it is not a parameter than can be easily changed. This is the biggest drawback of wet-chemical etching—the isotropic nature results in a significant enlargement of the traces as the dimensions get smaller. The advantage of wet-chemical etching is its low cost. Another significant advantage is the high selectivity. Selectivity is defined as the etch rate of the film (desired etch rate) divided by the etch rate of the resist (undesired etch rate):

$$S = \frac{R_{\text{film}}}{R_{\text{resist}}}. \tag{5.13}$$

Because the etch rate is entirely driven by chemical activity, it is possible to find chemicals that will etch the film and leave the resist virtually untouched. Most materials commonly used in electronics have well-established etchants with high selectivity with the resist. A comprehensive list of wet etchants for various materials can be found in the references listed at the end of this section. However, there are some cases in which the resist simply will not stand up to the etchant. One example is the etching of silicon with HNA solution (hydrofluoric, nitric and acetic acid mixture). This is an aggressive etchant, and most polymer-based resists will be removed very quickly. Therefore, the commonly used technique is to use the resist to pattern a more chemically resistant hard mask such as silicon nitride. Then, the patterned hard mask is used to etch the underlying silicon with the HNA solution.

It should also be noted that not all wet chemicals produce isotropic etches. In crystalline materials, the etch rate can be strongly dependent on the crystal orientation. The etch rate R_{r} will then have a strong dependence on the angle θ. Such etches are known as anisotropic wet-chemical etches. One example of this is the etching of

silicon with KOH (potassium hydroxide). When the surface orientation of silicon is <100>, the etch rate drops to zero at 54° to the surface along certain planes. The end result will be an etch profile that has sidewalls sloping at 54° where the etch rates come to a stop. An example of this etch profile is shown in Figure 5.16. This type of etch is used for making components in which the etch needs to come to a stop at a sharp corner (such as in atomic force microscopy tips) and in micro electro mechanical systems (MEMS).

5.4.2 Plasma etching

Plasma etching is also referred to as dry etching because it is performed in a gas phase without the use of liquids. It is sometimes referred to as reactive ion etching (RIE), although the correct term should be ion-assisted chemical vapor etching. In plasma etching systems, the substrate is placed in a vacuum chamber on the cathode of the plasma generator and gases are introduced to produce the reaction. This is illustrated in Figure 5.17. One advantage of this system is that fairly safe gases can be fed into the chamber, which in a plasma state become dissociated into highly reactive species. An example of this is CF_4. This gas is fairly inert under normal conditions, but in a plasma it can generate many F atoms (free radicals), which are highly reactive and spontaneously attack silicon to produce SiF_4. Because SiF_4 is a gas, silicon will be readily turned to a gas in such a plasma reaction. In addition, ions in the plasma will bombard the cathode similar to sputtering. This action creates an additional source of energy that can accelerate the etch rate parallel to the ion trajectories. Because ions are directed at normal incidence to the cathode, this has the effect of accelerating the etch rate normal to the substrate. This will result in an etch profile that has minimal undercut and is strongly anisotropic, as illustrated in Figure 5.18.

An important requirement in plasma etching is for the by-products to be volatile; that is, they need to be able to evaporate away into the pumping system. Any nonvolatile by-products will remain on the substrate as a deposited thin film and may impede
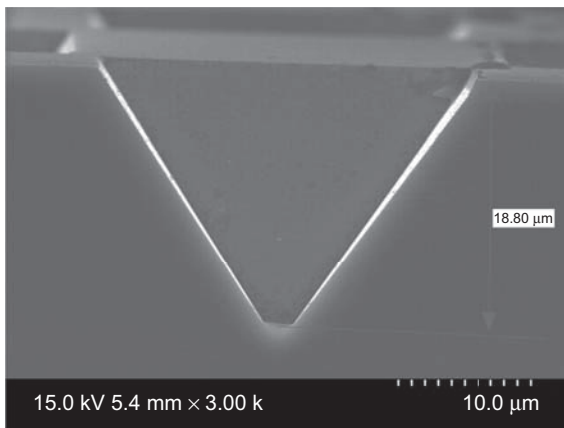

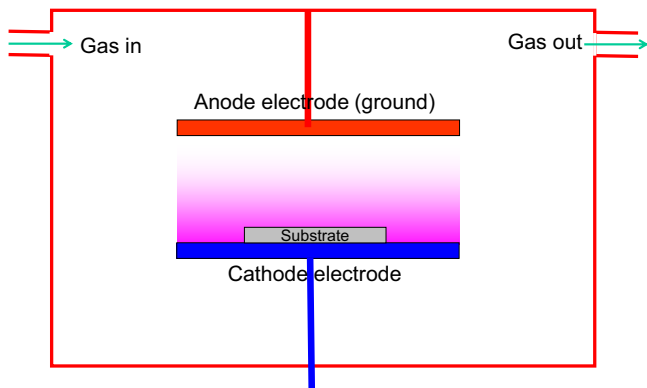**Figure 5.16** Profile of <100> silicon etched in KOH solution.

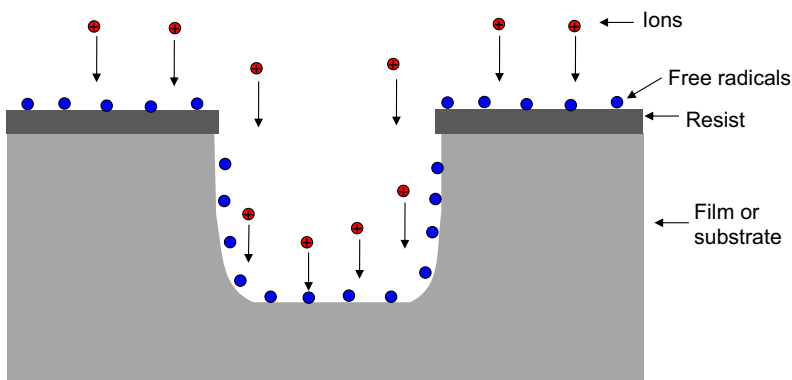**Figure 5.17** Parallel plate plasma etching configuration.

Figure 5.18 Interaction of the plasma with the substrate and mask.

the etching. Therefore, the gas chemistry has to be matched to the substrate being etched to produce the most volatile by-products. In addition to selecting the correct gas species, one can also control the substrate temperature, pressure, and plasma discharge powers to control the reaction and the by-products. The most common gases used in etching are fluorine-based or chlorine-based. Common fluorine-based gases are CF_4, SF_6, CHF_3, C_4F_8, etc. and chlorine-based gases are Cl_2, BCl_3, CCl_4, etc. Most semiconductors and metals can be etched with these gases. Unlike wet-chemical etching, it is not a trivial task to switch gases to etch different materials because several systems have to be specifically installed for each gas type, including piping, flow controllers, and exhausts.

Plasma etching is widely used for etching nanoscale features because it can produce vertical sidewall profiles with little or no undercut. An example is shown in Figure 5.19. This anisotropic feature can be further enhanced by allowing the deposition of certain passivating films to occur along vertical sidewalls during the plasma reaction. Along horizontal surfaces, these films will be removed by ion bombardment,
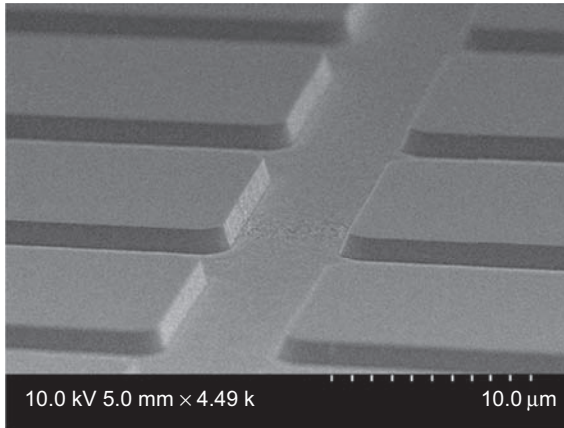
10.0 kV 5.0 mm × 4.49 k   10.0 µm

Figure 5.19 Example of a plasma etched substrate.

but they will remain on the vertical walls. By coating the sidewalls with these films, it is possible to produce very deep high-aspect-ratio features. Such etch processes are known as deep-RIE. Another aspect that makes plasma etching suitable for small features is the ability of gas species to diffuse into smaller features better than liquids. Plasma etching is also highly repeatable compared to a liquid bath.

The disadvantage of plasma etching, in addition to the higher cost of the system compared with wet-chemical etching, is the reduced selectivity. This arises because of the ion bombardment that can lead to sputter removal of the resist. Therefore, the resist gets consumed faster than in wet-chemical etching. As a result, ion-driven etches will have a poorer selectivity than free-radical-driven etches. Many dielectrics such as SiO_2 need ion bombardment to break bonds and make the molecules reactive. Without ion bombardment, these films will not etch. This will naturally lead to a lower selectivity. On the other hand, materials such as Si and certain metals can be etched by free-radicals alone, without the assistance of ions. These will have a higher selectivity. The resist thickness and selectivity determines the maximum etch depth that can be achieved. For example, if the resist thickness is 100 nm and the selectivity is 2, then the maximum etch depth that can be achieved will be 200 nm.

In a capacitively coupled plasma chamber, as shown in Figure 5.17, it is not possible to independently control the ion density and the free radical density. Both of these quantities are coupled. The only variable that can be controlled is the plasma discharge power, which will simultaneously increase both the ion density and the free radical density. When etching different materials, some only require ion bombardment, others only require free radicals, and some require both. Therefore, it is useful to have independent control of these parameters. This is accomplished through a capacitively coupled and inductively coupled plasma. In this system, the capacitively coupled part controls the ion density and free radical density, and the inductively coupled part controls only the free radicals. This is a very common configuration used in advanced plasma etching systems today.

Problems

1. Compare the following metals in terms of the temperature required to reach a vapor pressure of 1 torr and whether it will be a liquid or solid at this temperature: (a) titanium, (b) tungsten, (c) gold, (d) indium (use reference [2]).

2. Consider the sputter deposition of gold using an argon plasma.
   a. If the ion acceleration is 500 V and the ion current is 500 mA uniformly distributed over a 1-in square target area, then calculate the rate of atoms sputtered from the target, S. You need to look up the sputter yield of gold.
   ($S = 6.2 \times 10^{18}$ atoms/s)
   b. The substrate is held at a distance of $D = 6$ in from the target. Approximating the target as a point source, and assuming the sputtered atoms have an angular distribution of $\cos \theta$ where θ is the angle from the normal to the target surface, show that the deposition rate of the substrate along $\theta = 0$ is $R = \frac{S}{\pi D^2}$. Calculate this value.
   ($R = 8.6 \times 10^{15}$ atoms/cm^2/s)
   c. Using the above value, calculate the film growth rate on the substrate at $\theta = 0$. You need to use the atomic weight, density, and Avogadro's number.
   (14.5 Å/s)

3. In this problem, consider a PLD system using a 157-nm F_2 laser.
   a. If F_0 is the laser fluence (energy per pulse per unit area), α is the absorption coefficient, and R is the reflection coefficient, then show that the absorbed energy density is $U_a(z) = F_0(1 - R)\alpha e^{-\alpha z}$.
   b. The ablation depth is defined as the depth at which the absorbed energy density becomes equal to the energy density required to vaporize the material U_{vap}. Using this definition, show that the ablation depth can be written as $d_{vap} = \frac{1}{\alpha}\ln\left(\frac{F_0}{F_{th}}\right)$, where $F_{th} = \frac{U_{vap}}{(1 - R)\alpha}$.
   c. The following experimental values were measured for the ablation of fused silica:

F_0 (J/cm^2/pulse)	d_{vap} (nm/pulse)
2	40
10	135

   Calculate the threshold fluence F_{th} and the absorption coefficient α.
   ($F_{th} = 1$ J/cm^2/pulse; $\alpha = 1.7 \times 10^5$/cm)
   d. Look up the complex refractive index of fused silica at 157 nm from Ref. [5].
   e. Look up the explanation offered in Ref. [6] for the large difference between the two absorption constants.

4. Consider a projection lithography system that uses an i-line mercury lamp source, NA = 0.6 and $k_1 = 0.5$ with 4× reduction optics. Determine the smallest line width allowed on the photomask and the corresponding line width of the exposed line.
   (1.21 μm, 304 nm)

5. Consider a 193i lithography with SADP and NA = 1.3 capable of printing 22-nm-wide lines. Calculate the effective value of k_1 for this system and the depth of focus.
   (0.15, 49 nm)

6. Consider the Lloyd's mirror configuration in Figure 5.12. Assume the incident beam is collimated.

    a. Show that the interfering beams are incident at angles of $+\theta$ and $-\theta$ from the substrate normal.

    b. Derive Eqn (5.6).

    c. Assume the source is a 325-nm HeCd laser. The required pattern is a two-dimensional photonic crystal consisting of oval posts that has a pitch in the x-axis of 300 nm and pitch in the y-axis of 250 nm. Describe how this structure can be created using this system.

7. Calculate the wavelength of an electron that has been accelerated by 5 kV. Explain why the exposure resolution using this beam is typically significantly larger than this wavelength in most EBL systems.

    (17 pm)

8. Consider a 100-nm-thick metal film with a 250-nm-wide and 400-nm-thick photoresist line on top. If the metal is wet-etched until all of the metal is removed except the area under the resist, then calculate the width of the resulting metal line. Assume the etch selectivity is infinite.

    (50 nm)

9. The same structure as above is etched with a plasma instead of wet chemistry. The etch is anisotropic with a selectivity $S = 0.8$. Calculate the width of the resulting metal trace and the thickness of the remaining photoresist when the etch is completed.

    (250 nm, 275 nm)

References

[1] R.E. Honig, Vapor pressure data for the more common elements, RCA Rev. 18 (1957) 195−204.

[2] D.R. Lide (Ed.), CRC Handbook of Chemistry and Physics, 2005. Internet version.

[3] Y. Yamamura, H. Tawara, Energy dependence of ion-induced sputtering yields of monatomic solids, At. Data Nucl. Data Tables 62 (2) (March 1996).

[4] L. Martinu, D. Poitras, Plasma deposition of optical films and coatings: a review, J. Vac. Sci. Technol. A 18 (6) (November−December 2000).

[5] Palik, Handbook of the Optical Constants of Solids, CRC Press, 1998.

[6] P.R. Herman, et al., Processing applications with the 157-nm fluorine excimer laser, Proc. SPIE 2992, Excimer Lasers, Optics, Appl. 86 (March 31, 1997).

[7] S.M. George, Atomic layer deposition: an overview, Chem. Rev. 110 (2010) 111−131.

[8] D.B. Chrisey, G.K. Hubler (Eds), Pulsed Laser Deposition of Thin Films, Wiley-Interscience, ISBN-13: 978−0471592181.

[9] Special section on EUV sources for lithography, SPIE J. Micro/Nanolithogr., MEMS, MOEMS 11 (2) (April 2012).

[10] Special section on photomasks for extreme ultraviolet lithography, SPIE J. Micro/Nanolithogr., MEMS, MOEMS 12 (2) (April 2013).

Further reading

[1] C. Mack, Fundamental Principles of Optical Lithography: The Science of Microfabrication, John Wiley & Sons Ltd, NY, 2007, ISBN 978-0-470-01893-4.

[2] C.-S. Kim, S.-H. Ahn, D.-Y. Jang, Review: developments in micro/nanoscale fabrication by focused ion beams, Vacuum 86 (2012) 1014−1035.

[3] H. Schift, Nanoimprint lithography: an old story in modern times? A review, J. Vac. Sci. Technol. B 26 (2) (March/April 2008).

[4] K.R. Williams, K. Gupta, M. Wasilik, Etch rates for micromachining processing — Part II, J. Microelectromech. Syst. 12 (6) (December 2003).

[5] H. Jansen, H. Gardeniers, M. de Boer, M. Elwenspoek, J. Fluitman, A survey on the reactive ion etching of silicon in microtechnology, J. Micromech. Microeng. 6 (1996) 14−28.

Nanocharacterization

6

J.W. Haus
University of Dayton, Dayton, OH, USA

> *Measure what can be measured, and make measurable what cannot be measured.*
> **Galileo Galilei**

> *On looking back on this event, I am impressed by the great limitations of the human mind. How quick are we to learn, that is, to imitate what others have done or thought before. And how slow to understand, that is, to see the deeper connections. Slowest of all, however, are we in inventing new connections or even in applying old ideas in a new field.*
> **Fritz Zernike, 1953 Nobel Lecture on his observations of diffraction grating ghost images, explaining them and using them to develop phase contrast microscopy.**

6.1 Introduction

The first glimpse into the microscopic world was made possible by the development of the optical microscope. The history of the microscope is murky, with no one name attributed to the breakthrough invention. However, the historical record shows rapid progress (i.e., magnifying images beyond the power of a simple magnifying glass) in microscopy emerging from the Netherlands in the 1500s with the invention of the compound microscope. The book *Micrographia* by Robert Hooke published in 1665 is most often quoted for its detailed images and the definition of the "cell" from his biological observations. The microscope was adopted in many fields of science and was especially useful in biology and in advancing our medical knowledge of diseases. It enabled the direct observation of bacteria and organelles in cells. In the twentieth century, the new physics of X rays and electron waves were used to image objects with even greater resolution. As technology pushed the limits to a smaller scale, a myriad of new instruments were developed that are broadly called scanning probe techniques. Let us set the stage for understanding the advantages and limitations of nanocharacterization by first introducing the essential operational characteristics of microscopes. The principles introduced in the next section will then be examined when discussing twentieth-century instruments.

6.2   Basic optics

6.2.1   *Geometric optics using compound microscope*

The compound microscope consists of two parts: one containing objective lenses
(O) that magnify and focus the light from the object in an image plane (I) and the other
an eyepiece (E) that further magnifies the image and produces a virtual image of the
object (dashed outline). A sketch of the essential elements of a compound microscope
is found in Figure 6.1. Zacharias Janssen of Middleburg, Holland is credited with
inventing the compound microscope in 1590. In the figure, the compound microscope's
components are illustrated with single lenses; the actual parts contain several lenses.

   To better understand how a lens works, the basic principles of refraction and reflec-
tion are applied. Snell's law already discussed in the electrodynamics chapter provides
a geometric optic foundation to describe the magnification power of a lens. Without
going into detail, the object distance (p) and focal plane distance (q) for a thin lens
are given by the lensmaker's formula:

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f}, \tag{6.1}$$

where f is the focal length of the lens. The focal length satisfies the equation

$$\frac{1}{f} = \frac{(n - n_e)}{n_e}\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{6.2}$$

where n is the refractive index of the lens material and n_e is the refractive index of the
environment surrounding the lens. R_1 and R_2 are the radii of curvature of the lens
surfaces on the left and right, respectively. The radii have signs associated with them;
the radii are positive if their center of curvature is on the transmission side and negative
if they are on the reflection side of the lens. For instance, in a biconvex lens, $R_1 > 0$
and $R_2 < 0$. The object and image plane are illustrated in Figure 6.2.

   Simple geometric considerations of the object and image height yield an expression
for the lens system magnification:
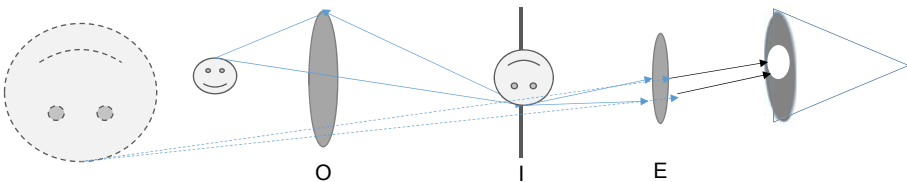
$$M = \frac{q}{p}. \tag{6.3}$$

Figure 6.1 Illustration of a compound microscope with objective (O) and eyepiece (E) lenses.
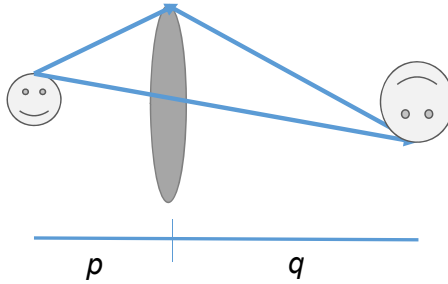The objective image plane is denoted by I.

Figure 6.2 The object placed a distance p in front of the lens and the image placed a distance q behind the lens.

This formula reveals how to obtain large magnification by designing a lens with a small working distance p, but practical limitations are imposed by the lens thickness and lateral size. For a compound microscope, the magnification is the product of the two lens groups (O and E):

$$M_\mathrm{m} = M_\mathrm{o}M_\mathrm{e}. \tag{6.4}$$

For instance, an eyepiece with $10\times$ magnification and an objective with $40\times$ magnification produce a compound microscope magnification of $400\times$.

6.2.2 *Wave propagation and diffraction*

In Chapter 2, Maxwell's equations for electrodynamics were introduced and discussed. It is crucial to our quantitative understanding of imaging and feature resolution. We use the Helmholtz form of the wave equation given by Eqn (2.28) and express it here in scalar form with material parameters equal to unity ($k_0 = \omega/c$)

$$\nabla^2 E(\vec{r}, \omega) + k_0^2 E(\vec{r}, \omega) = 0. \tag{6.5}$$

We consider an optical beam propagating close to the z-axis. When the transverse coordinates are Fourier transformed,

$$E(k_x, k_y, z, \omega) = \frac{1}{(2\pi)^2} \iint \mathrm{dx dy} e^{-\mathrm{i}(k_x x + k_y y)} E(x, y, z, \omega), \tag{6.6}$$

Equation (6.5) becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2} E(k_x, k_y, z, \omega) + \left(k_0^2 - k_x^2 - k_y^2 \right) E(k_x, k_y, z, \omega) = 0, \tag{6.7}$$

which can be solved for two counterpropagating plane waves. The solution we seek is propagating along the positive z-axis

$$E(k_x, k_y, z, \omega) = E(k_x, k_y, 0, \omega) e^{i\sqrt{(k_0^2 - k_x^2 - k_y^2)}z}. \tag{6.8}$$

For a linear system for which the input is at $z = 0$, the spatial frequency (k_x, k_y) response function is

$$H(k_x, k_y, z) = \frac{E(k_x, k_y, z, \omega)}{E(k_x, k_y, 0, \omega)} = e^{i\sqrt{(k_0^2 - k_x^2 - k_y^2)}z}, \tag{6.9}$$

which represents the plane-wave decomposition of light beam propagation a distance z in space.

The relation between the spatial expressions for the electric field and the response function is

$$E(x, y, z, \omega) = \iint dk_x dk_y e^{i(k_x x + k_y y)} E(k_x, k_y, 0, \omega) H(k_x, k_y, z). \tag{6.10}$$

The scalar electric field is expressed in spatial coordinates using the convolution form

$$E(x, y, z, \omega) = \iint dx' dy' E(x', y', 0, \omega) G(x - x', y - y', z). \tag{6.11}$$

The propagator $G(x - x', y - y', z)$ is the Fourier transform of the response function (k_x, k_y, z). It is also referred to as the spatial impulse response function:

$$G(x, y, z) = \iint dk_x dk_y H(k_x, k_y, z) e^{i(k_x x + k_y y)}. \tag{6.12}$$

The evaluation of Eqn (6.12) yields the result $(\rho = \sqrt{x^2 + y^2})$

$$G(x, y, z) = -\frac{ik_0}{2\pi\sqrt{z^2 + \rho^2}} \cdot \frac{z}{\sqrt{z^2 + \rho^2}} e^{ik_0\sqrt{z^2 + \rho^2}} \left(1 + \frac{i}{k_0\sqrt{z^2 + \rho^2}}\right). \tag{6.13}$$

The last factor becomes unity in the far field (i.e., $k_0 z \gg 1$). The second term is called the obliquity factor and is expressed as a cosine function $\left(\cos\Theta = \frac{z}{\sqrt{z^2 + \rho^2}}\right)$, where Θ is the off-axis angle. For small off-axis angles $\rho \ll z$, the paraxial approximation can be used for the propagator, yielding

$$G_p(x, y, z) = -\frac{ik_0}{2\pi z} e^{i(k_0 z - k_0 \rho^2/(2z))}. \tag{6.14}$$

The corresponding paraxial spatial frequency response function is $\left(\kappa = \sqrt{k_x^2 + k_y^2} \right)$

$$H_p(k_x, k_y, z) = e^{ik_0 z} e^{-i\kappa^2 z / 2k_0}. \tag{6.15}$$

H_p describes a spatial translation of a paraxial field. The spatial coordinate expression for the electric field is

$$E(x, y, z, \omega) = \iint dk_x dk_y e^{i(k_x x + k_y y)} E(k_x, k_y, 0, \omega) H_p(k_x, k_y, z). \tag{6.16}$$

In terms of the propagator, the above result can be written as

$$E(x, y, z, \omega) = \iint dx' dy' E(x', y', 0, \omega) G_p(x - x', y - y', z). \tag{6.17}$$

This is the Fresnel diffraction formula. The paraxial result is equivalent to solving the wave equation using the slowly varying envelope approximation (SVEA),

$$E(\mathbf{r}, \omega) = E_e(\mathbf{r}, \omega) e^{ik_0 z}, \tag{6.18}$$

where $E_e(\mathbf{r}, \omega)$ is the envelope function that satisfies the SVEA equation

$$\frac{\partial E_e(\mathbf{r}, \omega)}{\partial z} - \frac{i}{2k_0} \nabla_\perp^2 E_e(\mathbf{r}, \omega) = 0, \tag{6.19}$$

where the operator $\nabla_\perp^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the transverse Laplacian. Exploring the connection between the SVEA equation and the paraxial approximation is left as a problem.

6.2.2.1 Fresnel and Fraunhofer diffraction

The propagator in Eqn (6.14) has a quadratic (x, y) coordinate dependence in the exponential factor, which enables closed-form expressions to be derived for many apertures. A further approximation is made by expanding the quadratic expression in the exponential factor of the propagator

$$\frac{k_0 \left[(x - x')^2 + (y - y')^2 \right]}{2z} = \frac{k_0 \left[x^2 + y^2 - 2xx' - 2yy' + x'^2 + y'^2 \right]}{2z}. \tag{6.20}$$

The last two terms in brackets are restricted in size by the aperture dimension, say d. These two terms contribute to a negligible change of the phase in the integrand when the following inequality holds:

$$\frac{k_0 d^2}{2z} \ll 1. \tag{6.21}$$

Neglecting these two terms, the results derived from the expression are examples of Fraunhofer diffraction,

$$E(x, y, z, \omega) = -\frac{ik_0}{2\pi z} e^{ik_0 z} \iint dx' dy' E(x', y', 0, \omega) e^{i(K_x x' + K_y y')}, \tag{6.22}$$

where we define $K_x = \frac{k_0 x}{z}, K_y = \frac{k_0 y}{z}$.

We consider two important examples to illustrate Fraunhofer diffraction.

Example 1: Fraunhofer diffraction through a rectangular aperture

Consider a rectangular aperture for which the boundary is defined by the coordinates: $x = \pm a$ and $y = \pm b$. The aperture image is illustrated in Figure 6.3(a). When the aperture is illuminated with a uniform plane wave $E(x, y, 0, \omega) = E_0$, the boundaries define the limits of the integrals in Eqn (6.22),

$$E(x, y, z, \omega) = -\frac{ik_0}{2\pi z} e^{ik_0 z} \int_{-a}^{a} \int_{-b}^{b} dx' dy' E_0 e^{i(K_x x' + K_y y')}. \tag{6.23}$$

Evaluation of the integrals gives

$$E(x, y, z, \omega) = -\frac{i\pi k_0}{2z} 4ab E_0 e^{ik_0 z} \text{sinc}\left(\frac{K_x a}{\pi}\right) \text{sinc}\left(\frac{K_y b}{\pi}\right), \tag{6.24}$$
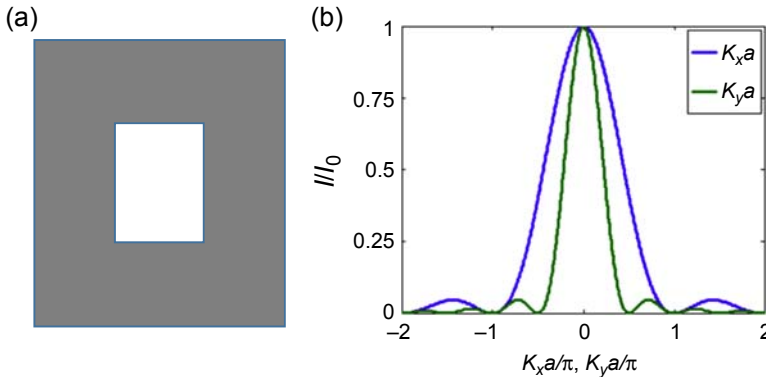
Figure 6.3 (a) A rectangular hole in an opaque screen. (b) Fraunhofer diffraction pattern for the aperture for a slice across the axes $K_y = 0$ and $K_x = 0$. For this figure, $b/a = 2$.

where we define the sinc function as $\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$. The Fraunhofer intensity diffraction pattern $(I = |E(x, y, z, \omega)|^2$ and $I_0 = |E(0, 0, z, \omega)|^2)$ rendered in Figure 6.3(b) has a strong central peak with small side lobes. For $a < b$, the spread of optical irradiance along the y-axis is narrower than its spread along the x-axis. This is a direct consequence of

the complementary uncertainty property of Fourier transforms between the variances in each space,

$$\langle \Delta x^2 \rangle_0 \langle \Delta x^2 \rangle_z \geq \left(\frac{2\pi z}{k_0} \right)^2, \tag{6.25}$$

where the averages are weighted by the intensity at the planes $z = 0$ and z. The brackets denote the average weighted by the intensity at the aperture and at the far-field screen. The minimum for the product is achieved under special circumstances of a focused Gaussian beam at $z = 0$ compared with the beam in the far-field regime. In any case, for the same profiles, but one aperture width larger than the other, as in Figure 6.3(b) with $b = 2a$, the comparable far-field angular spread of the wider beam is narrower.

Example 2: Fraunhofer diffraction through a circular aperture

A circular hole of radius a is cut in an otherwise opaque screen as shown in Figure 6.4(a). The hole is illuminated by a plane wave. The integral is written in polar coordinates (r, φ) as

$$E(r, \varphi, z, \omega) = -\frac{ik_0}{2\pi z} e^{ik_0 z} \int_0^a \int_0^{2\pi} d\varphi' r' dr' E_0 e^{iKr'\cos(\varphi'-\varphi)}. \tag{6.26}$$

Defining: $K = \sqrt{K_x^2 + K_y^2}$, with $K_x = \frac{k_0 r \cos \varphi}{z}, K_y = \frac{k_0 r \sin \varphi}{z}$. The integral is evaluated as

$$E(r, \varphi, z, \omega) = -\frac{ik_0}{z} e^{ik_0 z} \int_0^a r' dr' E_0 J_0(Kr') = -\frac{ik_0}{z} a^2 e^{ik_0 z} E_0 \frac{1}{aK} J_1(Ka). \tag{6.27}$$
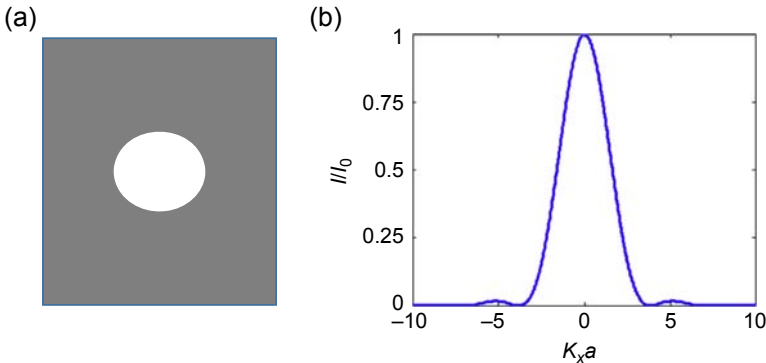
Figure 6.4 (a) A circular hole in an opaque screen. (b) Fraunhofer diffraction pattern for a circular aperture. The first zero of the diffraction pattern lies near $Ka = 3.83$.

Using the integral relations, $J_0(Kr') = \frac{1}{2\pi} \int_0^{2\pi} d\varphi' e^{iKr'\cos(\varphi'-\varphi)}$ and $\frac{a}{K} J_1(Ka) = \int_0^a r' dr' J_0(Kr')$,

where $J_0(x)$ and $J_1(x)$ are Bessel functions of zeroth and first order, respectively. The diffraction

pattern has circular symmetry with a maximum at the center and secondary intensity maxima in concentric rings. A cross-section through the diffraction pattern is shown in Figure 6.4(b).

6.2.3 Gaussian beams

Solving the paraxial equation or equivalently the SVEA equation for an initial Gaussian beam is instructive and has wide applications to optical technology. A focused Gaussian beam shape at $z = 0$ is expressed as

$$E(x, y, 0, \omega) = E_0 e^{-r^2/2w_0^2}, \tag{6.28}$$

where w_0 is the Gaussian beam width at $z = 0$. Applying the Fourier transform, the plane-wave decomposition of the Gaussian function is

$$E(k_x, k_y, 0, \omega) = \frac{w_0^2}{2\pi} E_0 e^{-\kappa^2 w_0^2/2}. \tag{6.29}$$

Applying this expression in Eqn (6.16), the integral is

$$E(x, y, z, \omega) = \iint dk_x dk_y e^{i(k_x x + k_y y)} \frac{w_0^2}{2\pi} E_0 e^{-\kappa^2 w_0^2/2} H_p(k_x, k_y, z). \tag{6.30}$$

The integral is evaluated with the explicit result

$$E(x, y, z, \omega) = \frac{w_0^2}{2\pi(w_0^2 + iz/k_0)} e^{ik_0 z} E_0 e^{-r^2/2(w_0^2 + iz/k_0)}. \tag{6.31}$$

The result can be recast into the following form:

$$E(x, y, z, \omega) = \frac{w_0^2}{w(z)^2} e^{i\varphi} e^{ik_0 z} E_0 e^{-r^2 \left(\frac{1}{w(z)^2} + \frac{ik_0}{R(z)}\right)/2}. \tag{6.32}$$

Using the following definitions for the beam width and phase front radius of curvature

$$w(z)^2 = w_0^2 \sqrt{1 + (z/z_R)^2}, \tag{6.33}$$

$$R(z) = \frac{(z^2 + z_R^2)}{z}. \tag{6.34}$$
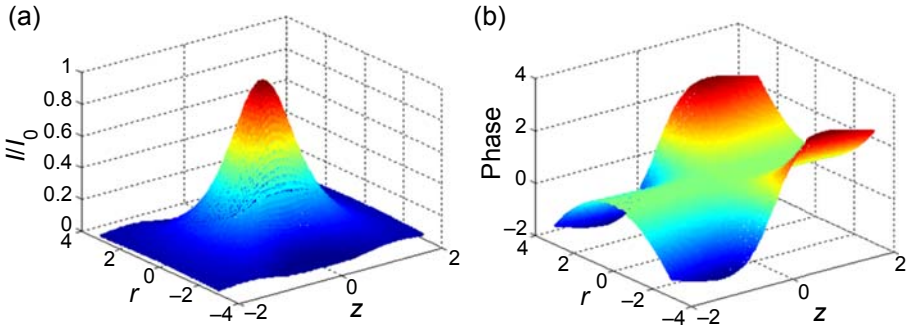
Figure 6.5 (a) The Gaussian beam (x, z) plane. The z coordinate is scaled to the Rayleigh range (z_R) and x is scaled to the beam width w_0. (b) The phase of the electric field in the (x, z) plane illustrating the curvature of the phase front.

The phase φ is called the Guoy phase and is expressed as

$$\tan \varphi = -\frac{z}{z_R}. \tag{6.35}$$

The definition

$$z_R = k_0 w_0^2, \tag{6.36}$$

defines the variable known as the Rayleigh range, which is the distance over which the square beam width broadens by $\sqrt{2}$. Note that $R(z)$ is infinite at $z = 0$ and changes sign as z passes from negative to positive. It asymptotically approaches $\pm\infty$ as z $\rightarrow \pm\infty$. The properties of a Gaussian beam and the phase front are illustrated in Figure 6.5. The beam width is a minimum and the intensity is a maximum at the $z = 0$; the beam spreads symmetrically on both sides of the center point. On the other hand, the phase is flat at $z = 0$ and the curvature has opposite signs on both sides. The curvature is negative for $z < 0$ and positive for $z > 0$.

6.2.4 Resolution

Increasing magnification has essential limitations imposed by the wave properties of light. As features are magnified to dimensions approaching the order of the wavelength, their size will be blurred. In terms of a point object, the size of the diffraction spot cannot be further reduced. This limit was quantified by Ernst Abbe, who estimated that the resolvable size of a feature is given by

$$\mathcal{R}_A = \frac{\lambda}{2n_e \sin \theta}, \tag{6.37}$$

where λ is the wavelength of light in vacuum and $2n_e \sin\theta$ is the numerical aperture (NA), which is proportional to the sine of the maximum angle of the light cone formed by a point on the object to the exit lens. This formula is called the Abbe criterion; resolution is limited, but it also shows strategies that can be used to improve the resolution—namely, it can be improved by using a shorter wavelength, by increasing the NA using a larger lens, or by immersing the lens in a higher index medium.

Another widely applied resolution formula was derived by Lord Rayleigh (John W. Strutt), which is derived by distinguishing two point-like features that are separated by a minimum angle, as illustrated in Figure 6.6(a). As noted in Figure 6.4, the first zero of the Fraunhofer diffraction pattern appears at $Ka = 1.22\pi = 3.83$. Using Figure 6.6(b) as a guide, two incident beams at different angles are incident on an aperture. The diffraction angle of a plane wave along the x-axis is $\alpha_x = \frac{x}{z}$; thus, two diffracted waves for which the intensity maximum appears in the image plane at position x_1 and x_2 are angularly separated by $\Delta\alpha_x = \alpha_{x2} - \alpha_{x1}$. By the Rayleigh criterion, the separation between two beams is distinguishable when $\Delta\alpha_x = 1.22\pi/k_0a$. The addition of two diffraction patterns is plotted in Figure 6.6(a) when the separation between the beams is just that given by the Rayleigh criterion. Naturally, the Rayleigh criterion was developed for incoherent sources and the intensities of the two beams are added when they are independent sources. For two coherent beams, the field amplitudes would be added before constructing the intensity.

Seen from the source side, the Rayleigh criterion is a condition on the distinguishability of two point-like objects. In Figure 6.7, the two point-like objects are separated by a distance x_0 and they lie in the focal plane of a lens for which the focal length is f. The minimum angle as given by the Rayleigh criterion is

$$\alpha_{\min} = \frac{x_0}{f} = \frac{1.22\pi}{k_0 a} = 0.61\lambda/a. \tag{6.38}$$

Because the NA for this system is NA $= a/f$, the Rayleigh criterion for resolving two point-like objects is

$$\mathscr{R}_R = x_0 = 0.61\lambda/\text{NA}. \tag{6.39}$$
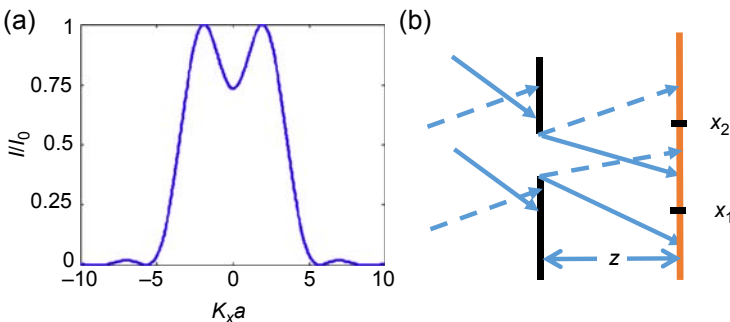

Figure 6.6 (a) Fraunhofer diffraction pattern of two plane waves for which the angular separation is 1.22π. (b) A schematic of the diffraction pattern formation at the image plane from two plane wave sources that are incoherently superposed.
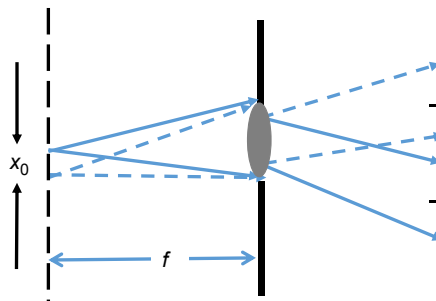
Figure 6.7 Resolution of two incoherent sources separated by a distance x_0 in the back focal plane of a lens of focal length f.

The distinction between the two resolution limits (Abbe Eqn (6.37) and Rayleigh Eqn (6.39)) is small. These criteria are a guide to estimate the ultimate achievable resolution of an optical system.

6.2.5 Depth of field

Another important concept is used to determine how tolerant an optical system's image sharpness is to variations in the object's position or the image plane's position relative to the focal plane of a system of lenses. The depth of field is a sharpness criterion for how well objects within a certain longitudinal range will be in focus. Figure 6.8 is a simple illustration of the depth of field. Light focused by a lens and the NA is defined by $D/2f$, where D is the clear aperture diameter of the lens. The focal plane is indicated by a vertical dashed line. A cylinder of radius $\mathscr{R}_R$ is drawn around the focus and intersects the boundary rays. The length of the cylinder is the depth of field. By geometrical considerations, the depth of field is given as

$$D_f = 1.22\lambda/\mathrm{NA}^2. \tag{6.40}$$

The depth of field is proportional to the wavelength and sensitive to the NA in comparison with the resolution criterion. For a wavelength of 500 nm and a NA of 0.1, the lateral resolution is $\mathscr{R}_R = 3.05$ μm and the longitudinal depth of field is $D_f = 61$ μm. A larger NA improves the resolution, but at the same time the depth of field shrinks; for instance, NA $= 1.2$ yields $\mathscr{R}_R = 0.25$ μm and $D_f = 0.42$ μm. The high resolution
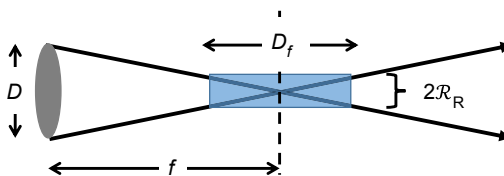
Figure 6.8 Diagram for the determination of the depth of field.

comes at a cost of blurring or obscuring features around the object point of interest. Ideally, one would like to have both high resolution and high depth of field. The electron microscope has the capability to meet both conditions, as the reader will discover in the following section.

6.3 Electron microscopy

An important alternative approach to characterizing nanoscale features is to use electrons instead of photons to create images. Electrons are ultimately described as quantum waves using the de Broglie electron momentum relation discussed in the chapter on quantum mechanics, but it is rewritten here as $p = h/\lambda$, where h is Planck's constant and λ is the electron's wavelength. Applying high voltages to accelerate electrons, their wavelengths are several orders of magnitude smaller than usable wavelengths for optical microscopy. The electron's acceleration across an applied voltage difference of V imparts a kinetic energy to the electron given by

$$K = eV. \tag{6.41}$$

In the nonrelativistic limit, the kinetic energy is related to the electron momentum by

$$K = \frac{p^2}{2m_0} \tag{6.42}$$

and m_0 is the free mass of the electron. It is left as a homework problem to derive the momentum energy for relativistic electrons.

Ernst Ruska is credited with reporting the first demonstration of an electron microscope in 1931 using magnetic lenses to focus the electrons for his doctoral research; he was finally recognized for this achievement with a Nobel Prize in 1986. It is interesting to note that in undertaking the development of the electron microscope, Ruska mentions in his Nobel lecture that he was not aware of the de Broglie relation; he was simply motivated by the small size of the electron for achieving high resolution.

In this section, we will discuss two types of electron microscopy: transmission electron microscopy (TEM) and scanning electron microscopy (SEM). Both techniques are widely used across many fields of technology. They have become an indispensable characterization tool for nanoscale systems. There are other useful electron microscopy instruments, which are perhaps not as widely available, but the interested reader can study the literature to learn more about those techniques.

6.3.1 Scanning electron microscopy

The SEM design has elements to focus the electron beam into a controllable spot size on the specimen and to raster scan the spot over the sample surface. The SEM elements
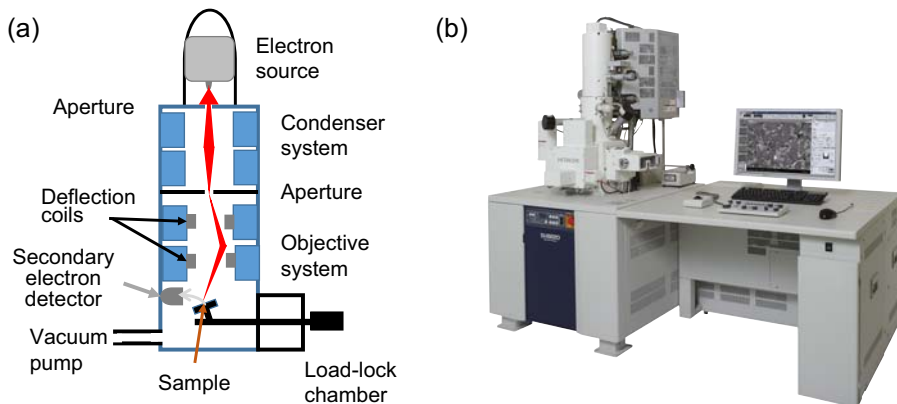
Figure 6.9 (a) SEM illustration with major elements. (b) Image of a high-resolution SEM system.

are shown in Figure 6.9(a) and an image of a machine is shown in Figure 6.9(b). At the top of the column, electrons are emitted from a source and accelerated to a small aperture by a potential difference *V*. For instance, a thermionic emission source may use a heated cathode to emit electrons; the electron beam is collimated through an aperture and accelerated in several stages of electron lenses that focus the electron beam into a small spot on the sample surface. Electron lenses deflect electrons using either electrostatic or magnetic fields. In an SEM, the electrons are imparted kinetic energy by passing through high voltages that can range from approximately 250 V to 50 keV. The electron beam spot size ranges from 0.5 to 5 nm in diameter. A set of deflection coils raster scan the beam across the surface and point the high-energy electron beam to (*x, y*) positions on the sample.

A closer view of the electron beam scattering at the surface is shown in Figure 6.10(a). The primary electron beam reaches the sample to produce reflected (ballistic) backscattered electrons and secondary electrons. Secondary electrons are emitted from the surface after scattering of the primary beam under the surface. After the primary beam is scattered within the sample, its energy is imparted to excite atomically bound electrons in the material. Therefore, the secondary electrons have relatively low energy and are deflected by a low voltage to a detector, and the electron yield is used for imaging the surface topology. Figure 6.10(b) further elucidates the interaction of a primary electron beam with a sample surface and the detection scheme. Backscattered electrons have high energy and follow a straight line trajectory. The secondary electrons having low energy are deflected by the electric field to a detector that measures the electron yield. Once in the detector, the electrons are accelerated to a scintillator, which emits photons. In turn, the photons are multiplied in a photomultiplier to create a large measurable signal. The secondary electrons are produced when the primary electrons cascade in energy as they pass through the surface and into the volume. As illustrated in Figure 6.10(b), the electron beam blooms from a small point to a larger volume because of the scattering events randomly redirecting the electron's
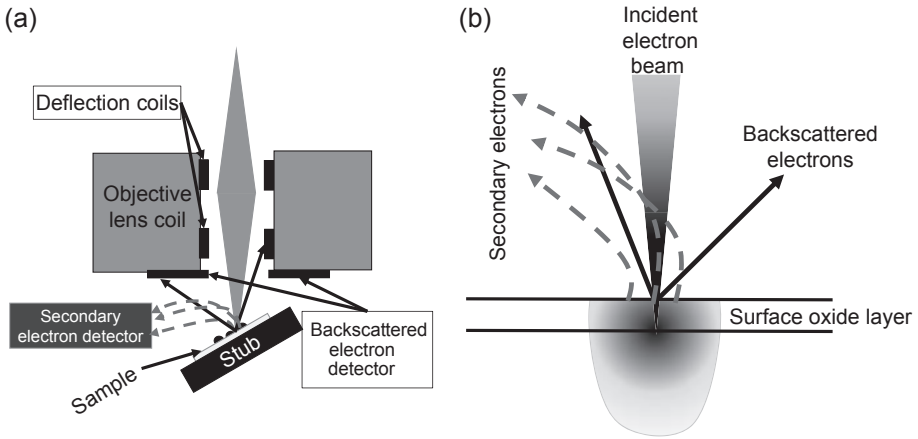
Figure 6.10 Specimen stage is highlighted and electrons scatter from the surface or penetrate into the substrate. (a) Primary electrons incident on the sample surface. (b) Primary electrons in the incident electron beam that penetrate the surface create a cascade of secondary electrons in the material. The secondary electrons that escape are collected in the detector.

momentum. At each inelastic collision, the primary electrons slow down while at the same time ionizing the atoms around them. The primary electrons can also knock out inner core electrons from the atoms, which will result in characteristic X-ray emission from the surface. The X rays, although not indicated in Figure 6.10, have energy spectra that can be used to determine the abundance of atomic constituents near the sample's surface. The backscattered electrons also provide quantitative information about the elements in the material; their scattering intensity is correlated with the Z number of the scattering atom.

Sample preparation for SEM is simple. Insulating samples are first coated with a thin conducting film by sputtering techniques; materials such as gold or carbon are commonly used to coat the surface with a conducting layer. The conductive coating is necessary to prevent charge buildup on the surface, which will cause blooming of the electron beam and subsequent distortion of the images. The prepared samples are mounted on a pedestal called a stub with conducting tape to hold them in place. A load lock chamber indicated in Figure 6.9(a) is a common and useful SEM feature. It is a small-volume box that can isolate the sample between the atmospheric pressure on the outside and vacuum maintained inside of the SEM. A sealed door and a gate valve are used to load and then transfer the sample to the inside. A load lock reduces the time for making measurements on different samples because the entire chamber is not opened to the atmosphere and then pumped to vacuum level again, which would require much more time.

We end this section showing two examples of SEM images as shown in Figure 6.11. It is striking that the depth of field can be so large for SEM images. This is a consequence of the small NA for electron beams in SEMs. At low magnification shown for the specimen from nature in Figure 6.11(a), the details of the eye are clear and

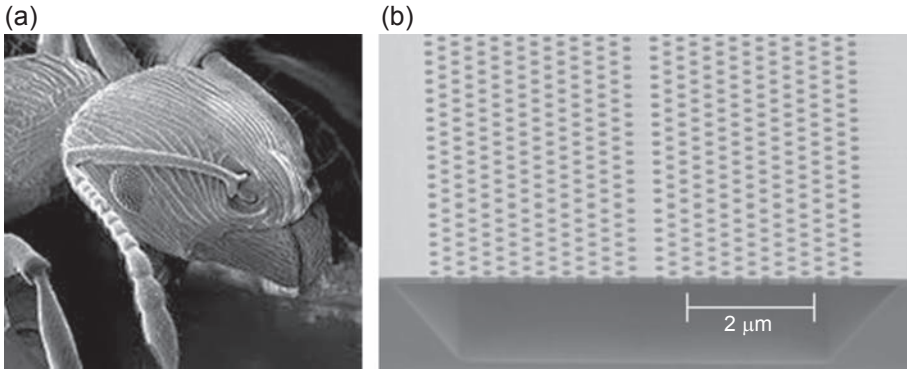**Figure 6.11** Examples of SEM images. (a) An image of an ant (http://usgsprobe.cr.usgs.gov/images/ant.gif). (b) An image of a photonic crystal waveguide. *(With permission from Nanophotonics 2 (2013) 39−55,* http://dx.doi.org/10.1515/nanoph-2012-0039.*)*

the antenna and leg in the foreground are equally in sharp focus. At higher magnification in Figure 6.11(b), the nanoscale holes are clearly imaged and they have a sharp appearance over many periods. The side view gives a glimpse at the cavity under the top membrane. This image is an example of a free-standing waveguide (center) surrounded by air on the top and bottom; the optical mode at specific wavelengths is confined laterally by the presence of the periodic array of air holes. The periodic array of holes is an example of a two-dimensional photonic crystal; more details on photonic crystals will be discussed in Chapter 11.

6.3.2   *Transmission electron microscopy*

For electrons to pass through a sample, it has to be very thin; this requires a sample thickness in the range of 50 nm. If suitable, TEM specimens are prepared slicing films from the sample of interest using microtome techniques to slice a sheet off of the specimen and placing the sheet on a conducting grid. Further thinning of the sample depends on its physical characteristics so that other artifacts are not created during the sample preparation process. The sample preparation process can be complicated and involve, for instance, diamond or ultrasonic cutting, mechanical thinning, cutting, dimple grinding, and ion milling. Preparing samples for TEM imaging that preserve the nanoscale features can be a demanding task and requires meticulous attention to each step.

TEM works on the same principle as an optical microscope. The side-by-side analogy between the two is shown in Figure 6.12(a). The sample is placed after the condenser lens, where illumination covers the region to be imaged. The electron beam passes through the sample undeflected, and the objective and projector lens areas magnify the image as in a compound microscope. The physical differences due to the replacement of photons by electrons endow the TEM images with very high resolution. The TEM electron-accelerating voltage differences can be much higher than for
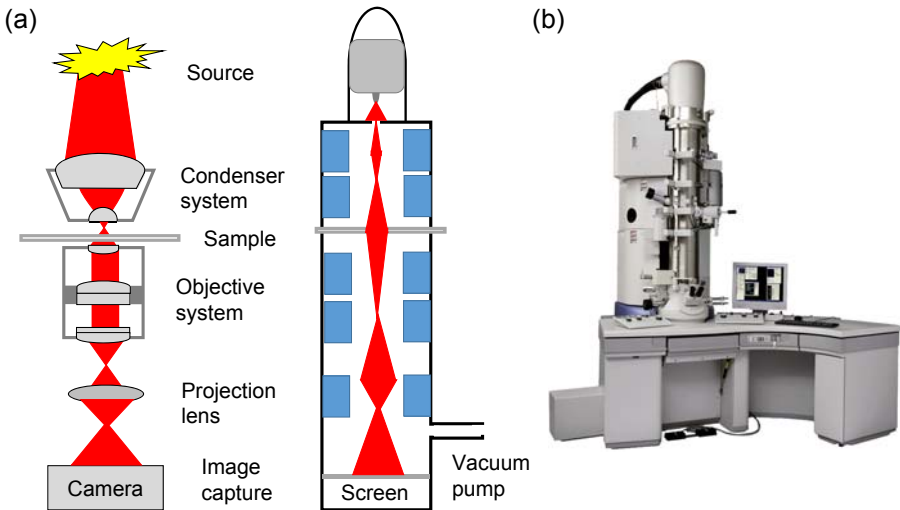
Figure 6.12 (a) Comparison of the elements of an optical microscope and a TEM instrument. (b) Image of a TEM instrument.

high-resolution SEM; TEM voltages can vary from several 100 keV to more than 1 MeV, exposing sample details in the images down to the atomic scale. To see how well TEM can image, consider a 100-keV electron beam for which the de Broglie wavelength is $\lambda = 3.89$ pm (using the nonrelativistic form of energy/momentum relation). The Rayleigh resolution and depth of field, Eqns (6.38) and (6.39) for the electron beam with an NA of 0.01 is $\mathcal{R}_R = 0.24$ nm and $D_R = 47.4$ nm.

The high-resolution TEM images of silver nanoparticles in Figure 6.13 demonstrate the atomic-level detail that is possible to resolve with this instrument. The individual atomic sites are captured in these pictures. From images of this quality, crystal structure can be determined and defects near the surfaces can be examined.

6.4 Scanning probe techniques

The development of the scanning tunneling microscope (STM) by Binnig and Rohrer and its and rapid acceptance by the scientific community opened the way to a myriad of new measurement tools capable of resolving features at the nanoscale. Among the new scanning probe tools is the atomic force microscope (AFM), which was invented in 1986 by Rohrer and collaborators, the same year that Binnig and Rohrer received the Nobel Prize (they shared the award with Ruska, who was recognized for his initial demonstration and seminal contributions to electron microscopy). The new scanning probe concept quickly led to the development of further scanning probe instruments, such as the magnetic force microscope, the ultrasonic AFM, etc. A selection of scanning probe techniques are listed in Table 6.1.
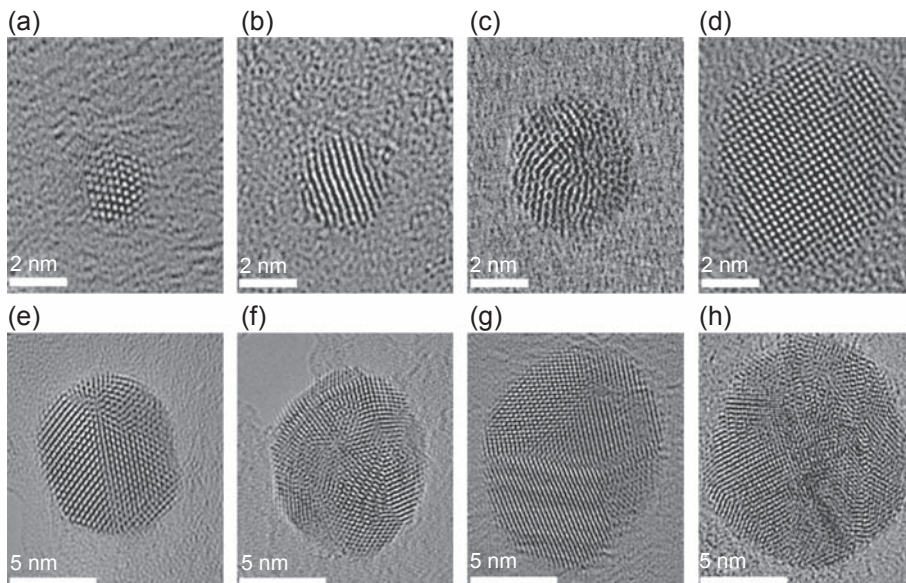
Figure 6.13 TEM images of silver nanoparticles. The nanoparticle diameter's progression in each image is 2, 3, 4.5, 6, 7.5, 9, 10.5 and 12 nm, respectively.
With permission from J.A. Scholl, et al., Nature 483 (2012) 421−427. http://dx.doi.org/10.1038/nature10904.

Around the same period as the development of the STM, Pohl filed a patent on a near-field optical technique that obtained higher resolution by defining an aperture size that is much less than a wavelength; subsequently, two groups published papers demonstrating spatial feature resolution exceeding the Rayleigh/Abbe condition by

Table 6.1 Scanning probe techniques

STM	Scanning tunneling microscope	AFM	Atomic force microscope
NSOM	Near-field scanning optical microscope	AFAM	Atomic force acoustic microscope
MFM	Magnetic force microscope	SEFM	Scanning electrostatic force microscope
SFAM	Scanning force acoustic microscope	FFM	Friction force microscope
SMM	Scanning magnetic microscope	SThM	Scanning thermal microscope
SEcM	Scanning electrochemical microscope	SKpM	Scanning Kelvin probe microscope
SCM	Scanning capacitance microscope	SICM	Scanning ion conductance microscope

a factor of 10. This technique has become widely known as either near-field scanning optical microscopy (NSOM) or scanning near-field optical microscopy (SNOM). NSOM has a considerably longer history than other SPM techniques dating back to 1928, when Synge published a result that recognized the resolution advantages that could be obtained by near-field measurements.

   In the rest of this chapter, we will explore basic properties of three scanning probe techniques that are widely in use as characterization tools in many laboratories. We will explain the essential features of the instruments. There are some common features to all scanning probe techniques that can be mentioned here. The probes are placed within a few nanometers of the surface; therefore, the instrument must have nanometer control of the tip placement.

6.4.1   Scanning tunneling microscope

The STM has the highest resolution of any scanning probe technique and can image features on an atomic scale (~ 0.1 nm). It consists of a metal tip that is controllably placed close to the surface under study and electronics to scan the tip over a region of interest. STM achieves a remarkable level of resolution because of several factors. First, it is based on electronic quantum tunneling current, which is exponentially sensitive to small changes in the height of the tip above the surface. Secondly, the ability to make tips that are very sharp, even containing one atom, provides exquisite sensitivity to the lateral and longitudinal positioning of the tip. Finally, piezoelectric materials respond with nanometer scale length changes under the application of a voltage; electronic control of piezomaterials are used to position the tip-sample distance to a fraction of a nanometer.

   A schematic of an STM is shown in Figure 6.14(a). The tip is held by a piezoelectric sleeve that is voltage controlled to differential displacements of better than 1 nm. The voltage controls move the tip with nanometer accuracy to a spatial coordinate (x, y, z). There is an applied voltage between the tip and the sample to drive a tunneling current
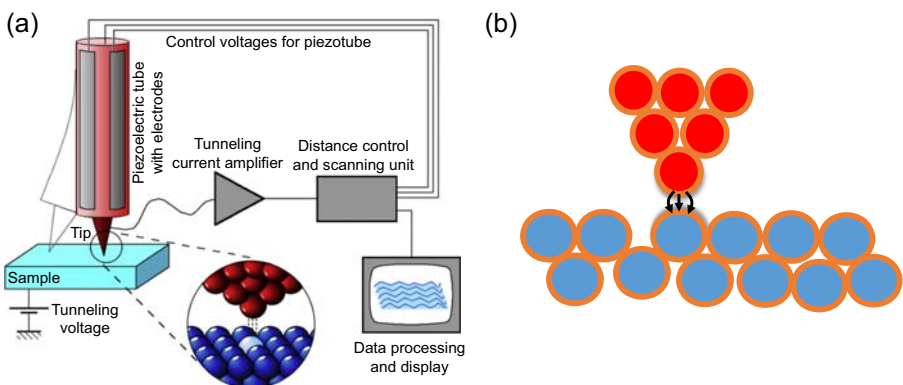

Figure 6.14 (a) Schematic of the STM apparatus. *(Figure from Michael Schmid, TU Wien.)* (b) Caricature of the tip−sample interaction. The electron cloud is depicted as surrounding the atoms.
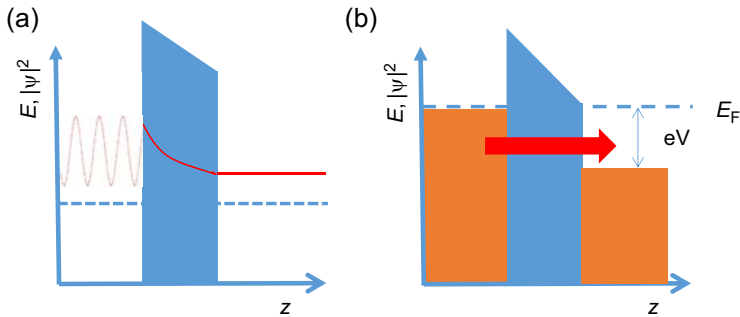
Figure 6.15 (a) The tunneling of the wave function through the barrier for an energy below the barrier. (b) An applied voltage drives current across the tunneling gap from the filled electron states in metal 1 to the empty electron states in metal 2.

between them. The current between the tip and the sample is measured and set to a value. As the tip is raster scanned across the surface (x, y), the height of the tip (z) is adjusted by changing the applied voltage on the piezoelectric actuator to keep the tunneling current constant. The voltage applied to adjust the tip height is recorded at each position (x, y), and the result is processed to render a three-dimensional image of the surface.

A tip interaction with the sample surface is illustrated in Figure 6.14(b). The electron cloud around the atoms is represented as its skin. The electron wave functions extend into the space between the tip and the sample and a current passes. A supertip is made when there is a single atom at the tip; this atom precisely defines the position of the tip and exquisite information on the local electronic states on the sample is obtained.

The quantum mechanical action of the tip–sample interaction is illustrated in Figure 6.15. A barrier is drawn between two materials in Figure 6.15(a). The electron wave impinging on the barrier from the left side is largely reflected, but has some extension in and through the barrier. Each material has a work function defined as the minimum energy for an electron to escape from the material. For metals, the minimum is from the Fermi level to the vacuum states.

STM images provide a direct observation of the local electron density near the surface. The tip is also used to manipulate single atoms on the surface to create structures made using a few atoms. These are so-called adatoms or "adsorbed atoms" that are bound to the surface. The adatom is moved by placing the tip above it and applying a tip voltage until the adatom overcomes the energy barrier to move to a neighboring equilibrium position. One famous and interesting image with atomic-scale resolution using this technique is reproduced in Figure 6.16. It has the shape of a corral of 48 iron atoms placed on a copper surface. Inside of the corral, the electronic density has spatially periodic oscillations revealing the true quantum nature of the electronic density.

6.4.2 Surface profiling and the AFM

There are many available techniques to determine the surface topography of a sample. Among the most widely used are profilometry, white light interferometry, and confocal
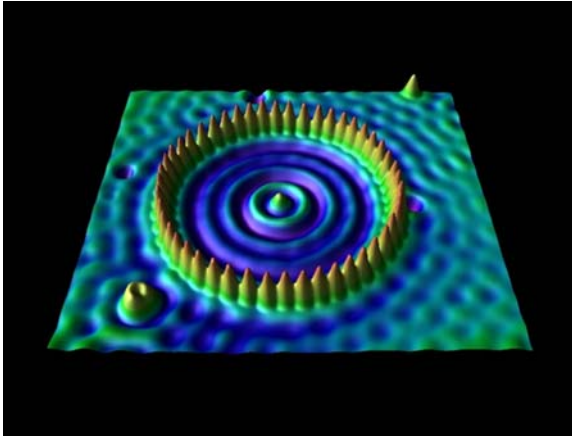
Figure 6.16 A corral made from 48 iron atoms placed on a copper (111) surface. The ripples are electronic wave functions interfering inside of the corral.
Reproduced with permission from Cover image from Physics Today, 46 (11). Copyright 1993, American Institute of Physics. Image from Don Eigler, IBM Almaden Research Center. The color image was created by Dominique Brodbent for the cover of Physics Today.

microscopy. The AFM is a descendant of the profilometer and the STM, which represents a significant advance in metrology. Its functionality is not just an improved tip version of the profilometer, but because of the small size, nanoscale physical forces play an important role in its operation.

In brief, a profilometer is a surface contact instrument with a stylus that is dragged across the sample's surface to measure its three-dimensional shape. The stylus tip size is typically a few microns in diameter, and that ultimately limits the lateral resolution measurable with the instrument. Despite this limitation, it is a very useful tool in the nanofabrication environment because its vertical resolution accuracy can be less than 1 nm. As the surface is scanned, a three-dimensional height profile is constructed from the data. Artifacts occur because of the size of the stylus; this is illustrated in the upper drawing in Figure 6.17, such as rounding a steep vertical step or a depression for which the width is smaller than the stylus diameter. The recorded artifacts in Figure 6.17 are illustrated in the lower panel of the drawing.

White light interferometry is a noncontact optical scanning method that can record subnanometer sample height variations. A schematic showing the essential elements of the setup is given in Figure 6.18(a). A white light source is collimated and focused by a microscope objective to a point. Two glass plates with flat, parallel surfaces inserted in the path; the second one splits the incident beam into two beams with the reflected one as a reference wave reflected from a mirror spot on the top of the first plate. The two waves recombine at the beam splitter plate and follow an optical path to the camera. This two-path interference setup is called a Mirau interferometer. As the (x, y) positions across the sample are scanned, the vertical position is accurately scanned by piezoelectric positioners to determine the maximum coherence position. Thus, the short coherence length of white light illumination is adapted using the two-path
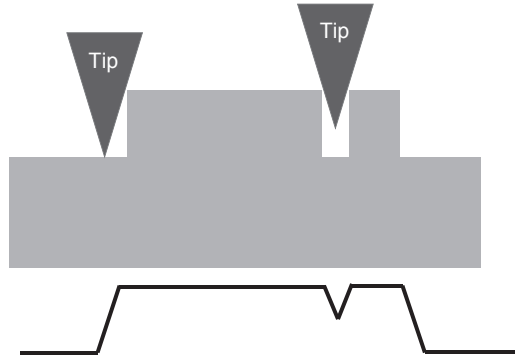
Figure 6.17 Examples of two types of artifacts when a stylus or tip is scanned across a surface. Top drawing shows an edge slope and a depression limitation due to the tip's geometric shape.

interferometer setup to measure an interference maximum. There are no ambiguities in the position of the interference maximum, which will occur when coherent light illumination is used. The vertical resolution is approximately 0.1 nm, whereas the lateral resolution is determined by the optical spot size, which is determined by the Abbe or Rayleigh condition of the instrument; in general, it is approximately 0.5 µm.
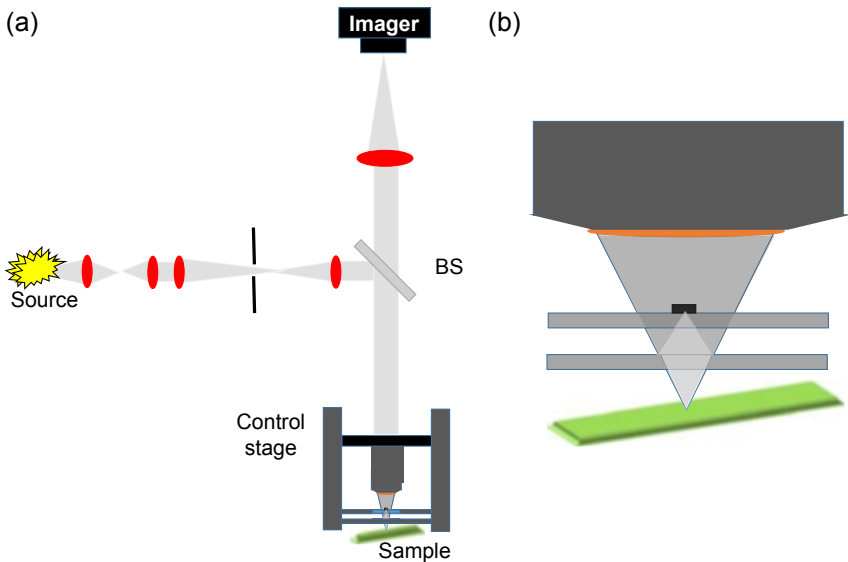
Figure 6.18 (a) Schematic of a white light interferometer. The sample is scanned vertically and horizontally to determine the surface profile. BS denotes a beam splitter. *(Illustration made after one appearing in Applied Optics 26 (1987) 2810.)* (b) A close-up of the interferometer arrangement. A beam is split at the top second glass plate surface and the two beams are recombined. BS denotes a beam splitter that deflects the source illumination to the interferometer.
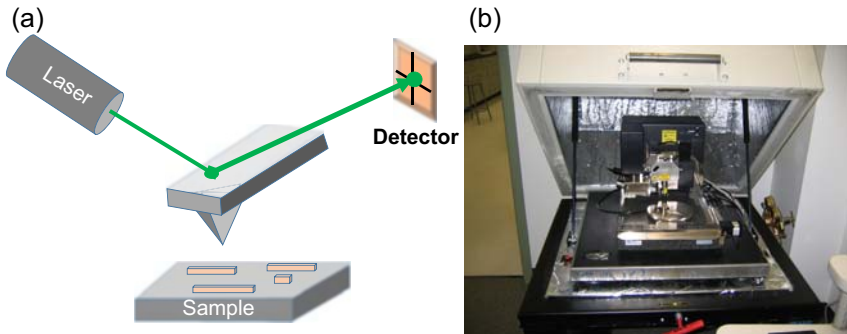
Figure 6.19 (a) A schematic diagram of the essential elements of an AFM. The cantilever has a sharp tip that is scanned over the surface, and the deflections of the tip are translated to laser beam deflections. *(Image from Opensource Handbook of Nanoscience and Nanotechnology.)* (b) Photograph of an AFM instrument. *(Image by Zureks.)*

A schematic of the essential elements of an AFM is shown in Figure 6.19(a). The sensing component is a sharp tip with a size of approximately 10 nm that is placed near the end of a cantilever. Various tip shapes are used, including conical, square, and trapezoidal. The shape of the cantilever determines its mechanical properties, especially the force constants, which may lie between 0.1 and 50 Nt/m, and the resonant frequency, which may lie between 1 kHz and 1 MHz. Cantilever shapes are varied from rectangular to V-shaped. The cantilever is attached to a piezoelectric scanner that controls the (x, y, z) position of the tip. Deflection of the cantilever as it is moved across the surface is measured by a laser system and a quad-detector, which determines the direction that the laser beam is deflected. A photo of a commercial AFM is shown in Figure 6.19(b). It is compact, and the head is placed on a plate that is isolated from environmental vibrations. The lid is closed during operation to further eliminate any air motion or acoustic noise that would add noise to the measurement.

The tip interacts with the surface via the van der Waals force, which has a repulsive action in the near field and attractive action at longer separations. Two AFM operational regions are indicated in Figure 6.20. In the repulsive force region, the tip has a strong interaction with the surface. Two AFM measurement modes are based on the repulsive force effect: the contact mode and tapping mode. For the contact mode, the force is maintained at a constant magnitude by using feedback electronics. The tip is held within 1-nm distance and the change of the cantilever's position is monitored is the laser beam's deflection. This mode gives the highest resolution, and imaging is rapid. The disadvantage may be damage to the surface. The tapping mode uses the oscillation of the cantilever at its resonant frequency, and the surface tapping of the tip determines the end oscillation deflection. By adjusting the feedback to maintain constant oscillation amplitude, the surface structure is mapped.

The attractive force is weaker and modulates over longer distances. For this reason, imaging done when the cantilever operates at these distances is called a noncontact mode. The cantilever deflection is monitored and adjusted by electronic feedback loops and the image is formed as the tip is scanned across the surface. The smaller
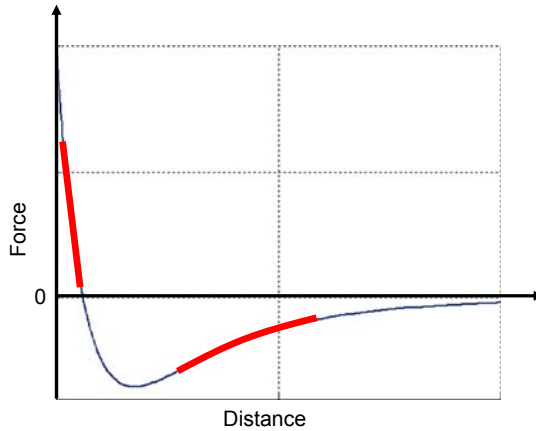
Figure 6.20 A van der Waals force curve is repulsive at short distances and transitions to an attractive force at long distances.

tip−surface interaction of this mode leads to a lower sensitivity and lower image resolution. AFM measurements suffer from the same types of artifacts previously discussed for the profilometer. Given the much smaller footprint of the AFM tip though, the dimensions of the artifacts are much finer.

The AFM can be modified to image samples made from many different materials and biological samples. In contrast to the STM, it does not require a conducting surface, and customized tips are available for different modalities of operation, including conductivity, electrical, and magnetic measurements. This has made AFM an excellent platform for making a wide variety of measurements.

6.4.3 Near-field scanning optical microscopy

NSOM (also called SNOM) is an optical technique that achieves high resolution with mechanical elements that are similar to the AFM and STM. For instance, it has a tip that also scans the surface and maintains a constant height by using piezoelectric steering elements. A simplified schematic of an NSOM apparatus is shown in Figure 6.21(a) with light illumination from above and light collection below the sample surface. A laser source illuminates the sample through an optical fiber. The end of the fiber is tapered and coated with metal; the light squeezes through a small, subwavelength aperture at the tip of the fiber. The NSOM fiber construction and electronics are similar to other SPM techniques, but with the metal tip replaced by a metal film-coated, tapered optical fiber. The fiber tip position is controlled using piezoelectric elements attached to the fiber and the detector measures variations in the scattered light due to tiny features on the surface. The small aperture at the fiber tip typically has a diameter of approximately $d \sim 50$ nm, as shown in Figure 6.21(b).

The size of the aperture, which is much smaller than the wavelength, determines the ultimate resolution of the NSOM. In the plane at the fiber tip end the aperture's Fourier
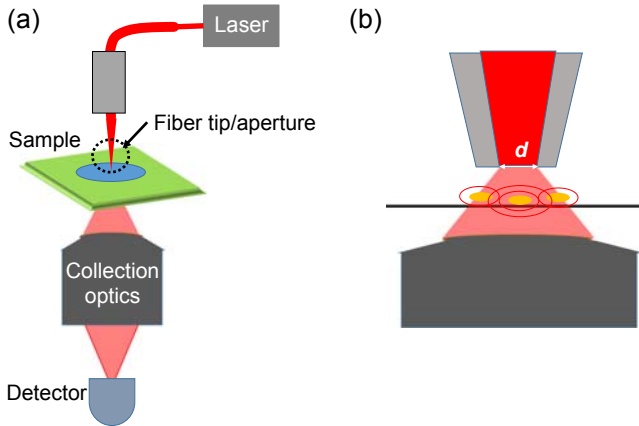
Figure 6.21 (a) The essential elements of an NSOM instrument. (b) A close-up of the fiber tip with light scattering features and collection optics below the surface.

transform is largely determined by evanescent modes, since the characteristic wave-number of the aperture satisfies $\frac{2\pi}{d} > \frac{2\pi}{\lambda}$. Therefore, diffraction of light emerging from the fiber tip is severe because the evanescent modes decay exponentially with separation from the aperture plane. The spreading of the beam is illustrated by a widening cone between the tip and the sample surface. Subwavelength resolution is maintained by keeping the fiber tip sufficiently close to the surface to minimize diffractive beam spreading.

In Figure 6.21(b), the scattered light transmitted through the sample is collected by a sensitive detector. The smallest aperture size is determined by a compromise between the desired resolution and the signal-to-noise ratio. A smaller aperture diameter can improve the resolution, but at the cost of reducing the light radiance that can squeeze through the tiny aperture. A qualitative understanding of the transmission can be gained by Bethe's famous treatment of the problem in 1944 by replacing the small hole with an oriented electric and magnetic dipole pair to satisfy (perfect metal) boundary conditions. This is a version of the electrostatic limit discussed in Chapter 2. The dipoles radiate electromagnetic energy with a flux that is proportional to $\left(\frac{d}{\lambda}\right)^4$, and this determines how small of a hole can be tolerated. The detectable signal is reduced by four orders of magnitude in going from $d = \lambda$, and $d = 0.1\lambda$ is severe.

NSOMs are used in several different operational modes. Three modalities are illustrated in Figure 6.22, and this is by no means an exhaustive list. Figure 6.22(a), as discussed previously, is a transmission mode with light passing through the small aperture of a tapered fiber, and scattered light from the surface or near-surface volume irregularities is collected in the far field by optical means. For opaque samples, the reflected light can also be collected by designing collection optics above the sample. In Figure 6.22(b), the sample is illuminated from below at an angle that exceeds the total internal reflection; in this modality, the scattered light is collected through the small aperture at the end of the tapered fiber. The fiber must be in the near field for
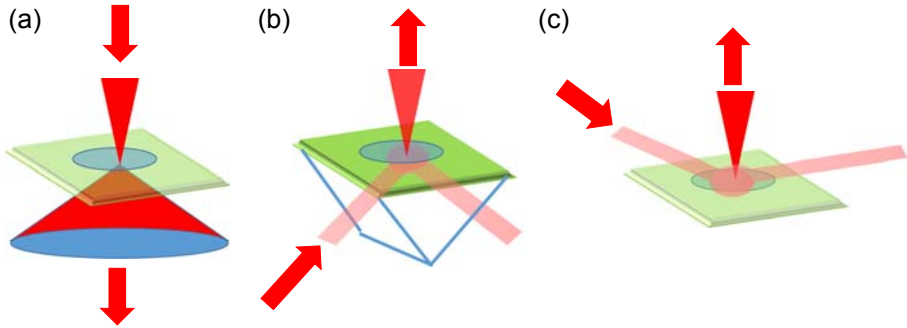
Figure 6.22 Three NSOM measurement modes: (a) fiber transmission with far-field collection optics, (b) evanescent wave illumination from below with tapered fiber collection, and (c) reflected light illumination with tapered fiber collection.

location-sensitive detection of the scattered light. A third modality is shown in Figure 6.22(c); the sample is illuminated from above by light and a near-field tip collected the scattered light from local irregularities on the surface. Other measurement modalities are also useful, but they are not discussed here. The interested reader can consult the references for extensive details on NSOM modalities and applications.

Problems

1. Consider the Helmholtz wave equation, Eqn (6.5).
   a. Use the decomposition of the field in SVEA to derive Eqn (6.19) using the approximation $k_0 |\partial E_e(r, \omega)/\partial z| \gg |\partial^2 E_e(r, \omega)/\partial z^2|$.
   b. Directly solve Eqn (2.19) using Fourier decomposition and show the result is equivalent to the integrand in Eqn (2.16).
2. Solve the Fraunhofer diffraction equation for plane-wave illumination of an elliptical aperture for which the border is defined by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

   where a and b are major/minor axes constants. Plot various aperture shapes in real space and compare with Fraunhofer diffraction patterns. Explain the feature that the larger dimension of the aperture has a smaller spread of diffracted power.
3. Consider a Gaussian beam of width w_0 and wavelength λ at its focal plane. Calculate the variance at the waist and in the far-field regime and show that the product satisfies the minimum value given in Eqn (6.25).
4. An electron has a kinetic energy given by K, so that the total energy is $E = m_0 c^2 + K$. The rest mass energy is $m_0 c^2$. Using the relativistic energy−momentum relation

$$E^2 = m_0^2 c^4 + p^2 c^2.$$

   a. Derive the relativistic expression for the electron's wavelength as a function of the kinetic energy K using deBroglie's momentum−wavelength equation.

    b. Show that it reduces to the nonrelativistic result (derived using Eqn (6.42) for sufficiently small kinetic energy. What ratio of energies determines whether the electron needs to be treated as a relativistic particle?

    c. For the following applied electron acceleration voltages, compare the electron wavelength for the nonrelativistic and relativistic expressions using a plot from $K = 10$ keV to 10 MeV.

5. List characteristics of the optical microscope, the TEM, the SEM, and AFM techniques. From your perspective, what are the pros and cons of each method. Think in terms of sample preparation, cost, resolution, working distance, depth of focus, three- or two-dimensional imagery, experimental artifacts, local environment for the sample, and extraction of additional information about the sample.

6. Consider an SEM.

    a. What are the effects of increasing the electron acceleration voltage?

    b. Suppose the SEM Rayleigh resolution is 1 nm for a 10-keV electron energy. Find the NA and determine the depth of the field.

    c. The electron energy is increased to 40 keV. Find the resolution and depth of field assuming the same NA.

Further reading

Books
[1] T.-C. Poon, P.P. Banerjee, Contemporary Optical Image Processing with Matlab, Elsevier, Amsterdam, 2001.
[2] J.W. Goodman, Introduction to Fourier Optics, McGraw-Hill, New York, 1996.
[3] H. Stark, Applications of Optical Transforms, Academic Press, Miami, 1982.
[4] L. Novotny, B. Hecht, Principles of Nano-Optics, second ed., Cambridge Press, Cambridge, 2012.
[5] D.B. Williams, C.B. Carter, Transmission Electron Microscopy, A Textbook for Materials Science, Plenum Press, New York, 1996.
[6] A. Zayats, D. Richards (Eds.), Nano-Optics and Near-Field Optical Microscopy, Artech House, Boston, 2009.

Articles
[1] J.A. Scholl, A.L. Koh, J.A. Dionne, Quantum Plasmon Resonances of Individual Metallic Nanoparticles, 483, 2012, 421, http://dx.doi.org/10.1038/nature10904.
[2] D.W. Pohl, W. Denk, W. Lanz, Optical stethoscopy: image recording with resolution λ/20, Appl. Phys. Lett. 44 (1984) 651.
[3] D.W. Pohl, Scanning near-field optical microscopy (SNOM), Adv. Opt. Electron Microsc. 12 (1991) 243.
[4] H.A. Bethe, Theory of diffraction by small holes, Phys. Rev. 66 (1944) 163−182.

Effective medium theories

M.A. Vincenti, D. de Ceglia
National Research Council, AMRDEC, Redstone Arsenal, AL, USA

7.1 Introduction

Effective medium theories provide macroscopic models of inhomogeneous media based on analytical, numerical, and sometimes experimental techniques. A description of composite materials in terms of effective medium approximations is a valuable and versatile tool to investigate, predict, and design the electromagnetic response of natural and structured materials. Effective medium models equip the macroscopic Maxwell's equations with very simple constitutive relations, eliminating the complexity of simulating light−matter interactions at the constituents' level. When approaching an electromagnetic problem with an effective medium theory, of extreme importance is the definition of its limits of validity. Pushing any effective medium theory beyond these limits may lead to reasonable but only partially correct results or to wrong predictions. Effective medium models usually depend on the electric and magnetic properties of the constituent materials, the volume fraction of each constituent, and in some case the geometry of the structure at the constituent level. The fundamental limitation of these models resides in the fact that the starting point of any approach for homogenizing structured materials is always the assumption that the wavelength of the field is much larger than the characteristic scale of the inhomogeneity. For this reason the electromagnetic parameters of the effective medium are usually given in terms of static solutions of the problem. Depending on the size, permittivity, and permeability of the constituents, as well as the index of the hosting medium, the limitations of the model may be more or less strict.

In this chapter we review several effective medium approaches, with particular attention to the Maxwell Garnett and the Bruggeman theories. We then discuss the theoretical derivation of these methods and the link with Mie-theory descriptions of small spheres and core−shell spherical particles. We then adapt these theories to the case of nonspherical inclusions, e.g., prolates and oblates. Quasi-static numerical approaches are then presented, one based on a capacitor model and the other on the Bergman−Milton spectral theory. The Nicolson−Ross−Weir technique based on full-wave solution of Maxwell's equation is then presented. Simple examples of the use of effective medium theory in the field of artificial materials are then discussed, namely, multilayer and wire media. Finally, we discuss the effects of retardation with an emphasis on spatial dispersion and the possibility to engineer artificial magnetic or bianisotropic materials.

7.2 Maxwell Garnett theory

This is the classical approach for homogenizing media with small inclusions dispersed in a continuous *host* medium or *matrix*. The basic structure is a two-phase medium with separated grains of a guest material, the inclusions with relative permittivity ε_i, hosted by a background medium, and the host with relative permittivity ε_h. We restrict the analysis to the case of nonmagnetic and isotropic materials. If the inclusions are small enough, then a quasi-static approximation can be adopted. For positive-permittivity inclusions the following rule of thumb is considered to be conservative: the particle size should not exceed one-tenth of the effective wavelength, which is the wavelength measured in the effective medium. However, in case of metallic or negative-permittivity inclusions, the limits of validity may be stricter, especially near the localized surface plasmon resonances (see Chapter 8). In the absence of any information about the shape of the inclusions, the most natural approach is to assume that they are small spheres. The idea behind the Maxwell Garnett homogenization approach is exemplified in Figure 7.1.

   If the material is excited by an external electric field, in the quasi-static approximation such a field can be considered to be constant on the length scale of each sphere. We indicate the external static field as $\mathbf{E}_e$. First we focus on the response of each isolated sphere to this excitation. Since the sphere is very small, it acts as a point source with an electric dipole moment proportional to the applied field. In other words, the response of an isolated sphere in the host medium is $\mathbf{p}_h = \varepsilon_0 \varepsilon_h \alpha \mathbf{E}_e$, where ε_0 is the vacuum permittivity, $\mathbf{p}_h$ is the induced dipole moment, $\alpha = 3V \frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + 2\varepsilon_h}$ is the static electric polarizability of the sphere, and V is the sphere's volume. The field inside the sphere, $\mathbf{E}_i = 3\varepsilon_h/(\varepsilon_i + 2\varepsilon_h)\mathbf{E}_e$, is uniform and parallel to the external field. The polarizability of the sphere is isotropic since both the permittivity and shape of the inclusions are assumed to be isotropic. The next step is to create an effective model of the distribution of small spheres (transition from the center to the right panel in Figure 7.1). The spheres are reduced to electric point dipoles, and the field radiated by each dipole is now influenced by the presence of all the other dipoles. At this point the information required is the number of dipoles per unit volume N. The definition of the effective permittivity is based on the average, or macroscopic, constitutive relation that links

Figure 7.1 On the left is the microstructure under investigation, which is a two-phase medium with very small guest inclusions of unknown shapes and dielectric constant ε_i dispersed in a continuous host medium with permittivity ε_h. On the central panel is the Maxwell Garnett "view" of the material on the left, in which the inclusions are described as small spheres of dielectric constant ε_i. On the right is Maxwell Garnett homogenization, with effective permittivity ε_{MG}.

the average electric field $\langle \mathbf{E} \rangle$ to the average displacement field $\langle \mathbf{D} \rangle$. The average operator integrates over sufficiently large volumes in order to provide an accurate description of average fields in the original medium. Hence one can write

$$\langle \mathbf{D} \rangle = \varepsilon_0 \varepsilon_{MG} \langle \mathbf{E} \rangle, \tag{7.1}$$

where ε_{MG} is the effective (relative) Maxwell Garnett permittivity that models the original mixture, as shown in Figure 7.1. One can also see the average medium response as the average response of the host medium plus the average response of the dipoles, so that

$$\langle \mathbf{D} \rangle = \varepsilon_0 \varepsilon_h \langle \mathbf{E} \rangle + \langle \mathbf{P} \rangle. \tag{7.2}$$

The average dipole response is $\langle \mathbf{P} \rangle = N\mathbf{p}$, where the dipole moment $\mathbf{p} \neq \mathbf{p}_h$ is now calculated in the presence of all the other dipoles. The evaluation of $\mathbf{p}$ is classically performed by evaluating the local electric field $\mathbf{E}_L$, which is the field locally "felt" by each dipole. This field is the average field $\langle \mathbf{E} \rangle$ augmented by a contribution due to the average polarization that surrounds each dipole, also known as the Lorentz field. In order to find the field $\mathbf{E}_L$ acting on a single dipole, a simple model of the mixture is adopted, in which a fictitious spherical boundary separates a macroscopic background with average polarization $\langle \mathbf{P} \rangle$ from a microscopic spherical cavity surrounding the dipole at the center of the sphere. This situation is schematized in Figure 7.2. The field due to the polarized background can easily be calculated by considering that the charge density $\nabla \cdot \mathbf{P}$ is zero everywhere, except on the cavity boundary where a surface charge distribution of density $-\langle \mathbf{P} \rangle \cdot \widehat{\mathbf{n}}$ induces an additional electric field at the dipole location, i.e., at the sphere's center.

In this simple picture the local field can be straightforwardly written as

$$\mathbf{E}_L = \langle \mathbf{E} \rangle + \frac{\langle \mathbf{P} \rangle}{3\varepsilon_0 \varepsilon_h} \tag{7.3}$$

and the dipole moment as

$$\mathbf{p} = \varepsilon_0 \varepsilon_h \alpha \langle \mathbf{E} \rangle + \alpha \frac{\langle \mathbf{P} \rangle}{3}. \tag{7.4}$$
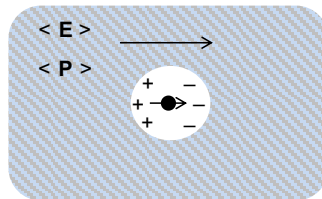

Figure 7.2 Classical definition of local electric field $\mathbf{E}_L$. A fictitious spherical cavity separates a homogeneous background from a microscopic phase in which a dipole is included. The field at the dipole location, the center of the sphere, is the local field $\mathbf{E}_L$.

We can now retrieve the Maxwell Garnett permittivity in terms of the polarizability α and the number density N,

$$\varepsilon_{MG} = \varepsilon_h \left(1 + \frac{N\alpha}{1 - \frac{N\alpha}{3}} \right). \tag{7.5}$$

For very diluted media $1 - \frac{N\alpha}{3} \approx 1$, and the effective permittivity is simply $\varepsilon_{MG} \approx \varepsilon_h(1 + N\alpha)$. The same expression can easily be obtained when the local field is $\mathbf{E}_L \approx \langle \mathbf{E} \rangle$. This approximation is fully justified in diluted mixtures where the interaction between dipoles is weak.

The following form of the relation in Eqn (7.5)

$$\frac{N\alpha}{3} = \frac{\varepsilon_{MG} - \varepsilon_h}{\varepsilon_{MG} + 2\varepsilon_h} \tag{7.6}$$

is known as the Clausius−Mossotti formula, Maxwell's formula, or the Lorentz−Lorenz formula.

Substitution of the expression of the polarizability α in the Clausius−Mossotti relation gives the Rayleigh formula

$$\frac{\varepsilon_{MG} - \varepsilon_h}{\varepsilon_{MG} + 2\varepsilon_h} = f \frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + 2\varepsilon_h}, \tag{7.7}$$

that relates the effective permittivity to the constituents' permittivities and to the parameter $f = NV$, which is the volume fraction of the inclusions (in this case spheres) in the medium. The so-called Maxwell Garnett formula derives from Eqn (7.7) and it is written as follows:

$$\varepsilon_{MG} = \varepsilon_h \left[1 + 3f \frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + 2\varepsilon_h - f(\varepsilon_i - \varepsilon_h)} \right]. \tag{7.8}$$

This simple formula represents the classical approach to homogenizing composite media, and it is widely used in many applications. It is interesting to notice that the only necessary parameters for retrieving the Maxwell Garnett permittivity are the permittivities of inclusions and host medium and the volume fraction of the inclusions. The formula does not require that the spheres are of the same size and located at specific positions (e.g., periodic arrays). The only requirement is that the wavelength in the medium should be much larger than the size of the inclusions. The Maxwell Garnett theory predicts that $\varepsilon_{MG} = \varepsilon_h$ for $f \to 0$, and $\varepsilon_{MG} = \varepsilon_i$ for $f \to 1$. Although the formula does not require small values of f, predictions for large values of f are questionable because of the higher-order multipole effects triggered by decreased interparticle distances. Higher-order terms cannot be handled with the simple approach described here, which is strictly based on the quasi-static and dipole approximations. However, multipole effects, which turn out to be important for values of f larger than

0.3−0.5, can be included in the Maxwell Garnett formula with the introduction of fictitious depolarization factors. Another important limitation of the Maxwell Garnett theory is the asymmetric behavior of the central formula in Eqn (7.7) or (7.8). The roles of the inclusion phase and the host phase are not interchangeable, so that the validity of the Maxwell Garnett formula is limited to mixtures in which there is a clear determination of a host medium phase and a small-inclusions phase. For denser mixtures and in any microstructure characterized by the lack of a clear distinction between host and inclusion media, different approaches are more suitable, including the Bruggeman theory that is presented in the next paragraph.

An alternative and interesting derivation of the Maxwell Garnett formula is based on the optical theorem and Mie theory. The microstructure under investigation is of the same type illustrated in Figure 7.1, a two-phased mixture with small inclusions of relative permittivity ε_i immersed in a host of relative permittivity ε_h. The microstructure is now modeled as a random distribution of unit cells dispersed in a homogeneous background of relative permittivity ε_{MG}, equal to the effective permittivity that we are trying to retrieve. In each unit cell we have a core−shell spherical particle, with a core of relative permittivity ε_i and radius R_i and a shell of relative permittivity ε_h and radius R_h. The approach is illustrated in Figure 7.3. The particles may have different sizes, but the volume fraction of the core medium with respect to the entire core−shell particle must be equal to the volume fraction of the inclusions in the original inhomogeneous structure. In other words, the requirement is that $f = R_i^3 / R_h^3$. In order to have a macroscopic permittivity equal to ε_{MG} in the random distribution of core−shell particles, the only possibility is that the particles must be "invisible" for an electromagnetic radiation of wavelength much larger than the inclusions.

This requires that the extinction cross-section of the core−shell particles must be zero. The optical theorem relates the extinction coefficient C_{ext} of a particle illuminated by radiation impinging from a specific direction ϑ_0 and with a determined polarization direction $\widehat{\mathbf{e}}_p$ to the scattering amplitude evaluated in the same direction of the input beam. The expression of the extinction cross-section is

$$C_{ext} = \frac{4\pi}{k^2} \mathrm{Re}\left[\left(\mathbf{S} \cdot \widehat{\mathbf{e}}_p \right)_{\vartheta = \vartheta_0} \right], \tag{7.9}$$
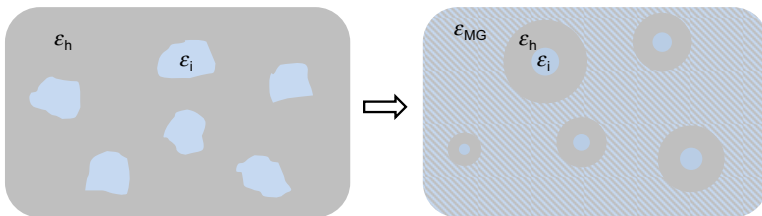

Figure 7.3 On the left is the microstructure under investigation, which is equal to that reported on the left of Figure 7.1. On the right the inclusions are described as small, coated spheres with a core having a relative permittivity ε_i and a shell of relative permittivity ε_h. The host medium has an effective permittivity ε_{MG}.

where $k = \frac{2\pi}{\lambda_0}\sqrt{\varepsilon_{MG}}$ is the wavenumber in the effective medium, and $\mathbf{S}$ is the vector-scattering amplitude. Our "invisibility" requirement translates into the condition $C_{ext} = 0$, or

$$\left(\mathbf{S}\cdot\widehat{\mathbf{e}}_p\right)_{\vartheta=\vartheta_0} = 0. \tag{7.10}$$

It is important to notice that under this condition the distribution of coated spheres would allow a plane wave to propagate without phase front deformation, so that the usual Fresnel formulas can be straightforwardly applied to the boundaries of the mixture. Using the Mie theory, which is discussed in more detail in Chapter 8, it is possible to relate the scattering amplitude to the properties of the particle (permittivity distribution and size). The expression of the extinction coefficient

$$\left(\mathbf{S}\cdot\widehat{\mathbf{e}}_p\right)_{\vartheta=\vartheta_0} = \frac{1}{2}\sum_1^\infty (2n+1)\mathrm{Re}\{a_n + b_n\} \tag{7.11}$$

depends on the scattering coefficients a_n and b_n, whose expressions are given in terms of Riccati−Bessel functions and their derivatives (see the details in Chapter 8). The expansion of the Eqn (7.11) to the first order, corresponding to the dipole or quasi-static approximation, reads as follows:

$$\left(\mathbf{S}\cdot\widehat{\mathbf{e}}_p\right)_{\vartheta=\vartheta_0} \approx \mathrm{i}(kR_h)^3\frac{(\varepsilon_h - \varepsilon_{MG})(\varepsilon_i + 2\varepsilon_h) + f(\varepsilon_i - \varepsilon_h)(\varepsilon_{MG} + 2\varepsilon_h)}{(\varepsilon_h + 2\varepsilon_{MG})(\varepsilon_i + 2\varepsilon_h) + 2f(\varepsilon_i - \varepsilon_h)(\varepsilon_h - \varepsilon_{MG})}. \tag{7.12}$$

The error in Eqn (7.12) due to higher-order multipoles is of the order $(kR_h)^5$.

The condition $\left(\mathbf{S}\cdot\widehat{\mathbf{e}}_p\right)_{\vartheta=\vartheta_0} = 0$ now translates into an explicit relation between the Maxwell Garnett effective permittivity, the constituents' permittivities, and the volume fraction. This relation

$$\frac{\varepsilon_{MG} - \varepsilon_h}{\varepsilon_{MG} + 2\varepsilon_h} = f\frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + 2\varepsilon_h} \tag{7.13}$$

is identical to that obtained in Eqn (7.7) using the Lorentz−Lorenz/Clausius−Mossotti formula and the concept of average and local fields.

A natural extension of the Maxwell Garnett theory to multiphase mixtures is given by the following adaptation of the Rayleigh formula in Eqn (7.7). It reads as follows:

$$\frac{\varepsilon_{MG} - \varepsilon_h}{\varepsilon_{MG} + 2\varepsilon_h} = \sum_{m=1}^M f_m\frac{\varepsilon_m - \varepsilon_h}{\varepsilon_m + 2\varepsilon_h}. \tag{7.14}$$

Here f_m and ε_m are the volume fraction and the relative permittivity of the m-th phase, respectively. Similar restrictions apply for this formula: all the inclusions

must be much smaller than the effective wavelength, and the effective permittivity is more accurate for small values of f_m.

So far the inclusions have been modeled as spherical particles with the result of an isotropic effective response. Nevertheless, the Maxwell Garnett theory can be extended to nonspherical inclusions. For arbitrarily shaped particles, the electrostatic polarizability can be retrieved by using numerical techniques (finite-element or finite-difference methods to solve the Poisson's equation). However, simple Maxwell Garnett formulations can be retrieved for mixtures of ellipsoids.

If the ellipsoids are fully aligned, the effective permittivity will be anisotropic. In this case the following approximation is usually adopted:

$$\frac{\langle \mathbf{E}_i \rangle}{\langle \mathbf{E}_h \rangle} \approx \frac{\breve{\mathbf{E}}_i}{\breve{\mathbf{E}}_h} = \sum_j \frac{\cos^2 \vartheta_j}{1 + \left(\frac{\varepsilon_i}{\varepsilon_h} - 1\right) L_j} \tag{7.15}$$

where $\langle \mathbf{E}_{i,h} \rangle$ is the average field in the inclusion and the host medium of the mixture and $\breve{\mathbf{E}}_{i,h}$ is the same field for the case of an isolated inclusion. ϑ_j is the angle between the driving (electrostatic) field and the j-th semiaxis of the ellipsoid, whereas L_j ($j = x$, y, z) is the particle geometrical factor in the x, y and z directions. Their expression is

$$L_j = \frac{R_x R_y R_z}{2} \int_0^\infty \left[\left(R_j^2 + q\right)f(q)\right]^{-1} dq, \tag{7.16}$$

where $f(q) = \sqrt{(q + R_x^2)^2(q + R_y^2)^2(q + R_z^2)^2}$ and R_j are the ellipsoid's semiaxes. The geometrical factors are related by $\sum_j L_j = 1$. For spheres $L_j = 1/3$. The revolution of an ellipse around its minor (major) axis generates prolate (oblate) ellipsoids. For such particles, closed-form expressions of the geometrical factors in Eqn (7.15) can be found. In particular, for prolates $R_z > R_x = R_y$, $L_z = \frac{1 - e^2}{2e^3}\left(\ln\frac{1+e}{1-e} - 2e\right)$, and $L_x = L_y = \frac{1}{2}(1 - L_z)$, where $e = \sqrt{1 - R_x^2/R_z^2}$ is the particle eccentricity. For oblates ($R_z < R_x = R_y$), $L_z = \frac{1 + e^2}{e^3}(e - \tan^{-1} e)$, and $L_x = L_y = \frac{1}{2}(1 - L_z)$, where $e = \sqrt{\frac{R_x^2}{R_z^2} - 1}$ is the particle eccentricity.

We now follow the usual Maxwell Garnett approach in which the average field in the mixture is $\langle \mathbf{E} \rangle = f\langle \mathbf{E}_i \rangle + (1 - f)\langle \mathbf{E}_h \rangle$, and the average displacement field is $\langle \mathbf{D} \rangle = \varepsilon_0[\varepsilon_i f\langle \mathbf{E}_i \rangle + \varepsilon_h(1 - f)\langle \mathbf{E}_h \rangle]$. The expression for the relative permittivities along the principal directions can be recast as a generalization of the homogenization formula for spherical inclusions given in Eqn (7.8):

$$\varepsilon_{MG,j} = \varepsilon_h\left(1 + f\frac{\varepsilon_i - \varepsilon_h}{\varepsilon_h + L_j(1 - f)(\varepsilon_i - \varepsilon_h)}\right). \tag{7.17}$$

For mixtures of ellipsoids randomly aligned, the effective medium becomes isotropic with relative permittivity given by

$$\varepsilon_{MG} = \frac{\varepsilon_h(1-f) + \frac{f\varepsilon_i}{3}\sum_j \frac{\varepsilon_h}{\varepsilon_h + L_j(\varepsilon_i - \varepsilon_h)}}{1 - f + \frac{f}{3}\sum_j \frac{\varepsilon_h}{\varepsilon_h + L_j(\varepsilon_i - \varepsilon_h)}}. \tag{7.18}$$

7.3 Bruggeman theory

The Maxwell Garnett formulas given in the previous paragraph represent a valid homogenization model for mixtures with a well-defined host medium and inclusions, and they result more accurate for relatively small values of the inclusion volume factor f. For aggregate mixtures with random distributions of two or more constituents effective medium theories based on a statistical formulation are more suitable. The classic theory for this class of inhomogeneous mixtures is the Bruggeman theory. We now consider a two-phase microstructure of the type illustrated in Figure 7.4, where the constituent with permittivity ε_i has volume fill factor f, and the constituent with permittivity ε_h has volume fill factor $1 - f$.

This mixture is now modeled as a continuous medium hosting a distribution of small spherical inclusions of two different dielectric permittivities. The probabilities of finding spheres with permittivity ε_i and ε_h are f and $1 - f$, respectively, which correspond to the volume fill factors of the two phases in the original mixture. We now assume that the host medium of the Bruggeman mixture has the unknown effective permittivity ε_{Br}, as indicated on the right side of Figure 7.4, and invokes the transparency or "invisibility" condition for the distribution of the spherical inclusions. In the quasi-static approximation, this condition can be retrieved by setting the scattering amplitude for small spherical particles of permittivity ε_p and radius R_p to zero, i.e.,

$$\left(\mathbf{S} \cdot \widehat{\mathbf{e}}_p\right)_{\vartheta = \vartheta_0} \approx i(kR_p)^3 \frac{\varepsilon_p - \varepsilon_{Br}}{\varepsilon_p + 2\varepsilon_{Br}} = 0. \tag{7.19}$$
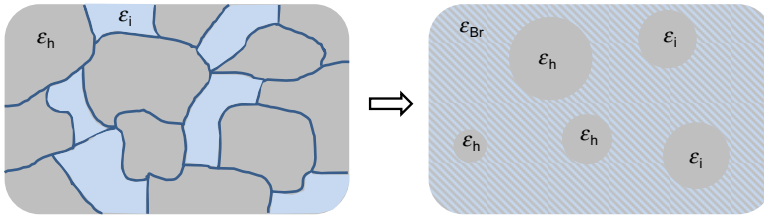
Figure 7.4 On the left is the two-phase inhomogeneous material, a space-filling random mixture of two phases. One phase has dielectric constant ε_i and fill factor f and the other phase has dielectric constant ε_h and fill factor $1 - f$. On the right is the Bruggeman theory model of the structure, with a background medium of effective permittivity ε_{Br} hosting very small spheres whose dielectric constant is ε_i with probability f and ε_h with probability $1 - f$.

In our distribution of spheres ε_p can be either ε_i with probability f or ε_h with probability $1 - f$. The resulting "averaged transparency" condition reads as follows:

$$f\frac{\varepsilon_i - \varepsilon_{Br}}{\varepsilon_i + 2\varepsilon_{Br}} + (1 - f)\frac{\varepsilon_h - \varepsilon_{Br}}{\varepsilon_h + 2\varepsilon_{Br}} = 0. \tag{7.20}$$

Equation (7.20) is the basic form of the Bruggeman theory. The formula is symmetric since the roles of inclusions and host materials are interchangeable. There are two solutions of Eqn (7.20) for the effective permittivity ε_{Br}, one of which is usually unphysical because it violates causality. The formula for spherical inclusions can be easily extended to multiphase aggregates by adding more terms in the relation Eqn (7.20), yielding

$$\sum_{m=1}^{M} f_m \frac{\varepsilon_m - \varepsilon_{Br}}{\varepsilon_m + 2\varepsilon_{Br}} = 0, \tag{7.21}$$

where f_m is the fill factor of the m-th constituent of the mixture, and M is the number of phases.

Shape effects of the inclusions can be included in the Bruggeman theory. The relative permittivity of a multiphase distribution of randomly oriented ellipsoids with equal shape and dielectric constants ε_m dispersed in a Bruggeman-type mixture of background permittivity ε_h has been derived by Polden and van Stanten and reads as follows:

$$\varepsilon_{Br} = \varepsilon_h \left\{ 1 - \frac{1}{3} \sum_{m=1}^{M} \left[f_m(\varepsilon_m - \varepsilon_h) \sum_{j=1}^{3} \frac{1}{\varepsilon_{Br} + (\varepsilon_m - \varepsilon_{Br})L_j} \right] \right\}^{-1}. \tag{7.22}$$

The Bruggeman formula for fully aligned ellipsoidal inclusions is the generalization of the formula given in Eqn (7.21), taking into account the shape effect, and reads as follows:

$$\sum_{m=1}^{M} f_m \frac{\varepsilon_m - \varepsilon_{Br,j}}{\varepsilon_m + K_j\varepsilon_{Br,j}} = 0, \tag{7.23}$$

where $\varepsilon_{Br,j}$ is the effective permittivity in the principal directions, and $K_j = (1 - L_j)/L_j$ is the screening parameter.

For a two-phase mixture with constituents having permittivities ε_i and ε_h and filling ratios of f and $1 - f$, respectively, the relation Eqn (7.23) reduces to

$$f\frac{\varepsilon_i - \varepsilon_{Br,j}}{\varepsilon_i + K_j\varepsilon_{Br,j}} + (1 - f)\frac{\varepsilon_h - \varepsilon_{Br,j}}{\varepsilon_h + K_j\varepsilon_{Br,j}} = 0. \tag{7.24}$$

For comparison, the Maxwell Garnett formula for fully aligned ellipsoids given in Eqn (7.17) can be written in terms of screening parameters and yields

$$\frac{\varepsilon_{\text{MG},j} - \varepsilon_{\text{h}}}{\varepsilon_{\text{MG},j} + K_j \varepsilon_{\text{h}}} = f \frac{\varepsilon_{\text{i}} - \varepsilon_{\text{h}}}{\varepsilon_{\text{i}} + K_j \varepsilon_{\text{h}}}. \tag{7.25}$$

Here we can highlight another limitation of the Maxwell Garnett theory that concerns particles with very small depolarization factors. A low depolarization factor implies an elongated shape of the particle with the result of stronger particle–particle interactions. In this situation, which is similar to the scenario of a mixture with large inclusions' fill factor, the Maxwell Garnett formula (Eqn (7.25)) is not accurate, and the Bruggeman prediction, (Eqn (7.24)), should be adopted. Another scenario in which the Bruggeman theory provides a more realistic electromagnetic description is for mixtures with large differences in the permittivities of the constituents—e.g., metal-dielectric mixtures, where a percolation phenomenon above a threshold of the metallic phase occurs. This threshold is the critical metal filling factor above which there is formation of long-range connectivity between metal grains, and the optical response of the mixture changes abruptly. In the Bruggeman theory for isotropic spherical, metallic inclusions dispersed in a dielectric host ($|\varepsilon_{\text{i}}| \gg \varepsilon_{\text{h}}$) the effective permittivity reduces to $\varepsilon_{\text{h}}/(1 - 3f)$. The zero in the denominator of this expression gives the critical fill factor $f_{\text{c}} = 1/3$ for metal-based mixtures. In other words, for $f \leq f_{\text{c}}$ the mixture behaves like an insulator, whereas for $f > f_{\text{c}}$ the mixture acts like a conductor. It is worth mentioning that the Maxwell Garnett theory predicts an unrealistic percolation threshold ($f_{\text{c}} = 1$).

7.4 Quasi-static numerical approaches

As in the effective medium theories discussed in the previous sections, these numerical techniques are based on quasi-static approximation, i.e., one assumes that the typical size of the inclusions is much smaller than the effective wavelength. Quasi-static approaches have been applied extensively to metamaterials, in particular to periodic and subwavelength arrangements of plasmonic resonators. Here we provide a brief overview of the basic implementation. We consider a mixture that can be modeled as an array of scatterers with relative permittivity ε_{i} immersed in a host medium with permittivity ε_{h} and divide the volume in rectangular-prism unit cells, similar to a three-dimensional metamaterial. In order to understand the macroscopic bulk response of this system we focus on the behavior of the unit cell centered at the origin of the Cartesian coordinates, i.e., the rectangular domain $\left[-\frac{a}{2}, \frac{a}{2}\right] \times \left[-\frac{b}{2}, \frac{b}{2}\right] \times \left[-\frac{c}{2}, \frac{c}{2}\right]$, as illustrated in Figure 7.5.

Here the constituent materials are assumed to be nonmagnetic. For simplicity we also assume scatterers are symmetric with respect to the planes xy, xz, and yz. Under these circumstances the only nonzero effective permittivities will be the diagonal terms $\varepsilon_{\text{eff}}^{xx}$, $\varepsilon_{\text{eff}}^{yy}$, and $\varepsilon_{\text{eff}}^{zz}$. When an external, static electric field $\mathbf{E}_0$ is applied to this unit cell, the local electric potential $\varphi(\mathbf{r})$ is found by solving the Poisson's equation

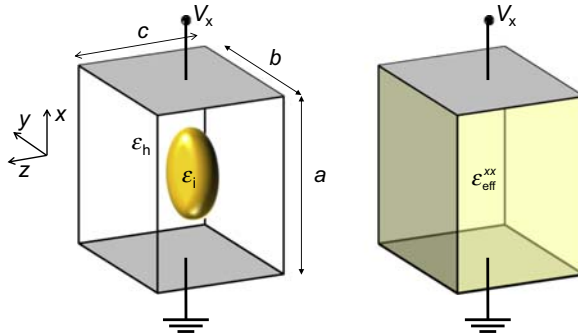$$\nabla \cdot [\varepsilon_{\omega}(\mathbf{r}) \nabla \varphi(\mathbf{r})] = 0, \tag{7.26}$$

Figure 7.5 On the left is the unit cell of a material with inclusions with permittivity ε_i in a host of permittivity ε_h. A static voltage is applied between the two virtual parallel plates at $x = -\frac{a}{2}, \frac{a}{2}$ in order to retrieve the xx term of the effective permittivity $\varepsilon_{\text{eff}}^{xx}$. On the right is the homogenized version of the unit cell, a capacitor filled with a homogeneous medium with permittivity $\varepsilon_{\text{eff}}^{xx}$.

where $\varepsilon_\omega(\mathbf{r})$ is the frequency-dependent absolute permittivity at position $\mathbf{r}$ in the unit cell. The boundary conditions needed to solve Eqn (7.26) depend on the permittivity component to retrieve. If we are interested in calculating $\varepsilon_{\text{eff}}^{xx}$ then it is useful to apply a field $\mathbf{E}_0 = E_{0,x}\hat{\mathbf{x}}$ along the x direction. At this point the unit cell may be thought of as a small capacitor with a static voltage $V_x = \varphi\left(-\frac{a}{2}, y, z\right) - \varphi\left(\frac{a}{2}, y, z\right) = E_{0,x}a$ applied between two virtual plates located at $x = -\frac{a}{2}, \frac{a}{2}$. This scenario is depicted in Figure 7.5. The boundary conditions at $y = -\frac{b}{2}, \frac{b}{2}$ and $z = -\frac{c}{2}, \frac{c}{2}$ are $\frac{\partial \varphi(\mathbf{r})}{\partial y} = 0$ and $\frac{\partial \varphi(\mathbf{r})}{\partial z} = 0$, respectively. Using these boundary conditions, the problem in Eqn (7.26) can be solved numerically with different approaches. The most popular are based on finite-differences and finite-element methods. Once the potential $\varphi(\mathbf{r})$ is known, we impose that the inhomogeneous unit cell under investigation shows the same capacitance of a parallel-plate capacitor filled with a homogeneous medium with relative permittivity $\varepsilon_{\text{eff}}^{xx}$. This capacitance is simply $C_{\text{eff},x} = \varepsilon_0 \varepsilon_{\text{eff}}^{xx} bc/a$. Figure 7.5 clarifies this equivalence. On the other hand, the capacitance of the inhomogeneous unit cell is given by $C_x = q_0/V_x$, where q_0 is the charge per unit length on the virtual plate at $x = a/2$. The charge is found straightforwardly by integrating the surface charge density $\sigma(y,z) = -\varepsilon_0 \varepsilon_h \frac{\partial \varphi(\mathbf{r})}{\partial x}\Big|_{x=a/2}$ over the surface of the virtual plate located at $x = a/2$. By equating $C_x = C_{\text{eff},x} = \varepsilon_0 \varepsilon_{\text{eff}}^{xx} bc/a$, the quasi-static effective permittivity is retrieved, and it reads as follows:

$$\varepsilon_{\text{eff}}^{xx} = -\varepsilon_h \frac{\int_{y=-b/2}^{b/2} \int_{z=-c/2}^{c/2} \frac{\partial \varphi(\mathbf{r})}{\partial x}\Big|_{x=a/2} dydz}{E_{0,x}bc}. \tag{7.27}$$

A similar procedure can be followed to retrieve $\varepsilon_{\text{eff}}^{yy}$ and $\varepsilon_{\text{eff}}^{zz}$ by applying static voltages across virtual parallel plates perpendicular to the y and z directions, respectively. An alternative definition of the electrostatic effective permittivity can be derived by equating the energy stored by the two capacitors in Figure 7.5, yielding

$$\varepsilon_{\text{eff}}^{xx} = \frac{\int_{x=-a/2}^{z/2} \int_{y=-b/2}^{b/2} \int_{z=-c/2}^{c/2} \varepsilon_\omega(\mathbf{r})[\nabla\varphi(\mathbf{r}) \cdot \nabla\varphi(\mathbf{r})]\mathrm{d}x\mathrm{d}y\mathrm{d}z}{\mathbf{E}_{0,x}^2 abc}. \tag{7.28}$$

For constituents with chromatic dispersion, such as metal or semiconductor inclusions, the procedure described above must be repeated for each frequency so that a frequency-dependent, complex permittivity tensor is obtained. For plasmonic metamaterials, or more generally metal-dielectric mixtures, a faster and more insightful approach for obtaining electrostatic homogenizations is based on the Bergman−Milton spectral theory. Here, metallic inclusions of permittivity $\varepsilon_i(\omega)$ are embedded in a dielectric host of permittivity ε_h. The first step of this method is to define the Bergman's spectral parameter $s(\omega) = \varepsilon_h[\varepsilon_h - \varepsilon_i(\omega)]^{-1}$ and rewrite the frequency-dependent permittivity distribution $\varepsilon_\omega(\mathbf{r})$ in the form $\varepsilon_\omega(\mathbf{r}, \omega) = [\varepsilon_i(\omega) - \varepsilon_h][\theta(\mathbf{r}) - s(\omega)]$, where the structure function $\theta(\mathbf{r})$ is equal to 1 in the inclusion volume and 0 elsewhere. If the potential distribution is written as the sum of the applied potential $\varphi_0(\mathbf{r}, \omega)$ plus the induced potential $\psi(\mathbf{r}, \omega)$, i.e., $\varphi(\mathbf{r}, \omega) = \varphi_0(\mathbf{r}, \omega) + \psi(\mathbf{r}, \omega)$, the Poisson's equation can be recast as follows:

$$\nabla \cdot [\theta(\mathbf{r})\nabla\psi(\mathbf{r}, \omega)] - s(\omega)\nabla^2\psi(\mathbf{r}, \omega) = -\nabla \cdot [\theta(\mathbf{r})\nabla\varphi_0(\mathbf{r}, \omega)], \tag{7.29}$$

where the Laplace equation $\nabla^2\varphi_0(\mathbf{r}, \omega) = 0$ for the external potential has been used. The boundary conditions are $\psi(\mathbf{r}, \omega) = 0$ on the edges where the voltage is applied and $\hat{\mathbf{n}} \cdot \nabla\psi(\mathbf{r}, \omega) = 0$ on the other edges of the unit cell. Next the eigenmodes of the system, which correspond to the surface plasmon modes supported by the metallic inclusions, are set as a basis to solve this equation. These modes are the source-free solutions of the problem in Eqn (7.29), and they are found by solving the following generalized eigenvalue problem:

$$\nabla \cdot [\theta(\mathbf{r})\nabla\varphi_n(\mathbf{r})] = s_n\nabla^2\varphi_n(\mathbf{r}). \tag{7.30}$$

By integrating Eqn (7.29) it follows that $s_n = \iiint \theta(\mathbf{r})\nabla\varphi_n \cdot \nabla\varphi_n \mathrm{d}V / \iiint \nabla\varphi_n \cdot \nabla\varphi_n \mathrm{d}V$, so that the eigenvalues will be real numbers and $0 \le s_n \le 1$. The potential $\varphi(\mathbf{r}, \omega)$ can now be expressed by using the following eigenmode expansion:

$$\varphi(\mathbf{r}, \omega) = \varphi_0(\mathbf{r}, \omega) + \sum_n \frac{s_n}{s_n - s(\omega)}(\varphi_n, \varphi_0)/(\varphi_n, \varphi_n)\varphi_n(\mathbf{r}), \tag{7.31}$$

where $(\rho, \tau) = \iiint \nabla\rho^* \cdot \nabla\tau \mathrm{d}V$. Once $\varphi(\mathbf{r}, \omega)$ is known, the quasi-static permittivity can easily be calculated by applying the definitions from Eqn (7.27) or (7.28).

The main advantage of this approach is the intimate connection with the physical properties of the microscopic structure of the mixture. Useful information about the optical properties of the structure can be retrieved once the eigenmodes are known. For example, the characteristic frequency of the source-free modes is a complex quantity $\omega_n + i\gamma_n$ that can be found by solving the equation $s_n = \varepsilon_h[\varepsilon_h - \varepsilon_i(\omega)]^{-1}$. The real part is associated with the frequency at which the mode can be excited; the imaginary part is related to the characteristic decay time of the mode. It is worth mentioning that for simple geometries, e.g., small spherical inclusions, ellipsoids, or the split-ring resonators, there are only a few physically meaningful eigenmodes, while complex geometries such as fractals or random mixtures generally support a larger number of such modes.

7.5 Nicolson–Ross–Weir method

This is a homogenization method based on the inversion of Fresnel formulas relative to the transmission and reflection coefficients through slabs of homogeneous media. The technique was conceived to estimate the complex permittivity and permeability of an unknown material from the measured transmission and reflection spectra of a finite-thickness sample. It was originally proposed in the time domain for pulsed measurement systems and then adapted to higher resolution, frequency domain systems. The transmission and reflection spectra may be retrieved with experiments or with numerical simulations. The idea behind this method is sketched in Figure 7.6. A slab of thickness d of the unknown natural or artificial mixture is modeled as a slab of a homogeneous medium with effective (relative) permittivity ε_{eff} and permeability μ_{eff}. It is supposed that the thickness of the homogeneous slab is equal to d. The (complex) reflection and transmission coefficients R and T under normal incidence plane wave excitation are somehow known, via an experiment, a theoretical prediction, or a numerical simulation.
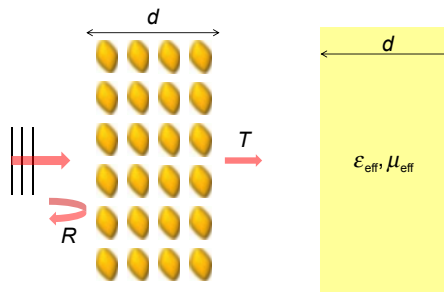
Figure 7.6 On the left is a slab of a composite medium with thickness d illuminated at normal incidence. R and T are the complex reflection and transmission coefficients. On the right is the homogenized slab with effective (relative) permittivity ε_{eff} and permeability μ_{eff}.

The Fresnel reflection (R) and transmission (T) coefficients may be written in the following form:

$$R = \frac{\Gamma\left(1 - e^{-2ikn_{\text{eff}}d}\right)}{1 - \Gamma^2 e^{-2ikn_{\text{eff}}d}},$$

(7.32)

$$T = \frac{\left(1 - \Gamma^2\right)e^{-ikn_{\text{eff}}d}}{1 - \Gamma^2 e^{-2ikn_{\text{eff}}d}},$$

(7.33)

which are dependent on the effective refractive index of the slab $n_{\text{eff}} = \sqrt{\varepsilon_{\text{eff}}\mu_{\text{eff}}}$, free-space wavenumber $k = \omega/c$, and the reflection coefficient Γ across the first interface between the input medium and the semi-infinite homogeneous slab with relative parameters (ε_{eff}, μ_{eff}). At normal incidence $\Gamma = \frac{\eta_{\text{eff}} - \eta_0}{\eta_{\text{eff}} + \eta_0}$, $\eta_{\text{eff}} = \sqrt{\mu_{\text{eff}}/\varepsilon_{\text{eff}}}$ is the effective intrinsic impedance of the homogeneous slab, and η_0 is the intrinsic impedance of the input/output medium. The inversion of Eqns (7.32) and (7.33) leads to the following expression of the effective impedance:

$$\eta_{\text{eff}} = \pm\eta_0\sqrt{\frac{(1 + R)^2 - T^2}{(1 - R)^2 - T^2}}$$

(7.34)

and the following expression for the quantity $Q = e^{-ikn_{\text{eff}}d}$

$$Q = \frac{T}{1 - R\frac{\eta_{\text{eff}} - \eta_0}{\eta_{\text{eff}} + \eta_0}}.$$

(7.35)

From Eqns (7.34) and (7.35) one can write the effective refractive index as follows:

$$n_{\text{eff}} = \frac{i}{kd}\log(Q) = \frac{1}{kd}\{i\text{Log}|Q| - [\text{Arg}(Q) + 2m\pi]\}.$$

(7.36)

Here $\log(Q)$ is the complex, multiple-valued logarithm of (Q), $\text{Log}|Q|$ is the ordinary real logarithm of $|Q|$, $\text{Arg}(Q)$ is the argument in the principal branch ($m = 0$), and $m = \pm1, \pm2,\ldots$ indicates the branch of $\log(Q)$. The effective parameters can finally be written as follows:

$$\varepsilon_{\text{eff}} = \frac{kn_{\text{eff}}}{\omega\eta_{\text{eff}}}, \quad \mu_{\text{eff}} = \frac{kn_{\text{eff}}\eta_{\text{eff}}}{\omega}.$$

(7.37)

This method has several limitations. The main one is that while the choice of sign for the effective impedance and refractive index does not alter the value of the effective parameters extracted via Eqn (7.37), there is an intrinsic ambiguity in the definition of Re(n_{eff}) in Eqn (7.36), owing to the multiple-valued complex logarithm $\log(Q)$ and the choice of the branch order m. This problem may be solved in very thin slabs in which

the effective wavelength is larger than $2d$. In this case it is possible to consider only the principal branch ($m = 0$) in Eqn (7.36) so that an unambiguous definition of the effective parameters is retrieved by using Eqn (7.37). However, this assumption implies an a priori knowledge or at least estimate of the effective refractive index. Moreover, the application of homogenization principles conceived for bulk natural or artificial mixtures may be misleading for the homogenization of very thin films. For structured films or artificial surfaces, more sophisticated techniques have been developed, based on the introduction of a nonlocal, effective surface susceptibility. In the case of electrically thick slabs, a common solution is to apply the retrieval method described above for two different values of slab thickness d_1 and d_2. The value of the effective refractive index should be unique once the frequency is fixed and must not depend on the sample thickness. As a consequence, the ambiguity is removed by identifying the correct branches that lead to the same value of $\mathrm{Re}(n_{\mathrm{eff}})$ for the two sample thicknesses ($d = d_1$ and $d = d_2$). The Nicolson−Ross−Weir method has been extended to the characterization of mixtures in the case of oblique plane wave incidence for the study of spatial dispersion effects in metamaterials.

7.6 Anisotropic mixtures: multilayers and wire media

In the previous paragraphs we described examples of anisotropic media realized with three-dimensional distributions of fully aligned inclusions (molecules or particles). Similar results may be obtained with *artificial molecules* or *metamolecules*, which are the basic constituents or unit cells of metamaterials. A strong anisotropy can be engineered in artificial materials with inhomogeneous distributions of inclusions in one or two dimensions. In Figure 7.7 we illustrate a multilayer (on the left) and a wire medium (on the right). Both structures have two phases with inclusions of permittivities ε_i immersed in a host medium with permittivity ε_h. The inclusion filling factor is f. The homogenization of these structures in the quasi-static approximation is reported here.
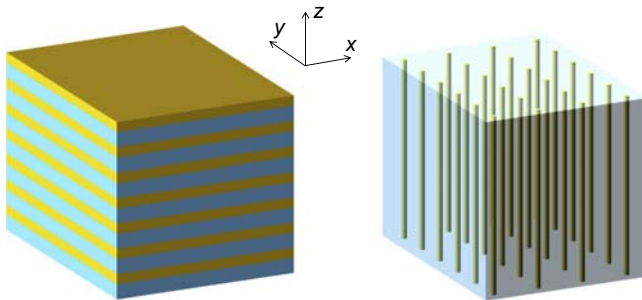
Figure 7.7 On the left is a one-dimensional multilayer composite medium with thin layer inclusions (yellow regions) in a host medium (blue regions). On the right is a wire medium with cylindrical inclusions (yellow regions) dispersed in a host medium (blue region).

We start with the homogenization of the multilayer. This material shows two different responses for waves polarized parallel or perpendicular to the plane of the layers, so that it is appropriate to assume an anisotropic uniaxial response with effective permittivity $\overline{\varepsilon} = \varepsilon_\parallel(\widehat{\mathbf{x}}\widehat{\mathbf{x}} + \widehat{\mathbf{y}}\widehat{\mathbf{y}}) + \varepsilon_\perp \widehat{\mathbf{z}}\widehat{\mathbf{z}}$, where $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$, and $\widehat{\mathbf{z}}$ are the unit vectors along the principal axes, as shown in Figure 7.7. In order to find $\varepsilon_\parallel$, one should consider a wave with the electric field polarized in the parallel direction, i.e., in the plane x-y. For this kind of wave, the electric field is tangential to the interfaces, hence it must be continuous so that the average static electric field in the multilayer can be written as

$$\left\langle \mathbf{E}_\parallel \right\rangle = \mathbf{E}_{\parallel,i} = \mathbf{E}_{\parallel,h}, \tag{7.38}$$

where $\mathbf{E}_{\parallel,i}$ and $\mathbf{E}_{\parallel,h}$ are the static electric fields in the inclusion and host layers. From Eqn (7.38) and the definition of the average displacement field,

$$\left\langle \mathbf{D}_\parallel \right\rangle = \varepsilon_0 \varepsilon_\parallel \left\langle \mathbf{E}_\parallel \right\rangle = \varepsilon_0 f \varepsilon_i \mathbf{E}_{\parallel,i} + \varepsilon_0 (1-f)\varepsilon_h \mathbf{E}_{\parallel,h} \tag{7.39}$$

directly follows the expression of the parallel permittivity:

$$\varepsilon_\parallel = f\varepsilon_i + (1-f)\varepsilon_h. \tag{7.40}$$

For waves polarized in the direction perpendicular to the layers, the boundary condition at the interfaces requires continuity of the displacement field, therefore the average displacement field is simply

$$\langle \mathbf{D}_\perp \rangle = \varepsilon_0 \varepsilon_\perp \langle \mathbf{E}_\perp \rangle = \mathbf{D}_{\perp,i} = \mathbf{D}_{\perp,h}, \tag{7.41}$$

where $\mathbf{D}_{\perp,i}$ and $\mathbf{D}_{\perp,h}$ are the static displacement fields in the inclusion and host layers, respectively. From the definition of the average electric field, $\langle \mathbf{E}_\perp \rangle = f\mathbf{E}_{\perp,i} + (1-f)\mathbf{E}_{\perp,h}$, where $\mathbf{E}_{\perp,i} = \frac{\mathbf{D}_{\perp,i}}{\varepsilon_0 \varepsilon_i}$ and $\mathbf{E}_{\perp,h} = \frac{\mathbf{D}_{\perp,h}}{\varepsilon_0 \varepsilon_h}$ are the static electric fields in the inclusion and host layers, in that order, we can derive the expression of the effective permittivity $\varepsilon_\perp$, which reads as follows:

$$\varepsilon_\perp = \frac{\varepsilon_i \varepsilon_h}{f\varepsilon_h + (1-f)\varepsilon_i}. \tag{7.42}$$

The permittivity tensor of the multilayer can be also retrieved by applying the Maxwell Garnett (Eqn (7.25)) or the Bruggeman (Eqn (7.24)) effective medium theories for ellipsoids, obtaining the same results as Eqns (7.40) and (7.42). The depolarization factors in the case of multilayers are $L_x = L_y = 0$ and $L_z = 1$, equivalent to virtual disk-like particles with eccentricity $e \to \infty$. The corresponding screening parameters are $K_x = K_y \to \infty$ and $K_z \to 0$.

The static homogenization of the wire medium follows similar criteria. The material shows two different responses for waves polarized parallel or perpendicular to the wires, so that it is appropriate to assume an anisotropic uniaxial response with effective permittivity $\overline{\varepsilon} = \varepsilon_\perp(\widehat{\mathbf{x}}\widehat{\mathbf{x}} + \widehat{\mathbf{y}}\widehat{\mathbf{y}}) + \varepsilon_\parallel \widehat{\mathbf{z}}\widehat{\mathbf{z}}$. Note that in this case the parallel component of the effective permittivity is related to waves polarized along the z direction, i.e., parallel to the wires' long axis. For these waves the electric field is continuous and the average displacement field is as in Eqn (7.39). It follows that $\varepsilon_\parallel = f\varepsilon_i + (1-f)\varepsilon_h$. This result may also be derived from the Maxwell Garnett formula (with $L_z = 0$). For the permittivity component $\varepsilon_\perp$, associated with waves polarized in the plane perpendicular to the wires, the most natural homogenization strategy is again the Maxwell Garnett formula. Indeed the cross-section view of the wire medium is very similar to the typical Maxwell Garnett mixture (see Figure 7.1), in which small spherical inclusions are dispersed in a homogeneous host. The only difference with respect to three-dimensional mixtures is that the depolarization factor for cylinders is with $L_x = L_y = 1/2$. The resulting expression for the quasi-static permittivity is obtained by applying Eqn (7.17), and it reads as follows:

$$\varepsilon_\perp = \varepsilon_h\left(1 + 2f\frac{\varepsilon_i - \varepsilon_h}{\varepsilon_i + \varepsilon_h - f(\varepsilon_i - \varepsilon_h)}\right). \tag{7.43}$$

7.7 Spatial dispersion effects

The recipes to obtain the effective permittivities reported in the previous paragraphs are strictly valid for nonmagnetic mixtures excited in the long-wavelength limit, where the quasi-static approach is a justified approximation. It should be mentioned that the extension of these techniques to mixtures with intrinsically magnetic materials is straightforward and it is similarly valid within the limits of the magneto-static approximation. In engineered materials such as metamaterials, the miniaturization of the period and inclusion sizes are, however, limited by fabrication technologies. Moreover, if the permittivity contrast between the inclusion(s) and the host media is large, these approximations may not be adequate even for nanometer-sized layers or wires and extremely subwavelength periodicities. Typical examples of this scenario are metal-dielectric mixtures or metamaterials with high (positive) permittivity inclusions. In these structures both *effective* and *intrinsic* nonlocal responses may occur. The *intrinsic* nonlocal behavior of metal-dielectric nanostructures is related to the hydrodynamic gas pressure acting on free electrons that adds a spatially dispersive term to the classical Drude model of a free-electron gas. These effects are relevant for very small metallic inclusions (nanoparticles, nanolayers, or nanowires) or for structures with subnanometer interparticle distances (nanogaps). At this scale, quantum size and quantum tunneling effects are similarly important and may induce significant deviations from purely classical approaches. The details on the intrinsic nonlocal contribution and the implications on the optical response of metallic nanostructures will be discussed more extensively in Chapter 8.

The *effective* nonlocal response of structured materials is a stronger effect due to retardation and becomes relevant when the characteristic length scale of the inhomogeneity is smaller than the effective wavelength but much larger than the crystal structure of the constituent materials. This is the typical regime in which metamaterials operate, especially those designed for optical frequencies. Simple corrections to the electrostatic approaches described above are available only for *very weak nonlocalities* and for very specific combinations of inclusion and material structures. For example, it is possible to add retardation effects in the electrostatic formulas for polarizabilities and effective permittivities. For a sphere, a dynamic depolarization factor can be introduced, and it may be written as follows:

$$L_d = L_0 - \frac{1}{3}k^2R^2 - i\frac{2}{9}k^3R^3 \tag{7.44}$$

where k is the host medium wavenumber and R is the sphere radius. The first term on the right-hand side of Eqn (7.44), equal to 1/3 for spheres, represents the static, Lorentz depolarization factor; the second term accounts for the dynamic depolarization due to retardation; and the third term is due to radiation damping.

The nonlocal effects in the bulk response of metal-dielectric multilayers are usually treated by using analytical approaches based on the Bloch theory and the transfer matrix method. Unfortunately, adapting these techniques to two-dimensional or three-dimensional nonlocal metamaterials with complicated inclusion shapes is not an easy task. For this reason, more sophisticated averaging and homogenization techniques are required for *mesoscopic* systems in which the effective response is nonlocal. An example of *highly nonlocal* artificial materials is a photonic crystal, i.e., a periodic structure with inhomogeneities on the wavelength scale. The bulk response of photonic crystals can be described by a multivalued dispersion function of type $\omega(\mathbf{k}_B)$, in which the frequency ω is related to the Bloch wavevector $\mathbf{k}_B$. A band structure is typically formed in these systems. Once the dispersion relation is known, it is possible to introduce an effective index based on the definition of phase velocity, the measurement, or on the numerical evaluation of reflection and transmission coefficients. Although useful to predict specific effects such as light refraction, these definitions are intrinsically ambiguous since they are multivalued and depend on propagation direction and polarization.

In nonlocal metamaterials, the nonlocality is closely related to the multipolar nature of the inclusions and to the fact that the periodicity size may be close to the effective wavelength. In these systems, the response shows some degree of *spatial dispersion*, i.e., the response depends not only on the field but also on its spatial derivatives. To understand the effects associated with a nonlocal response, one may first adopt a *phenomenological* approach and write the macroscopic-induced current density in the bulk of a nonlocal medium (here assumed infinitely extended) as

$$\mathbf{J}(\mathbf{r}, \omega) = \int \mathbf{R}(\mathbf{r} - r', \omega)\mathbf{E}(r', \omega)dV'. \tag{7.45}$$

The integration is performed in a spherical volume around the observation point $\mathbf{r}$, whose radius r_0 depends on the strength of the nonlocality in the medium under investigation. In natural dielectrics r_0 is on the scale of the atomic size, hence orders of magnitude smaller than the wavelength. In these materials, an impulsive response function of type $\mathbf{R}(\mathbf{r} - r', \omega) = \mathbf{R}_0 \delta(\mathbf{r} - r')$ may be assumed so that the resulting optical response is usually modeled as purely local. However, the discrete nature of metamaterials, whose "atomic" size is smaller but on the same order of magnitude of the wavelength, tends to increase the radius r_0 over which the function $\mathbf{R}$ is nonzero. Modeling the metamaterial with the nonlocal relation Eqn (7.45) is formally correct but impractical. A transformation of Eqn (7.45) in the Fourier space

$$\mathbf{J}(\mathbf{k}, \omega) = \mathbf{R}(\mathbf{k}, \omega)\mathbf{E}(\mathbf{k}, \omega) \tag{7.46}$$

leads to a simplified picture of the problem that is now put in terms of the wavevector $\mathbf{k}$. Assuming that the metamaterial operates in a regime of *weak spatial dispersion*, the response function $\mathbf{R}$ can be expanded in a Taylor series around the point $\mathbf{k} = 0$, and an approximated expression of the response can be retrieved by truncating the series. Retaining derivatives of the electric field $\mathbf{E}$ up to the second order, the induced current is approximated by

$$J_i(\mathbf{k}, \omega) \approx -i\omega[a_{ij} + b_{ijl}k_l + c_{ijlm}k_l k_m]E_j(\mathbf{k}, \omega), \tag{7.47}$$

where the tensors a_{ij}, b_{ijl}, and c_{ijlm} depend on the derivatives of the response function $\mathbf{R}(\mathbf{k}, \omega)$ evaluated at $\mathbf{k} = 0$. Antitransformation of Eqn (7.47) into the real space provides an expression for the displacement field, defined as $\mathbf{D}_i(\mathbf{r}, \omega) = \varepsilon_0 \delta_{ij} \mathbf{E}_j(\mathbf{r}, \omega) + \frac{i}{\omega} J_i(\mathbf{r}, \omega)$, as a function of spatial derivatives of the electric field. The expression of $\mathbf{D}$ in index notation reads as follows:

$$\mathbf{D}_i(\mathbf{r}, \omega) = \varepsilon_0 [\delta_{ij} + a_{ij} + b_{ijl}\partial_l + c_{ijlm}\partial_l \partial_m]E_j(\mathbf{r}, \omega). \tag{7.48}$$

Equation (7.48) together with the definition of the magnetic field as

$$\mathbf{H}_i(\mathbf{r}, \omega) = \mu_0^{-1}\mathbf{B}_i(\mathbf{r}, \omega), \tag{7.49}$$

with $\mathbf{B}$ indicating the induction field, describes the bulk response of a metamaterial with *weak spatial dispersion* up to the second-order field derivative. The main limitation of this model is that it is strictly valid for an infinitely extended medium, i.e., far from interfaces. In fact, the usual boundary conditions that require the continuity of the tangential electric and magnetic fields cannot be simply applied if the displacement field of one of the media composing the interface depends on the derivatives of the electric field, as in Eqn (7.48). Additional boundary conditions or the introduction of transition layers have been proposed to circumvent this problem.

Another approach to treat spatial dispersion relies on the *multipole theory* for macroscopic media. The metamaterial is modeled as a periodic arrangement of

multipoles susceptible to an applied electromagnetic field. Considering the magnetic dipole and the electric quadrupole contributions in addition to the electric dipole one, the induced current can be approximated by

$$\mathbf{J} \approx -i\omega\mathbf{P} + \frac{i\omega}{2}\nabla\cdot\mathbf{Q} + \nabla\times\mathbf{M}, \tag{7.50}$$

where $\mathbf{P}$ is the electric dipole polarization, $\mathbf{Q}$ is the electric quadrupole polarization tensor, and $\mathbf{M}$ is the magnetic dipole polarization vector. The electric part of the induced current is $\mathbf{J}_e = -i\omega\mathbf{P} + \frac{i\omega}{2}\nabla\cdot\mathbf{Q}$, whereas the magnetic part is $\mathbf{J}_m = \nabla\times\mathbf{M}$. The usual definition for the displacement field and the magnetic field reads as follows:

$$\begin{aligned} \mathbf{D} &= \varepsilon_0\mathbf{E} + \mathbf{P} - \frac{1}{2}\nabla\cdot\mathbf{Q} \\ \mathbf{H} &= \mu_0^{-1}\mathbf{B} + \mathbf{M} \end{aligned} \tag{7.51}$$

It is reasonable to assume that the electric field driving the multipoles is slowly varying in the unit cell; therefore, in analogy to Eqn (7.48), the multipole polarizations can be expanded as follows:

$$\begin{aligned} P_i &= \left[\alpha_{ij} + \beta_{ijl}\partial_l + \gamma_{ijlm}\partial_l\partial_m + \ldots\right]E_j \\ Q_{ij} &= \left[\alpha'_{ijk} + \beta'_{ijkl}\partial_l + \ldots\right]E_k \\ M_i &= \left[\alpha''_{ijk} + \beta''_{ijl}\partial_l + \gamma''_{ijlm}\partial_l\partial_m + \ldots\right]E_j \end{aligned} \tag{7.52}$$

Substitution of the multipole polarizations in Eqn (7.52) in Eqn (7.51) provides the expressions of the fields $\mathbf{D}$ and $\mathbf{H}$ as functions of the electric field and its first and second derivatives. It is interesting to note that the multipole theory approach leads to a formulation similar to that of the phenomenological approach (Eqns (7.48) and (7.49)), in which the main effect is spatial dispersion in the macroscopic response.

We now neglect second-order derivatives of the field in the definitions of $\mathbf{D}$ and $\mathbf{H}$ in Eqn (7.50), and we focus on the physical meaning of the first-order spatial dispersion effects. Using Serdyukov−Fedorov transformations, i.e., the validity of Maxwell's equations is preserved after the redefinition of the fields $\mathbf{D}$ and $\mathbf{H}$ as $D' = \mathbf{D} + \nabla\times\mathbf{T}$ and $H' = \mathbf{H} - i\omega\mathbf{T}$, it is possible, with the proper choice of the vector field $\mathbf{T}$, to rewrite the relations Eqn (7.51) in the form of bianisotropic constitutive relations:

$$\begin{aligned} \mathbf{D} &= \bar{\bar{\varepsilon}}\cdot\mathbf{E} + i\bar{\bar{\xi}}\cdot\mathbf{B} \\ \mathbf{H} &= \mu_0^{-1}\mathbf{B} + i\bar{\bar{\xi}}^T\cdot\mathbf{E} \end{aligned} \tag{7.53}$$

where $\bar{\bar{\xi}}$ is the magneto-electric coupling tensor. The effective medium described by Eqn (7.53) is (1) electro-dipole susceptible to the average electric field through the tensor $\bar{\bar{\varepsilon}}$, (2) electro-magneto susceptible to the vortex part of the electric field, that is the field $\mathbf{B}$, and (3) quadrupole susceptible to the average electric field through the parameter $\bar{\bar{\xi}}$. The dependence of $\mathbf{H}$ from the electric field $\mathbf{E}$ reflects the reciprocity of the medium. Moreover, the application of usual boundary conditions is natural with the constitutive relations in Eqn (7.53), hence they allow the integration of Maxwell's equations in the presence of interfaces without resorting to additional boundary conditions or the introduction of artificial transition layers. For isotropic media the response becomes chiral and the scalar ξ is known as a chirality parameter. In this case the material relations are

$$\mathbf{D} = \varepsilon\mathbf{E} + i\xi\mathbf{B}$$
$$\mathbf{H} = \mu_0^{-1}\mathbf{B} + i\xi\mathbf{E} \tag{7.54}$$

We now examine the effects of second-order derivatives of the field in the definitions Eqn (7.51). For simplicity we start from the isotropic relations in Eqn (7.53) and add the terms of type $\partial_l\partial_m\mathbf{E}_j$. It is possible to show that the only nonzero second-order differential operators are $\nabla\nabla\cdot$ and $\nabla\times\nabla\times$ in isotropic systems, therefore Eqn (7.53) changes to

$$\mathbf{D} = \varepsilon\mathbf{E} + i\xi\mathbf{B} + \beta\nabla\nabla\cdot\mathbf{E} + \gamma\nabla\times\nabla\times\mathbf{E}$$
$$\mathbf{H} = \mu_0^{-1}\mathbf{B} + i\xi\mathbf{E} \tag{7.55}$$

The transformation $D' = \mathbf{D} + \nabla\times\mathbf{T}$ and $H' = \mathbf{H} - i\omega\mathbf{T}$ with $\mathbf{T} = \gamma\nabla\times\mathbf{E}$ leads to the form

$$D' = \varepsilon\mathbf{E} + i\xi\mathbf{B} + \beta\nabla\nabla\cdot\mathbf{E}$$
$$H' = \mu^{-1}\mathbf{B} + i\xi\mathbf{E} \tag{7.56}$$

where the (artificial) magnetic permeability is $\mu^{-1} = \mu_0^{-1} - \omega^2\gamma$.

Some observations on *weak spatial dispersion* effects must be pointed out after an inspection of Eqn (7.56): (1) the electric susceptibility to the uniform (quasi-static) part of the electric field is due to zero-order spatial dispersion effects though the effective parameter ε; (2) first-order spatial dispersion produces bianisotropic effects, e.g., chirality, through the parameter ξ; (3) second-order spatial dispersion leads to artificial magnetism through the effective permeability μ and to the term $\beta\nabla\nabla\cdot\mathbf{E}$; and (4) when higher-order multipole can be neglected and $\beta = 0$, weak spatial dispersion effects in metamaterials can be treated with a local, macroscopic model in which the artificial bianisotropy and magnetism take into account the medium susceptibility to first- and second-order field derivatives. Artificial magnetism in materials can be obtained with particle shapes that produce uniform current loops, such as split-ring resonators, or with magnetic Mie-type resonances in high-permittivity particles.

Problems

1. Calculate the effective permittivity of a mixture made of small silver particles in air (assume a frequency of 1 GHz, silver conductivity 6.3×10^7 S/m. Plot the real and imaginary parts of the effective permittivity as functions of the silver filling factor f.

2. Draw, as a function of the inclusion's fill factor, the relative effective permittivity of a mixture of spherical inclusions with dielectric constant $\varepsilon = 10$, immersed in air.

3. Repeat the calculation of Problem 2 for randomly oriented, needle-like inclusions (ellipsoids with one semiaxis much larger than the other two) and randomly oriented, disc-like inclusions (ellipsoids with one semiaxes much smaller than the other two).

4. Consider a metal-dielectric, planar multilayer in the electrostatic approximation with fill factor $f = 0.5$. Assume for metal a complex, frequency-dependent dielectric constant $\varepsilon = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma}$, with $\omega_p = 2\pi\, 2.18 \times 10^{15}$ Hz and $\gamma = 2\pi\, 4.35 \times 10^{12}$ Hz, and for the dielectric a dispersion-free relative permittivity of 2.25. Determine the wavelength ranges in which the dispersion of the mixture for TM polarized fields is *hyperbolic*, i.e., when Real$[\varepsilon_\parallel] \times$ Real$[\varepsilon_\perp]$. How does the metal fill factor move the hyperbolic ranges in the wavelength domain?

Further reading

[1] A.H. Sihvola, Electromagnetic mixing formulas and applications, in: IEE Electromagnetic Waves Series, 47, 1999.

[2] G.A. Niklasson, C.G. Granqvist, O. Hunderi, Effective medium models for the optical properties of inhomogeneous materials, Appl. Opt. 20 (1981) 26.

[3] G. Bánhegyi, Comparison of electrical mixture rules for composites, Colloid Polym. Sci. 264 (1986) 1030.

[4] W. Cai, V. Shalaev, Optical Metamaterials: Fundamentals and Applications, Springer, 2010.

[5] D. Polder, J.H. van Santen, The effective permeability of mixtures of solids, Physica XII (1946) 257.

[6] Y.A. Urzhumov, G. Shvets, Quasistatic effective medium theory of plasmonic nanostructures, Proc. SPIE 6642 (2007) 66420X-01.

[7] G.W. Milton, The Theory of Composites, Cambridge University Press, 2002.

[8] M.I. Stockman, S.V. Faleev, D.J. Bergman, Localization versus delocalization of surface plasmons in nanosystems: can one state have both characteristics? Phys. Rev. Lett. 87 (2001) 167401.

[9] A.M. Nicolson, G.F. Ross, Measurement of the intrinsic properties of materials by time-domain techniques, IEEE Trans. Instrum. Meas. 19 (1970) 377.

[10] W.B. Weir, Automatic measurement of complex dielectric constant and permeability at microwave frequencies, Proc. IEEE 62 (1974) 33.

[11] S. Arslanagic, T.V. Hansen, N.A. Mortensen, A.H. Gregersen, O. Sigmund, R.W. Ziolkowski, O. Breinbjerg, A review of the scattering-parameter extraction method with clarification of ambiguity issues in relation to metamaterial homogenization, IEEE Antennas and Propag. Mag. 55 (2013) 91.

[12] F. Capolino, Theory and Phenomena of Metamaterials, Taylor and Francis, 2009.

Plasmonics

<div style="text-align:right">**8**</div>

D. de Ceglia, M.A. Vincenti
National Research Council, AMRDEC, Redstone Arsenal, AL, USA

8.1   Introduction

The term plasmonics refers to a major topical subject in nanophotonics that is devoted to the study of optical phenomena resulting from the interaction of electromagnetic fields with conduction electrons in metals. Two outcomes of electron−photon interactions at metal-insulator interfaces or in metallic nanostructures are subdiffraction field *confinement* and field *enhancement*. Plasmonics simulations are usually performed with well-established computational tools that incorporate Maxwell's equations with the conventional boundary conditions and a classical description of the optical properties of metals. The renewed and ever-increasing interest in plasmonic research is motivated by the rapid advancements in nanofabrication technologies and characterization tools, easier access to powerful and fast computational resources, and progress in related research fields, e.g., metamaterials science. The existing and potential applications of plasmonic-based nanostructures are innumerable and span a wide range of fields, including energy harvesting, sensing devices, electro-optical devices, biomolecule detection, drug delivery, and many more.

This chapter is structured as follows. In the first section, the optical properties of metals are explained in terms of a simple free-electron gas model. In the second section, the dispersion properties of surface plasmon polaritons are discussed for a metal-insulator interface and for a metallic film. We then introduce the concept of localized surface plasmon for a simple metallic nanoparticle, e.g., a metallic nanosphere. Next, the fundamentals of Mie theory are presented, and the concepts of scattering and absorption cross-sections are introduced for nanospheres and nanoshells.

8.2   Optical properties of metals

The simplest description of electromagnetic field interaction with metals is the free-electron gas model, the so-called Drude model. The classical Lorentz model was introduced in Chapter 2, and the Drude model is closely related to it. In this section we will treat free electrons without the hydrodynamic contributions where the electrons are treated as a fluid. The metal is assumed to be a gas of free electrons with density n per unit volume moving in a fixed lattice of positive ions. In the presence of an external electromagnetic field, the equation of motion for an electron is given by

$$m\dot{\mathbf{v}} + m\gamma\mathbf{v} \, = \, -\mathrm{e}(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{8.1}$$

where m, e, and $\mathbf{v}$ are the electron mass, charge, and velocity, respectively. The electron position $\mathbf{r}$ and velocity $\mathbf{v}$ are related by $\mathbf{v} = \dot{\mathbf{r}}$, and the time dependence of $\mathbf{v}$, $\mathbf{E}$, and $\mathbf{B}$ is implicit in the equation. In this model, the electron motion is damped by collisions with ions, whose frequency γ is on the order of 100 THz for noble metals. Electron−electron interactions and the effects of crystal lattice potential are neglected. The first term on the right-hand side of Eqn (8.1) is the Coulomb interaction between the electron charge and the electric field $\mathbf{E}$, while the second term is the Lorentz force due to the presence of the magnetic flux density $\mathbf{B}$. The relation between the microscopic dynamics described by Eqn (8.1) and the macroscopic response of the medium is given by introducing the polarization density $\mathbf{P} = -e\mathbf{r}n$. The macroscopic version of Eqn (8.1) can be written as

$$\ddot{\mathbf{P}} + \gamma\dot{\mathbf{P}} = \varepsilon_0\omega_{\mathrm{p}}^2(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{8.2}$$

where $\omega_{\mathrm{p}} = \left(\frac{ne^2}{\varepsilon_0 m}\right)^{1/2}$ is the plasma frequency of the electron gas. It is important to note that the mass m in the free-electron model is an optical effective mass. Such mass, which mainly depends on the band structure of the medium, may significantly differ from the electron rest mass. If we consider a time-harmonic excitation with time dependence $\mathrm{e}^{-i\omega t}$ and neglect the magnetic force $-e\mathbf{v} \times \mathbf{B}$, we obtain the frequency-domain expression (equal to Eqn (2.24)) of the polarization density

$$\mathbf{P}(\omega) = -\frac{\varepsilon_0\omega_{\mathrm{p}}^2}{\omega^2 + i\gamma\omega}\mathbf{E}(\omega), \tag{8.3}$$

The contribution of the magnetic force to the polarization density should not be neglected when a static magnetic field is applied or when one is interested in the nonlinear response for frequency mixing problems, e.g., second-harmonic generation. From the resulting constitutive relation between the displacement field $\mathbf{D}$ and the electric field,

$$\mathbf{D}(\omega) = \varepsilon_0\mathbf{E}(\omega) + \mathbf{P}(\omega) = \varepsilon_0\varepsilon_{\mathrm{f}}(\omega)\mathbf{E}(\omega) = \varepsilon_0\left(1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + i\gamma\omega}\right)\mathbf{E}(\omega) \tag{8.4}$$

it is clear that the plasma dielectric function is negative below the frequency $(\omega_{\mathrm{p}}^2 - \gamma^2)^{1/2}$. For media with damping coefficient γ much smaller than the plasma frequency, ω_{p} is approximately the zero-crossing frequency for the dielectric function $\varepsilon_{\mathrm{f}}(\omega)$. The expression of $\varepsilon_{\mathrm{f}}(\omega)$ resulting from Eqn (8.4) is known as the Drude dispersion model for an ideal gas of free-electrons. This model is a good approximation of the optical response of noble metals only below the threshold frequency for the onset of interband transitions. Above this threshold, which may occur at visible or near-infrared wavelengths for noble metals, the free-electron model in Eqn (8.4)

becomes inadequate, and the band structure of the material must be taken into account. The complex dielectric permittivity resulting from Eqn (8.4) is $\varepsilon_{\mathrm{f}} = \varepsilon_{\mathrm{f}}' + \mathrm{i}\varepsilon_{\mathrm{f}}'' = \left(1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \gamma^2}\right) + \mathrm{i}\left(\frac{\omega_{\mathrm{p}}^2\gamma}{\omega^3 + \omega\gamma^2}\right)$. From the expression of the induced current density $\mathbf{J}_{\mathrm{f}}(\omega) = -\mathrm{i}\omega\mathbf{P}(\omega) = \sigma(\omega)\mathbf{E}(\omega)$, the dynamic conductivity can be extracted and written as

$$\sigma(\omega) = \frac{\varepsilon_0 \omega_{\mathrm{p}}^2}{-\mathrm{i}\omega + \gamma}. \tag{8.5}$$

In the very low frequency regime, i.e., $\omega \ll \gamma$, the conductivity becomes purely real, and the electrons follow the electric field without a significant phase shift. In this regime the static conductivity $\sigma_0 = \varepsilon_0 \omega_{\mathrm{p}}^2/\gamma$ dominates the metal response to an applied field. At optical frequencies, the usual approach to include the effects of interband transitions consists in adding a number N of Lorentz oscillators (see Eqn (2.23) for bound electrons) to the free-electron response, so that the metal complex permittivity now reads as follows:

$$\varepsilon_m(\omega) = \varepsilon_{\mathrm{f}}(\omega) + \varepsilon_{\mathrm{b}}(\omega) = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \mathrm{i}\gamma\omega} - \sum_{j=1}^{N} \frac{f_j \omega_{\mathrm{p}}^2}{\left(\omega^2 - \omega_j^2\right) + \mathrm{i}\omega\gamma_j}, \tag{8.6}$$

where ω_j is the resonance frequency of the j-th oscillator, f_j its strength, and γ_j its damping coefficient. Equation (8.6) defines the so-called Drude−Lorentz dispersion model. All the parameters appearing in this model may be found by fitting the experimental values of the complex dielectric constants of metals available in the literature. A simplified version of the Drude−Lorentz model is obtained by considering $\varepsilon_{\mathrm{b}}(\omega)$ constant and by writing $\varepsilon_{\mathrm{m}}(\omega) = \varepsilon_{\infty} - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \mathrm{i}\gamma\omega}$, where the dielectric constant $\varepsilon_{\infty} = \varepsilon_{\mathrm{m}}(\omega \gg \omega_{\mathrm{p}})$ for very large frequencies takes into account the residual response of the positive background of the ion cores. Although this correction improves the free-electron Drude model [Eqn (8.4)], it does not provide an accurate description of the optical response in the frequency region where interband transitions occur. All the dispersion models introduced above may be implemented straightforwardly in finite-difference time-domain and pseudo-spectral time-domain propagators by introducing an additional differential equation for each oscillator appearing in Eqn (8.6). Frequency-domain solvers, e.g., the finite-element frequency-domain method and the finite-difference frequency-domain method, as well as semianalytical tools, such as the rigorous coupled-wave analysis, handle naturally the complex and frequency-dependent dielectric constants that are measured from experimental characterization tools (such as ellipsometry or thin film, normal and oblique incidence reflection, and transmission experiments).

8.3 Surface plasmon polaritons at metal−dielectric interfaces

We consider a smooth and flat interface between a metal with relative permittivity ε_m and a lossless dielectric with relative permittivity ε_d. The boundary between the two media is on the plane $z = 0$, so that $\varepsilon(z) = \varepsilon_m$ for $z < 0$ and $\varepsilon(z) = \varepsilon_d$ for $z > 0$. We are interested in time-harmonic solutions of Maxwell's equations corresponding to waves propagating in the x direction with a generically complex k_x wavenumber, i.e., the x-dependence of the solution is of the form $e^{-ik_x x}$ and $k_y = 0$. For this problem, the solutions can be either s-polarized or p-polarized. The nonvanishing fields of s-polarized modes, also known as TE modes, are (E_y, H_x, H_z), while the significant fields for p-polarized modes (TM) are (H_y, E_x, E_z). The dispersion relations $k_x(\omega)$ of modes supported by the interface may be found by applying the continuity of the tangential fields (see Chapter 2) at the interface $z = 0$ or by using the transverse resonance technique based on the transmission line equivalent model of the structure in the z-direction. The dispersion relations for s- and p-polarizations are

$$s: \frac{1}{k_{z,m}} + \frac{1}{k_{z,d}} = 0 \tag{8.7a}$$

$$p: \frac{k_{z,m}}{\varepsilon_m} + \frac{k_{z,d}}{\varepsilon_d} = 0 \tag{8.7b}$$

where $k_{z,m/d} = \sqrt{k_0^2 \varepsilon_{m/d} - k_x^2}$ is the z-component of the mode wavevector and $k_0 = \omega/c$ is the vacuum wavenumber. If we are interested in evanescent surface waves, i.e., modes with a field confined at the metal−dielectric interface and with evanescent decay at $|z| \to \infty$, then our solutions must satisfy the requirements $\mathrm{Re}(k_{z,m}) > 0$ and $\mathrm{Re}(k_{z,d}) > 0$. This means that s-polarized solutions, which follow the dispersion relation $k_{z,m} + k_{z,d} = 0$ [see Eqn (8.7a)], are not allowed. The dispersion relation for p-polarization can be derived by Eqn (8.7b) and reads as

$$k_x = k_0 \sqrt{\frac{\varepsilon_m \varepsilon_d}{\varepsilon_m + \varepsilon_d}}. \tag{8.8}$$

In Figure 8.1 the dispersion relation for p-polarization is plotted for two cases: an interface between air ($\varepsilon_d = 1$) and a loss-less, free-electron plasma ($\gamma = 0$ and $\varepsilon_m = 1 - \omega_p^2/\omega^2$) and an interface between air and a lossy plasma ($\gamma = 0.1\,\omega_p$ and $\varepsilon_m = \varepsilon_f$). In the loss-less case, the condition $\varepsilon_m = -\varepsilon_d$ represents a pole for k_x and defines the so-called surface plasmon resonance. This resonance occurs at a frequency $\omega = \omega_{sp} = \omega_p/\sqrt{1 + \varepsilon_d}$. The other relevant frequency is $\omega = \omega_p$ above which $\varepsilon_m > 0$. Three different regimes for p-polarized modes can be identified in the loss-less scenario: (i) for $\omega < \omega_{sp}$, i.e., $\varepsilon_m < -\varepsilon_d$, the mode is an evanescent surface wave, i.e., a surface plasmon polariton (SPP), bound at the interface (since $k_{z,m/d}$ are purely imaginary) and propagates in the x-direction with a real k_x; (ii) in the region
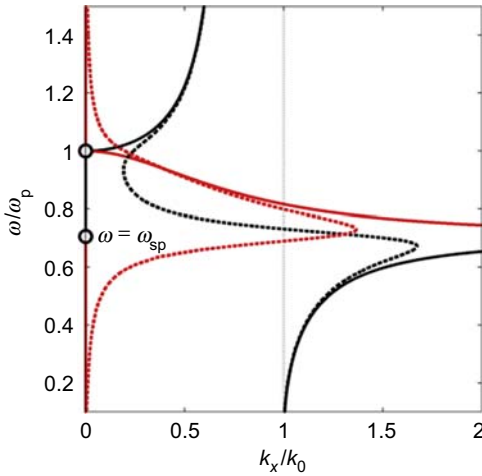
Figure 8.1 Dispersion of loss-less (solid lines) and lossy (dashed lines) surface plasmons propagating on a flat metal surface. Real (black lines) and imaginary (red lines) parts of k_x are normalized to the wavenumber in air k_0. The light line, $k_x = k_0$, is plotted in gray.

$\omega_p > \omega > \omega_{sp}$, the mode has a real $k_{z,m/d}$ and a purely imaginary k_x, hence propagation is not allowed in the x-direction; and (iii) for $\omega > \omega_p$, the mode becomes radiative since both k_x and $k_{z,m/d}$ are real. If the damping γ is present (dashed curves in Figure 8.1), k_x and $k_{z,m/d}$ become complex quantities and the dispersion relation remains fairly similar to the loss-less case only for very small or very large frequencies ω. Losses limit the propagation length of the surface plasmon polariton, $L_p = [2\mathrm{Im}(k_x)]^{-1}$, allow quasi-bound or leaky modes in the region $\omega_p > \omega > \omega_{sp}$, and limit $\mathrm{Re}(k_x)$ at the surface plasmon resonance ω_{sp}. It is important to notice that interfaces with dielectrics with larger ε_d show lower values of surface plasmon resonance and support surface plasmons with larger effective refractive index $n_{spp} = \mathrm{Re}(k_x)/k_0$. A measure of the field confinement at the interface is given by $[\mathrm{Im}(k_{z,m/d})]^{-1}$. For a surface plasmon of wavelength 600 nm supported at a silver−air interface, the evanescent tail of the electromagnetic field extends for ∼390 nm on the air side and ∼24 nm on the silver side.

8.4 Surface plasmon polaritons of metallic thin films

We now consider the p-polarized modes of a multilayer formed by a thin, metallic film with relative permittivity ε_m and thickness d surrounded by a dielectric medium with relative permittivity ε_d. The structure can be thought as a system of coupled wave-guides. Each isolated interface is a waveguide that supports the propagation of a surface-plasmon polariton with dispersion as in Eqn (8.8).

When the distance d between the two interfaces (i.e., the metal film thickness) is of the order of the field penetration depth in the metal, the two surface waves are coupled through their evanescent tails. Because of the symmetry of the structure, the modes supported by the film can be either symmetric or antisymmetric with respect to the center of the film (plane $z = 0$). The distribution of the electric field component normal to

Figure 8.2 Electric field distribution for the two nonradiative modes of a thin metal film.
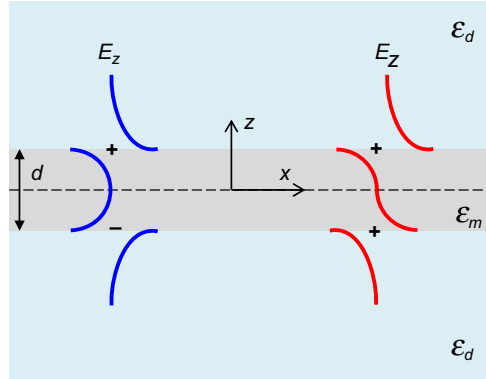
the metal−dielectric interfaces (E_z) is reported in Figure 8.2 for the two p-polarized, nonradiative modes of the metal film. The blue curve refers to the symmetric mode and the red curve to the antisymmetric one. Because of the field distribution, charges of opposite (same) sign face each other in the symmetric (antisymmetric) mode. The dispersion relations for the symmetric and the antisymmetric modes can be found in the usual way by setting the continuity of the tangential fields, and read as follows:

$$S: \ \varepsilon_m k_{z,d} + \varepsilon_d k_{z,m} \ \tanh\left(-\frac{ik_{z,m}d}{2}\right) = 0 \tag{8.9}$$

$$A: \ \varepsilon_m k_{z,d} + \varepsilon_d k_{z,m} \ \coth\left(-\frac{ik_{z,m}d}{2}\right) = 0. \tag{8.10}$$

The solutions in the complex k_x plane are usually found numerically by minimizing the complex functions $|S|$ and $|A|$. It is interesting to notice that for $d \rightarrow \infty$, i.e., when the film is very thick and two interfaces can be assumed as isolated, the dispersion relations of the two modes are equal and tend to the dispersion of the metal−dielectric surface-plasmon polariton, whose solution is given in Eqn (8.8). Nonradiative solutions for s-polarization are not allowed in any planar metal−dielectric structure, since the continuity of the electric field component E_y prevents charge accumulations at the interfaces. In Figure 8.3 we report the dispersion curves relative to the symmetric (blue curves) and antisymmetric (red curves) modes for three different values of film thickness d (20, 40, and 80 nm), assuming a lossy free-electron gas model with $\gamma = 0.1\ \omega_p$ for the film and air ($\varepsilon_d = 1$) for the surrounding medium.

For decreasing values of thickness d, the antisymmetric mode (represented by the red curves) increases its effective refractive index $\mathrm{Re}(k_x/k_0)$, while the symmetric mode branches (blue curves) move toward lower values of $\mathrm{Re}(k_x/k_0)$. The dispersion curves for the thickest film ($d = 80$ nm) are very close to the dispersion of the isolated metal−dielectric surface plasmon-polariton (black dashed curve). Antisymmetric modes are very well confined at metal surfaces and therefore display high propagation
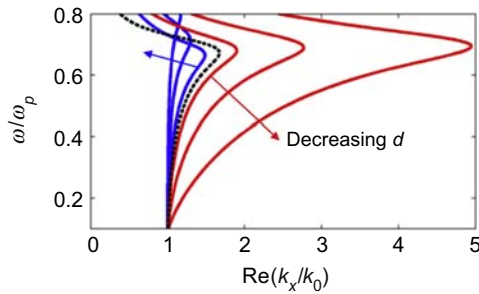
Figure 8.3 Dispersion curves for symmetric (blue curves) and antisymmetric (red curves) modes supported by thin metal films surrounded by air. Three values of film thickness d (20, 40, and 80 nm) are considered. The black curve indicates the dispersion of the surface plasmon polariton mode of the flat metal−air interface.

losses. Conversely, the dispersion of the symmetric mode is closer to the light line $(k_x = k_0)$ of the surrounding medium, resulting in a weaker confinement at the metal surfaces and lower propagation losses. For this reason, the symmetric and antisymmetric modes are sometimes referred to as long- and short-range surface plasmon polaritons, respectively.

8.5   Excitation of surface plasmon polaritons

As seen in the previous paragraph, the phase wavevector, $\mathrm{Re}(k_x)$, of surface plasmon polaritons guided by a metal−dielectric interface or a thin metallic film is larger than the wavenumber of the surrounding dielectric. This means that the field associated with SPPs decreases exponentially in the surrounding dielectric media. As a consequence, any light source, e.g., a plane wave or a finite cross-section beam, propagating in the dielectric and impinging on the interface with the metal at any incidence angle cannot couple to SPPs. As for dielectric slab waveguides, evanescent fields are required in order to excite surface waves optically. The common techniques to bring evanescent fields near metal interfaces and achieve phase-matching to SPPs are based on prisms and gratings. Prism coupling setups include the Kretschmann configuration and the Otto configuration. In the Kretschmann setup, a metal film is deposited on a prism's face, and a p-polarized light source excites the film from the prism side at oblique incidence. In this asymmetric system two SPPs are present: one pertaining to the air−metal interface and the other to the prism−metal interface. Evanescent fields on the air side, and excitation of the air−metal SPP, may be achieved only above the critical angle of incidence of the prism−air system, i.e., for internal incidence angles $\theta_i > \theta_c = \mathrm{asin}(1/n_\mathrm{p})$, where n_p is the prism refractive index. A quasi phase-matching to the air−metal SPP is obtained when $n_\mathrm{p}\sin(\theta_i) = n_\mathrm{spp} = \mathrm{Re}[\sqrt{\varepsilon_m/(\varepsilon_m + 1)}]$. The metal film should be semitransparent in order to allow tunneling of a portion of the light source from the prism side to the air side. Perfect phase matching is not possible, because the SPP propagating on the air−metal interface of this asymmetric multilayer is a leaky mode, i.e., its dispersion curve $k_x(\omega)$ lies within the prism light cone. In other words, the air−metal SPP undergoes not only the inherent absorption losses in the metal film but also radiation losses due to leakage into the prism. The SPP is observable as a dip in reflection

by varying the angle of incidence. The dip occurs at $\theta_i = \theta_{SPP} = \mathrm{asin}(n_{spp}/n_p)$, and it is due to the increased absorption in the metal film when the leaky SPP is excited. Zero reflection or perfect absorption is achievable under critical coupling conditions, i.e., when the damping due to radiation leakage equals the damping due to absorption losses. This condition can be reached by properly choosing the metal film thickness. In Figure 8.4 we show absorption as a function of frequency ω and angle of incidence θ_i in the Kretschmann configuration for a prism with index $n_p = 1.5$ and a metal film modeled as a lossy free-electron gas with $\gamma = 0.01\,\omega_p$ and thickness 80 nm. The absorption peak above the critical angle θ_c corresponds to the nonradiative branch of the SPP, while the absorption peak for $\omega > \omega_p$ is due to the radiative branch of the SPP dispersion (see the similarities with Figure 8.1).

In the Otto excitation scheme, a subwavelength air gap between the metal surface and the prism allows the generation of evanescent fields at the prism–air interface above the critical angle θ_c. If the air gap is thin enough, the evanescent tail reaches the metal surface and is able to couple to the air–metal SPP when $\theta_i = \theta_{SPP}$. In the Otto configuration, the thickness of the air gap is a critical parameter that influences the coupling strength to the SPP. This geometry is particularly suitable for monitoring the surface quality and, in general, for all applications in which direct contact with metal must be avoided.

Evanescent fields can also be obtained in gratings when a diffraction order assumes a wavevector larger than the incident grazing radiation. A periodic perturbation on a metallic surface with periodicity p introduces diffraction orders with parallel wavevector $k_0 n_i \sin(\theta_i) \pm 2\pi N/p$, with $N = 1, 2,...$ assuming light impinges on the structure from a medium with refractive index n_i at an angle θ_i with respect to the normal to the patterned surface. If one of these orders matches the SPP wavevector, i.e., if

$$k_0 n_{spp} = k_0 n_i \sin(\theta_i) \pm 2\pi N/p, \qquad\qquad (8.11)$$

the SPP is excited, and a minimum in reflection occurs.
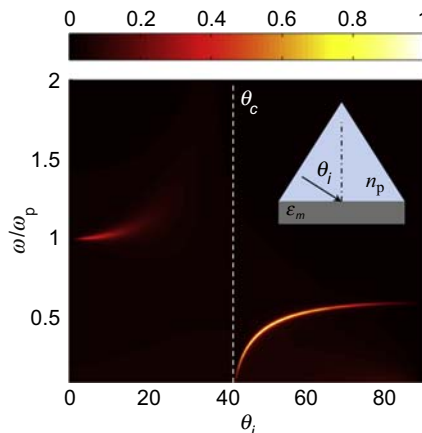
Figure 8.4 Absorption as a function of frequency ω, normalized to the plasma frequency ω_p, and angle of incidence θ_i in the Kretschmann configuration (inset).
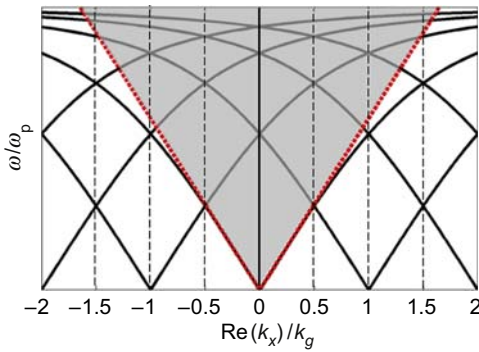
Figure 8.5 The effect of a periodic perturbation with period p on the SPP dispersion.

The scalar Eqn (8.11) is valid when the grating wavevector $k_g = 2\pi/p\hat{\mathbf{v}}$ is parallel to the plane of incidence so that it results collinear to the incident wavevector and the SPP wavevector. It is important to mention that the expression of n_{spp} for the SPP dispersion in Eqn (8.11) is not equal to the dispersion of the planar and unperturbed, metal–dielectric-interface SPP. In fact, the periodicity generally introduces a change in both the real (wavevector) and imaginary part (attenuation constant) of the complex SPP wavenumber. However, for small perturbations, i.e., shallow grooves or bumps, the SPP dispersion in the phase-matching Eqn (8.11) can be approximated by the planar SPP dispersion $n_{spp} = \mathrm{Re}(k_x/k_0) = \mathrm{Re}\left(\sqrt{\varepsilon_m \varepsilon_d/(\varepsilon_m + \varepsilon_d)}\right)$. The effect of a periodic perturbation with period p and wavevector k_g ($k_g = 2\pi/p$) on the SPP dispersion is illustrated in Figure 8.5. Radiative coupling of homogeneous plane waves to the grating SPP is possible when the dispersion branches fall within the light cone (gray area in Figure 8.5), usually inducing reflection minima and absorption maxima.

8.6 Localized surface plasmons

While planar, metal–dielectric interfaces act as waveguides and support the propagation of surface waves coupled to the electron plasma, metallic nanostructures, i.e., small particles or nanopatterned metallic surfaces, may host nonpropagating electromagnetic modes, also known as localized surface plasmons (LSPs). These modes induce resonances observable as peaks in the scattering and absorption cross-sections spectra. The spectral position of the resonances depends on several factors: the metal permittivity, the permittivity of the surrounding dielectric, and the shape and size of the metallic nanostructure. A very simple scenario consists of a metallic, spherical nanoparticle much smaller than the wavelength of the interacting electromagnetic field. The electrostatic approximation (i.e., frequency $\omega \to 0$) provides a good description of the problem in the near-field region of the particle (i.e., for $kr \ll 1$) and a simple prediction of the resonant frequency. For a spherical, metallic particle with radius R and relative permittivity ε_m in a homogeneous dielectric background

with relative permittivity ε_d, the solutions of the Laplace equation for the potential, $\nabla^2\varphi = 0$, inside and outside the particle are given by

$$\varphi_{\text{in}} = -\frac{3\varepsilon_d}{\varepsilon_m + 2\varepsilon_d}E_0 r \cos(\theta) \tag{8.12}$$

$$\varphi_{\text{out}} = -E_0 r \cos(\theta) + \frac{\mathbf{p}\cdot\mathbf{r}}{4\pi\varepsilon_0\varepsilon_d r^3}. \tag{8.13}$$

In these expressions $\mathbf{p} = \varepsilon_0\varepsilon_d\alpha\mathbf{E}_0$ is the dipole moment associated with the spherical nanoparticle, $\alpha = 4\pi R^3\frac{\varepsilon_m - \varepsilon_d}{\varepsilon_m + 2\varepsilon_d}$ is its polarizability, $\mathbf{r}$ is the position vector (with amplitude $r = |\mathbf{r}|$), $\mathbf{E}_0$ is the applied (static) electric field, and θ is the angle between $\mathbf{E}_0$ and $\mathbf{r}$. It is interesting to note that the expression of the polarizability has the same form as the Clausius−Mossotti relation in the framework of the effective medium approximation for solids (see Chapter 7); therefore, it is quite general and can be applied to nonmetallic spherical inclusions. The electric field distribution is evaluated as $\mathbf{E} = -\nabla\varphi$, and its expressions inside and outside the particle are, respectively,

$$\mathbf{E}_{\text{in}} = \frac{3\varepsilon_d}{\varepsilon_m + 2\varepsilon_d}\mathbf{E}_0 \tag{8.14}$$

$$\mathbf{E}_{\text{out}} = \mathbf{E}_0 + \frac{3\mathbf{n}(\mathbf{n}\cdot\mathbf{p}) - \mathbf{p}}{4\pi\varepsilon_0\varepsilon_d r^3} \tag{8.15}$$

where $\mathbf{n} = \mathbf{r}/r$ is the radial unit vector. From Eqns (8.14) and (8.15) it is clear that both the inside and outside fields diverge when the particle polarizability becomes singular, i.e., when $\varepsilon_m + 2\varepsilon_d = 0$. If the relative background permittivity ε_d is positive, the singularity is only possible for particles with negative relative permittivity ε_m, i.e., metallic nanoparticles operating below the metal's plasma frequency. However, the presence of losses in the particle prevents the fields from diverging, resulting in a resonance at the frequency that satisfies the condition $\text{Re}(\varepsilon_m) = -2\varepsilon_d$. The latter is sometimes referred to as the Fröhlich condition or Fröhlich resonance. It is interesting to note that in the electrostatic approximation, the localized surface plasmon resonance associated with the Fröhlich condition depends solely on the particle and surrounding medium permittivities. This behavior is at the basis of plasmonic sensing devices. In such sensors the presence or absence of organic/inorganic media around the particle alters the effective permittivity ε_d of the surrounding medium and may be detected as a frequency shift of the plasmonic resonance, i.e., the Fröhlich condition. Since $|\varepsilon_m|$ gets larger when the frequency decreases (normal chromatic dispersion), an increase in the surrounding permittivity ε_d results in a typical red-shift of the plasmonic resonance. An improved description of the problem, in particular in the intermediate and far-field regions, can be obtained by using the *quasi-static* approximation. Now the incident field is time- and space-dependent so that the wavelength λ and the frequency ω are finite. However, if the sphere is very small with respect to the incident wavelength ($R\ll\lambda$), one can still assume that spatial retardation effects across the particle are negligible and that the particle behaves like a point dipole. For the time-harmonic excitation, the

induced dipole moment will now read as $\mathbf{p}(t) = \varepsilon_0 \varepsilon_d \alpha \mathbf{E}_0 e^{-i\omega t}$, where the polarizability α is still defined as in the purely electrostatic problem $\left(\alpha = 4\pi R^3 \frac{\varepsilon_m - \varepsilon_d}{\varepsilon_m + 2\varepsilon_d} \right)$. The fields scattered by the excited particle can then be written in the usual way:

$$\mathbf{E}_s = \frac{1}{\varepsilon_0 \varepsilon_d} \left\{ k^2 \mathbf{n} \times (\mathbf{p} \times \mathbf{n}) + \left(\frac{1}{r} - ik \right) \left[\frac{3\mathbf{n}(\mathbf{n} \cdot \mathbf{p}) - \mathbf{p}}{r} \right] \right\} G(\mathbf{r}) \tag{8.16}$$

$$\mathbf{H}_s = -i\omega \left(\frac{1}{r} - ik \right) (\mathbf{p} \times \mathbf{n}) G(\mathbf{r}). \tag{8.17}$$

In the above expression $G(\mathbf{r}) = \frac{e^{ikr}}{4\pi r}$ is Green's function for the scalar Helmholtz equation $\nabla^2 G + k^2 G = -\delta(\mathbf{r})$, $k = \omega\sqrt{\varepsilon_d}/c$ is the wavenumber in the background medium, and the time dependence $e^{-i\omega t}$ is implicit. It is worth noting that the electrodynamic, quasi-static solution reported in Eqn (8.16) tends to the electrostatic one [Eqn (8.14)] in the near-field zone, i.e., for $kr \ll 1$. In the far-field zone, where $kr \gg 1$, the fields in Eqns (8.16) and (8.17) can be approximated by the simplified expressions

$$\mathbf{E}_{\text{far}} = \frac{k^2}{\varepsilon_0 \varepsilon_d} \mathbf{n} \times (\mathbf{p} \times \mathbf{n}) G(\mathbf{r}) \tag{8.18}$$

$$\mathbf{H}_{\text{far}} = \frac{1}{\eta_d} \frac{k^2}{\varepsilon_0 \varepsilon_d} (\mathbf{n} \times \mathbf{p}) G(\mathbf{r}) \tag{8.19}$$

where $\eta_d = \sqrt{\mu_0/(\varepsilon_0 \varepsilon_d)}$ is the impedance of the host dielectric medium. The power radiated into the far field is

$$P_{\text{dip}} = \frac{c^2 k^4 \eta_d}{12\pi \varepsilon_d} |\mathbf{p}|^2. \tag{8.20}$$

The scattering cross-section, C_{sca}, is calculated by normalizing the power P_{dip} radiated by the induced electric dipole to the input irradiance $I_{\text{in}} = 1/2\varepsilon_0 c \sqrt{\varepsilon_d} |\mathbf{E}_0|^2$ and is expressed by

$$C_{\text{sca}} = \frac{k^4}{6\pi} |\alpha|^2. \tag{8.21}$$

The absorption cross-section, C_{abs}, is instead the power absorbed by the particle normalized with respect to I_{in}, and it is given by

$$C_{\text{abs}} = k \text{Im}\{\alpha\}. \tag{8.22}$$

The sum of scattering and absorption cross-sections defines the so-called extinction cross-section $C_{ext} = C_{sca} + C_{abs}$.

Geometrical features like particle size and shape play an important role in establishing the spectral position of the localized plasmon resonance. For simple geometries, such as ellipsoidal particles, analytical expressions of polarizability and cross-sections can be retrieved in the electrostatic approximation. We consider an ellipsoid with a surface defined by $\frac{x^2}{R_1^2} + \frac{y^2}{R_2^2} + \frac{z^2}{R_3^2} = 1$, where R_1, R_2, and R_3 are the ellipsoid semiaxes in the x, y, and z directions, respectively, and we assume $R_1 \geq R_2 \geq R_3$. Setting up the electrostatic problem in elliptical coordinates and expanding the fields in ellipsoidal harmonics leads to the following expression for the anisotropic polarizability:

$$\alpha_j = 4\pi R_1 R_2 R_3 \frac{\varepsilon_m - \varepsilon_d}{3\varepsilon_d + 3L_j(\varepsilon_m - \varepsilon_d)}, \quad j = 1, 2, 3. \tag{8.23}$$

In the expression above, L_j is a geometrical factor given by the integral

$$L_j = \frac{R_1 R_2 R_3}{2} \int_0^\infty \left[\left(R_j^2 + q \right) f(q) \right]^{-1} dq \tag{8.24}$$

where $f(q) = \sqrt{(q + R_1^2)^2 (q + R_2^2)^2 (q + R_3^2)^2}$.

Only two of the geometrical factors are independent, since $L_1 + L_2 + L_3 = 1$. It is interesting to note that $L_1 = L_2 = L_3 = 1/3$ for a sphere, and for a generic ellipsoid $L_1 \leq L_2 \leq L_3$. Spheroids are special ellipsoids with two equal semiaxes. Cigar-shaped spheroids, known as *prolates*, have $R_2 = R_3$ and $L_2 = L_3$ and are obtained by rotating an ellipse about its major axis. *Oblates* (pancake-shaped spheroids) have $R_1 = R_2$ and $L_1 = L_2$ and are obtained by rotating an ellipse around its minor axis.

The polarizability of a coated ellipsoid, or core−shell ellipsoid, can be found in the electrostatic approximation by applying the proper boundary conditions at the interface between the inner ellipsoid, i.e., the core with semiaxes r_1, r_2, and r_3 and permittivity ε_1, and the outer ellipsoid with semiaxes R_1, R_2, and R_3 that defines the shell region with permittivity ε_2. The solution gives the polarizability

$$\alpha_j = \frac{4\pi R_1 R_2 R_3}{3}$$

$$\times \frac{(\varepsilon_2 - \varepsilon_d)\left[\varepsilon_2 + (\varepsilon_1 - \varepsilon_2)\left(L_j^{(1)} - fL_j^{(2)}\right)\right] + f\varepsilon_2(\varepsilon_1 - \varepsilon_2)}{\left[\varepsilon_2 + (\varepsilon_1 - \varepsilon_2)\left(L_j^{(1)} - fL_j^{(2)}\right)\right]\left[\varepsilon_d + (\varepsilon_2 - \varepsilon_d)L_j^{(2)}\right] + fL_j^{(2)}\varepsilon_2(\varepsilon_1 - \varepsilon_2)},$$

$$j = 1, 2, 3$$

$$\tag{8.25}$$

where $L_j^{(1)}$ and $L_j^{(2)}$ are the geometrical factors of the inner and outer ellipsoids and can be determined using Eqn (8.24), and $f = r_1 r_2 r_3 / (R_1 R_2 R_3)$ is the volume fraction of the particle occupied by the core. For a coated sphere, or core−shell spherical particle, the geometrical factors are $L_j^{(1)} = L_j^{(2)} = 1/3$ and $R_1 = R_2 = R_3 = R$, so that the polarizability becomes isotropic and its expression reduces to

$$\alpha = 4\pi R^3 \frac{(\varepsilon_2 - \varepsilon_d)(\varepsilon_1 + 2\varepsilon_2) + f(\varepsilon_1 - \varepsilon_2)(\varepsilon_d + 2\varepsilon_2)}{(\varepsilon_1 + 2\varepsilon_2)(\varepsilon_2 + 2\varepsilon_d) + 2f(\varepsilon_1 - \varepsilon_2)(\varepsilon_2 - \varepsilon_d)}. \tag{8.26}$$

Vanishing scattering and absorption cross-sections may be obtained in core−shell particles when $\alpha = 0$, a possibility that is forbidden in the case of spherical particles. This condition may be exploited to design epsilon near-zero metamaterials and cloaking devices. A spherical dielectric particle with positive permittivity $\varepsilon_1 > 1$ immersed in air ($\varepsilon_d = 1$) can be made invisible with a shell of permittivity $\varepsilon_2 < 0$, for example using a metal below its plasma frequency. The resulting core−shell system displays zero polarizability as well as zero scattering and absorption cross-sections when $(\varepsilon_2 - \varepsilon_d)(\varepsilon_1 + 2\varepsilon_2) + f(\varepsilon_1 - \varepsilon_2)(\varepsilon_d + 2\varepsilon_2) = 0$. In a realistic system, i.e., lossy and dispersive plasma, the chromatic dispersion in the metallic shell allows zero polarizability only at specific frequencies, and the metal losses prevent the polarizability from being exactly zero. However, minima of polarizability can be obtained in narrow bands, where scattering and absorption cross-sections are minimized.

For geometries like spheres and coated spheres, analytical solutions of the electrodynamic scattering problem may be obtained by expanding the incident, scattered, and internal fields in vector spherical harmonics and applying the continuity boundary conditions for tangential electric and magnetic fields. This procedure, known as the Mie or Lorenz−Mie theory, allows for expressing scattering, extinction, and absorption cross-sections in terms of coefficients, usually referred to as Mie or scattering coefficients, which depend on the wavelength, particle size, and background permittivity. Here we report the expressions of the scattering and extinction cross-sections, respectively, omitting their straightforward but lengthy derivation:

$$C_{\text{sca}} = \frac{2\pi}{k^2} \sum_1^\infty (2n + 1)\left(|a_n|^2 + |b_n|^2\right) \tag{8.27}$$

$$C_{\text{ext}} = \frac{2\pi}{k^2} \sum_1^\infty (2n + 1)\operatorname{Re}\{a_n + b_n\}. \tag{8.28}$$

The scattering coefficients for a sphere with radius R and permittivity ε_m in a background of permittivity ε_d are given in terms of Riccati−Bessel functions $\psi_n(\rho) = \rho j_n(\rho)$

and $\xi_n(\rho) = \rho[j_n(\rho) + iy_n(\rho)]$ and spherical Bessel functions of first $[j_n(\rho)]$ and second $[y_n(\rho)]$ kind as well as their derivatives $\psi'_n(\rho)$ and $\xi'_n(\rho)$ as follows:

$$a_n = \frac{m\psi_n(mx)\psi'_n(x) - \psi_n(x)\psi'_n(mx)}{m\psi_n(mx)\xi'_n(x) - \xi_n(x)\psi'_n(mx)} \tag{8.29}$$

$$b_n = \frac{\psi_n(mx)\psi'_n(x) - m\psi_n(x)\psi'_n(mx)}{\psi_n(mx)\xi'_n(x) - m\xi_n(x)\psi'_n(mx)} \tag{8.30}$$

where $m = \sqrt{\varepsilon_m/\varepsilon_d}$ and $x = kR$.

The poles of the scattering coefficients a_n and b_n are associated with the electric and magnetic multipole resonances, respectively. The electric polarizability of the sphere due to the electric dipole response can be written by using the only coefficient a_1 as

$$\alpha_e = \frac{i6\pi}{k^3}a_1 \tag{8.31}$$

Similarly, the sphere magnetic polarizability can easily be written as

$$\alpha_m = \frac{i6\pi}{k^3}b_1. \tag{8.32}$$

The difference between the quasi-static expression of the electric polarizability, $\alpha = 4\pi R^3 \frac{\varepsilon_m - \varepsilon_d}{\varepsilon_m + 2\varepsilon_d}$, and the Mie-theory-based expression in Eqn (8.31) is that the pole of the quasi-static polarizability does not display any dependence on the particle size. In fact, the radius R alters only the value of the quasi-static polarizability, but the particle resonance is fixed at the Fröhlich frequency at which $\varepsilon_m + 2\varepsilon_d = 0$. On the other hand, retardation, i.e., the effect of nonvanishing k, is included in the Mie-theory-based expression of the polarizability and turns out to be important for larger particle sizes. A better idea of the effects of retardation can be gained by expanding the Riccati−Bessel functions in the expression of the Mie polarizability α_e in Eqn (8.31). The resulting approximation reads as follows:

$$\alpha_e \approx V\frac{1 - \frac{(\varepsilon_m + \varepsilon_d)x^2}{10\varepsilon_d} + O(x^4)}{\frac{1}{3} + \frac{\varepsilon_d}{\varepsilon_m - \varepsilon_d} - \frac{(\varepsilon_m + 10\varepsilon_d)x^2}{30\varepsilon_d} - \frac{i2}{9}x^3 + O(x^4)} \tag{8.33}$$

where V is the particle volume. The term in x^2 in the numerator is due to the retardation of the external field across the sphere. Its effect is a shift in the plasmon resonance frequency. The term in x^2 in the denominator is due to the retardation of the depolarization field in the particle and gives an additional frequency shift of the plasmon resonance. The term in x^3 in the denominator is associated with radiation damping. The overall impact of retardation is a red-shift of the dipole resonance. The effect is mainly related to the fact that, in the dipole approximation, charge separation increases with the particle size and therefore produces a smaller restoring force and a decreased resonance frequency.

It is possible to exploit the Mie theory to include retardation effects in the expression of the dipole polarizability of core−shell particles. In this case, it is enough to calculate the Mie coefficients for the core−shell particle and use Eqn (8.31) to determine its polarizability. Here we give the recursive formulas of the scattering coefficients in the case of spherical core−shell particles in terms of Riccati−Bessel functions, $\psi_n(\rho) = \rho j_n(\rho)$, $\xi_n(\rho) = \rho[j_n(\rho) + i y_n(\rho)]$, $\chi_n(\rho) = -\rho y_n(\rho)$, and their derivatives (indicated with primes):

$$a_n = \frac{\psi_n(y)\left[\psi_n'(m_2 y) - A_n \chi_n'(m_2 y)\right] - m_2 \psi_n'(y)\left[\psi_n(m_2 y) - A_n \chi_n(m_2 y)\right]}{\xi_n(y)\left[\psi_n'(m_2 y) - A_n \chi_n'(m_2 y)\right] - m_2 \xi_n'(y)\left[\psi_n(m_2 y) - A_n \chi_n(m_2 y)\right]}$$

(8.34)

$$b_n = \frac{m_2 \psi_n(y)\left[\psi_n'(m_2 y) - B_n \chi_n'(m_2 y)\right] - \psi_n'(y)\left[\psi_n(m_2 y) - B_n \chi_n(m_2 y)\right]}{m_2 \xi_n(y)\left[\psi_n'(m_2 y) - B_n \chi_n'(m_2 y)\right] - \xi_n'(y)\left[\psi_n(m_2 y) - B_n \chi_n(m_2 y)\right]}$$

(8.35)

where

$$A_n = \frac{m_2 \psi_n(m_2 x)\psi_n'(m_1 x) - m_1 \psi_n(m_1 x)\psi_n'(m_2 x)}{m_2 \chi_n(m_2 x)\psi_n'(m_1 x) - m_1 \psi_n(m_1 x)\chi_n'(m_2 x)}$$

(8.36)

$$B_n = \frac{m_2 \psi_n(m_1 x)\psi_n'(m_2 x) - m_1 \psi_n(m_2 x)\psi_n'(m_1 x)}{m_2 \psi_n(m_1 x)\chi_n'(m_2 x) - m_1 \chi_n(m_2 x)\psi_n'(m_1 x)}.$$

(8.37)

In the above expressions, $m_{1/2} = \sqrt{\varepsilon_{1/2}/\varepsilon_d}$ is the index ratio of the core (m_1) and shell (m_2) regions, $x = kR_1$, and $y = kR_2$, R_1 is the inner radius, and R_2 is the outer radius of the shell. An interesting case is the particle with a dielectric core and a metallic shell. Such a system supports two distinct localized surface plasmon resonances for each multipole. The most important spectral features come from the two electric dipolar resonances, which dominate the scattering and absorption cross-sections in the case of very small core−shell particles, i.e., $y \ll 1$. Predictions of the resonance frequencies can be obtained in the case of Drude-like plasmas using the hybridization theory. When interband transitions cannot be neglected, e.g., at visible and near-infrared wavelengths for noble metals, the quasi-static expression in Eqn (8.26) or the more accurate, Mie-theory-based expression of the particle polarizability in Eqn (8.31) must be used.

8.7   Nonlocal response of metallic nanostructures

Modeling a metal as a Drude-like, free-electron gas as in Eqn (8.4) or the improved model that takes into account interband transitions [Eqn (8.6)] provides a satisfactory description of the metal response, even for subwavelength structures. However, when metallic nanostructures reach the nanometer size or when the

distances between metallic regions are smaller than a few nanometers, additional effects must be included in the description of the free-electron gas. The simplicity and the effectiveness of the Drude and Drude−Lorentz models reside in their local nature. These models predict that the response of the metal at a certain location in space is proportional to the field evaluated at that specific location. In other words, if one considers the equation of motion for free electrons and neglects the force associated with the magnetic field $-e\mathbf{v} \times \mathbf{B}$, i.e., $m\dot{\mathbf{v}} + m\gamma\mathbf{v} = -e\mathbf{E}$, the induced current density at position $\mathbf{r}$ is simply proportional to the applied field. For a time-harmonic excitation, the expression of the current density associated with the motion of free electrons is

$$\mathbf{J_f}(\mathbf{r}, \omega) = \sigma(\omega)\mathbf{E}(\mathbf{r}, \omega) = \frac{\varepsilon_0\omega_p^2}{-i\omega + \gamma}\mathbf{E}(\mathbf{r}, \omega). \tag{8.38}$$

In this description, screening charges are confined at the metal surface, and they are not allowed to penetrate into the metal bulk region. Such approximation is generally acceptable because the field penetration depth in the metal, of the order of a few tens of nanometers at visible wavelengths, is much larger than the electronic screening length. The latter quantity is strictly zero only in the local approximation [Eqn (8.38)], in which the metal response is independent from field derivatives. A direct consequence of the local approximation and the zero screening length is the unlimited amount of electromagnetic confinements at the corners of metallic objects (lightning-rod effect) or in the gap of metal−insulator−metal structures. A more accurate representation of the optical response of metals is provided by the hydrodynamic model of the free-electron gas. This model corrects the constitutive relation for metals by adding a dependence on the spatial derivative of the field and introduces a screening length larger than zero. The overall effect is to impose more realistic limitations to the field enhancement near metallic surfaces. We start with the equation of motion for a hydrodynamic gas:

$$m\dot{\mathbf{v}} + m\gamma\mathbf{v} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - \nabla p/n, \tag{8.39}$$

where m, e, and $\mathbf{v}$ are the electron mass, charge, and velocity, defined as in the ideal free-electron gas model of Eqn (8.1). In the hydrodynamic model we have an additional force, $-\nabla p/n$, due to the spatial differences of gas pressure p that drives the conduction electrons from higher density regions to lower density regions. The quantum pressure can be linked to the electron density via a polytropic relation:

$$p = p_0(n/n_0)^{\Gamma} \tag{8.40}$$

where the $p_0 \propto n_0 E_F$ is the Fermi pressure, i.e., the pressure of the quantum electron gas at zero temperature, n_0 is the equilibrium free-electron density, and E_F is the Fermi energy. For a three-dimensional gas the polytropic exponent Γ assumes the value 5/3.

Multiplying Eqn (8.39) by a factor $-en/m$ and expanding the full derivative of the velocity as $\dot{\mathbf{v}} = \frac{\partial v}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v}$ leads to the expression

$$- en\frac{\partial v}{\partial t} - en(\mathbf{v} \cdot \nabla)\mathbf{v} - en\gamma\mathbf{v} = \frac{ne^2}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + e\nabla p/m, \tag{8.41}$$

We now expand the fields and the time-varying quantities as the sum of static terms plus small perturbations oscillating at the excitation frequency (i.e., the frequency of the driving electromagnetic field) and its harmonics, i.e., $q = q_0 + q_\omega e^{-i\omega t} + q_{2\omega}e^{-i2\omega t}$, where $q = n, \mathbf{E}, \mathbf{B}, \mathbf{v}$. Such perturbative approach is justified when the input electromagnetic field induces very small perturbations of electron densities with respect to the equilibrium, which corresponds to the condition $n_{2\omega} \ll n_\omega \ll n_0$. It is now possible to linearize Eqn (8.41), considering only the terms oscillating at the driving frequency ω, and writing the time derivative as $-i\omega$ in order to obtain a time-harmonic, nonlocal, constitutive relation between the induced current density and the electromagnetic field. The gas pressure term reduces to $\frac{e}{m}\nabla p = \frac{5}{3}\frac{ep_0}{mn_0}\nabla n$ by using the relation in Eqn (8.40) between the pressure and the electron density. Considering the identity $\mathbf{J}_\omega = -en_0\mathbf{v}_\omega$, one obtains the frequency-domain relation

$$\omega^2 \mathbf{J}_\omega + i\omega\gamma\mathbf{J}_\omega = i\omega\frac{n_0 e^2}{m}\mathbf{E}_\omega - \frac{5}{3}\frac{p_0}{mn_0}\nabla(\nabla \cdot \mathbf{J}_\omega), \tag{8.42}$$

where the continuity equation, $\partial n/\partial t = (\nabla \cdot \mathbf{J})/e$, has been used to explicit the gas pressure in terms of the spatial derivative of the induced current density. Since $p_0 \propto n_0 E_F = n_0\left(\frac{1}{2}mv_F^2\right)$, Eqn (8.42) can be recast as

$$\beta^2\nabla(\nabla \cdot \mathbf{J}_\omega) + \omega^2\mathbf{J}_\omega + i\omega\gamma\mathbf{J}_\omega = i\omega\varepsilon_0\omega_p^2\mathbf{E}_\omega, \tag{8.43}$$

where β is proportional to the Fermi velocity v_F, and the expression of the free-electron plasma frequency is reintroduced. If one neglects the pressure term, i.e., if $\beta = 0$, the above expression corresponds to the classical local response of the Drude model, as in Eqn (8.38). For noble metals $v_F \sim 10^{-2}c$ and β is on the same order of magnitude of v_F. It is now clear that the pressure term introduces second-order spatial derivatives, hence adding to the metal optical response a nonlocal, or spatially dispersive, contribution. It is worth noticing that the inclusion of electron–electron interaction effects through the nonlocal term $\beta^2\nabla(\nabla \cdot \mathbf{J}_\omega)$ requires the introduction of an additional boundary condition, involving the current density $\mathbf{J}_\omega$, in order to solve Maxwell's equation at metal–dielectric interfaces. This requirement is absent in the local description, since the current density is simply proportional to the electric field. Although it is not the only possibility, a common choice is to consider the interface as a hard boundary where the electron density goes abruptly from n_0 on the metal side to 0 on the dielectric or free-space side. In this case, the proper boundary condition for the current density is $\mathbf{J}_\omega \cdot \hat{\mathbf{n}} = 0$, where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the metallic

interface. Equation (8.43) plus the usual Helmholtz equation and the boundary conditions can be solved numerically by using frequency domain solvers such as finite-element or finite-difference schemes. More sophisticated techniques to take the hydrodynamic gas pressure into account are based on different definitions of the metal−dielectric interface. A possible choice is a soft boundary with a smooth transition of the electron density as well as the plasma frequency from the metal side to the dielectric side of the interface. Regardless of the boundary definition, the main effect of the nonlocal term is a smearing of the charges at the metal surface on distances on the order of $\beta/\omega_{\mathrm{p}}$, i.e., few angstroms. In other words, in the local description the free-electron gas is assumed to be infinitely incompressible; therefore, surface charges form a Dirac delta distribution centered at the metal surface, and the normal electric field undergoes a step-like discontinuity. The introduction of the nonlocal gas pressure term allows the charges to distribute in a finite thickness of a few angstroms under the metal surface. Macroscopic effects due to nonlocal corrections are significant only for nanometer- and subnanometer-sized metallic nanoparticles or for nanometer and subnanometer gaps between metallic objects. Plasmonic resonances are blue-shifted when the nonlocality is included in the solution of Maxwell's equations. The reason can be intuitively understood by transforming the constitutive relation [Eqn (8.43)] into the k-space domain, i.e., using the transformation $\nabla \rightarrow ik$. The resulting nonlocal permittivity is

$$\varepsilon_{nl}(k,\omega) = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + i\omega\gamma - |k|^2\beta^2} \tag{8.44}$$

where the term $|k|^2\beta^2$ introduces a small perturbation of the permittivity that pushes plasmonic resonances at slightly higher frequencies. The k-space representation in Eqn (8.44) may be exploited to describe the permittivity for longitudinal waves in simple geometries with translational invariance, e.g., planar metallic slabs or metallo-dielectric stacks. For transverse waves the classic, local ($\beta = 0$) Drude model is valid. Other nonlocal effects in nanoplasmonic structures are related to the presence of additional resonances above the plasma frequency, which can be ascribed to the excitation of longitudinal waves. The other important impact of nonlocalities is on metallic sharp corners and tips, where the typical field divergence predicted by the local theory is removed by the nonlocal charge-screening effect. For the same reason, the nonlocality imposes a limitation to the field enhancement allowed in gaps between metallic walls when the gap size is reduced below the nanometer scale.

Problems

1. Calculate the phase velocity of a surface plasmon propagating on a flat silver−air interface at $\lambda = 532$ nm, and compare it to the phase velocity of light in air. Assume the real part of the dielectric constant of silver is $\varepsilon = -9.30 + i0.87$ at $\lambda = 532$ nm. Determine the decay length of the surface plasmon along the propagation direction.

2. Consider the surface plasmon of Problem 1, and determine the angle of incidence of a TM-polarized plane wave required to excite it in the Kretschmann configuration. Assume the prism glass has a refractive index of 1.5. If the surface is patterned with a shallow perturbation, what is the periodicity required to excite the surface plasmon at normal incidence with the first diffraction order of the grating?

3. Calculate the decay lengths (i.e., the length at which the field amplitude reduces by a factor 1/e) of a gold−air surface plasmon polariton at 1064 nm in the direction perpendicular to the propagation direction. For gold, assume a complex dielectric constant of −43.8 + i4.2 at 1064 nm. Repeat the same calculation at 532 nm, assuming a dielectric constant of −4.3 + i2.3 for gold.

4. Determine the resonance wavelength shift per refractive index unit change for silver nanoparticles in water ($n = 1.33$). Use the electrostatic approximation and assume, for silver, a Drude permittivity with plasma frequency of 2.18 PHz and damping frequency of 4.35 THz.

5. Write a Matlab or Mathematica script to plot the polarizability tensor components as functions of frequency (or wavelength) for a silver nanodisk with diameter 100 nm and height 25 nm. Assume the surrounding medium is glass ($n = 1.5$), and approximate the nanoparticle as a spheroid with Drude permittivity (plasma frequency 2.18 PHz and damping frequency 4.35 THz).

6. Write a Matlab or Mathematica script to plot the scattering, absorption, and extinction cross-sections as functions of wavelength (ranging from 300 to 3000 nm) for a gold nanorod illuminated by a plane wave polarized along the long axis of the rod. The nanorod has circular cross-section with radius 20 nm and height of 120 nm and is surrounded by air. Approximate the nanorod as spheroids with Drude-like permittivity (plasma frequency 2.18 PHz and damping frequency 6.46 THz). What is the maximum extinction wavelength? How does this wavelength shift for larger or smaller rod heights?

7. Calculate the resonant wavelengths (i.e., maximum extinction cross-section wavelengths) of a core−shell spherical nanoparticle made of a silica core and a gold shell surrounded by air. The silica core has a permittivity of 2.25 and a radius of 30 nm, whereas the gold shell has Drude permittivity with plasma frequency 2.18 PHz, damping frequency 6.46 THz, and thickness 10 nm (external radius 40 nm). How are these resonant wavelengths influenced by changes of the shell's thickness? Create a three-dimensional plot (or a color map) of the extinction cross-section as a function of the wavelength and the shell's thickness.

Further reading

[1]   S.A. Maier, Plasmonics: Fundamentals and Applications, Springer, 2007.

[2]   H. Raether, Surface Plasmons, Springer-Verlag, 1986.

[3]   C.F. Bohren, D.R. Huffman, Absorption and Scattering of Light by Small Particles, John Wiley and Sons, 1983.

[4]   A.D. Rakic, A.B. Djurisic, J.M. Elazar, M.L. Majewski, Optical properties of metallic films for vertical-cavity optoelectronic devices, Appl. Opt. 37 (1998) 5271.

[5]   J. Jin, The Finite Element Method in Electromagnetics, Wiley-IEEE Press, 2014.

[6]   A. Taflove, S.C. Hagness, Computational Electrodynamics: The Finite-Difference Time-Domain Method, Artech House, 2005.

[7]   J.J. Burke, G.I. Stegeman, T. Tamir, Surface-polariton-like waves guided by thin, lossy metal films, Phys. Rev. B 33 (1986) 5186.

[8]   J.A. Dionne, L.A. Sweatlock, H.A. Atwater, A. Polman, Planar metal plasmon waveguides: frequency-dependent dispersion, propagation, localization, and loss beyond the free electron model, Phys. Rev. B 72 (2005) 075405.

[9] A.D. Boardman, Electromagnetic Surface Modes, John Wiley and Sons, 1982.

[10] S. Raza, G. Toscano, A.-P. Jauho, M. Wubs, N.A. Mortensen, Unusual resonances in nanoplasmonic structures due to nonlocal response, Phys. Rev. B 84 (2011) 121412(R).

[11] M. Scalora, M.A. Vincenti, D. de Ceglia, V. Roppo, M. Centini, N. Akozbek, J.W. Haus, Second- and third-harmonic generation in metal-based structures, Phys. Rev. A 82 (2010) 043828.

[12] S.J. Oldenburg, R.D. Averitt, S.L. Westcott, N.J. Halas, Nanoengineering of optical resonances, Chem. Phys. Lett. 288 (1998) 243.

Metamaterials

<div style="text-align:right">**9**</div>

J. Sun, N.M. Litchinitser
University at Buffalo, The State University of New York, NY, USA

9.1   Introduction

We know from basic physics and chemistry that all materials consist of atoms and molecules. The atoms and their relative compositions determine many physical properties of matter, including its optical properties. For example, in metals, valence electrons are weakly bound to the nuclei; as a result, they can freely move inside of the material, which enables conduction. Another example is the property of birefringence in crystals, which originates from anisotropy in the binding forces between the atoms, forming a crystal.

For thousands of years, people studied various materials to use them in a wide range of applications, spanning from architecture, transportation, and medicine to electronics, optics, and space exploration. Eventually, we have reached a stage when naturally available materials can no longer meet the demands of fast-developing fundamental science and applications. For instance, the best resolution of a conventional optical microscope, which once revolutionized the entire field of natural sciences by allowing the visualization of microscale objects, is on the order of 100 nm because of the fundamental diffraction limit. Therefore, to image nanoscale structures, one has to use electron-based microscopy that enables orders of magnitude higher resolution, but can only be used for conducting materials. The crucial question that scientists today are asking is whether it is possible to overcome the limitations imposed by the natural materials and components used in optical microscopy and develop an all-optical technique for nanoscale imaging. Although several near-field techniques have been developed for the point-by-point nanoscale imaging of planar structures, far-field, large-scale, high-resolution imaging remained a challenge until now. The emergence of metamaterials is likely to provide a viable solution to this long-standing problem.

Metamaterials are rationally designed artificial materials that gain their properties from their structures and carefully engineered unit cells (meta-atoms) rather than from the properties of their constitutive materials. In Greek language, *meta* means "beyond." It should be noted that the optical properties of another class of engineered materials—photonic crystals—are also somewhat defined by their structures. However, there is a fundamental difference between these two classes of engineered materials: the length scales of their unit cells. Although the period and the unit cell size of photonic crystals are comparable to the wavelength of light, the size of the meta-atom is much smaller than the wavelength of light. Thus, metamaterials can be considered as effective media, whereas photonic crystals cannot be.

In the following sections, we will show several examples of metamaterials enabling very unusual electromagnetic (EM) properties and functionalities, such as negative, near-zero, and indefinite permittivity or permeability indices; backward waves and backward phase matching in nonlinear optics; imaging below the diffraction limit; perfect absorption; and cloaking.

9.2   Types of metamaterials

The prediction of a possibility of negative refractive index and flat lenses based on negative index materials (NIMs) stimulated the development of the entire field of metamaterials. NIMs were first studied in detail by Russian physicist Victor Veselago in 1968 [1]. An important property of the wave propagation in NIMs is that wave and Poynting vectors (or phase and energy velocities) are antiparallel. Today, this property is often considered to be the most general definition of NIM. Using Maxwell's equations for the medium with negative dielectric permittivity and magnetic permeability (and, as a result, negative refractive index), Veselago predicted that the right-handed triplet of vectors **E** (electric field), **H** (magnetic field), and k (wave vector) in conventional, positive index material (PIM) changes to the left-handed triplet in NIM, leading to the negative refraction of a light beam [1]. In other words, when an EM wave propagates from PIMs to NIMs, negative refraction occurs at the interface, such that both incident and refracted waves are on the same side of the normal to the interface. To date, negative refraction was predicted and realized in three different physical systems: double-negative or "left-handed" metamaterials (with negative permittivity and permeability) [2−9], so-called hyperbolic metamaterials with anisotropic permittivity or permeability [10−13], and near the bandgap edge in photonic crystals [14−16].

9.2.1   Double-negative metamaterials

Following the original Veselago's work [1], we consider EM wave propagation in NIMs using the set of Maxwell equations. We assume a plane wave with the frequency ω is incident from air with refractive index $n_1 = 1$ onto the double-negative NIM with refractive index $n_2 < 0$. The electric field **E** and magnetic field **H** can be described by $\mathbf{E} = \mathbf{E}_0 \exp[\mathrm{i}(kr - \omega t)]$ and $\mathbf{H} = \mathbf{H}_0 \exp[\mathrm{i}(kr - \omega t)]$, respectively, where k is the wave vector. Substituting **E** and **H** fields into Maxwell's equations and taking into account the constitutive relations

$$\nabla \times \mathbf{E} = \mathrm{i}\frac{\omega\mu}{c}\mathbf{H}, \tag{9.1a}$$

$$\nabla \times \mathbf{H} = -\mathrm{i}\frac{\omega\varepsilon}{c}\mathbf{E}, \tag{9.1b}$$

$$D = \varepsilon\mathbf{E}, \tag{9.1c}$$

$$B = \mu\mathbf{H}, \tag{9.1d}$$

we obtain

$$k \times \mathbf{E} = \frac{\omega\mu}{c}\mathbf{H}, \tag{9.2a}$$

$$k \times \mathbf{H} = -\frac{\omega\varepsilon}{c}\mathbf{E}, \tag{9.2b}$$

where ε and μ are relative dielectric permittivity and magnetic permeability, respectively. From Eqns (9.2a) and (9.2b), one can see that $\mathbf{E}$, $\mathbf{H}$, and k form a left-handed triplet when both ε and μ are negative, which explains why the NIMs sometimes are referred to as "left-handed materials."

Because the Poynting vector is determined by $\mathbf{S} = \mathbf{E} \times \mathbf{H}^*$, the wave vector of the beam is unparallel to the direction of the Poynting vector in NIMs, which means that the directions of the phase velocity and the energy velocity are opposite to each other. The refractive index is the square root of the $\varepsilon\mu$ and, in this case, we take the negative value of the square root $n = \pm\sqrt{\varepsilon\mu}$. According to Snell's law, the refraction angle is also negative: $\sin\theta_r = \sin\theta_i/n$, which means that the incident beam and the refracted beam are on the same side of the normal.

Let us briefly discuss why one has to choose the negative value of the square root in $n = \pm\sqrt{\varepsilon\mu}$ when $\varepsilon < 0$ and $\mu < 0$. Assuming that both ε and μ are complex numbers, that $\varepsilon = \varepsilon_r + i\varepsilon_i$, and that $\mu = \mu_r + i\mu_i$, the refractive index can be written as $n = n_r + ik$. Then, the relation between them should be:

$$\begin{aligned}
n^2 = (n_r + ik)^2 &= (\varepsilon_r + i\varepsilon_i)(\mu_r + i\mu_i) \\
n_r^2 - k^2 + 2in_rk &= (\varepsilon_r\mu_r - \varepsilon_i\mu_i) + i(\varepsilon_r\mu_i + \varepsilon_i\mu_r) \\
n_r^2 - k^2 &= \varepsilon_r\mu_r - \varepsilon_i\mu_i \\
2n_rk &= \varepsilon_r\mu_i + \varepsilon_i\mu_r
\end{aligned} \tag{9.3}$$

where $\varepsilon_r < 0$, $\varepsilon_i > 0$ and $\mu_r < 0$, $\mu_i > 0$. Therefore, $(\varepsilon_r\mu_i + \varepsilon_i\mu_r)$ is negative. For materials with no gain, the extinction value k should be positive. According to $2n_rk = \varepsilon_r\mu_i + \varepsilon_i\mu_r < 0$, the value of n is negative under the condition of $\varepsilon_r < 0$, $\varepsilon_i > 0$, and $\mu_r < 0$, $\mu_i > 0$.

The phenomenon of negative refraction can also be explained using equi-frequency contours (EFCs). Let us consider a transverse magnetic (TM) plane wave with frequency ω and the wave vector k incident on an NIM with $\varepsilon < 0$ and $\mu < 0$, such that the wave vector k is along the zx-plane and the $\mathbf{H}$ field is along the y-axis. According to Eqn (9.2b), the $\mathbf{E}$ field inside of the NIM is given by

$$\mathbf{E} = \frac{c}{\omega\varepsilon}\left(k_z\mathbf{H}_0\widehat{u}_{xx} - k_x\mathbf{H}_0\widehat{u}_{zz}\right)\exp\left[i\left(k_x\widehat{u}_{xx}x + k_z\widehat{u}_{zz}z - \omega t\right)\right], \tag{9.4}$$

$$\mathbf{H} = \mathbf{H}_0\widehat{u}_{yy}\exp\left[i\left(k_x\widehat{u}_{xx}x + k_z\widehat{u}_{zz}z - \omega t\right)\right], \tag{9.5}$$

where $\widehat{u}_{xx}$, $\widehat{u}_{yy}$, and $\widehat{u}_{zz}$ are the unit vectors along the x-, y-, and z-axes, respectively. Then, substituting Eqns (9.4) and (9.5) into Eqn (9.2a), we obtain

$$\frac{c}{\omega\varepsilon}\left(k_z^2\mathbf{H}_0 + k_x^2\mathbf{H}_0\right)\widehat{u}_{yy}\exp\left[i\left(k_x\widehat{u}_{xx}x + k_z\widehat{u}_{zz}z - \omega t\right)\right]$$
$$= \frac{\omega\mu}{c}\mathbf{H}_0\widehat{u}_{yy}\exp\left[i\left(k_x\widehat{u}_{xx}x + k_z\widehat{u}_{zz}z - \omega t\right)\right] \qquad (9.6)$$

Finally, the dispersion relation for the EFC can be written as

$$\frac{k_z^2}{\varepsilon\mu} + \frac{k_x^2}{\varepsilon\mu} = \frac{\omega^2}{c^2} \qquad (9.7)$$

The lateral component of the wave vector k_z can be determined from Snell's law as $k_z = k_i\sin\theta_i = k_r\sin\theta_r$. Now, let us compare wave propagation in a PIM $(n_2 = \sqrt{\varepsilon\mu})$ and an NIM $(n_2 = -\sqrt{\varepsilon\mu})$.

The wave vector k_x^+ in the PIM is given by

$$k_x^+ = |n_2|\sqrt{\left(\frac{\omega^2}{c^2} - \frac{k_z^2}{\varepsilon\mu}\right)}. \qquad (9.8)$$

The wave vector k_x^- in the NIM is given by

$$k_x^- = -|n_2|\sqrt{\left(\frac{\omega^2}{c^2} - \frac{k_z^2}{\varepsilon\mu}\right)} = -k_x^+. \qquad (9.9)$$

The refraction angle of the wave vector $\theta_r = a\mathrm{tan}(k_x/k_z)$ in the NIM is negative $(\theta_r^- = -\theta_r^+)$ with respect to the PIM, as shown in Figure 9.1(a). In addition, the wave vector corresponding to the refracted wave $k_r^-(-k_x^+, k_z)$ in the NIM bends to the opposite side of the normal to the interface, as compared with that in the PIM.

The beam's refraction direction is determined by its Poynting vector.

From Eqns (9.8) and (9.9), we conclude that, in the PIM $(n_2, \varepsilon > 0)$, the Poynting vector's direction is the same with the wave vector's direction, whereas in the NIM $(n_2, \varepsilon < 0)$, the Poynting vector's direction is antiparallel to the direction of the wave vector, as shown in Figure 9.1(b). Therefore, if the energy velocity is pointing forward, then the phase velocity is pointing backward. Once again, this property is often taken as the most general definition of NIM.

As discussed above, one of the ways to realize NIM is to find or design a material with simultaneously negative permittivity and permeability. Although there is no fundamental physics law that would rule out such a combination of material parameters, materials with simultaneously negative ε and μ have not yet been found in nature. The magnetic susceptibility of the most natural materials is very small in comparison
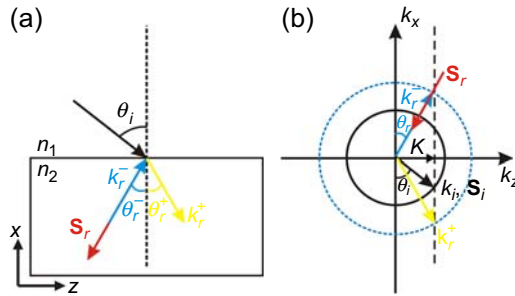
Figure 9.1 (a) Negative refraction in an NIM ($n_2 < 0$). The incident angle is θ_i and the refraction angle is θ_r. (b) The schematic of the EFC: the black circle is the EFC of air and the blue dot circle is the EFC of the NIM with negative refractive index. k_i and $\mathbf{S}_i$ are the wave vector and the Poynting vector of the incident beam, respectively. k_x^- and $\mathbf{S}_r$ are the wave vector and the Poynting vector of the refracted beam, respectively. k_x^+ is the wave vector of the refracted beam if the medium was the PIM $|n_2|$.

with the dielectric susceptibility, which limits the interaction of atoms to the electric component of the EM wave and leaves the magnetic component largely unexploited. As a result, μ is close to 1 for many naturally existing materials.

The reason for this difference in the strength of electric and magnetic field coupling to atoms is that the magnetization of any (nonferromagnetic) material is a relativistic effect of the order $\sim v^2/c^2 \sim \alpha^2 \ll 1$, where v is the velocity of the electrons in the atoms and $\alpha \cong 1/137$ is the fine-structure constant. The emergence of metamaterials offered a unique solution to this problem: each unit cell (the meta-atom) of a metamaterial can be engineered to facilitate electric and magnetic field coupling to such an artificial atom. Moreover, the meta-atoms can be designed to enable positive, negative, or near-zero ε and μ at any desirable frequency. With respect to NIMs, it should be noted that although materials with negative permittivity are not rare (e.g., metals below their plasma frequency), negative permeability is not commonly found in nature, especially at optical frequencies. Therefore, one of the first challenges to be solved using the metamaterials approach is the design of a meta-atom with negative permeability.

Let us start with a detailed discussion of how such a meta-atom can be realized. Figure 9.2(a) shows a so-called split-ring resonator (SRR) that enables magnetic response through the excitation of a circulating current, as follows [17].

The SRRs are places in the yz-plane; their dimensions—a (side length) and d (gap size)—are optimized to provide a resonant response at a particular frequency. Along the x direction, the distance between the neighboring SRR is l, and each column of SRR along x can be treated as a tightly wound solenoid, as shown in Figure 9.3. When an EM wave passes through the SRR array with its **H** field penetrating through the SRR, a circulating current I is induced in the SRR. The ring acts as a coil with certain inductance, and the split provides a certain capacitance. A single SRR can be described by an effective RLC circuit, where the resistance originates from the metal itself, as shown in Figure 9.2. The source in this circuit is the induced electromotive
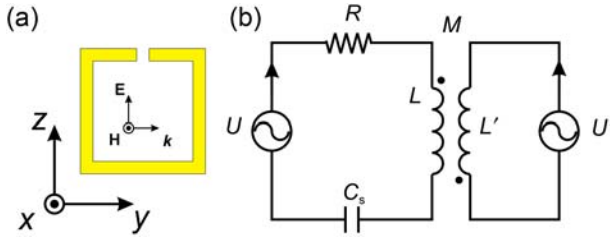
Figure 9.2 (a) The incident EM wave polarization with respect to the orientation of the SRR to excite the magnetic resonance. (b) The effective circuit of an SRR.
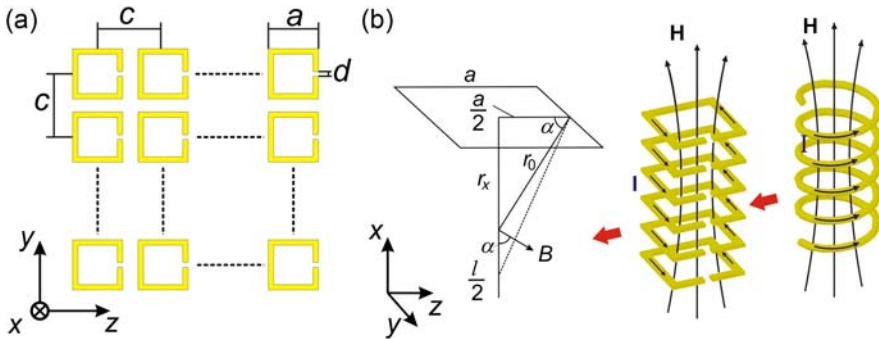
Figure 9.3 (a) SRR array top view: yz-plane. (b) One column (x direction) in the SRR array, which can be analyzed by the tightly wound solenoid model. The left inset shows the schematic of the B calculation.

force from the **H** field of the EM wave. Finally, a mutual inductance is added into this effective circuit to describe the neighboring SRRs' effects on an individual SRR. Considering all of these effects, the SRR can be represented by a circuit, as shown in Figure 9.2(b) [18].

First, let us use the tightly wound solenoid model to calculate the inductance of SRR in the effective circuit. According to the Biot–Savart law, the magnetic flux density that penetrates SRR is given by

$$
B = \int_{-\frac{l}{2}}^{\frac{l}{2}} \frac{\mu_0 I}{2\pi} \frac{a^2}{r_0^2 \sqrt{\left(\frac{a}{2}\right)^2 + r_0^2}} N \, \mathrm{d}r_x
\tag{9.10}
$$

where $r_0 = \sqrt{\left(\frac{a}{2}\right)^2 + r_x^2} = \frac{a}{2\cos\alpha}$, $r_x = r_0 \sin\alpha$. r_x is the distance to the SRR along the x axis. If there are N SRRs per unit length, then N should be $1/l$. By taking derivatives, we can obtain

$$
\mathrm{d}r_x = \frac{a}{2\cos^2\alpha} \mathrm{d}\alpha
\tag{9.11}
$$

By putting Eqn (9.11) into Eqn (9.10), the integral can be rewritten as

$$B = \int_{\alpha_1}^{\alpha_2} \frac{\mu_0 I}{\pi l} \frac{1}{\sqrt{\frac{1}{4} + \frac{1}{4\cos^2\alpha}}} d\alpha \qquad (9.12)$$

If we neglect the edge effect,[1] then the **H** field at the central axis of the SRR is constant. Then we calculate the integral from $\alpha_1 = -\pi/2$ to $\alpha_2 = \pi/2$. By using variable substitution

$$\beta = \frac{\pi}{2} - \alpha. \qquad (9.13)$$

Equation (9.12) can be transformed into

$$B = \frac{2\mu_0 I}{\pi l} \int_0^\pi \frac{1}{\sqrt{1 + \frac{1}{\sin^2\beta}}} d\beta, \qquad (9.14)$$

where $f(\beta) = \int_0^\pi \frac{1}{\sqrt{1 + \frac{1}{\sin^2\beta}}} d\beta$ is the elliptic integral of the first kind, which is equal to $\pi/2$. Therefore, the magnetic flux density in one SRR can be written as

$$B = \frac{\mu_0 I}{l}. \qquad (9.15)$$

The magnetic flux passing through the SRR is

$$\Phi = a^2 \mu_0 I / (l). \qquad (9.16)$$

From Eqn (9.16), we can obtain the inductance L as

$$L = \frac{a^2 \mu_0}{l}. \qquad (9.17)$$

All other neighboring SRRs will also have an effect on the target SRR, which can be described by mutual inductance M. If the total number of SRRs in the array at the yz-plane is n, then the magnetic flux induced by all of the neighboring SRRs will be

---

[1] The edge effect of the solenoid refers to the diverging magnetic field at the two ends of the solenoid. Here, we assume that the magnetic field is uniform, even near the ends, by neglecting the edge effect.

$\Phi d = (n-1)LI$. Then, the magnetic flux through the target SRR is given by $\Phi_L = (a^2/nc^2)\Phi_d$. Therefore, the mutual inductance M can be written as

$$M = \frac{\Phi_L}{I} = \lim_{n\to\infty} \frac{(a^2/nc^2)\Phi_d}{I} = \lim_{n\to\infty} \frac{a^2(n-1)L}{nc^2} = \frac{a^2}{c^2}L = A_m L, \quad (9.18)$$

where A_m is the filling ratio of SRR in one unit. Applying Kirchhoff's law to the equivalent circuit in Figure 9.2(b), we obtain

$$U = IR - \frac{I}{i\omega C_s} - i\omega LI - (-i\omega M)I. \quad (9.19)$$

Using Faraday's law, we can write U as

$$U = i\omega\mu_0 a^2 \mathbf{H}_0. \quad (9.20)$$

Then, the current in the SRR is given by

$$I = \frac{i\omega\mu_0 a^2 \mathbf{H}_0}{R - \frac{1}{i\omega C_s} - i\omega L - (-i\omega M)} = \frac{-\mathbf{H}_0 l}{1 - A_m - \frac{1}{\omega^2 LC_s} + i\frac{R}{\omega L}}. \quad (9.21)$$

The dipole moment of each SRR is

$$\overline{M}_d = \frac{1}{c^2 l}a^2 I = \frac{Ba^2}{\mu_0 c^2}. \quad (9.22)$$

According to $\overline{B} = \mathbf{H}_0\mu_0$, the effective permeability of the SRR can be written as

$$\mu_{eff} = \frac{\overline{B}/\mu_0}{\overline{B}/\mu_0 - \overline{M}_d} = 1 - \frac{A_m}{1 - \frac{1}{\omega^2 LC_s} + i\frac{R}{\omega L}}, \quad (9.23)$$

where the resonance frequency ω_0 and plasma frequency ω_{mp} are

$$\omega_{m0} = \frac{1}{\sqrt{LC_s}}, \quad \omega_{mp} = \frac{\omega_{m0}}{\sqrt{1-A_m}} = \frac{1}{\sqrt{LC_s(1-A_m)}}, \quad (9.24)$$

respectively.

Equation (9.23) can be simplified as

$$\mu_{eff} = 1 - \frac{A_m\omega^2}{\omega^2 - \omega_{m0}^2 + i\omega\gamma_m}, \quad (9.25)$$

which shows a Lorentz dispersion relation. In the resonance region, $\omega_0 \sim \omega_{mp}$, μ_{xx} is negative. Such engineered effective amagnetic permeability can be used in realizing the NIM.

We considered a square SRR as an example of how an "artificial magnetic meta-atom" enables magnetic resonance in a certain frequency range. Although the square-shaped SRR is just a particular example of many other SRRs' shapes reported in the literature [19−21], the basic idea enabling the magnetic resonance in all of those geometries is the same. By using the SRR structure, negative permeability was successfully realized in microwave and terahertz (THz) frequency ranges [2,17,22].

The negative permittivity can be achieved by thin metal wires, by which the effective mass of the electrons was greatly enhanced. This resulted in a much lower plasma frequency, so that the structure had a desired dielectric permittivity at microwave frequencies [23].

$$\varepsilon_{\text{eff}} = 1 - \frac{\omega_{\text{p}}^2}{\omega\left(\omega + i\varepsilon_0 a^2 \omega_{\text{p}}^2 \big/ \pi r^2 \sigma\right)} \tag{9.26}$$

By combining the SRR and the thin metal wires, the first NIM was realized in 2000 at the University of California−San Diego [2]. Negative refraction was demonstrated using the refraction angle scanning experiment shown in Figure 9.4 [9].

Next, we describe a so-called retrieval method based on the S parameters (Fresnel coefficients) that gives a reliable, direct way to describe the dispersion relations of the
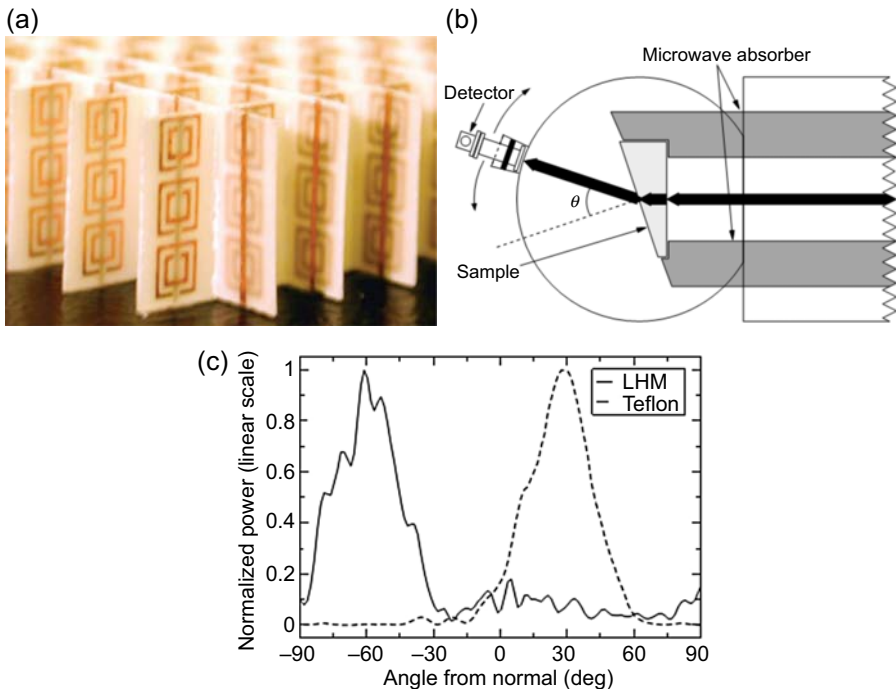
Figure 9.4 (a) The first NIM in the world. (b) The measurement setup of the negative refraction in the NIM in the microwave range and (c) the measurement results [9].

permittivity, permeability, and the refractive index [24–27]. According to this method, a metamaterial with its unit cell size much smaller than the working wavelength is considered to be an effective medium layer. Figure 9.5(a) shows a 2.5-mm cubic unit cell of the SRR and metal wire, which is the standard design of the NIM. Figure 9.5(b) and (c) shows the results of numerical simulation for the S parameters, as defined in Eqn (9.27). Considering that the sample is the metamaterial consisting of a single layer of the unit cell, S parameters including both magnitudes and phases can be written as

$$S_{11} = \frac{R_{01}\left(1 - e^{i2nk_0d}\right)}{1 - R_{01}^2 e^{i2nk_0d}},$$

$$(9.27)$$

$$S_{21} = \frac{\left(1 - R_{01}^2\right)e^{i2nk_0d}}{1 - R_{01}^2 e^{i2nk_0d}},$$
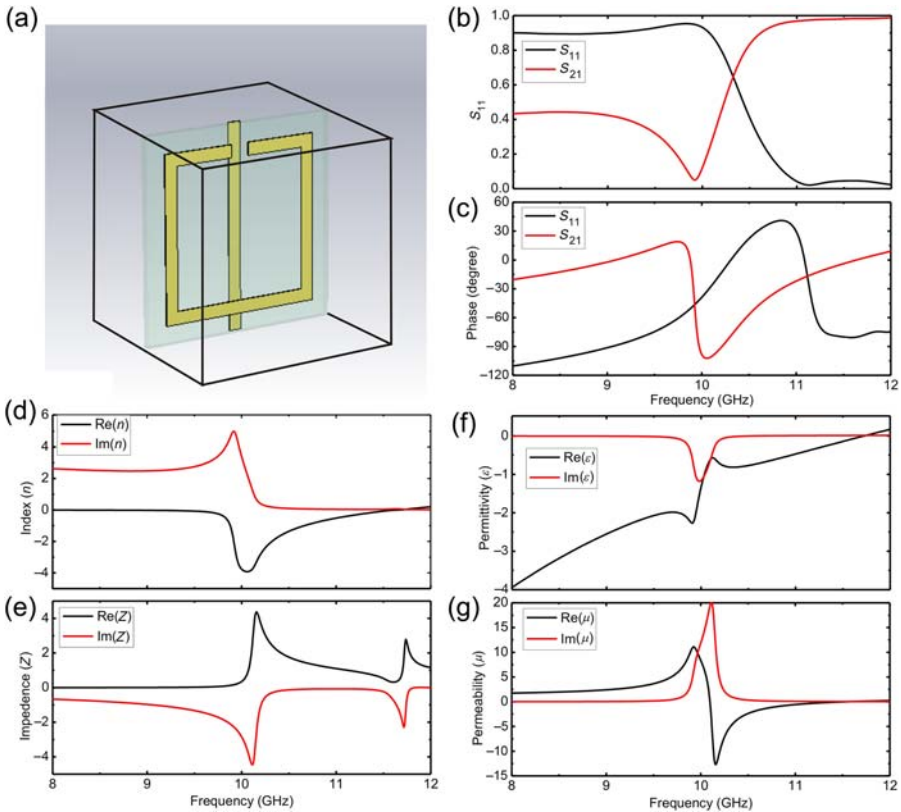
Figure 9.5 (a) The unit cell of the NIM consisting of an SRR and a metal wire. (b) Magnitudes of the S parameters. (c) Phases of the S parameters. (d–g) The computational results from the retrieval method: (d) effective refractive index of the metamaterial, (e) impedance of the metamaterial, (f) effective permittivity, and (g) effective permeability. The refractive index is negative at the frequency range where the negative permittivity range and the negative permeability range overlap [26].

where $R_{01} = \frac{z-1}{z+1}$. Then, the refractive index, impedance, permittivity, and permeability can be calculated by

$$z = \pm \sqrt{\frac{(1+S_{11})^2 - S_{21}^2}{(1-S_{11})^2 - S_{21}^2}},$$

$$e^{ink_0 d} = X \pm i\sqrt{1-X^2}$$

(9.28)

where $X = 1/2S_{21}(1 - S_{11}^2 + S_{21}^2)$. Figure 9.5(d–g) shows the results of the retrieval procedure. The designed plasma frequency is approximately 11.8 GHz. Below this frequency, the permittivity is negative, whereas the permeability exhibits a Lorentz resonance at approximately 10 GHz, so that the permeability is also negative. Therefore, the negative refractive index occurs in the frequency range of approximately 10 GHz, as shown in Figure 9.5(d).

In addition to the previously described SRR-metal wire-based NIM structure, other types of NIMs have also been demonstrated in the microwave frequency range. However, it was shown challenging to extend these ideas to the visible and near-infrared frequency range not only because of the difficulties of nanofabrication, but also because the properties of metals change significantly at high frequencies [28]. Therefore, alternative designs had to be proposed for metamaterials operating in the visible and infrared frequency ranges.

One of the first designs of NIMs at optical frequencies was based on the pairs of metal nanowires or metal plates [29–31]. Figure 9.6(a) illustrates the relation between normal SRRs and cut-wire pairs. As discussed above, the SRR can be viewed as an LC (or RLC) circuit, in which the inductance L originates from the ring and the capacitance C is provided by the split. Therefore, its resonance occurs near the resonant frequency of such LC circuit given by $\omega_{LC} = 1/\sqrt{LC}$. To further increase the resonance frequency, SRR is transformed into the U shape by opening the split, which may decrease capacitance C. Moreover, a second serial capacitance can be added to the circuit by
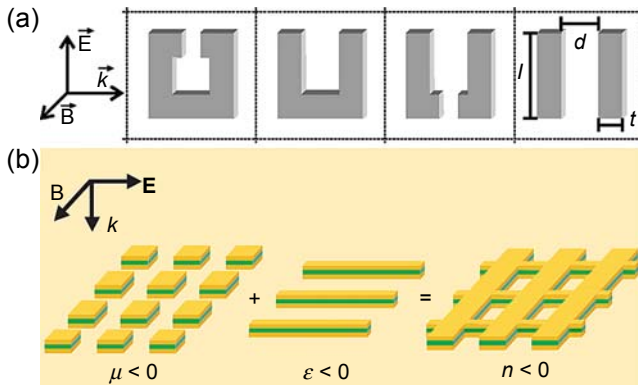
Figure 9.6 (a) Schematic of the transition from SRRs (left) to cut-wire pairs (right) as "magnetic atoms" for optical metamaterials [29]. (b) A combination of magnetic and electric meta-atoms results in a "fishnet" NIM structure operating at optical frequencies [7].
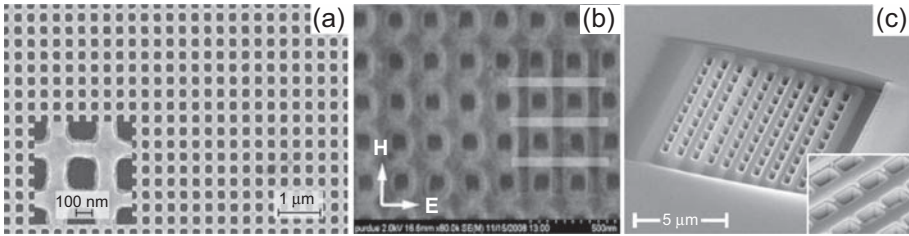
Figure 9.7 NIMs in optical frequency range: fishnet structure (a) Fishnet NIM working in 780 nm [4], (b) Fishnet NIM working in 570 nm [32] and (c) Fishnet prism working in infrared (around 1600 nm) [33].

breaking the bottom arm of the resultant U shape, which further reduces the net capacitance of the circuit. Further broadening the bottom split may also decrease the capacitance; finally, the SRR degenerate into a pair of cut wires. The magnetic resonance then originates from the antiparallel current in the wire pair with an opposite sign charge accumulating at the corresponding ends of the wires. This resonance provides $\mu < 0$. In addition, an electric resonance with $\varepsilon < 0$ results in excitation of a parallel current oscillation in the metal wires.

However, it was realized that achieving the overlap of the regions where ε and μ are both negative with only nanowire pairs was very challenging. Therefore, other structures were proposed and developed soon after the first experimental demonstration of the NIM.

One successful design is the so-called "double-fishnet" structure, which consists of a pair of metal fishnets separated by a dielectric spacer (Figure 9.7) [4,32,33]. To date, the fishnet structure has been realized as a negative refractive index material in both thin film (metal-dielectric-metal) and semibulk configurations [33]. Figure 9.8 shows the history of the NIM from microwave to optical range [34].
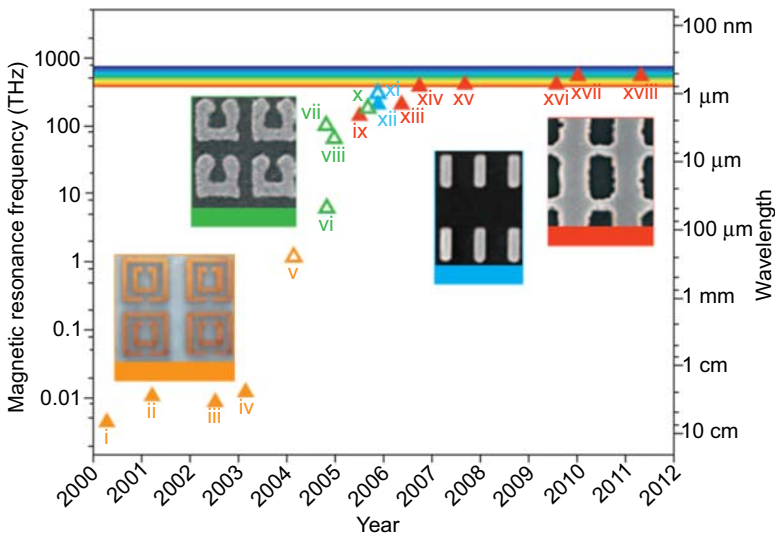
Figure 9.8 Development of the NIM [34].

In summary, NIMs were the first class of metamaterials studied in detail. Because the realization of negative index of refraction requires simultaneous negative permittivity and permeability, which cannot be found in nature, NIMs can only be built using so-called meta-atoms, or artificial structures that allow for overcoming the limitations imposed on light−matter interactions by natural atoms and molecules. The demonstration of NIMs not only gave rise to a new research direction in electromagnetics and materials science, but it also opened a possibility of realization of several unique applications.

9.2.2 Hyperbolic metamaterials

Hyperbolic materials, also called indefinite materials, are a kind of strongly anisotropic material. The term indefinite material refers to a medium whose the permittivity and permeability tensor elements (along principal axes) are of opposite signs, resulting in a strong anisotropy. As a result of such strong anisotropy, the EFC is hyperbolic [10], which can also give rise to an omnidirectional negative refraction for a certain polarization of the EM wave [11,13,35].

In the NIM, permittivity and permeability are simultaneously negative, which enables a negative refractive index. In this way, the k, $\mathbf{E}$, and $\mathbf{H}$ vectors of the EM wave are defined by the left-hand rule. Both the refractive vector k_r, $\mathbf{S}_r$ and the incident vector k_i, $\mathbf{S}_i$ are in the same side of the normal to the surface. The wave vector k_r is along $\mathbf{S}_r$, but with an opposite direction. If we only consider the energy negative refraction, which means only the refractive Poynting vector $\mathbf{S}_r$ is negatively refracted and is in the same side with the incident vector k_i, $\mathbf{S}_i$ whereas the refractive wave vector k_r shows a positive refraction, then the strict rule of the simultaneously negative permittivity and permeability is no longer needed. Anisotropic permittivity or permeability can enable the negative refraction.

As shown in Figure 9.9(a), when an EM wave transmits into an anisotropic medium, the electric-field vector $\mathbf{E}$ of the refractive light is usually not parallel to
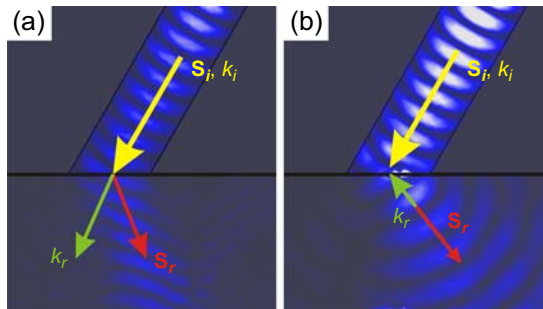
Figure 9.9 (a) Negative ray refraction in an anisotropic material. The refraction wave vector k_r is not parallel to the refraction Poynting vector $\mathbf{S}_r$, and the k, $\mathbf{E}$, and $\mathbf{H}$ vectors of the EM wave are defined by the right-hand rule. (b) Negative wave refraction: the wave vector k_r is along $\mathbf{S}_r$, yet with a conversed direction. The k, $\mathbf{E}$, and $\mathbf{H}$ vectors of the EM wave are defined by the left-hand rule [36].

the electric-displacement vector **D**. As a result, the Poynting vector, **S**$_r$, pointing in the direction of the energy flow, and the wave vector, k_r, directed along the wavefront normal, are not parallel except for the case of normal incidence. However, the phases of the waves in both media move with the same phase velocity along the interface so that the field is continuous over the interface. This requires the components of the wave vectors parallel to the interface to be equal in both media. Consequently, it is possible for the refracted beam to experience positive refraction with respect to k_r and negative refraction with respect to **S**$_r$. Therefore, the negative refraction in anisotropic materials is the negative (group) refraction of **S**$_r$, the direction of which is different from that of k_r. The wavefront propagating away from the interface and the vectors k, **E**, and **H** of the EM wave are determined by the right-hand rule.

In Section 9.2.1, we used EFC to demonstrate the negative refraction in the NIM with the isotropic negative refractive index. Similarly, we can also use EFC to analyze the refraction behavior in anisotropic media, which have anisotropic permittivity but isotropic permeability or anisotropic permeability and isotropic permittivity. The anisotropic items are generally described by tensors:

$$\tilde{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & 0 & 0 \\ 0 & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{pmatrix}, \tag{9.29}$$

$$\tilde{\mu} = \mu_0 \begin{pmatrix} \mu_{xx} & 0 & 0 \\ 0 & \mu_{yy} & 0 \\ 0 & 0 & \mu_{zz} \end{pmatrix}, \tag{9.30}$$

where the ε_{xx}, ε_{yy}, ε_{zz} or μ_{xx}, μ_{yy}, μ_{zz} are not equal to each other.

Let us first consider a nonmagnetic uniaxial crystal with its optics axis along the z-axis: $\varepsilon_{xx} = \varepsilon_{yy} \neq \varepsilon_{zz}$. A TM wave with frequency ω and the wave vector k is incident on the crystal. The general format of the **E** and **H** fields of the TM wave can be written as

$$\mathbf{E} = \mathbf{E}_0 \exp[i(kr - \omega t)] \tag{9.31}$$

$$\mathbf{H} = \mathbf{H}_0 \exp[i(kr - \omega t)]. \tag{9.32}$$

In the principal axis, suppose that the wave vector k lies in the zx-plane and the **H** is along the y-axis. According to Maxwell's equations, the TM wave in the crystal can be explicitly written as

$$\mathbf{E} = \frac{c}{\omega\varepsilon}\left(\frac{k_z \mathbf{H}_0}{\varepsilon_{xx}}\widehat{u}_{xx} - \frac{k_x \mathbf{H}_0}{\varepsilon_{zz}}\widehat{u}_{zz}\right)\exp\left[i\left(k_x\widehat{u}_{xx}x + k_z\widehat{u}_{zz}z - \omega t\right)\right], \tag{9.33}$$

$$\mathbf{H} = \mathbf{H}_0 \widehat{u}_{yy} \exp\left[i\left(k_x \widehat{u}_{xx} x + k_z \widehat{u}_{zz} z - \omega t\right)\right],$$ (9.34)

where $\widehat{u}_{xx}$, $\widehat{u}_{yy}$, and $\widehat{u}_{zz}$ are the unit vectors along the x-, y-, and z-axes, respectively. Then, we can write down the $\mathbf{S}_r$ as

$$\mathbf{S}_r = \frac{1}{2}\mathbf{E} \times \mathbf{H}^* = \frac{1}{2}\frac{c}{\omega}\left(\frac{k_x \mathbf{H}_0^2}{\varepsilon_{zz}}\widehat{u}_{xx} + \frac{k_z \mathbf{H}_0^2}{\varepsilon_{xx}}\widehat{u}_{zz}\right).$$ (9.35)

Thus, the EFC equation can also be derived as

$$\frac{k_x^2}{\varepsilon_{zz}} + \frac{k_z^2}{\varepsilon_{xx}} = \frac{\omega^2}{c^2},$$ (9.36)

which is a quadratic function. As shown by Eqns (9.35) and (9.36), $\mathbf{S}_r$ is parallel to the normal of the EFC.

A crystal with $\varepsilon_{xx} > 0$, $\varepsilon_{zz} > 0$, and $\varepsilon_{xx} \neq \varepsilon_{zz}$ has an EFC (Eqn (9.36)) that is elliptical (Figure 9.10(a)). For a required angle between the normal to the crystal surface and the optics axis, negative refraction can occur for certain incident angles. Accordingly, negative refraction can be observed in a birefringence crystal with a large difference between n_o (ordinary refraction index) and n_e (extraordinary refraction index) [37−39].

Materials with strong anisotropy have a hyperbolic EFC, induced by indefinite permittivity ($\varepsilon_{xx} < 0$, $\varepsilon_{zz} > 0$). Figure 9.10(b) shows the negative refraction resulting
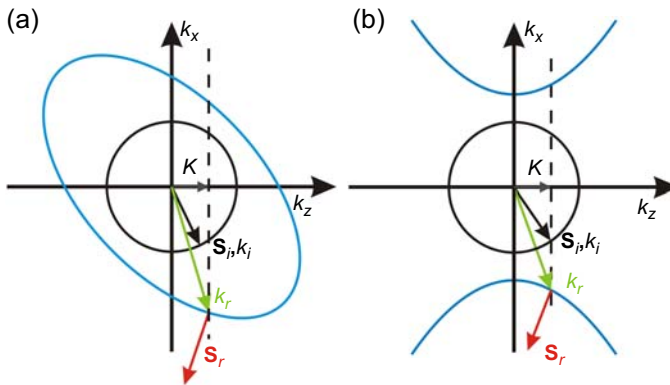
Figure 9.10 The scheme of negative refraction in anisotropic media. A light is incident onto the interface from the free space (black circle) and the uniaxial media (blue circle). The wave vector in both media is continuous along the interface ($k_{iz} = k_{rz} = K$). The Poynting vector ($\mathbf{S}_r$) is along the direction of the normal on the EFC. (a) The elliptical EFC of a birefringence crystal, resulting in a positive k_r refraction and negative $\mathbf{S}_r$ refraction for certain incident angles, under the condition of a proper angle between the optics axis and the surface. (b) The hyperbolic EFC of an indefinite medium, resulting in an all-angle negative refraction to $\mathbf{S}_r$ [36].

from the hyperbolic EFC. The normal to the hyperbolic EFC (S_r) and the incident ray (S_i) are on the same side of the normal to the medium surface for all incident angles. This is called the all-angle negative refraction.

Indefinite media are often discussed in a context of metamaterials, which achieve novel properties that are not attainable in nature using the designed artificial structures. Well-studied structures for indefinite metamaterials include binary composites (metal/dielectric or two semiconductors), such as those with a multilayered structure [13,40], a cylindrical multilayered structure [41,42], and a metallic wires array [11,35,43−45]. In these structures, the negative element of the permittivity tensor is obtained along the conductive layers or the wires, whereas the other permittivity tensor elements are always positive along the other perpendicular directions. In addition, artificially made metamaterials, such as a fishnet structure, have also been shown to enable indefinite permittivity and permeability [46] (Figure 9.11).

In 2007, Hoffman et al. [13] designed and fabricated an 8.1-μm-thick multilayered structure composed of interleaved $Al_{0.48}In_{0.52}As/In_{0.53}Ga_{0.47}As$ on an InP substrate. This multilayered structure enables strong anisotropic permittivity ($\varepsilon_\parallel \neq \varepsilon_\perp$): InGaAs at the free-carrier density (n_d) of 7.5×10^{18} cm^{-3} results in a plasma resonance that gives a negative $\varepsilon_\perp$ in the infrared range. Meanwhile, the $\varepsilon_\parallel$ at the corresponding frequency is positive (Figure 9.12). Consequently, Hoffman et al. observed infrared negative refraction behavior in the plasma resonance regime (8.8−11.8 μm).
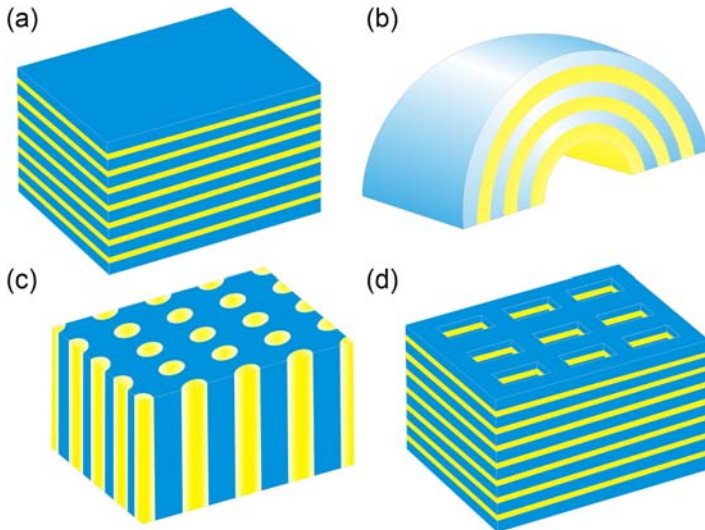
Figure 9.11 Examples of various hyperbolic metamaterials: (a) metal/dielectric multilayered structures, (b) cylindrical metal/dielectric multilayered structures, (c) metal wires array in a dielectric matrix, and (d) fishnet structure [51].
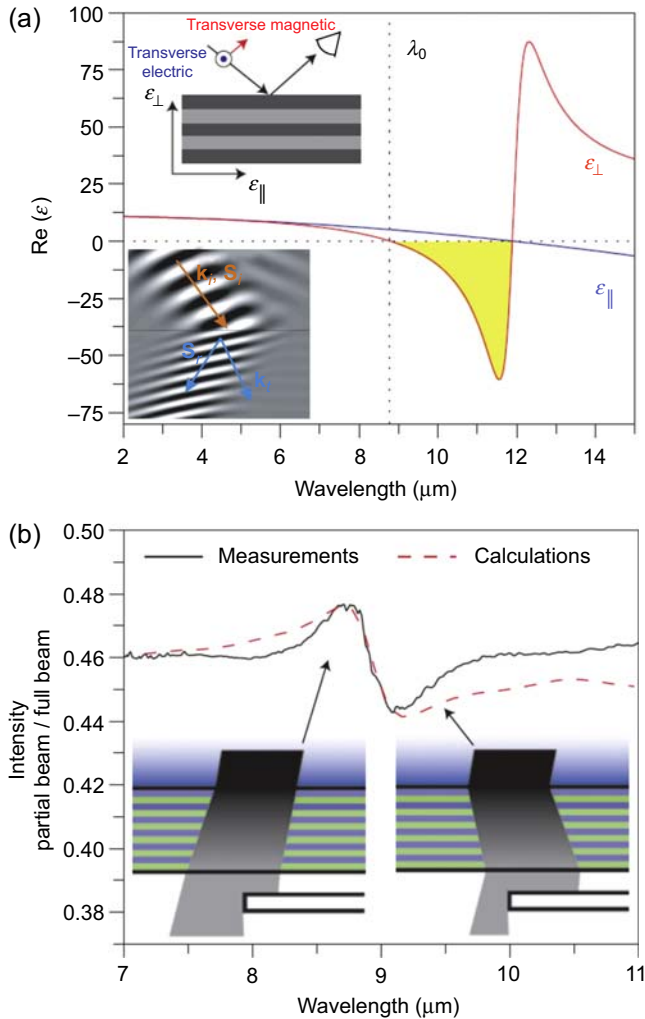
Figure 9.12 (a) Dielectric function, $\varepsilon_{\parallel}$ and $\varepsilon_{\perp}$, of a multilayered semiconductor structure with $n_d = 7.5 \times 10^{18}$ cm^{-3}. In the waveband of $\lambda < \lambda_0$, $\varepsilon_{\parallel} > 0$, and $\varepsilon_{\perp} > 0$, positive refraction may occur, whereas in the waveband of $\lambda > \lambda_0$, $\varepsilon_{\parallel} > 0$, and $\varepsilon_{\perp} < 0$, negative refraction occurs. The left top inset shows the relative orientation of the dielectric function ($\varepsilon_{\parallel}$ and $\varepsilon_{\perp}$), the polarization of the incident EM wave, and the layered structure. The left bottom inset shows the full numerical calculations of the negative refraction at the air−metamaterial interface.
(b) Experimental demonstration of negative refraction. The sample was placed with its InP substrate pointing upward. Behind the sample is a blade. TM polarized light was incident onto the InP substrate with an incident angle of $50°$. More light is measured for $\lambda < \lambda_0$ because the transmitted beam shifts away from the blade. In contrast, for $\lambda > \lambda_0$, less light is transmitted because the beam shifts behind the blade [13].

Zhang's group [11,35] fabricated a metamaterial of silver nanowire arrays in an anodic aluminum oxide (AAO) matrix. The metamaterial with a wire diameter of 60 nm and a center-to-center distance of 110 nm exhibited an anisotropic permittivity. $\varepsilon_{\parallel}$, parallel to the silver nanowire, is described by the Drude model with a plasma frequency (ω_p) of 1.5×10^{16} rad/s, and $\varepsilon_{\perp}$, perpendicular to the nanowire, exhibits a resonant behavior (due to localized surface plasmon polaritions) accompanied by a strong dispersion and a very large imaginary part [35]. Using the effective medium approximation model, the metamaterial of the AAO matrix filled with silver nanowires (filling ratio $= 0.27$) can be described by an indefinite dielectric tensor in the band above 500 nm. In experimental research, Zhang et al. fabricated two samples with thicknesses of 4.5 and 11 μm. A beam shifting experiment was used to demonstrate the negative refraction in the two samples by probing light at the wavelengths of 660 and 780 nm, respectively. Because the probe light with the wavelength (660 and 780 nm) is far away from the resonance wavelength of $\varepsilon_{\perp}$ (424 nm), the system has intrinsically low loss (Figure 9.13).

According to the symmetry of Maxwell's equations, anisotropic permeability can also produce a negative refraction by forming a hyperbolic EFC, which is a similar case as shown in Eqn (9.36). In most of the cases, materials with optical materials possess no magnetic response, so that $\mu = 1$. Therefore, one has to use metamaterial to generate a magnetic response to the **H** field in one orientation, whereas there is no response in other orientations. The SRR, which was first proposed by Pendry, is the most common design to achieve anisotropic permeability [17]. As discussed in Section 9.2.1, a resonant magnetic response can be produced when the **H** field of an EM wave penetrates the SRR loop. The permeability along the x direction can be approximated by the modified Lorentzian-oscillator function [47,48]:

$$\mu_{xx} = 1 - \frac{A_m\omega^2}{\omega^2 - \omega_{m0}^2 + i\omega\gamma_m}, \qquad (9.37)$$
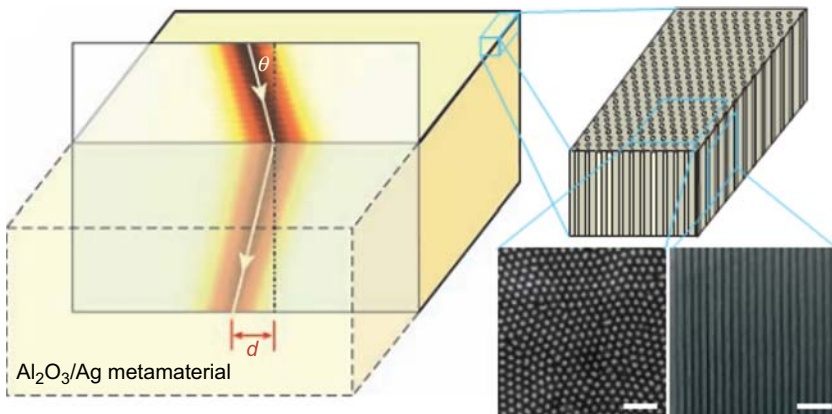
Figure 9.13 A visible beam experiencing a negative refraction when transmitted from air to a metamaterial composed of silver nanowire arrays. The right inset shows the scanning electron microscopy photograph of the front and facet views of the AAO matrix filled with silver nanowires [11].

where A_m is the oscillator amplitude, ω_{m0} is the resonance center frequency, and γ_m is the damping constant. When the **H** field of an EM wave is parallel to the loop of the SRR, no magnetic resonance is induced and thus, $\mu_{yy} = \mu_{zz} = 1$. At the frequency range from plasma frequency (ω_p) to ω_{m0}, the $\mu_{xx} < 0$ and $\mu_{yy} = \mu_{zz} = 1$. Therefore, the permeability of the SRR has an indefinite form:

$$\mu = \mu_0 \begin{pmatrix} \mu_{xx} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9.38}$$

According to Maxwell's equations, EFC can be written as

$$\frac{k_x^2}{\mu_{zz}} + \frac{k_z^2}{\mu_{xx}} = \frac{\omega^2}{c^2}\varepsilon. \tag{9.39}$$

Consequently, the magnetic metamaterial composed by SRR arrays exhibits strong anisotropic permeability in the resonant band, where the EFC is hyperbolic. This results in an all-angle negative refraction for the EM wave with a certain polarization [49,50] (Figure 9.14).

On the basis of this theory, negative refraction and slab focus behavior were observed in SRR array structures in the microwave range.

Some naturally occurring materials [51], such as graphite or graphite-like material (MgB$_2$) [52,53] and crystal with perovskite-layer crystal structure (cuprate and
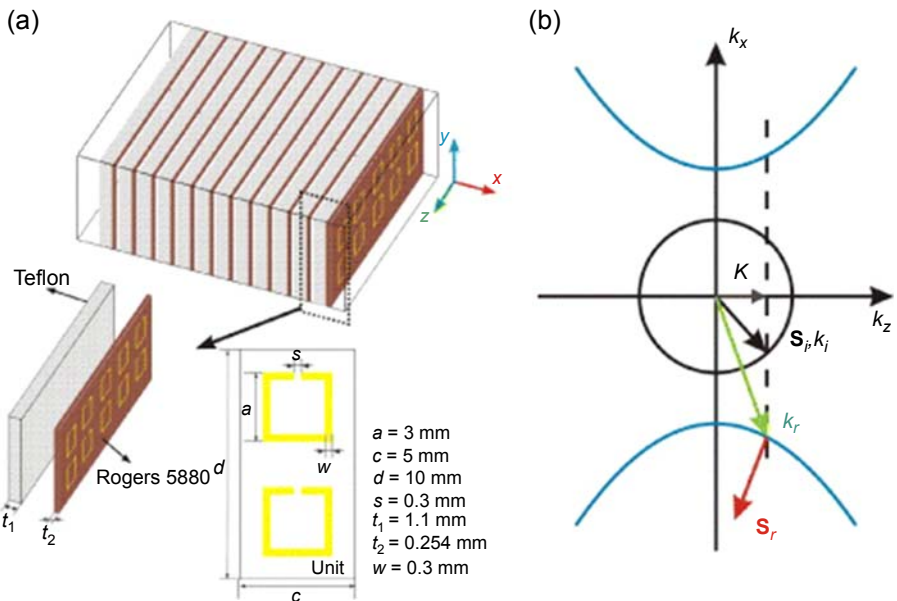
Figure 9.14 (a) An SRR array metamaterial with indefinite permeability to the EM wave with **H** field polarized in the xz-plane. (b) The EFC of the SRR array [49].

ruthenate) [54−56], as shown in Figure 9.15, possess very similar structure to that of the layered indefinite metamaterials; therefore, they possess the indefinite property resulting from this multilayered structure. As mentioned before, natural crystals with birefringence may show negative refraction in a limited condition because their anisotropy is not strong enough. In contrast to the birefringent crystals, these indefinite crystals always have two-dimensional conductivity mechanisms, which make them act as conductors or superconductors in macroscale. In the single crystalline state, the dielectric property in the direction along the conductive layer is entirely different from those in the direction perpendicular to the layer, which results in an indefinite dielectric tensor in the crystal.

Reference [51] provides a comprehensive review on the indefinite dielectric properties of materials existing in nature. The indefinite properties of these crystals can be demonstrated by characterizing their anisotropic dielectric spectra in different orientations, which can be done by the standard optical characterization methods, such as ellipsometry and reflectance spectra measurements. The measurement results of these materials listed in Figure 9.15 can also be found in the literature. A summary of the results reveals that the indefinite properties from these natural materials can cover the range from terahertz to ultraviolet frequencies, which unlocks nearly unlimited possibilities for device applications. On the other hand, as homogenous materials, intrinsic indefinite materials hold an advantage of no wave scattering caused by the inner structures in the artificially engineered materials and eliminate the need for the complicated design and fine fabrication techniques inevitable in manmade materials, thereby opening a new route to practical applications of their unique properties.

In summary, hyperbolic media are the most widely experimentally studied metamaterial configurations that have a strong potential for practical applications. Their indefinite dielectric or magnetic properties enable a hyperbolic EFC to have the all angle
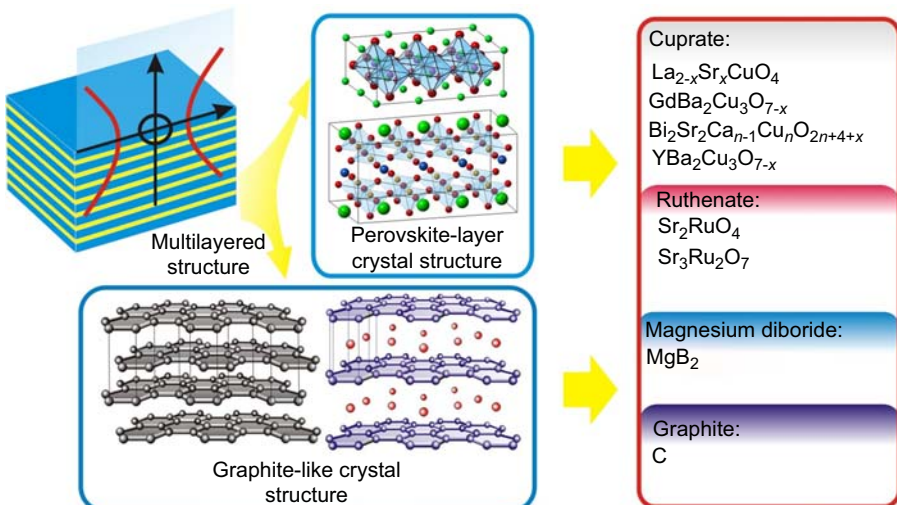
Figure 9.15 Indefinite materials in nature [51].

negative refraction. In addition, the hyperbolic EFC can also enable high density of states and imaging beyond diffraction limit using a so-called hyperlens, which will be introduced later in this chapter.

9.2.3   Zero-index materials

Recently, zero or near-zero refractive index metamaterials attracted significant attention because of their unique properties, which include a zero-phase delay. This delay implies a possibility of quasi-infinite phase velocity and infinite wavelength. The EM wave inside of the zero-index material is static in the spatial domain (i.e., the phase difference between any two arbitrary points is equal to zero). At the same time, it is dynamic in the time domain, thereby allowing energy transport. As a result, every point within the metamaterial experiences a quasiuniform phase; the shape of the wavefront at the output of such metamaterial depends only on the shape of the exit surfaces of the metamaterials. This property provides great flexibility in the design of the phase patterns of the light beams.

The concept of zero-index material was first proposed in 2002 as a method to show how a metamaterial allows us to control the direction of emission of a source located inside of the material to collect all of the energy in a small, angular domain around the normal [57]. According to Snell's law, $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$, $n_1 = 0$ results in the critical angle of the total reflection equal to 0. If the source is outside of the zero-index medium, then no light can be transmitted into the zero-index medium because of the total reflection. If the source is inside of the zero-index medium, no matter what the kind of source, then the direction of the output beam will be perpendicular to the surface of the zero-index medium.

This case can also be explained using the EFC. As shown in Figure 9.16, the blue circle corresponds to the EFC of the air and the red circle corresponds to the near zero index material, the radius of which is much smaller than that of the air EFC and will be zero for a zero-index material. As we discussed before, both the wave vector k_r and the Poynting vector $\mathbf{S}_r$ are defined by the EFC; one can find that the k_r and $\mathbf{S}_r$ are limited into a very narrow region, resulting in a refraction angle $\theta_r = 0$, as shown in
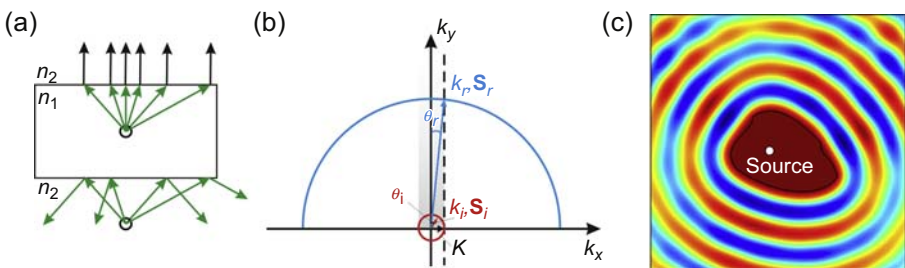
Figure 9.16 (a) Refraction of light in the near zero index material when source is placed inside or outside of the material. (b) The EFC corresponding to the near zero index material. (c) Illustration of the fact that the direction of emission is defined by the shape of the near zero index material if a source is placed inside.

Figure 9.16. Therefore, no matter what kind of source is placed inside of the near zero index material, light will be emitted out according to the shape of the zero-index material surface, as shown in Figure 9.16(c) [58].

From the equation $n = \sqrt{\varepsilon\mu}$, we know that $n = 0$ can be achieved by three ways: $\varepsilon = \mu = 0$; $\varepsilon = 0$, $\mu \neq 0$; or $\varepsilon \neq 0$, $\mu = 0$. Let us consider a plane wave propagating along the y direction with the $\mathbf{H}$ field pointing along the z direction in epsilon-zero material (ENZ; i.e., $\varepsilon = 0$, $\mu \neq 0$). From Maxwell's equations,

$$\nabla \times \mathbf{E} = i\frac{\omega\mu}{c}\mathbf{H}, \tag{9.40}$$

$$\nabla \times \mathbf{H} = -i\frac{\omega\varepsilon}{c}\mathbf{E}, \tag{9.41}$$

$\nabla \times \mathbf{H} = 0$ can be obtained when $\varepsilon = 0$, which means that the magnetic field in an ENZ material is constant. Then, by solving Eqn (9.40), we can find that

$$\frac{d\mathbf{E}_x}{dy} = \frac{\omega\mu}{c}\mathbf{H}_z, \tag{9.42}$$

which shows a linear change along its propagation direction.

Similarly, in $\varepsilon \neq 0$, $\mu = 0$ material, the electric field is constant, whereas the $\mathbf{H}$ field linearly changes along the propagation direction,

$$\frac{d\mathbf{H}_z}{dy} = -\frac{\omega\varepsilon}{c}\mathbf{E}_x. \tag{9.43}$$

For the $\varepsilon = \mu = 0$ case, both the $\mathbf{E}$ and $\mathbf{H}$ fields are constant inside of the zero-index material.

It is not easy to realize the zero-index material with both permittivity and permeability equal to zero. However, several realizations of zero-index material with either zero permittivity or zero permeability (with the other parameter non-zero) have been demonstrated.

The design strategy is similar to that used for the realization of the NIMs. For example, if a metal rod that has a Drude model dispersion is used to supply the negative permittivity in the microwave range, then one can take the near-zero permittivity around the plasma frequency of the Drude model and $\varepsilon = 0$ at the plasma frequency. Figure 9.17(a) shows a zero-index metamaterial with a mesh structure in the microwave range [57]. Similar to the utilization of the metal rod, the mesh structure was used to decrease the plasma frequency of the metal so that the zero-index frequency could be moved to the microwave range.

In the optical range, the plasma frequency of metals can be shifted by combining metals with dielectric materials. One example of this frequency shift can be seen in multilayered structures with alternating metal and dielectric layers, as shown in Figure 9.17(b). By using a proper ratio between the metal and the dielectric, the
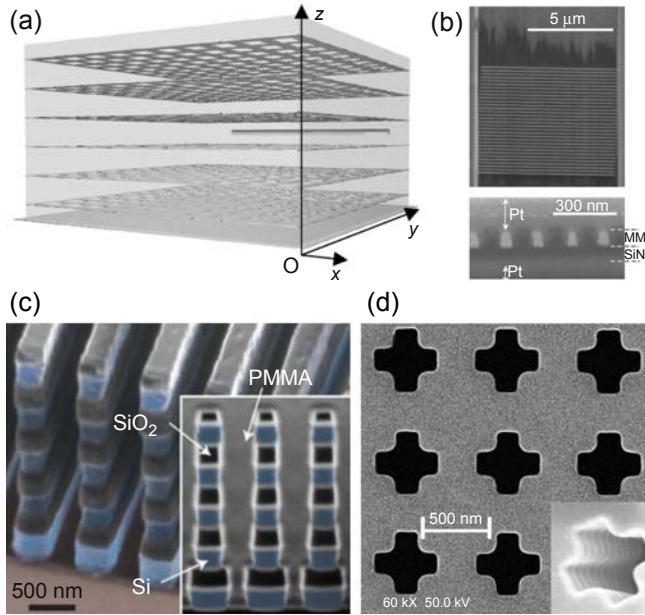
Figure 9.17 Different kinds of near zero index material: (a) ENZ material in the microwave range [57], (b) ENZ consisting of a multilayered structure in the optical range [59], (c) ENZ made of a woodpile structure in the infrared range [60], and (d) zero permeability material in the microwave range [61].

permittivity along the layers can be tuned to zero at visible wavelengths [59]. In addition to this, an all-dielectric woodpile structure can also enable zero-index material in the infrared range (Figure 9.17(c)) [60].

In the fishnet structure, there is also a frequency that causes refractive index changes from positive to negative. Therefore, the fishnet structure can also be used to compose a zero-index material with $n = 0$, as shown in Figure 9.17(d) [61].

The wavelength inside of the zero-index material is infinitively large; therefore, it enables many interesting properties, such as strong field enhancement [62], super coupling [63], phase front control [58], tunneling effect [64,65], and super cloaking [66].

Another interesting phenomenon enabled in zero or near zero index metamaterials is anomalous absorption near zero index transition in graded-index structures, which are so-called transition metamaterials with dielectric permittivity and magnetic permeability gradually changing among positive, zero, and negative values. The initial studies predicted strong, polarization-sensitive, anomalous field enhancement near the zero refractive index point under the oblique incidence of the plane wave on a realistic, lossy transition metamaterial layer, which would potentially enable various applications, including subwavelength transmission and low-intensity nonlinear optical devices [67−74].

A theoretical model describing the wave propagation in transition metamaterials can be derived from Maxwell's equations. Let us consider TM or TE polarized waves such that either magnetic or electric field is polarized along the y-axis and both μ and ε profiles are inhomogeneous with respect to x. Wave propagation is governed by the following equation:

$$\frac{\partial^2 \mathbf{H}_y}{\partial x^2} - \frac{1}{\varepsilon}\frac{\partial \varepsilon}{\partial x}\frac{\partial \mathbf{H}_y}{\partial x} + \frac{\omega^2}{c^2}\left(\varepsilon\mu - \sin^2\theta_0\right)\mathbf{H}_y = 0 \tag{9.44}$$

$$\frac{\partial^2 \mathbf{E}_y}{\partial x^2} - \frac{1}{\mu}\frac{\partial \mu}{\partial x}\frac{\partial \mathbf{E}_y}{\partial x} + \frac{\omega^2}{c^2}\left(\varepsilon\mu - \sin^2\theta_0\right)\mathbf{E}_y = 0. \tag{9.45}$$

Figure 9.18(a) shows the schematic of a transition layer between the PIM (left) and the NIM (right). The profiles have regions where both material parameters linearly decrease from a positive index region to a negative index region and two regions where the parameters are constant. Because there is a negative index region now beyond the point where the refractive index is zero, the wave that is transmitted beyond this point is a propagating wave (Figure 9.18(b)).

The dimensionless wave equation can then be written as

$$\frac{\partial^2 \Phi}{\partial \varsigma^2} - \frac{1}{\upsilon}\frac{\partial \upsilon}{\partial \varsigma}\frac{\partial \Phi}{\partial \varsigma} + \left(h^2 k_0^2\right)\left(\varepsilon\mu - \sin^2\theta_0\right)\Phi = 0, \tag{9.46}$$

where $k_0 = (\omega/c_0)\sqrt{\varepsilon_0\mu_0}$, $\varsigma = x/h$, $\upsilon = \varepsilon$, or μ for a TM or TE polarized wave, respectively. The incident field can be written in the form $\mathbf{H}_y(\mathbf{E}_y) = \Phi(x)\exp(ik_0\sin\theta_0 z)$, and the explicit forms of the profiles are

$$\varepsilon(x) = \varepsilon_0(1 - x/h), \quad \mu(x) = \mu_0(1 - x/h) \tag{9.47}$$

where the materials' parameter h is a measure of the transition gradient.

Consider the case of a TE polarized wave. In the case of normal incidence (i.e., when $\sin\theta_0 = 0$), the solution is simply given by

$$\Phi = C_1 \exp\left(\frac{-ik_0(1 - \varsigma)^2}{2}\right) + C_2 \exp\left(\frac{ik_0(1 - \varsigma)^2}{2}\right). \tag{9.48}$$
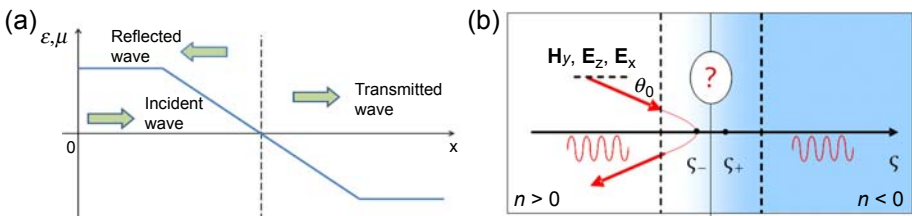
Figure 9.18 (a) Schematic of a transition metamaterials layer between PIM (left) and NIM (right). (b) Diagram of wave behavior in the transition layer with a linear graded profile.

From Eqn (9.48), it is clear that no unusual behavior of the electric and magnetic field components is expected at normal incidence; however, it was found that the wave propagation at oblique incidence is far more surprising.

First of all, from Eqn (9.46), it is clear that there are two points possible at which the wave can be totally internally reflected. These so-called turning points are given by

$$\varsigma_\pm = 1 \pm \sin\theta_0. \tag{9.49}$$

The analytical solution to Eqn (9.46) can be written in terms of a confluent hypergeometric function, U, that describes the behavior of the magnetic field within the entire layer as

$$
\begin{aligned}
\Phi(\varsigma) = {} & C_1 \exp\left(\frac{-ihk_0(1-\varsigma)^2}{2}\right) ihk_0(1-\varsigma)^2 \\
& \times U\left(1 - \frac{ih^2k_0^2\sin^2\theta_0}{4hk_0}, 2, ihk_0(1-\varsigma)^2\right) \\
& + C_2 \exp\left(\frac{ihk_0(1-\varsigma)^2}{2}\right) ihk_0(1-\varsigma)^2 \\
& \times U\left(1 + \frac{ih^2k_0^2\sin^2\theta_0}{4hk_0}, 2, -ihk_0(1-\varsigma)^2\right).
\end{aligned}
\tag{9.50}
$$

Near the singular point, the solution can be approximated as

$$
\begin{aligned}
\Phi(\varsigma) \approx {} & (C_1 - C_2)\left(1 + \frac{h^2k_0^2\sin^2\theta_0(1-\varsigma)^2}{2}\ln(hk_0(1-\varsigma))\right) \\
& - \frac{i}{2}(C_1 + C_2)\left(1 - \frac{\pi h^2k_0^2\sin^2\theta_0}{4hk_0}\right)hk_0(1-\varsigma)^2,
\end{aligned}
\tag{9.51}
$$

where the values of the coefficients C_1 and C_2 are obtained from boundary conditions. The first term in this expression has a logarithmic singularity. The second term corresponds to a regular solution of Eqn (9.46). The logarithmic term is well defined for $\varsigma < 1$; however, for $\varsigma > 1$, the definition of the logarithm is not unique and depends on the choice of the path around the zero-index point.

Following the approach used in Refs [67,75] and Maxwell's equations, we obtain

$$
\begin{aligned}
E_x = {} & \frac{ic}{\omega\varepsilon}\frac{\partial H_y}{\partial z} = -\frac{c}{\omega\varepsilon}k_0\sin\theta_0\Phi\exp(ik_0\sin\theta_0 z - \omega t) \\
= {} & -\sqrt{\frac{\varepsilon_0}{\mu_0}}H_0(1-\varsigma)^{-1}\sin\theta_0\exp(ik_0\sin\theta_0 z - \omega t) + O\left((1-\varsigma)^2\right), \tag{9.52}
\end{aligned}
$$

$$\mathbf{E}_z = -\frac{ic}{\omega\varepsilon}\frac{\partial \mathbf{H}_y}{\partial x} = -\frac{ic}{\omega\varepsilon}\frac{1}{h}\frac{\partial \Phi}{\partial \varsigma}\exp(ik_0\sin\theta_0 z)$$

$$= -ik_0 h\mu_0\mathbf{H}_0\sin^2\theta_0\ln(hk_0(1-\varsigma))\exp(ik_0\sin\theta_0 z - \omega t) + O\left((1-\varsigma)^2\right),$$

$$(9.53)$$

where the value of the constant $\mathbf{H}_0$ is the value of the field at $\varsigma = 1$. Although the y component of the magnetic field $\mathbf{H}_y$ is continuous, the x component of the electric field $\mathbf{E}_x$ is singular at $\varsigma = 1$; the z component $\mathbf{E}_z$ experiences a jump when the value of ε changes sign from positive to negative.

Figure 9.19 shows the results of full numerical simulations confirming the above predictions. The magnetic field forms a standing wave pattern in the PIM region because of the interference of incident and reflected waves. Beyond the turning point, the wave decreases as it approaches the singular point at $\varsigma = 1$. As it passes the singular point, it converts back into a propagating wave once again. The magnitude of the electric field component $\mathbf{E}_x$ shows the predicted resonant enhancement at the point $\varsigma = 1$, where it is predicted to be singular.

The most remarkable phenomenon occurring in transition metamaterials is anomalous absorption near the zero-index transition. To demonstrate this absorption, we calculate the difference of the longitudinal components of the Poynting vector (averaged over rapid field oscillations) before and after the transition of the wave through the point $n = 0$, as given by

$$\Delta\mathbf{S}_x = c\pi\mathbf{E}_0^2(h/\lambda)\varepsilon_0\sin^2\theta_0.$$

$$(9.54)$$

Note that although the model discussed so far assumed losses (imaginary parts of ε and μ) to be infinitesimally small, at the point $\varsigma = 1$ ($n = 0$), the real parts of ε and μ are zero; therefore, their contribution of the small, imaginary parts become significant and can no longer be neglected. Consequently, the dissipation of energy Q occurs due to these losses at $\varsigma = 1$ and $Q = \mathbf{S}_x$. As a result, the transition through the point where ε and μ change sign even with infinitesimal loss is accompanied by a finite dissipation of
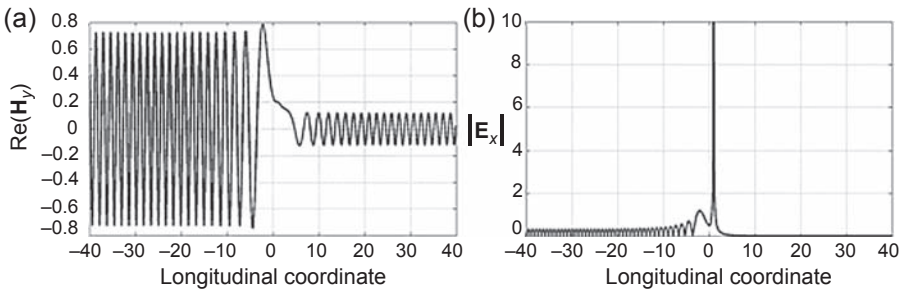
Figure 9.19 (a) Real part of the magnetic field component $\mathbf{H}_y$ as a function of longitudinal coordinate. (b) Absolute value of $\mathbf{E}_x$ as a function of longitudinal coordinate.

incident wave energy. The specific mechanism of dissipation would be determined by a physical model of a metamaterial.

These results may have significant implications for the design of particular applications of graded-index metamaterials. For example, with respect to perfect absorber application, a particular design could make use of a transition layer to enhance its absorptive properties. However, there are cases where graded index interfaces can make unwanted appearances on structures such as layered semiconductors, cloaking devices, and superlenses. In these structures, the possibility of resonant absorption if the index goes through zero could be decreased by properly designing the metamaterial structure.

In conclusion, most natural transparent optical materials possess a refractive index that is larger than one. The metamaterial approach provides a feasible way to create materials with low and even zero refractive index. Both zero-index metamaterials and transition metamaterials show very unusual properties. In the zero-index metamaterial, the wavelength can become infinitively large. If the source is inside of the zero-index medium, no matter what the kind of source, then the direction of the output beam will be perpendicular to the surface of the zero-index medium. On the other hand, it is hard for a wave to transmit into the zero-index material from the outside because of the total reflection. However, if the wave propagates in a graded-index material, such as the transition metamaterials, the strong field enhancement around the zero-index region is predicted.

9.3 Nonlinear effects in metamaterials

Nonlinear optics is a fascinating branch of science that investigates the light−matter interactions in media, in which dielectric polarization responds nonlinearly to the electric or magnetic field of the light [76−78]. Although this field has been developing for decades, the nonlinear optical materials available to date are still limited by either slow material response time in such phenomena as saturable absorption, photorefractive effect, and thermal nonlinear phenomena or by relatively low and generally band-limited nonlinear susceptibilities responsible for ultrafast nonlinear processes [79−81]. For decades, researchers have been exploring ways of creating materials with a large, fast, and broadband nonlinear response that, if found, would revolutionize nonlinear optics, leading to low-power, compact, and ultrafast applications of nonlinear optical phenomena. The emergence of metamaterials is likely to cause a paradigm shift in the development of such materials. Indeed, metamaterials were predicted to enable a plethora of novel linear and nonlinear light−matter interactions, including magnetic nonlinear response, backward phase-matching, and the nonlinear mirror, recently demonstrated at microwave frequencies [1,82−86]. In particular, it is expected that the nonlinear optical properties of metamaterials can be tailored and controlled at the level of individual meta-atoms such that the effective nonlinear susceptibility of the metamaterial exceeds those of the constituent materials. This expectation was also validated in the microwave frequency range [87,88]. Although these first steps prove the feasibility of realizing the theoretically

predicted remarkable nonlinear properties of metamaterials, the challenge of making nonlinear optical metamaterials competitive for practical applications is their high losses and narrow bandwidth. However, all-dielectric (silicon-based) metamaterials based on Mie resonances have recently been demonstrated to facilitate low loss and broadband performance, opening a practical route to the demonstration of strong nonlinear interactions in all-dielectric based metamaterials in the near-infrared frequency range, from frequency mixing and harmonics generation to bistability and optical switching [83].

Before we discuss some of the unique nonlinear phenomena in metamaterials, it would be useful to briefly review the key physics behind the nonlinear optical material response and resulting nonlinear light−matter interactions. Let us start with an introduction to the basics of light−matter interactions in different media. Any material can be considered as a collection of charged particles, electrons, and ions. In conductors, when an external electric field is applied, positive and negative charges move in the opposite directions. In dielectrics, the charged particles are bound together so that the charge cannot move freely in the externally applied field. Instead, they are displaced from their original positions and with respect to each other. This leads to the formation of induced dipole moments. In the regime of linear optics, the oscillation of electrons is proportional to the strength of the electric field of light. However, the nonlinear response is related to the anharmonic motion of bound electrons in the presence of an electric field.

The position of electrons is governed by the oscillator equation,

$$e\mathbf{E} = m\left(\frac{d^2x}{dt^2}\right) + 2\Gamma\left(\frac{dx}{dt}\right) + \Omega^2 x - \left(\xi^{(3)}x^2 + \xi^{(3)}x^3 + \ldots\right) \tag{9.55}$$

Here, x is the displacement from the mean position, Ω is the resonance frequency, and Γ is the damping constant. The term from the right-hand side represents the force on the electron due to applied field, which leads to the driving oscillations.

Considering just the harmonic term, we can thereby get an equivalent equation for x as follows:

$$\mathbf{E}(t) = \mathbf{E}_0 \cos(\omega t) = \frac{1}{2}\mathbf{E}_0\left[e^{i\omega t} + e^{-i\omega t}\right] \tag{9.56}$$

$$x = -\frac{e\mathbf{E}}{2m}\frac{e^{-i\omega t}}{\Omega^2 - 2i\Gamma\omega - \omega^2} + \text{c.c.} \tag{9.57}$$

If N is the number of electric dipoles per unit volume, the polarization induced in the medium is given by $P = -N_{ex}$. Thus, based on x, we can find the polarization P for this system as

$$P = -\frac{\chi\varepsilon_0\mathbf{E}}{2}e^{-i\omega t} + \text{c.c.}, \tag{9.58}$$

where χ is the susceptibility.

The motion of charged particles in a dielectric medium is linear in a limited range of values of **E**. However, as the intensity increases, then the response becomes nonlinear, and this nonlinearity can be accounted for with the anharmonic terms in the oscillator equation, Eqn (9.55). The polarization P (or magnetization M) in such a case will be expressed based on an anharmonic term as

$$P = \varepsilon_0 \left(\chi^{(1)} \mathbf{E} + \chi^{(2)} \mathbf{E}^2 + \chi^{(3)} \mathbf{E}^3 + \dots \right), \tag{9.59}$$

where $\chi^{(1)}$ represents a linear susceptibility and the higher order terms are the nonlinear susceptibilities of the medium.

Recently, an analytical description of a nonlinear metamaterial in terms of effective nonlinear susceptibilities based on a perturbative solution to the nonlinear oscillator model was proposed in Refs [89,90]. An application of this approach for a varactor-loaded split-ring resonator (VLSRR) medium can be summarized as follows.

Figure 9.20 shows the orientation of the VLSRR-based unit cell with respect to the incident field and its equivalent representation as an effective *RLC* circuit. The response of the SRR can be expressed in terms of the charge across the capacitive gap, which is described by the following nonlinear oscillator equation:

$$\ddot{q} + \gamma \dot{q} + \omega_0^2 V_D(q) = -\omega_0^2 A \mu_0 \dot{\mathbf{H}}_y \tag{9.60}$$

where $q(t)$ is the normalized charge, $V_D(q)$ is the voltage across the effective capacitance, ω_0 is the linear resonant frequency of the unit cell, γ is the damping factor, A is the area of the circuit, μ_0 is the permeability of vacuum, and $\mathbf{H}_y(t)$ is the incident magnetic field.

The voltage V_D can be expanded in a Taylor series in terms of the normalized charge according to $V_D(q) = q + aq^2 + bq^3$, where the Taylor coefficients a and b depend on the particular mechanism of nonlinearity. The perturbation solution to Eqn (9.57) leads to the following expressions [89] for the linear and the second-order effective susceptibilities:

$$\chi_y^{(1)}(\omega) = \frac{F\omega^2}{D(\omega)} \tag{9.61}$$
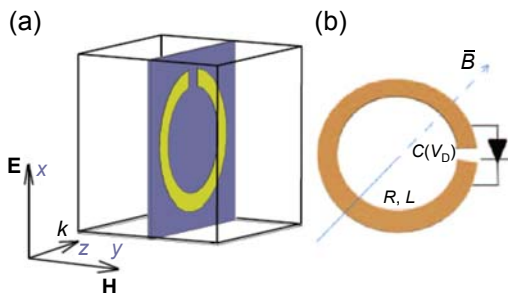
Figure 9.20 (a) Configuration of the unit cell. (b) Equivalent effective circuit model [90].

$$\chi^{(2)}_{yyy}(\omega_r; \omega_n, \omega_m) \ = \ -ia \frac{\omega_0^4(\omega_n + \omega_m)\omega_n\omega_m\mu_0 AF}{D(\omega_n)D(\omega_m)D(\omega_n + \omega_m)}, \qquad (9.62)$$

where $\omega_r \equiv \omega_n + \omega_m$, indices n and m take values between $\pm\Lambda$, and Λ is the total number of distinct waves incident on the medium. The denominator is defined as $D(\omega) \equiv \omega_0^2 - \omega^2 - i\gamma\omega$, $F \equiv \omega_0^2 NA^2 C_0\mu_0$ is the amplitude factor in the expression for the linear susceptibility, and the arguments of the nonlinear susceptibility are written in conventional notations signifying that the first term is the sum of the subsequent arguments.

Next, the microscopic equation of motion (Eqn 9.57) for a single inclusion can be converted into the macroscopic one for the effective medium polarization that, in combination with Maxwell's equations, provides a complete description of the metamaterials. For a dilute medium, the magnetization can be written as $M(t) = Nm(t)$, where N is the volume density of moments and $m(t)$ is the magnetic-dipole moment of the effective circuit enclosing the effective area S. Then, the equation for the magnetization can be written as

$$\ddot{M}_y + \gamma\dot{M}_y + \omega_0^2 M_y \ = \ -F\ddot{H}_y - \alpha M_y \int M_y \mathrm{d}t, \qquad (9.63)$$

where $\alpha \equiv \frac{2\omega_0^2 a}{NAC_0}$, assuming that only a second-order nonlinear response in the expansion of V_D was included.

It should be noted that $\chi^{(1)}$ in Eqn 9.56 is, in general, a tensor element. The terms $\chi^{(2)}$ and $\chi^{(3)}$ are usually relatively small; the overall impact of these terms on the optical response of the material to light is insignificant. Thus, these terms were mostly ignored until the advent of ultrafast lasers that could produce optical pulses by which the electric fields could be generated with sufficiently large intensity so that these terms become important. In fact, these terms have important consequences in optical systems. For instance, $\chi^{(2)}$, the second-order nonlinear susceptibility, only occurs in non-centrosymmetric crystals and gives rise to several nonlinear effects, including second-harmonic generation (SHG), optical rectification, sum and difference frequency generation, and parametric amplification. However, $\chi^{(3)}$, the third-order nonlinear susceptibility, occurs in all materials and gives rise to such phenomena as self-phase modulation, spatial solitons, cross-phase modulation, optical-phase conjugation, self-focusing through the optical Kerr effect, third-harmonic generation, and temporal solitons. Thus, remarkable optical properties have been demonstrated through nonlinear optics through the use of ultrafast pulsed lasers. These nonlinear optical processes, once a subject of scientific curiosity, are prevalent in everyday devices such as solid-state green and blue lasers, multiphoton microscopes, and the supercontinuum light sources used in metrology applications.

However, the observation of nonlinear optical processes in naturally occurring materials has relied on high-intensity laser beams. There have been some attempts to tailor the response in these materials by manipulating the structure of the material. For example, quasi-phase matching in SHG [91] used layered structures that effectively

increase the interaction lengths by correcting the relative phase. In an analogous manner, nanostructured optical metamaterials have the potential to increase the nonlinear response by modifying the material structure to enhance the interaction between light and matter. Recently, there have been several reports related to wave-mixing in metamaterials using the second-order nonlinearity. Enhancement of SHG has been demonstrated by several groups [92−97]. Husu et al. recently demonstrated that the second-order nonlinear response can be tailored by controlling the interactions between metal nanoparticles [98]. Wegener's group used nonlinear optical spectroscopy to demonstrate that the fundamental SRR in split rings serves as the nonlinear source [99]. In addition, novel propagation dynamics have also been realized. SHG in negative index metamaterials has been demonstrated to result in backward propagation of the second-harmonic light back toward the source [100−104]. Smith's group has demonstrated a nonlinear-optical mirror in which the SHG in NIM is generated in the backward direction (i.e., toward the source) [86]. In addition, there has been an interest in coherent optical amplification [105].

In addition to second-order nonlinearities, there has been continued interest in third-order nonlinearities. Four-wave mixing in NIMs has been investigated [106]. Soliton propagation in nonlinear and NIMs has demonstrated new propagation dynamics [105−116]. Katko et al. have demonstrated phase conjugation using SRRs [117]. Modulation instability in structures with a saturable nonlinearity has also been demonstrated [118,119].

The nonlinear properties of conventional materials are limited by the properties of naturally existing materials that are determined by their constituent components—atoms and molecules. The rapidly growing field of metamaterials opens unprecedented opportunities to overcome those limitations. It was previously shown that the unique optical properties of metamaterials, such as magnetism at optical frequencies, negative index of refraction, backward waves, strong anisotropy, or chirality, enabled several new regimes of light−matter interaction in the linear optical regime. However, it became clear in recent years that the nonlinear properties of metamaterials can be engineered as well, which opens new perspectives for modern nonlinear optics and enables devices with entirely novel functionalities [120].

The development of nonlinear metamaterials is likely to affect the nonlinear optics and metamaterials fields by enabling entirely new phenomena [121]. In particular, the novel optical properties facilitated by the metamaterials field prompt us to reconsider many nonlinear light−matter interactions, including SHG, parametrical amplification, ultra-short pulse dynamics, and soliton propagation [122−128]. From an applications viewpoint, nonlinear metamaterials make possible creating dynamically tunable materials, optical switches, filters, beam deflectors, focusing/defocusing reflectors, and reconfigurable structures with controlled transparency, refractive index, and nonlinear response.

In natural materials, nonlinear optical response conventionally depends on the intensity of the electric field and is described only by the nonlinear properties of dielectric permittivity, whereas the nonlinear magnetic response is neglected. The metamaterials can "turn on" the interaction with the magnetic component of the EM wave using carefully engineered linear and nonlinear components of magnetic

permeability [129]. It was recently shown that metamaterials containing SRRs loaded with varactor possess a nonlinear magnetic response and can be used as building blocks for nonlinear metamaterials with light-controllable properties. This approach provides a range of new functionalities, such as the focusing, defocusing, and deflection of light and, most importantly, the possibility of controlling and manipulating the properties of materials with unprecedented speed, enabling new approaches to all-optical processing, tunable lenses and waveguides, and reconfigurable cloaking devices.

Nonlinear wave-mixing is one of the best-studied nonlinear processes in metamaterials [86,99−106,130,131]. In particular, many new regimes of nonlinear interactions were predicted in NIMs, including backward phase-matching, unconventional Manley−Rowe relations, spatially distributed nonlinear feedback, and cavity-less optical parametrical oscillations. These unusual properties potentially enable such novel functionalities as quadratic nonlinear mirror and SHG-based lenses.

Here we consider SHG as one of the most fundamental parametric nonlinear processes. The material is assumed to possess a negative refractive index at the fundamental field (FF) frequency ω_1 and a positive refractive index at the second harmonic (SH) frequency $\omega_2 = 2\omega_1$. The electric fields of the FF and that of the SH waves are taken in the following form: $\overline{\mathbf{E}}_{1,2} = \mathbf{E}_{1,2} \exp(ik_{1,2}z - i\omega_{1,2}t) + \text{c.c.}$, respectively. Following Ref. [132], comprehensive studies of the SHG process for continuous wave and pulse propagation, as well as the propagation of complex light beams such as optical vortices or solitons with transverse field distribution, can be performed using the following model:

$$\nabla^2_\perp \mathbf{E}_1 + 2ik_1 \frac{\partial \mathbf{E}_1}{\partial z} + \frac{2ik_1}{v_1^g} \frac{\partial \mathbf{E}_1}{\partial z} - \frac{1}{c^2} \left(\frac{\omega_1 \mu_1 \alpha_1'}{2} + \gamma_1 \alpha_1 + \frac{\gamma_1' \omega_1 \varepsilon_1}{2} \right) \frac{\partial^2 \mathbf{E}_1}{\partial t^2}$$

$$+ \frac{8\pi \omega_1^2 \mu_1 \chi^{(2)} \mathbf{E}_1^* \mathbf{E}_2 e^{i\Delta kz}}{c^2} = 0$$

$$(9.64)$$

$$\nabla^2_\perp \mathbf{E}_2 + 2ik_2 \frac{\partial \mathbf{E}_2}{\partial z} + \frac{2ik_2}{v_2^g} \frac{\partial \mathbf{E}_2}{\partial z} - \frac{1}{c^2} \left(\frac{\omega_2 \mu_2 \alpha_2'}{2} + \gamma_2 \alpha_2 + \frac{\gamma_2' \omega_2 \varepsilon_2}{2} \right) \frac{\partial^2 \mathbf{E}_2}{\partial t^2}$$

$$+ \frac{4\pi \omega_2^2 \mu_2 \chi^{(2)} \mathbf{E}_1^2 e^{-i\Delta kz}}{c^2} = 0$$

$$(9.65)$$

where $\nabla^2_\perp$ is the transverse Laplace operator; $\chi^{(2)}$ is the effective second-order nonlinear susceptibility; $\alpha_{1,2}' = \frac{\partial[\omega_{1,2}\varepsilon(\omega_{1,2})]}{\partial\omega^2}$, $\alpha_{1,2}' = \frac{\partial^2[\omega_{1,2}\varepsilon(\omega_{1,2})]}{\partial\omega^2}$, $\gamma_{1,2} = \frac{\partial[\omega_{1,2}\mu(\omega_{1,2})]}{\partial\omega}$, $\gamma_{1,2}' = \frac{\partial^2[\omega_{1,2}\mu(\omega_{1,2})]}{\partial\omega^2}$, $\varepsilon_{1,2}$, and $\mu_{1,2}$ are dielectric permittivity and magnetic permeability for fundamental and SH waves, respectively; $v_{1,2}^g$ is the group velocities of pump and SH waves; $\Delta k = k_2 - 2k_1$ is phase mismatch; and c is the speed of light in free space. Note that Eqns (9.61) and (9.62) describe SHG for both positive and NIMs, but the signs for the wave vectors $k_{1,2}$, magnetic permeabilities $\mu_{1,2}$, dielectric permittivities $\varepsilon_{1,2}$, and group

velocities $v_{1,2}^g$ should be chosen according to the phase-matching geometry and material properties at the corresponding frequency.

For simplicity, we consider the steady-state process with perfect phase matching for plane waves, assuming that dispersion for permeability and permittivity are negligible. For the SHG process in an NIM slab with a finite length L, decomposing the complex amplitudes into real amplitudes and phases as $\mathbf{E}_{1,2} = \mathbf{E}_{1,2}e^{i\varphi_{1,2}}$ and using the boundary conditions $\mathbf{E}_1(z = 0) = \mathbf{E}_1^0$, $\mathbf{E}_2(z = L) = 0$, the solution can be written in the following form:

$$\mathbf{E}_1(z) = \frac{C}{\cos[Cg(L-z)]}, \tag{9.66}$$

$$\mathbf{E}_2(z) = C\tan[Cg(L-z)] \tag{9.67}$$

where $CgL = \cos^{-1}[C/\mathbf{E}_1^0]$, $g = \frac{4\pi\omega^2\mu_2}{c_2k_2}\chi^{(2)}$.

Figure 9.21(a) illustrates the directions of wave vectors and Poynting vectors for the FF and SH in a metamaterials that possesses NIM characteristics at the FF frequency and PIM characteristics at the SH frequency. This is due to the backward phase-matching in a PIM−NIM system, signifying that the energy flows of FF and SH waves are counterdirectional with respect to each other. Figure 9.21(b) shows the field distributions of the FF and SH waves. Importantly, the difference between energy flows is constant at each point of such metamaterial slab, whereas in conventional materials, the sum of the energy flows is unchanged with propagating distance. Furthermore, as could be seen from Eqn (9.58), the intensity of the SHG is a function of the slab thickness, such that a long enough lossless metamaterials slab would act as a 100% nonlinear mirror.

Experimental realization of propagating SHG at microwave wavelengths was demonstrated using the experimental setup shown in Figure 9.21(c) [86]. The nonlinear metamaterial was realized using the VLSRRs placed in an aluminum waveguide. SHG was studied in three configurations, including the reflected SH phase matching in a negative-index spectral range, transmitted SH quasiphase matching, and simultaneous quasiphase matching of both the reflected and transmitted SH waves near a zero-index spectral range. In addition, experimental measurements of three- and four-wave mixing phenomena in an artificially structured nonlinear magnetic metacrystal at microwave frequencies have been reported [133].
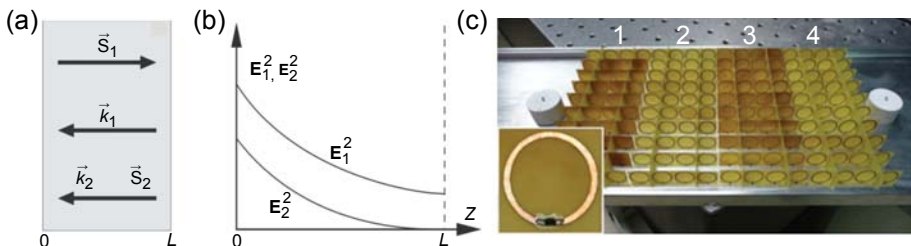
Figure 9.21 The SHG process: (a) phase-matching geometry, (b) FF and second-harmonic intensity distribution [123], and (c) photo of experimental setup for observing SHG in NIM [86].

Observation of SHG and third-harmonic generation at optical wavelengths was inspired by the theoretical work by Pendry, who predicted that SRRs would provide artificial magnetism up to optical frequencies [134]. Essentially, the light field can induce a circulating current in the ring leading to a large magnetic-dipole moment close to the magnetic-resonance frequency. In addition, it was predicted that metamaterials composed of SRRs would enable enhanced nonlinear-optical effects due to the combination of resonance effects and local-field enhancements. First experiments on SHG and third-harmonic generation at a surface of a nonlinear magnetic metamaterial composed of nanoscale gold SRRs were performed by Wegener's group and demonstrated that much larger signals are detected when magnetic-dipole resonances are excited than when purely electric-dipole resonances exist [97,99,135−137].

In parallel with the studies for continuous wave nonlinear optical effects, several fascinating phenomena have been discovered in the pulsed SHG regime [125−127]. In particular, it has been shown that in positive index, nonmetallic materials, we generally find qualitative agreement with previous reports regarding the presence of a double-peaked second-harmonic signal, which comprises a pulse that walks off and propagates at the nominal group velocity one expects at the second-harmonic frequency and a second pulse that is "captured" and propagates under the pump pulse. The origin of the double-peaked structure resides in a phase-locking mechanism that characterizes not only SHG but also $\chi^{(3)}$ processes and third-harmonic generation. A similar phase-locking phenomenon in NIMs. A spectral analysis of the pump and the generated signals reveals that the phase-locking phenomenon causes the forward moving, phase-locked second-harmonic pulse to experience the same negative index as the pump pulse, although the index of refraction at the second-harmonic frequency is positive. It should be mentioned that optical NIMs are particularly challenging to optimize for practical applications because of absorption losses. In the case of phase locking, the pump impresses its dispersive properties to its harmonics, which in turn experience no absorption as long as the material is somewhat transparent to the pump.

Another nonlinear wave missing process, optical parametric amplification (OPA), was shown to be of significant importance in metamaterials. One of the main obstacles, delaying practical applications of optical metamaterials, is significant losses. The losses originate from several sources, including the resonant nature of the metamaterial's magnetic response, intrinsic absorption of the metallic constitutive components, and losses from surface roughness. Therefore, developing efficient loss-compensating techniques is of paramount importance. One promising approach to loss compensation is the technique based on a three-wave mixing process that takes place in nonlinear media exhibiting second-order susceptibility and induces the OPA. OPA refers to a process of amplification of a light signal through mixing with the pump light in a nonlinear material, in which the photon flux in the signal wave grows through coherent energy transfer from a higher frequency, intense pump wave. A photon from an

incident pump laser is divided into two photons, one of which is a photon at the signal frequency. The strong pump field with angular frequency and wavenumber and a weak signal generate a difference frequency idler. The OPA process requires both momentum and energy conservation. Although most OPA devices to date have been realized in conventional PIMs, an OPA process in NIMs was predicted to have several advantages, including the possibility of optical parametric oscillations without a cavity, compactness, and simplicity in design and alignment. Just like in the case of the SHG process, the difference between PIM-based and NIM-based OPAs is that the wave vector and the Poynting vector are anti-parallel (i.e., they move in opposing directions). Thus, an OPA with counterdirected energy flows can be realized with all three waves having co-directed wave vectors. Therefore, if the pump and idler frequencies correspond to the PIM and the signal wave frequency belongs to the NIM, the energy flow of the signal wave will be antiparallel to that of the pump and the idler, which will result in an effective feedback mechanism without any external mirrors or gratings. Such an OPA allows for the compensation of metamaterial absorption with the parametric amplification process and allows energy transfer to the strongly absorbing wave from the pump wave [123,124].

In addition, properly engineered optical metamaterial structures can be used to realize strong localization of light and, as a result, the enhancement of nonlinear interactions. Various approaches to field localization and enhancement were investigated, including graded and near-zero refractive index metamaterials and periodic structures consisting of alternating layers of PIM and NIM [112,138,139]. For instance, D'Aguanno et al. predicted that significantly improved conversion efficiencies of SHG may exist because of the presence of NIM layers. Furthermore, the efficiency stays relatively high even in the case of strongly absorbing structures.

Recently, several nonlinearly controlled devices were proposed [140−144], including nonlinear cloaking devices; reconfigurable nonlinear light concentrators; and the light-tunable reflection, shaping, and focusing of EM waves in metamaterials. The microwave light-tunable metamaterials mirror consists of an array of broadside-coupled SRRs, such that each SRR contains a pair of varactors (one in each ring) and the biasing of the varactors is achieved by photodiodes (Figure 9.22). This device enables reconfigurable mirrors or lenses in which switching between focusing or defocusing can be achieved by adjusting the light illumination profile. The advantages of this approach include fast and remote control of the device performance, the possibility of extension to higher dimensions, and the possibility of realizing such devices at higher frequencies, at least up to the terahertz range.

In summary, we discussed several important examples of novel nonlinear optical effects in microwave and optical metamaterials. In particular, it was shown that nonlinear optical properties of metamaterials can be controlled at the level of individual meta-atoms such that the effective nonlinear susceptibilities of the metamaterials significantly differ from those of the constituent materials. These new degrees of design freedom open unlimited opportunities for light-controlled optical functionalities and devices.
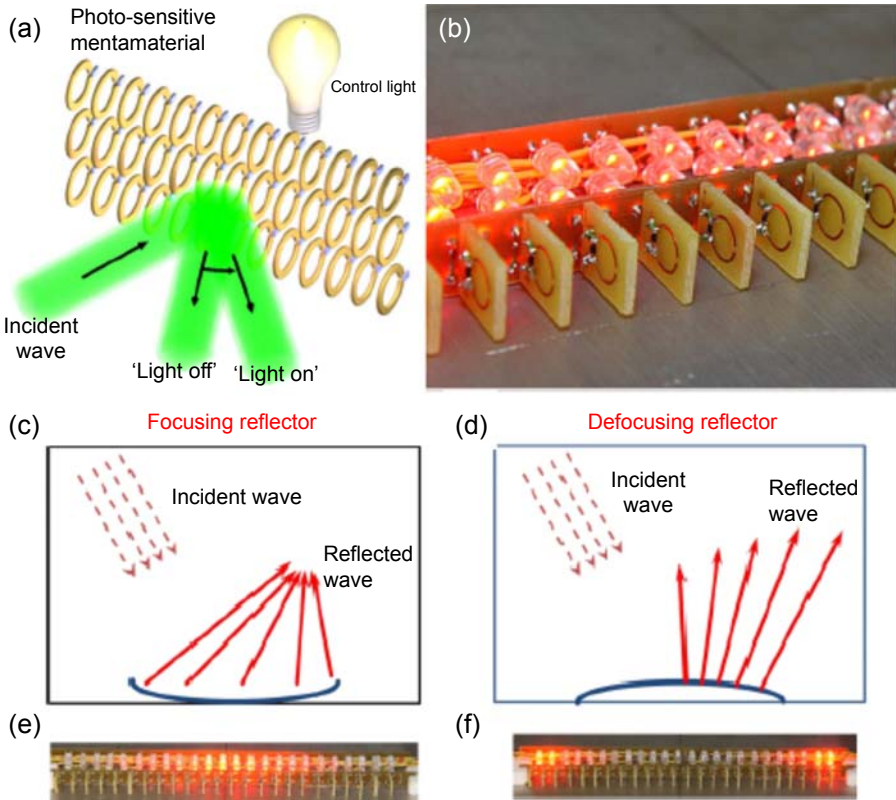
Figure 9.22 (a) Light-tunable metamaterials that reflect incident EM wave at different angles depending on the control light illumination using an array of light-emitting diodes. (b) Image of the microwave light-tunable metamaterials mirror made of an array of broadside-coupled SRRs. (Each SRR contains a pair of varactors, one in each ring. The biasing of the varactors is achieved by photodiodes.) (c and d) The schematics of (c) focusing and (d) defocusing reflectors and their corresponding performance: (e and f), respectively [144].

9.4   Metamaterials applications

9.4.1   Perfect absorbers

EM absorbers are specifically chosen or designed materials that can inhibit the reflection or transmission of EM radiation. Once EM absorbers are mentioned, one can easily associate this concept to stealth technology. First proposed in the 1950s, stealth technology has been well developed and widely used in the military. Applications include the spy plane, fighter plane, bomber, and warship, which have greatly challenged the ability of radar technology to detect nearby transport. However, for both military and civil use, high absorbance in wide bandwidth is always required for a good absorber, which is also an abnormal property that cannot often be obtained from natural materials.

A perfect matching absorbance metamaterial absorber is a high-loss material with impedance perfectly matched to the environment medium [145]. Let us consider the metamaterial absorber as a slab of thickness d of magneto-dielectric medium described by both the permittivity $\varepsilon(\omega) = \varepsilon_0 \varepsilon_r(\omega)$ and the magnetic permeability $\mu(\omega) = \mu_0 \mu_r(\omega)$. According to the Fresnel equation, reflectivity can be written as

$$R_{\mathrm{TE}} = |r_{\mathrm{TE}}|^2 = \left| \frac{\cos\theta - \mu_r^{-1} \sqrt{n^2 - (\sin\theta)^2}}{\cos\theta + \mu_r^{-1} \sqrt{n^2 - (\sin\theta)^2}} \right| \tag{9.68}$$

$$R_{\mathrm{TM}} = |r_{\mathrm{TM}}|^2 = \left| \frac{\varepsilon_r \cos\theta - \sqrt{n^2 - (\sin\theta)^2}}{\varepsilon_r \cos\theta + \sqrt{n^2 - (\sin\theta)^2}} \right| \tag{9.69}$$

where θ is the incident angle and $n = \sqrt{\varepsilon_r \mu_r}$ is the refractive index of the metamaterial. Considering the normal incidence case (i.e., $\theta = 0$), we obtain

$$R_{\mathrm{TE}} = R_{\mathrm{TM}} = \left| \frac{\mu_r - n}{\mu_r + n} \right|^2 = \left| \frac{Z - Z_0}{Z + Z_0} \right|^2, \tag{9.70}$$

where $Z = \sqrt{\frac{\mu}{\varepsilon}}$ and $Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}}$ are the impedance of the metamaterial and the free space, respectively. Therefore, the absorbance is

$$A = 1 - R - T \tag{9.71}$$

$$A = 1 - \left| \frac{Z - Z_0}{Z + Z_0} \right|^2 - T \tag{9.72}$$

T would be the energy transmitted into the metamaterial. Considering the refractive index $n = \sqrt{\varepsilon_r \mu_r} = n + \kappa \mathrm{i}$, the transmittance of the wave reaching the second surface of the metamaterial slab can be written as

$$T = T_0 e^{2k_0 n \kappa d} \tag{9.73}$$

In the perfect matched case, $\mu_r = \varepsilon_r$, then $Z = Z_0$, $n = \sqrt{\varepsilon_r \mu_r} = \varepsilon_r = \mu_r$. If the loss of the material (the imaginary part of the ε_r or μ_r) is high enough and the thickness of the slab d is large enough, then T extends to 0. Then, from Eqn (9.71), we achieve $A = 1$, which corresponds to perfect absorption. Because the wave is absorbed before it reaches the second surface of the slab, only transmittance and reflectance at the first surface of the slab needs to be considered. If the material is not of sufficient loss and thickness, then there will be transmission and reflection at the second boundary.

To further demonstrate this, we consider an artificial magneto-dielectric material with a thickness of d as a perfect absorber. Its permittivity and permeability can be described by

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - i\omega\gamma}, \tag{9.74}$$

$$\mu(\omega) = \mu + \frac{\omega_{mp}^2}{\omega_{m0}^2 - \omega^2 - i\omega\gamma_m}, \tag{9.75}$$

where ω_p is the plasma frequencies, ω_0 is the center frequencies of the oscillator, γ is the damping frequencies, and ε_∞ is the static permittivity at infinite frequency. As a perfect absorber, $\varepsilon(\omega) = \mu(\omega)$, the parameters in ω_{mp}, ω_{m0}, γ_m, and μ_∞ in the magnetic oscillator Eqn (9.74) are equal to those corresponding parameters in Eqn (9.75). Here we take $\omega_p = \omega_{mp} = 2\pi \times 1.25$ THz, $\omega_0 = \omega_{m0} = 2\pi \times 1.0$ THz, $\gamma = \gamma_m = 2\pi \times 0.1$ THz, and $\varepsilon_\infty = \mu_\infty = 1.0$.

The permittivity and permeability of such a magneto-dielectric Lorentz oscillator layer described by Eqns (9.74) and (9.75) are shown in Figure 9.23. The permittivity and the permeability are exactly the same, enabling perfect impedance matching and, therefore, no reflection ($R = 0$). At the resonance $\omega = \omega_0$, the loss is extremely high. No transmission occurs (i.e., $T = 0$) and absorption is equal to 1 ($A = 1$). Therefore, the medium works as a perfect absorber at ω_0. If we put a metal film under this medium, then the transmittance and reflectance at ω_0 are still zero. At other frequency ranges, transmission is zero ($T = 0$) and reflection is equal to 1 ($R = 1$) because all of the transmitted waves are reflected by the metal film.

The task for metamaterials is how to realize a material for which the permittivity and permeability are equal to each other and with high loss. In the NIM part, Lorentz-type permeability was produced by using SRR. In the resonance range, there is a peak of the imaginary part of the permeability. Then, the next step is to produce a Lorentz-type permittivity that has the same value as the permeability. In 2008, the first experimental demonstration of such an metamaterial perfect absorber (MPA) was realized in the giga-hertz microwave range, which utilized a metamaterial two-dimensional electric ring resonator (ERR) structure over a cut-wire medium separated by a dielectric layer [146].

Figure 9.24 shows a single unit cell of the metamaterial absorber consisting of two distinct metallic elements. The ERR, which was composed of two standard SRRs sharing one side of the ring, is used to supply the electric resonance. This ERR combined with the cut wire induces magnetic coupling when the **H** field is polarized perpendicular to the cut wire. A magnetic flux is created by circulating charges perpendicular to the propagation vector. The magnetic response can be tuned by changing the geometry of the cut wire and the separation between the cut wire and ERR without changing the ERR. Therefore, we can decouple ε and μ, individually tune each resonance so that $\varepsilon = \mu$, and thus create an impedance near the free space value. Then, 100% absorbance can be realized. This structure experimentally shows a very excellent performance of 99.9972% absorbance.
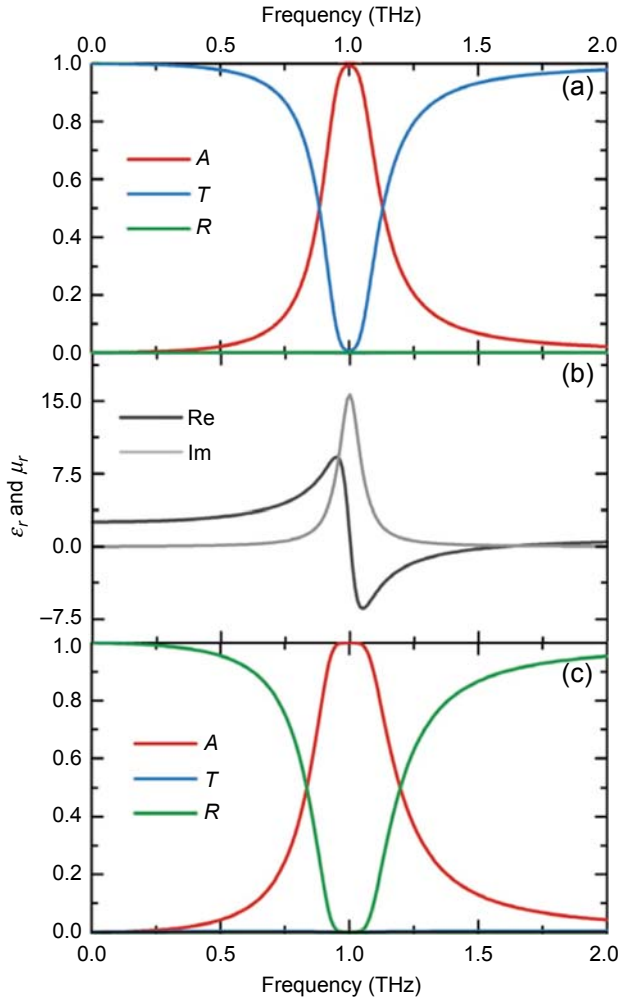
Figure 9.23 (a) Absorbance (red), transmittance (blue), and reflectance (green) for a magneto-dielectric medium of thickness d. (b) Real (dark gray) and imaginary (light gray) portions of the permittivity and permeability of the magneto-dielectric material. (c) Absorbance (red), transmittance (blue), and reflectance (green) for a magneto-dielectric medium backed by a metallic ground plane [145].

After this, perfect metamaterial absorbers were also realized in terahertz and infrared ranges using a similar design [147,148]. However, the perfect metamaterial absorbers have a serious problem in the narrow bandwidth. The high loss of the metamaterial is usually obtained by resonance, which occurs in a narrow frequency range. The general way to solve this problem is to put more resonators into the metamaterial layer so that there are more resonances in an operating frequency range, as shown in Figure 9.25 [149].
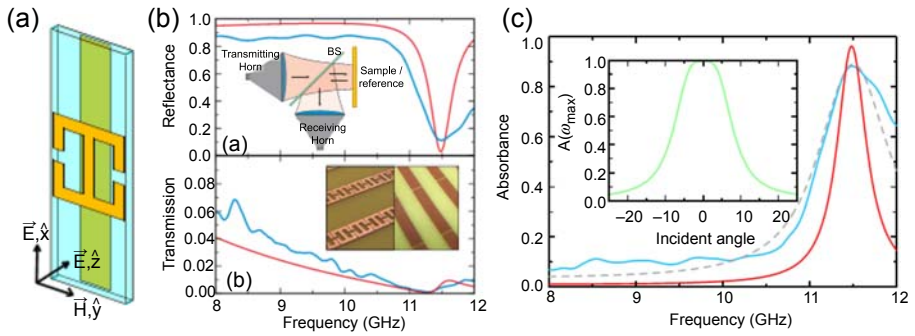
Figure 9.24 (a) The structure of the perfect absorber based on metamaterial. (b) The reflectance and transmittance measurement of the metamaterial absorber. (c) The calculated absorbance based on the measurement results [146].
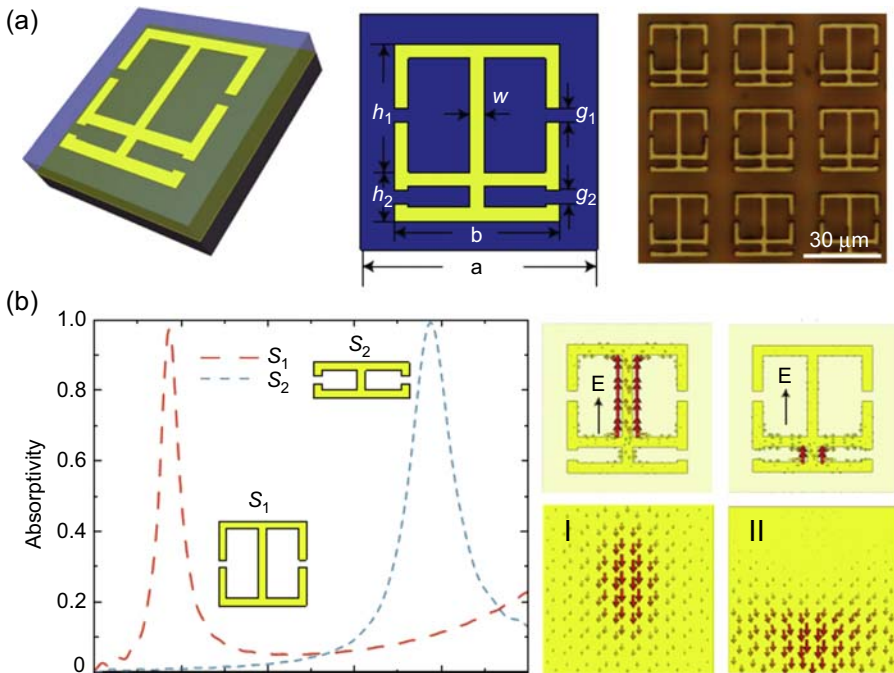
Figure 9.25 Design of the perfect absorber with wide bandwidth. (a) Structures of the perfect absorber. (b) Mechanism of wide absorbance bandwidth due to multiple resonators [149].

In summary, we discussed the basic theory of the metamaterials-based perfect absorber. A metamaterial with equal permittivity and permeability enables a perfect impedance matching with the surrounding medium, thus eliminating reflection from the absorber, whereas high losses enable total absorption of the incident wave.

In addition, the working bandwidth of the metamaterial absorber can be greatly broadened by using a multiresonator structure.

9.4.2 Transformation optics-enabled metamaterials devices

The immense potential of refractive index engineering in metamaterials can be further exemplified by recent progress in the field of transformation optics that enabled novel opportunities in the design of graded-index structures [150−161]. First applied to the development of the cloak, the transformation method is now considered a very general and powerful design tool that offers unparalleled opportunities for controlling light propagation through careful refractive index engineering.

The basic idea of the transformation method [150,162,163] is that, to guide waves along a certain trajectory, either the space should be deformed, assuming that material properties remain the same, or the material properties should be properly modified. The latter approach is typically used. Under a coordinate transformation, the form of Maxwell's equations should remain invariant, whereas new ε and μ would contain the information regarding the coordinate transformation and the original material parameters. To design a particular property/light trajectory, it is possible to design an anisotropic material with prescribed components of permittivity and permeability tensors calculated through a particular coordinate transformation.

The general design approach using the transformation method includes two main steps. In the first step, a coordinate transformation of the space with the desired property is built. In the next step, a set of material properties is calculated that would realize this property of the transformed space in the original space using the following equations:

$$
\begin{aligned}
\varepsilon^{i'j'} &= \left| \det\left(\Lambda_i^{i'} \right) \right|^{-1} \Lambda_i^{i'} \Lambda_j^{j'} \varepsilon \\
\mu^{i'j'} &= \left| \det\left(\Lambda_i^{i'} \right) \right|^{-1} \Lambda_i^{i'} \Lambda_j^{j'} \mu
\end{aligned}
\qquad i,j = 1,2,3,
\tag{9.76}
$$

where it is assumed that the original space is isotopic, that transformations are time invariant, and that $\Lambda_\alpha^{\alpha'} = \frac{\partial x^{\alpha'}}{\partial x^\alpha}$ are the elements of the Jacobian transformation matrix.

Recently, various functionalities and novel device applications enabled by this approach have been proposed, including light concentrators and lenses, gradient-index waveguides and bends, black holes, and even illusion devices [156−161]. Some of these devices are illustrated in Figure 9.26. Unfortunately, many of these devices, including the first cloaking devices, were inherently narrowband. Indeed, because the refractive index of the cloaking shell was designed to vary from 0 to 1, the phase velocity of light inside of the shell is greater than the velocity of light in a vacuum. Although this condition itself does not contradict any law of physics, it implies that the material parameters must be dispersive. Therefore, one of the remaining challenges is the realization of broadband transformation optics-based applications.
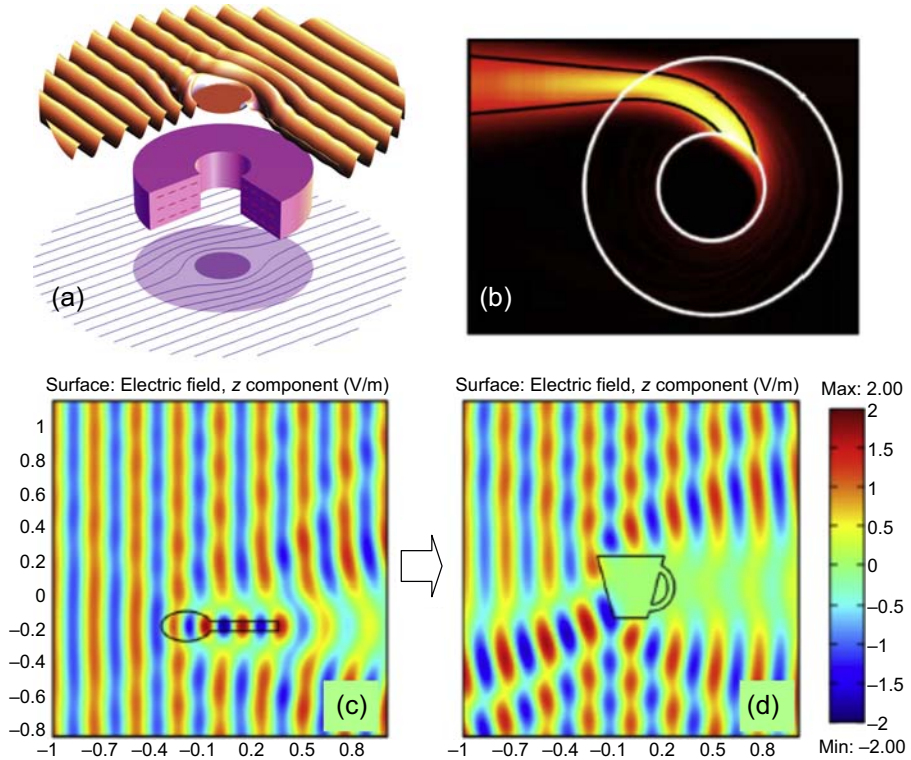
Figure 9.26 Examples of gradient-index structure designed using a transformation optics approach (a) cloak [159], (b) optical black hole [160], (c) and (d) illusion optics [161].

To address this challenge, the transformation method was applied to design a broadband cloaking device that functions in the wavelength multiplexing manner shown in Figure 9.27 [164]. The basic idea can be understood as follows. Because the anisotropic constituent materials of a cloak for one wavelength cannot be transparent at other frequencies, cloaks for all of the wavelengths being considered have to share the same outer boundary—the physical boundary of the device. However, the inner boundary and transformation for each operating wavelength are unique, as shown in Figure 9.27(a)−(c).
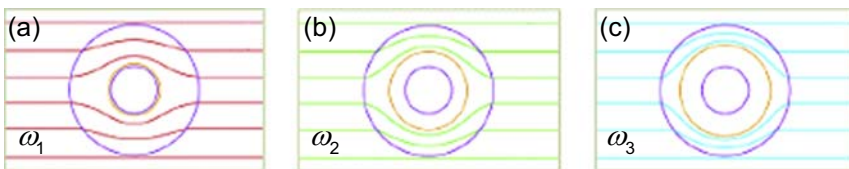
Figure 9.27 Simulated schematic of a cloaking system for multiple wavelengths or a finite bandwidth, with $\omega_1 > \omega_2 > \omega_3$, shown in (a−c), respectively [164].

Recently, two very different designs of optical cloaks that address the narrowband issue were proposed and experimentally demonstrated. One of them is a "carpet" cloak that is made of a dielectric and operates in a broad range of wavelengths [165−168]. The basic idea behind this approach is illustrated in Figure 9.28. The carpet cloak compresses an object in only one direction into a conducting sheet. When the object is placed under a curved reflecting surface with the carpet cloak on top of it, the object appears as if it was the original flat reflecting surface, so it is hidden under a "carpet." This approach avoids both material and geometry singularities. The cloak region is obtained by varying the effective refractive index in a two-dimensional space. This index profile is designed using quasiconformal mapping. The carpet cloaking approach solves several problems associated with other approaches. The approach can be implemented using nonresonant elements (e.g., conventional dielectric materials) and therefore allows for low-loss, broadband cloaking at optical wavelengths. Indeed, this approach was already demonstrated in a broad range of optical wavelengths extending from 1400 to 1800 nm.

In yet another approach, the metamaterial requiring anisotropic dielectric permittivity and magnetic permeability was emulated by a specially designed tapered waveguide [169]. It was shown that the transformation optics approach allows us to map a planar region of space filled with an inhomogeneous, anisotropic metamaterial within an equivalent region of empty space with curvilinear boundaries (a tapered waveguide).

This approach leads to low-loss, broadband performance in the visible wavelength range. The cloak consists of a double convex glass lens coated on one side with a gold film placed with the gold-coated side down on top of a flat, gold-coated glass slide. The air gap between these surfaces can be used as an adiabatically changing waveguide. It is well known that the modes of the waveguide have cutoff wavelengths. It turns out that, for a particular mode in such a waveguide, the cutoff radius is given by the same expression as that of the radius of the corresponding Newton ring. As a result, no photon launched into the waveguide can reach an area within the radius from the point of contact between the two gold-coated surfaces. This approach leads to low-loss, broadband cloaking performance. Importantly, the cloak that has already been realized with this approach was 100 times larger than the wavelength of light.

Over the last 10 years, enormous progress in the field of metamaterials has been made. As a result, fascinating theoretical predictions were transformed into an even more amazing reality of materials with properties far exceeding those available in nature. The new material properties and device functionalities enabled by metamaterials
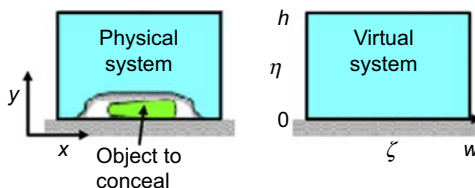
Figure 9.28 Basic idea of "carpet" cloaking [165].

technology have a strong potential for contributing to nearly all areas of fundamental and applied science and technology, including imaging, defense, telecommunications, optical computing, and sensing.

9.4.3   High-resolution imaging applications: hyperlens and superlens

The resolution of traditional optical lenses ($\varepsilon > 0$, $\mu > 0$) is generally constrained by the diffraction limit. This is because the evanescent wave that contains the subwavelength information decays exponentially with distance and is undetectable to image by conventional lenses (Figure 9.29(a)). So far, the information carried by evanescent waves can only be gathered by scanning near-field optical microscopy techniques. By studying the wave propagation NIMs, Pendry showed that the evanescent wave in NIM is magnified exponentially with distance and the subwavelength details are retained for imaging, which could break through the diffraction limit [170,171]. However, because of the decay of the evanescent field outside of the NIM lens (superlens), the image of this superlens still requires the near-field detection [172,173] (Figure 9.29(b)). Jacob indicated that an ideal optical media is the one that is able to transfer the information carried by an evanescent wave into a portion of the propagating spectrum. Via this conversion, propagating waves would be detected and processed in the far field by conventional optical devices (Figure 9.29(c)). After the theoretical computation, Jacob verified that the indefinite medium has the ideal capability of a hyperlens [174].

Consider an EM wave propagating in vacuum along the x-axis. The electric component of the field is given by a two-dimensional Fourier expansion:

$$\mathbf{E}(r,t) = \sum_{\sigma,k_y,k_z} \mathbf{E}_\sigma(k_y, k_z) \times \exp(ik_x x + ik_y y + ik_z - i\omega t), \tag{9.77}$$

Following the Maxwell's equations, the x component of the wave vector k_x can be written as

$$k_x = +\sqrt{\omega^2 c^{-2} - k_z^2 - k_y^2}, \quad \omega^2 c^{-2} > k_z^2 + k_y^2. \tag{9.78}$$

(a)   (b)   (c)
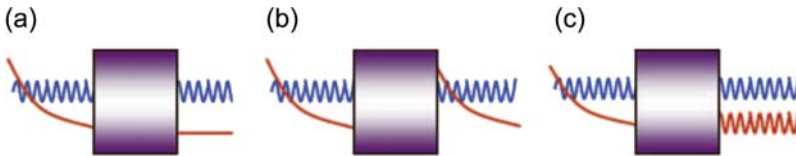
Figure 9.29 (a) An EM wave transmitted in conventional media in propagation mode. The evanescent part decays exponentially with distance, which is only detectable in the near field. (b) A "superlens" amplifies the evanescent waves but does not change their decaying character. (c) An ideal device would convert evanescent waves to propagating waves for ease of detection and processing in the far-field [174].

However, for larger values of the transverse wave vector,

$$k_x = +i\sqrt{k_z^2 + k_y^2 - \omega^2 c^{-2}}, \quad \omega^2 c^{-2} < k_z^2 + k_y^2. \tag{9.79}$$

Thus, these evanescent waves decay exponentially with distance x and subwavelengthed detail is missing.

In the case of NIM, the wave vector is in a reversed direction, which means that the value of the square root in Eqn (9.79) should be negative:

$$k_x = -i\sqrt{k_z^2 + k_y^2 - \omega^2 c^{-2}}, \quad \omega^2 c^{-2} < k_z^2 + k_y^2, \tag{9.80}$$

If we put it in to Eqn (9.77), the amplitude of the **E** field shows an exponential increase with the propagation. This way, the detailed information carried by the evanescent wave is preserved and magnified inside of the NIM. However, after it comes out of the NIM, the evanescent wave remains decayed exponentially in the positive material, as shown in Figure 9.29(b).

In the case of an indefinite medium (**E** polarized in the xz-plane), according to the hyperbolic EFC,

$$\frac{k_x^2}{\varepsilon_{zz}} + \frac{k_z^2}{\varepsilon_{xx}} = \omega^2 c^{-2}, \tag{9.81}$$

with $\varepsilon_{xx} < 0$. Then,

$$k_x = \varepsilon_{zz}\sqrt{\omega^2 c^{-2} - \frac{k_z^2}{\varepsilon_{xx}}}, \tag{9.82}$$

which is constantly greater than zero. This ensures a propagation mode of the evanescent wave in the indefinite medium, which corresponds to the portion of the EFC with $k_z > \omega/c$ in Figure 9.30.

The first hyperlens was realized in 2007 [42,176]. It was a cylindrical magnifying hyperlens composed by 16 layers of interleaved Ag/Al_2O_3, which was able to show anisotropy between the radial and tangential direction of the cylinder. Therefore, the EFC of such a concentric column structure in Figure 9.31(a) is

$$\frac{k_r^2}{\varepsilon_\theta} + \frac{k_\theta^2}{\varepsilon_r} = \omega^2 c^{-2}, \tag{9.83}$$

where ε_r is the radial permittivity and ε_θ is the tangential permittivity. The strong anisotropy between the tangential and radial directions ($\varepsilon_r < 0$, $\varepsilon_\theta > 0$) enables propagation of the evanescent wave along the radial direction.

In the experiment, a subdiffraction-limited object of "ON" split (split width = 40 nm, average distance = 200 nm) was inscribed into the chrome layer located on the inner
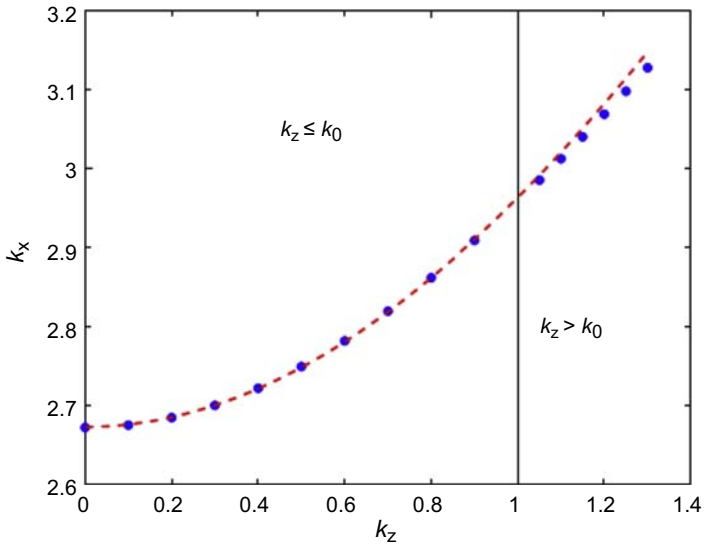
Figure 9.30 Hyperbolic EFC of an indefinite medium, $k_z < \omega/c$ for propagation mode and $k_z > \omega/c$ for evanescent mode [175].
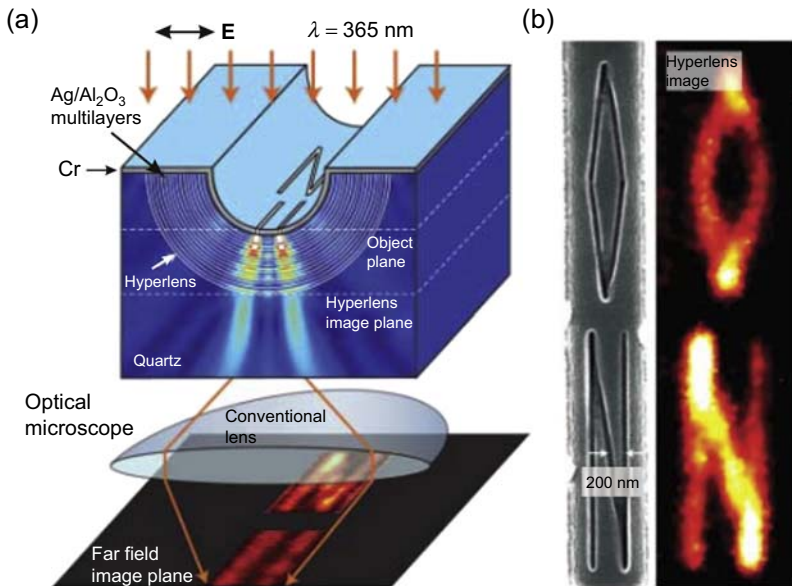
Figure 9.31 Optical hyperlens based on anisotropic permittivity. (a) Schemes of the high-resolution image. (b) Subdiffraction-limited object of "ON" split was inscribed into the chrome layer located on the inner surface, and the image formed at the far field [42].

surface. A 365-nm light illumination on the split formed an object. The object was imaged with a hyperlens and clearly focused with an optical microscope, which demonstrates the preservation of the subdiffraction image into the far field.

In conclusion, superlens and hyperlens are the two most promising applications of the negative index metamaterials: NIMs and hyperbolic materials. Although the NIM-based superlens magnifies evanescent waves, the hyperlens converts evanescent waves into propagating waves. Both of the two give a feasible way to break the diffraction limit of conventional optics and thus realize the nanoscale imaging and nanolithography using light waves.

Acknowledgments

References

[1]   V.G. Veselago, The electrodynamics substances with simultaneously negative values of ε and μ, Sov. Phys. Usp. 10 (1968) 509−514.

[2]   D.R. Smith, W.J. Padilla, D.C. Vier, et al., Composite medium with simultaneously negative permeability and permittivity, Phys. Rev. Lett. 84 (2000) 4184−4187.

[3]   A.A. Houck, J.B. Brock, I.L. Chuang, Experimental observations of a left-handed material that obeys Snell's law, Phys. Rev. Lett. 90 (2003) 137401−137404.

[4]   G. Dolling, M. Wegener, C.M. Soukoulis, S. Linden, Negative-index metamaterial at 780 nm wavelength, Opt. Lett. 30 (2005) 53−55.

[5]   C.M. Soukoulis, S. Linden, M. Wegener, Negative refractive index at optical wavelengths, Science 315 (2007) 47−49.

[6]   V.M. Shalaev, Optical negative-index metamaterials, Nat. Photonics 1 (2007) 41−48.

[7]   G. Dolling, C. Enkrich, M. Wegener, et al., Simultaneous negative phase and group velocity of light in a metamaterial, Science 312 (2006) 892−894.

[8]   R.W. Ziolkowski, E. Heyman, Wave propagation in media having negative permittivity and permeability, Phys. Rev. E 64 (2001) 056625−056639.

[9]   R.A. Shelby, D.R. Smith, S. Schultz, Experimental verification of a negative index of refraction, Science 292 (2001) 77−79.

[10] D.R. Smith, D. Schurig, Electromagnetic wave propagation in media with indefinite permittivity and permeability tensors, Phys. Rev. Lett. 90 (2003) 077405−077409.

[11] J. Yao, Z.W. Liu, Y.M. Liu, et al., Optical negative refraction in bulk metamaterials of nanowires, Science 321 (2008) 930.

[12] D.R. Smith, P. Kolinko, D. Schurig, Negative refraction in indefinite media, J. Opt. Soc. Am. B 21 (2004) 1032−1043.

[13] A.J. Hoffman, L. Alekseyev, S.S. Howard, et al., Negative refraction in semiconductor metamaterials, Nat. Mater. 6 (2007) 946−950.

[14] C.Y. Luo, S.G. Johnson, J.D. Joannopoulos, et al., All-angle negative refraction without negative effective index, Phys. Rev. B 65 (2002) 201104−201107.

[15] E. Cubukcu, K. Aydin, E. Ozbay, et al., Negative refraction by photonic crystals, Nature 423 (2003) 604−605.

[16] A. Berrier, M. Mulot, M. Swillo, et al., Negative refraction at infrared wavelengths in a two-dimensional photonic crystal, Phys. Rev. Lett. 93 (2004) 073902−073905.

[17] J.B. Pendry, A.J. Holden, D.J. Robbins, et al., Magnetism from conductors and enhanced nonlinear phenomena, IEEE Trans. Microwave Theory Tech. 47 (1999) 2075−2084.

[18] H. Chen, L. Ran, J. Huangfu, et al., Equivalent circuit model for left-handed metamaterials, J. Appl. Phys. 100 (2006) 024915−024920.

[19] T.M. Grzegorczyk, J. Lu, X. Chen, J. Pacheco Jr., Properties of left-handed meta-materials: transmission, backward phase, negative refraction and focusing, IEEE Trans. Microwave Theory Tech. 53 (9) (2005) 2956−2967.

[20] C.R. Simovski, B. Sauviac, Toward creating isotropic microwave composites with negative refraction, Radio Sci. 39 (2004).

[21] H. Chen, L. Ran, J. Huangfu, X. Zhang, K. Chen, Left-handed materials composed of only s-shaped resonators, Phys. Rev. E 70 (2004) 057605.

[22] T.J. Yen, W.J. Padilla, N. Fang, D.C. Vier, D.R. Smith, J.B. Pendry, D.N. Basov, X. Zhang, Terahertz magnetic response from artificial materials, Science 303 (2004) 1494−1496.

[23] J.B. Pendry, A.J. Holden, W.J. Stewart, I. Youngs, Extremely low frequency plasmons in metallic mesostrucures, Phys. Rev. Lett. 76 (1996) 4773−4776.

[24] D.R. Smith, S. Schultz, Determination of effective permittivity and permeability of metamaterials from reflection and transmission coefficients, Phys. Rev. B 65 (2002) 195104−195108.

[25] P. Markos, C.M. Soukoulis, Transmission studies of left-handed materials. Phys. Rev. B, 65: 033401−033404.

[26] D.R. Smith, D.C. Vier, T. Koschny, C.M. Soukoulis, Electromagnetic parameter retrieval from inhomogeneous metamaterials, Phys. Rev. E 71 (2005) 036617−036627.

[27] X. Chen, T.M. Grzegorczyk, B.I. Wu, et al., Robust method to retrieve the constitutive effective parameters of metamaterials, Phys. Rev. E 70 (2004) 016608−016614.

[28] J. Zhou, Th Koschny, M. Kafesaki, E.N. Economou, J.B. Pendry, C.M. Soukoulis, Saturation of the magnetic response of split-ring resonators at optical frequencies, Phys. Rev. Lett. 95 (2005) 223902.

[29] G. Dolling, C. Enrich, M. Wegener, J.F. Zhou, C.M. Soukoulis, Cut-wire pairs and plate pairs as magnetic atoms for optical metamaterials, Opt. Lett. 30 (2005) 3198−3200.

[30] S. Zhang, W. Fan, N.C. Panoiu, K.J. Malloy, R.M. Osgood, S.R.J. Brueck, Optical negative-index bulk metamaterials consisting of 2D perforated metal-dielectric stacks, Opt. Express 12 (2006) 6778−6787.

[31] V.M. Shalaev, W.S. Cai, U.K. Chettiar, et al., Negative index of refraction in optical metamaterials, Opt. Lett. 30 (2005) 3356−3358.

[32] S.M. Xiao, U.K. Chettiar, A.V. Kildishev, V.P. Drachev, V.M. Shalaev, Yellow-light negative-index metamaterials, Opt. Lett. 34 (2009) 3478−3480.

[33] J. Valentine, S. Zhange, T. Zentgraf, et al., Three-dimensional optical metamaterial with a negative refractive index, Nature 455 (2008) 376−379.

[34] C.M. Soukoulis, M. Wegener, Past achievement and future challenges in the development of three-dimensional photonic metamaterials, Nat. Photonics 5 (2011) 523−530.

[35] Y.M. Liu, G. Bartal, X. Zhang, All-angle negative refraction and imaging in a bulk medium made of metallic nanowires in the visible region, Opt. Express 16 (2008) 15439−15448.

[36] J. Sun, J. Zhou, Negative refraction in anisotropic materials, Chin. Sci. Bull. 57 (2012) 231−238 [Chinese version].

[37] X.L. Chen, M. He, Y.X. Du, et al., Negative refraction: an intrinsic property of uniaxial crystals, Phys. Rev. B 72 (2005) 113111−113114.

[38] W.Q. Zhang, F. Yang, Negative refraction at various crystal interfaces, Opt. Commun. 281 (2008) 3081−3086.

[39] R.A. Brazhe, R.M. Meftakhutdinov, Negative optical refraction in crystals with strong birefringence, Tech. Phys. 52 (2007) 793−795.

[40] X. Yang, J. Yao, J. Rho, X. Yin, X. Zhang, Experimental realization of three-dimensional indefinite cavities at the nanoscale with anomalous scaling laws, Nat. Photonics 6 (2012) 450−454.

[41] Q. Meng, X. Zhang, L. Cheng, P. Cao, Y. Li, H. Zhang, G. Wang, Deep subwavelength focusing of light by a trumpet hyperlens, J. Opt. 13 (2011) 075102−075105.

[42] Z.W. Liu, H. Lee, Y. Xiong, C. Sun, X. Zhang, Far-field optical hyperlens magnifying sub-diffraction-limited objects, Science 315 (2007) 1686.

[43] J. Yao, X. Yang, X. Yin, G. Bartal, X. Zhang, Three-dimensional nanometer-scale optical cavities of indefinite medium, PNAS 108 (2011) 11327−11331.

[44] J. Yao, K.T. Tsai, Y. Wang, Z. Liu, G. Bartal, Y.L. Wang, X. Zhang, Imaging visible light using anisotropic metamaterials slab lens, Opt. Express 17 (2009) 22380−22385.

[45] W.T. Lu, S. Sridhar, Superlens imaging theory for anisotropic nanostructured metamaterials with broadband all-angle negative refraction, Phys. Rev. B 77 (2008) 233101−233104.

[46] S.S. Kruk, D.A. Powell, A. Minovich, D.N. Neshev, Y.S. Kivshar, Spatial dispersion of multilayer fishnet metamaterials, Opt. Express 20 (2012) 15100−15105.

[47] H. Liu, M.Y. Liu, S.M. Wang, et al., Coupled magnetic plasmons in metamaterials, Phys. Status Solidi B 246 (2009) 1397−1406.

[48] D.R. Smith, J.J. Mock, A.F. Starr, Gradient index metamaterials, Phys. Rev. E 71 (2005) 036609−036614.

[49] J. Sun, L. Kang, R. Wang, et al., Low loss negative refraction metamaterial using a close arrangement of split-ring resonator arrays, New J. Phys. 12 (2010) 083020−083027.

[50] D.R. Smith, D. Schurig, J.J. Mock, et al., Partial focusing of radiation by a slab of indefinite media, Appl. Phys. Lett. 84 (2004) 2244−2246.

[51] J. Sun, N.M. Litchinitser, J. Zhou, Indefinite by nature: from ultraviolet to terahertz, ACS Photonics 1 (2014) 293−303.

[52] J. Sun, J. Zhou, B. Li, F. Kang, Indefinite permittivity and negative refraction in natural material: Graphite, Appl. Phys. Lett. 98 (2011) 101901−101903.

[53] V. Guritanu, A.B. Kuzmenko, D.V.D. Marel, Anisotropic optical conductivity and two colors of MgB_2, Phys. Rev. B 73 (2006) 104509−104519.

[54] J. Kunert, M. Backer, M. Falter, D. Schroeder-Obst, Comparison of CSD-YBCO growth on different single crystal substrates, J. Phys. Conf. Ser. 97 (2008) 012148−012152.

[55] B. Hu, G.T. McCandless, M. Menard, et al., Surface and Bulk Structure Properties of Single Crystalline $Sr_3Ru_2O_7$, 2010. ArXiv cond-mad: 1003.5221.

[56] Y. Maeno, H. Hashimoto, K. Yoshida, et al., Superconductivity in a layered perovskite without copper, Nature 372 (1994) 532−534.

[57] S. Enoch, G. Tayeb, P. Sabouroux, N. Guérin, P. Vincent, A metamaterial for directive emission, Phys. Rev. Lett. 89 (2002) 213902.

[58] A. Alu, M.G. Silveirinha, A. Salandrino, N. Engheta, Epsilon-near-zero (ENZ) metamaterials and electromagnetic sources: tailoring the radiation phase pattern, Phys. Rev. B 75 (2007) 155410.

[59] R. Mass, J. Parsons, N. Engheta, A. Polman, Experimental realization of an epsilon-near-zero metamaterial at visible wavelengths, Nat. Photonics 7 (2013) 907−912.

[60] P. Moitra, Y. Yang, Z. Anderson, I. Kravchenko, D. Briggs, J. Valentine, Realization of an all-dielectric zero-index optical metamaterial, Nat. Photonics, 7 (2013) 791−795.

[61] H. Suchowski, K. O'Brien, Z.J. Wong, A. Salandrino, X. Yin, X. Zhang, Phase mismatch−free nonlinear propagation in optical zero-index materials, Science 342 (2013) 1223−1226.

[62] M.A. Vincenti, D. de Ceglia, A. Ciattoni, M. Scalora, Singularity-driven second- and third-harmonic generation at ε-near-zero crossing points, Phys. Rev. A 84 (2011) 063826.

[63] M. Silveririnha, N. Engheta, Design of matched zero-index metamaterials using nonmagnetic inclusions in epsilon-near-zero media, Phys. Rev. B 75 (2007) 075119.

[64] M. Silveirinha, N. Engheta, Tunneling of electromagnetic energy through sub-wavelength channels and bends using ε-near-zero-materials, Phys. Rev. Lett. 97 (2006) 157403.

[65] R. Liu, Q. Cheng, T. Hand, et al., Experimental demonstration of electromagnetic tunneling through an epsilon-near-zero metamaterial at microwave frequencies, Phys. Rev. Lett. 100 (2008) 023903.

[66] J. Hao, W. Yan, M. Qiu, Super-reflection and cloaking based on zero index metamaterial, Appl. Phys. Lett. 96 (2010) 101109.

[67] N.M. Litchinitser, A.I. Maimistov, I.R. Gabitov, R.Z. Sagdeev, V.M. Shalaev, Metamaterials: electromagnetic enhancement at zero-index transition, Opt. Lett. 33 (2008) 2350−2352.

[68] K. Kim, D.-H. Lee, H. Lim, Resonant absorption and mode conversion in a transition layer between positive-index and negative-index media, Opt. Express 16 (2008) 18505−18513.

[69] M. Dalarsson, P. Tassin, Analytical solution for wave propagation through a graded index interface between a right-handed and a left-handed material, Opt. Express 17 (2009) 6747−6752.

[70] I. Mozjerin, E.A. Gibson, E.P. Furlani, I.R. Gabitov, N.M. Litchinitser, Electromagnetic enhancement in lossy optical transition metamaterials, Opt. Lett. 35 (2010) 3240.

[71] E.A. Gibson, M. Pennybacker, A.I. Maimistov, I.R. Gabitov, N.M. Litchinitser, Resonant absorption in transition metamaterials: parametric study, J. Opt. 13 (5) (2011) 024013.

[72] E.A. Gibson, I.R. Gabitov, A.I. Maimistov, N.M. Litchinitser, Transition metamaterials with spatially separated zeros, Opt. Lett. 36 (2011) 3624−3626.

[73] F. Alali, N.M. Litchinitser, Gaussian beams in near-zero transition metamaterials, Opt. Commun. 291 (2013) 179−183.

[74] Z.A. Kudyshev, I.R. Gabitov, A.I. Maimistov, R.Z. Sagdeev, N.M. Litchinitser, Second harmonic generation in transition metamaterials, J. Opt. 16 (2014) 114011.

[75] V.L. Ginzburg, The Propagation of Electromagnetic Waves in Plasma, Pergamon, 1970.

[76] N. Bloembergen, Nonlinear Optics, fourth ed., World Scientific, 1996.

[77]   Y.R. Shen, Principles of Nonlinear Optics, Wiley, 1984.

[78]   R.W. Boyd, Nonlinear Optics, Academic Press, 1992.

[79]   J.B. Khurgin, G. Sun, Plasmonic enhancement of the third order nonlinear optical phenomena: figures of merit, Opt. Express 21 (2013) 27460−27480.

[80]   R.W. Boyd, G.L. Fischer, Nonlinear Optical Materials, in: Encyclopedia of Materials: Science and Technology, 2001, pp. 6237−6244.

[81]   M.G. Kuzyk, Nonlinear optics: fundamental limits of nonlinear susceptibilities, Opt. Photonics News 14 (2003) 26.

[82]   N.I. Zheludev, Y.S. Kivshar, From metamaterials to metadevices, Nat. Mater. 11 (2012) 917−924.

[83]   M. Lapine, I.V. Shadrivov, Y.S. Kivshar, Colloquium: nonlinear metamaterials, Rev. Mod. Phys. 86 (2014) 1093.

[84]   N.M. Litchinitser, V.M. Shalaev, Metamaterials: transforming theory into reality, J. Opt. Soc. Am. B 26 (2009) 161−169.

[85]   N.M. Litchinitser, I.R. Gabitov, A.I. Maimistov, V.M. Shalaev, Negative refractive index metamaterials in Optics, in: E. Wolf (Ed.), Progress in Optics, vol. 51, 2008, pp. 1−68 (Chapter 1).

[86]   A. Rose, D. Huang, D.R. Smith, Controlling the second harmonic in a phase-matched negative-index metamaterials, Phys. Rev. Lett. 107 (6) (2011) 063902.

[87]   I.V. Shadrivov, P.V. Kapitanova, S.I. Maslovski, Y.S. Kivshar, Metamaterials controlled with light, Phys. Rev. Lett. 109 (2012) 083902.

[88]   Y.S. Kivshar, Tunable and nonlinear metamaterials: toward functional metadevices, Adv. Nat. Sci. Nanosci. Nanotechnol. 5 (2014) 013001−013008.

[89]   E. Poutrina, D. Huang, D.R. Smith, Analysis of nonlinear electromagnetic metamaterials, N. J. Phys. 12 (2010) 093010.

[90]   E. Poutrina, D. Huang, Y. Urzhumov, D.R. Smith, Nonlinear oscillator metamaterial model: numerical and experimental verification, Opt. Express 19 (2011) 8312−8319.

[91]   P.A. Franken, J.F. Ward, Optical harmonics and nonlinear phenomena, Rev. Mod. Phys. 35 (1963) 23−39.

[92]   T. Kanazawa, et al., Enhancement of second harmonic generation in a doubly resonant metamaterial, Appl. Phys. Lett. 99 (3) (2011) 024101.

[93]   Z.Y. Wang, et al., Second-harmonic generation and spectrum modulation by an active nonlinear metamaterial, Appl. Phys. Lett. 94 (3) (2009) 134102.

[94]   N.I. Zheludev, V.I. Emel'yanov, Phase matched second harmonic generation from nanostructured metallic surfaces, J. Opt. A 6 (2004) 26−28.

[95]   H. Merbold, A. Bitzer, T. Feurer, Second harmonic generation based on strong field enhancement in nanostructured THz materials, Opt. Express 19 (2011) 7262−7273.

[96]   W.L. Schaich, Second harmonic generation by periodically-structured metal surfaces, Phys. Rev. B 78 (2008) 8.

[97]   M.W. Klein, et al., Second-harmonic generation from magnetic metamaterials, Science 313 (2006) 502−504.

[98]   H. Husu, et al., Metamaterials with tailored nonlinear optical response, Nano Lett. 12 (2012) 673−677.

[99]   F.B.P. Niesler, et al., Second-harmonic optical spectroscopy on split-ring-resonator arrays, Opt. Lett. 36 (2011) 1533−1535.

[100] L. Chen, C.H. Liang, X.J. Dang, Second-harmonic generation in nonlinear left-handed metamaterials, Acta Phys. Sin. 56 (2007) 6398−6402.

[101] V. Roppo, et al., Second harmonic generation in a generic negative index medium, J. Opt. Soc. Am. B 27 (2010) 1671−1679.

[102]   I.V. Shadrivov, A.A. Zharov, Y.S. Kivshar, Second-harmonic generation in nonlinear left-handed metamaterials, J. Opt. Soc. Am. B 23 (3) (2006) 529−534.

[103]   A.K. Popov, V.V. Slabko, V.M. Shalaev, Second harmonic generation in left-handed metamaterials, Laser Phys. Lett. 3 (2006) 293−297.

[104]   D. de Ceglia, et al., Enhancement and inhibition of second-harmonic generation and absorption in a negative index cavity, Opt. Lett. 32 (2007) 265−267.

[105]   I.R. Gabitov, B. Kennedy, A.I. Maimistov, Coherent amplification of optical pulses in Metamaterials, IEEE J. Sel. Top. Quantum Electron. 16 (2010) 401−409.

[106]   S.M. Gao, S.L. He, Four-wave mixing in left-handed materials, J. Nonlinear Opt. Phys. Mater. 16 (2007) 485−496.

[107]   I.V. Shadrivov, Y.S. Kivshar, Spatial solitons in nonlinear left-handed metamaterials, J. Opt. A: Pure Appl. Opt. 7 (2005) S68−S72.

[108]   A.B. Kozyrev, D.W. van der Weide, Nonlinear left-handed transmission line metamaterials, J. Phys. D 41 (10) (2008) 173001.

[109]   X.Y. Dai, et al., Frequency characteristics of the dark and bright surface solitons at a nonlinear metamaterial interface, Opt. Commun. 283 (2010) 1607−1612.

[110]   N.A. Zharova, I.V. Shadrivov, A.A. Zharov, Nonlinear transmission and spatiotemporal solitons in metamaterials with negative refraction, Opt. Express 13 (2005) 1291−1298.

[111]   W.N. Cui, et al., Self-induced gap solitons in nonlinear magnetic metamaterials, Phys. Rev. E 80 (5) (2009) 036608.

[112]   M. Scalora, et al., Gap solitons in a nonlinear quadratic negative-index cavity, Phys. Rev. E 75 (6) (2007) 066606.

[113]   N.N. Rosanov, et al., Knotted solitons in nonlinear magnetic metamaterials, Phys. Rev. Lett. 108 (4) (2012) 133902.

[114]   M. Marklund, et al., Solitons and decoherence in left-handed metamaterials, Phys. Lett. A 341 (2005) 231−234.

[115]   R. Noskov, P. Belov, Y. Kivshar, Oscillons, solitons, and domain walls in arrays of nonlinear plasmonic nanoparticles, Scientific Rep. 2 (8) (2012) 873.

[116]   Y.M. Liu, et al., Subwavelength discrete solitons in nonlinear metamaterials, Phys. Rev. Lett. 99 (4) (2007) 153901.

[117]   A.R. Katko, et al., Phase conjugation and negative refraction using nonlinear active metamaterials, Phys. Rev. Lett. 105 (4) (2010) 123905.

[118]   Y.J. Xiang, et al., Modulation instability in metamaterials with saturable nonlinearity, J. Opt. Soc. Am. B 28 (4) (2011) 908−916.

[119]   Y.J. Xiang, et al., Modulation instability induced by nonlinear dispersion in nonlinear metamaterials, J. Opt. Soc. Am. B 24 (12) (2007) 3058−3063.

[120]   A.I. Maimistov, I.R. Gabitov, Nonlinear optical effects in artificial materials, Eur. Phys. J. Spec. Top. 147 (2007) 265−286.

[121]   I.V. Shadrivov, et al., Nonlinear magnetic metamaterials, Opt. Express 16 (2008) 20266.

[122]   I.V. Shadrivov, et al., Second-harmonic generation in nonlinear left-handed metamaterials, J. Opt. Soc. Am. B 23 (2006) 529−534.

[123]   A.K. Popov, V.M. Shalaev, Negative-index metamaterials: second-harmonic generation, Manley−Rowe relations and parametric amplification, Appl. Phys. B 84 (2006) 131−137.

[124]   A.K. Popov, V.M. Shalaev, Compensating losses in negative-index metamaterials by optical parametric amplification, Opt. Lett. 31 (2006) 2169−2171.

[125]   V. Roppo, et al., Anomalous momentum states, non-specular reflections, and negative refraction of phase-locked, second-harmonic pulses, Metamaterials Congress 2 (2008) 135−144.

[126] M. Scalora, et al., Dynamics of short pulses and phase matched second harmonic generation in negative index materials, Opt. Express 14 (2006) 4746−4756.

[127] V. Roppo, et al., Role of phase matching in pulsed second-harmonic generation: walk-off and phase-locked twin pulses in negative-index media, Phys. Rev. A 76 (2007) 033829.

[128] P.Y.P. Chena, B.A. Malomed, Single- and multi-peak solitons in two-component models of metamaterials and photonic crystals, Opt. Commun. 283 (2010) 1598−1606.

[129] I.V. Shadrivov, et al., Metamaterials controlled with light, Phys. Rev. Lett. 109 (2012) 083902.

[130] A.K. Popov, et al., Four-wave mixing, quantum control, and compensating losses in doped negative-index photonic metamaterials, Opt. Lett. 32 (2007) 3044−3046.

[131] S. Gao, S. He, Four-wave mixing in left-handed materials, J. Nonlinear Optic. Phys. Mater. 16 (2007) 485.

[132] M. Scalora, et al., Generalized nonlinear Schrodinger equation for dispersive susceptibility and permeability: application to negative index materials, Phys. Rev. Lett. 95 (2005) 013902−013904. Erratum, Phys. Rev. Lett. 2005, 95, 239902(E).

[133] D. Huang, et al., Wave mixing in nonlinear magnetic metacrystal, Appl. Phys. Lett. 98 (2011) 204102.

[134] J.B. Pendry, et al., Magnetism from conductors and enhanced nonlinear phenomena, IEEE Trans. Microwave Theory Tech. 47 (1999) 2075−2084.

[135] M.W. Klein, et al., Experiments on second- and third-harmonic generation from magnetic metamaterials, Opt. Express 15 (2007) 5238−5247.

[136] N. Feth, et al., Second-harmonic generation from complementary split-ring resonators, Opt. Lett. 33 (2008) 1975−1977.

[137] F.B.P. Niesler, et al., Second-harmonic generation from split-ring resonators on a GaAs substrate, Opt. Lett. 34 (2009) 1997−1999.

[138] D. de Ceglia, et al., Enhancement and inhibition of second-harmonic generation and absorption in a negative index cavity, Opt. Lett. 32 (2007) 265−267.

[139] G. D'Aguanno, et al., Second-harmonic generation at angular incidence in a negative-positive index photonic band-gap structure, Phys. Rev. E 74 (2006) 026608.

[140] M. Lapine, Magnetoelastic metamaterials, Nat. Mater. 11 (2012) 33012.

[141] M. Lapine, et al., Metamaterials with conformational nonlinearity, Sci. Rep. 1 (2011) 138.

[142] N.A. Zharova, et al., Nonlinear control of invisibility cloaking, Opt. Express 20 (2012) 14954−14959.

[143] A. Pandey, N.M. Litchinitser, Nonlinear light concentrators, Opt. Lett. 37 (2012) 5238−5240.

[144] I.V. Shadrivov, et al., Metamaterials controlled with light, Phys. Rev. Lett. 109 (4) (2012) 083902.

[145] C.M. Watts, X. Liu, W.J. Padilla, Metamaterial electromagnetic wave absorbers, Adv. Mater. 24 (2012) OP98−OP120.

[146] N.I. Landy, S. Sajuyigbe, J.J. Mock, D.R. Smith, W.J. Padilla, Perfect metamaterial absorber, Phys. Rev. Lett. 100 (2008) 207402−207405.

[147] H. Tao, N.I. Landy, C.M. Bingham, X. Zhang, R.D. Averitt, W.J. Padilla, A metamaterial absorber for the terahertz regime: design, fabrication and characterization, Opt. Express 16 (2008) 7181−7188.

[148] H. Tao, C.M. Bingham, A.C. Strikwerda, D. Pilon, D. Shrekenhamer, N.I. Landy, K. Fan, X. Zhang, W.J. Padilla, R.D. Averitt, Highly flexible wide angle of incident terahertz metamaterial absorber: design, fabrication, and characterization, Phys. Rev. B 78 (2008) 241103−241106(R).

[149] H. Tao, C.M. Bingham, D. Pilon, et al., A dual band terahertz metamaterial absorber, J. Phys. D: Appl. Phys. 43 (2010) 225102.

[150] J.B. Pendry, D. Schurig, D.R. Smith, Controlling electromagnetic fields, Science 312 (2006) 1780.

[151] U. Leonhardt, Optical conformal mapping, Science 312 (2006) 1777.

[152] U. Leonhardt, T.G. Philbin, General relativity in electrical engineering, New J. Phys. 8 (2006) 247.

[153] D. Schurig, J.J. Mock, B.J. Justice, S.A. Cummer, J.B. Pendry, A.F. Starr, D.R. Smith, Metamaterial electromagnetic cloak at microwave frequencies, Science 314 (2006) 977−980.

[154] W. Cai, U.K. Chettiar, A.V. Kildishev, V.M. Shalaev, Optical cloaking with metamaterials, Nat. Photonics 1 (2007) 224.

[155] W. Cai, U.K. Chettiar, A.V. Kildishev, V.M. Shalaev, G. Milton, Nonmagnetic cloak with minimized scattering, Appl. Phys. Lett. 91 (2007) 111105.

[156] M. Rahm, D. Schurig, D.A. Roberts, S.A. Cummer, D.R. Smith, J.B. Pendry, Design of electromagnetic cloaks and concentrators using form-invariant coordinate transformations of Maxwell's equations, Photonics Nanostruct. Fundam. Appl. 6 (2008) 87.

[157] A.V. Kildishev, V.M. Shalaev, Engineering space for light via transformation optics, Opt. Lett. 33 (2008) 43.

[158] W.X. Jiang, T.J. Cui, X.Y. Zhou, X.M. Yang, Q. Cheng, Arbitrary bending of electromagnetic waves using realizable inhomogeneous and anisotropic materials, Phys. Rev. E 78 (2008) 066607.

[159] W. Cai, V.M. Shalaev, Optical Metamaterial Fundamentals and Applications, Springer-Verlag, NY, 2010.

[160] E.E. Narimanov, A.V. Kildishev, Optical black hole: broadband omnidirectional light absorber, Appl. Phys. Lett. 95 (2009) 041106.

[161] Y. Lai, J. Ng, H.Y. Chen, D. Han, J. Xiao, Z.-Q. Zhang, C.T. Chan, Illusion optics: the optical transformation of an object into another object, Phys. Rev. Lett. 102 (2009) 253902.

[162] L.S. Dolin, On the possibility of comparison of three-dimensional electromagnetic systems with nonuniform anisotropic filling, Izv. VUZov, Radiofizika 4 (1961) 964.

[163] J. Ward, J.B. Pendry, Refraction and geometry in Maxwell's equations, J. Mod. Opt. 43 (1996) 773.

[164] A.V. Kildishev, W. Cai, U.K. Chettiar, V.M. Shalaev, Transformation optics: approaching broadband electromagnetic cloaking, New J. Phys. 10 (2008) 115029.

[165] J. Li, J.B. Pendry, Hiding under the carpet: a new strategy for cloaking, Phys. Rev. Lett. 101 (2008) 203901.

[166] J. Valentine, J. Li, T. Zentgraf, G. Bartal, X. Zhang, An optical cloak made of dielectrics, Nat. Mater. 8 (2009) 568.

[167] R. Liu, C. Ji, J.J. Mock, J.Y. Chin, T.J. Cui, D.R. Smith, Broadband ground-plane cloak, Science 323 (2009) 366.

[168] L.H. Gabrielli, J. Cardenas, C.B. Poitras, M. Lipson, Silicon nanostructure cloak operating at optical frequencies, Nat. Photonics 3 (2009) 461.

[169] I.I. Smolyaninov, V.N. Smolyaninova, A.V. Kildishev, V.M. Shalaev, Anisotropic metamaterials emulated by tapered waveguides: application to optical cloaking, Phys. Rev. Lett. 102 (2009) 213901.

[170] J.B. Pendry, Negative refraction makes a perfect lens, Phys. Rev. Lett. 85 (2000) 3966.

[171] S.A. Cummer, Simulated causal subwavelength focusing by a negative refractive index slab, Appl. Phys. Lett. 82 (2003) 1503−1505.

[172]   N. Fang, H. Lee, C. Sun, X. Zhang, Sub-diffraction limited optical imaging with a silver superlens, Science 308 (2005) 534−537.

[173]   K. Aydin, I. Bulu, E. Ozbay, Focusing of electromagnetic waves by a left-handed metamaterial flat lens, Opt. Express 13 (2005) 8753−8759.

[174]   Z. Jacob, L.V. Alekseyev, E. Narimanov, Optical hyperlens: far-field imaging beyond the diffraction limit, Opt. Express 14 (2006) 8247−8256.

[175]   A. Fang, T. Koschny, C.M. Soukoulis, et al., Optical anisotropic metamaterials: negative refraction and focusing, Phys. Rev. B 79 (2009) 245127−245133.

[176]   H. Lee, Z.W. Liu, Y. Xiong, et al., Development of optical hyperlens for imaging below the diffraction limit, Opt. Express 15 (2007) 15886−15891.

This page intentionally left blank

A dynamical, classical oscillator model for linear and nonlinear optics

10

M. Scalora
Charles M. Bowden Research Center, AMRDEC, RDECOM, AL, USA

In this chapter we will build and develop a self-consistent, classical oscillator model to describe linear and nonlinear optical interactions like refraction and frequency conversion in both centrosymmetric and noncentrosymmetric materials. In addition to being quite ubiquitous in all of physics, the classical oscillator model of matter is an enormously pedagogical tool that serves as a natural springboard to the description and understanding of quantum systems and leads to a rather detailed portrayal of all the dynamical factors that contribute to most linear and nonlinear optical phenomena. The method is endowed with causality as well as a natural degree of self-consistency that includes linear and nonlinear material dispersions, elements that are usually necessary to understand many of the subtleties of the interaction of light with matter. By way of examples, using this classical approach we will examine harmonic generation in bulk materials and in metal-based nanostructures. In centrosymmetric materials like metals (materials composed of molecules that lack a center of symmetry), second harmonic generation (SHG) arises mostly from nearly free, conduction electrons (nearly free because they are confined by the metal walls) and is due to a combination of spatial symmetry breaking (interfaces), the magnetic portion of the Lorentz force, and, to a lesser extent, the interaction of third harmonic (TH) and pump photons. By the same token, the third order nonlinearity ($\chi^{(3)}$) gives rise to most of the TH signal, while to a small degree the interaction of pump and SH photons also produces cascaded, TH photons. The classical oscillator model will be pivotal in these systems as well, where a combination of free (Drude) and bound (Lorentz) electrons suffices to describe most linear and nonlinear optical phenomena.

It is well known that the nonlinear polarization of a medium may be written according to a well-established order where the electric dipole contribution is much larger than the combination of electric quadrupole and magnetic dipole, which in turn is much larger than the combination of electric octupole and magnetic quadrupole, and so on. Using this classification, the lowest order contributions to the nonlinear polarization of a generic medium may be written as

$$P_i^{(\mathrm{NL})} = \chi_{ijk}^{(2),\mathrm{ed}} E_i E_j + \chi_{ijkl}^{(3),\mathrm{ed}} E_j E_k E_l + \chi_{ijk}^{(2),\mathrm{md}} E_j B_j + \chi_{ijkl}^{(2),\mathrm{eq}} E_j \nabla_k E_l + \dots,$$

$$(10.1)$$

where the subscripts i,j,k,l are Cartesian coordinates; $\chi_{ijk}^{(2),ed}$ and $\chi_{ijkl}^{(3),ed}$ are the tensor components of second and third order nonlinear coefficients; the superscripts ed, md, eq stand for electric dipole, magnetic dipole, and electric quadrupole, respectively; $E_{i,j,k}$ and $B_{i,j,k}$ are the Cartesian components of the electric and magnetic fields; and the usual summation convention has been assumed on the right-hand side. Rather than using Eqn (10.1) as a starting point, the classical, local oscillator model—Figure 10.1—seems like a more natural initial step, and so one may simply begin with an equation of motion for a charge assumed to be under the action of internal forces (damping, harmonic and anharmonic restoring forces—electron on a spring, also known as the Lorentz model of the atom) and external forces due to the applied electromagnetic fields. The charge moves against a much more massive nucleus, which for simplicity is assumed to be at rest. A possible, quite basic way to describe nonlinear optical processes is then to modify the Lorentz model, described by Eqn (2.20), by introducing appropriate nonlinear terms. For example, neglecting for the moment electric quadrupole and higher order contributions, to the lowest order we may write

$$m^*\ddot{\mathbf{r}}(t) + \gamma\, m^*\dot{\mathbf{r}}(t) + k\, \mathbf{r}(t) - m^*b(\mathbf{r}(t)\cdot\mathbf{r}(t))\mathbf{r}(t) \;=\; \mathrm{e}(\mathbf{E}(\mathbf{r}, t) + \dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r}, t)),$$
(10.2)

where m^* is the effective mass of the oscillator; k is the spring constant associated with a linear restoring force; b is a coefficient associated with a nonlinear, third order restoring force; e is the charge; and γ is the damping coefficient that denotes a certain rate at which the oscillator re-emits energy into its surroundings. Implicit in Eqn (10.2) are two assumptions: (1) the second order bulk nonlinear response is null, i.e., the medium is centrosymmetric; and (2) the medium is also isotropic. The anisotropy may be reintroduced by assuming different spring constants, effective masses, and damping coefficients in the separate spatial directions.
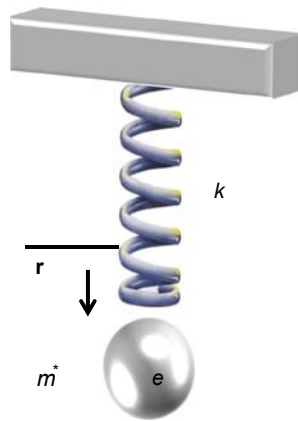
Figure 10.1 Lorentz model of the atom, consisting of an electron bound to the atom by a spring that provides a restoring force. The charge e is subject to external electric and magnetic forces and to internal linear and nonlinear restoring forces. **r** is the displacement from equilibrium.

Before delving into the more complex aspects of wave and/or pulse propagation phenomena, it is worthwhile to explore some of the consequences relating to the form of Eqn (10.2). For instance, one may ask the following: how does one arrive at a third order nonlinear optical coefficient from Eqn (10.2) to describe third harmonic generation (THG)? For THG, incident pump photons of frequency ω are generally converted to photons at 3ω. Therefore, both pump and TH photons will be present simultaneously from the very start in the environment that envelopes the oscillator. As a result, at the electron position one may write the total field present as a superposition of fundamental frequency (FF) and TH fields as follows:

$$
\begin{aligned}
\mathbf{E} &= \mathbf{E}_{\omega}\, e^{-i\omega t} + \mathbf{E}_{3\omega}\, e^{-3i\omega t} + c.c. \\
\mathbf{H} &= \mathbf{H}_{\omega}\, e^{-i\omega t} + \mathbf{H}_{3\omega}\, e^{-3i\omega t} + c.c.
\end{aligned}
\tag{10.3}
$$

Here $\mathbf{H}_{\omega}$, $\mathbf{E}_{\omega}$, $\mathbf{H}_{3\omega}$, $\mathbf{E}_{3\omega}$ are complex field amplitudes that for the moment are assumed to be nearly constant in time. That is to say, we are assuming incident pulses are at least several tens of optical cycles in duration or slowly varying in time as compared to the optical period. As a result, a general, simplified solution for the electron displacement from its equilibrium position may also be written as a superposition of FF and TH amplitudes that vary relatively slowly in time as compared to the optical cycle, represented by the carrier frequency ω, so that

$$
\mathbf{r} = \mathbf{r}_{\omega} e^{-i\omega t} + \mathbf{r}_{3\omega} e^{-3i\omega t} + c.c. \tag{10.4}
$$

Substituting Eqns (10.3) and (10.4) into Eqn (10.2), retaining lowest order terms, neglecting the magnetic term, and equating terms that oscillate at the same frequency, one obtains

$$
\begin{aligned}
\mathbf{r}_{\omega} &= \frac{e}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}\mathbf{E}_{\omega} + \frac{b}{\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}|\mathbf{r}_{\omega}|^2 \mathbf{r}_{\omega} + \dots \\
\mathbf{r}_{3\omega} &= \frac{e}{m^*\left(\omega_0^2 - 9\omega^2 + 3i\gamma\omega\right)}\mathbf{E}_{3\omega} + \frac{b}{\left(\omega_0^2 - 9\omega^2 + 3i\gamma\omega\right)}(\mathbf{r}_{\omega}\cdot\mathbf{r}_{\omega})\mathbf{r}_{\omega} + \dots
\end{aligned}
\tag{10.5}
$$

where $\omega_0^2 = \frac{k}{m^*}$ is the oscillator's resonance frequency. It is understood that each nonlinear contribution is much smaller than the corresponding, leading linear term. Solutions may now be found for the amplitudes $\mathbf{r}_{\omega}$ and $\mathbf{r}_{3\omega}$ in Eqn (10.5). For instance, the first of Eqn (10.5) may be recast as follows:

$$
\mathbf{r}_{\omega} = \frac{e}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)\left(1 - \frac{b}{\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}|\mathbf{r}_{\omega}|^2\right)}\mathbf{E}_{\omega}. \tag{10.6}
$$

Using the binomial expansion on the nonlinear denominator in Eqn (10.6) leads to

$$\mathbf{r}_\omega = \frac{e}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}\mathbf{E}_\omega\left(1 + \frac{b}{\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}|\mathbf{r}_\omega|^2 + \ldots\right), \qquad (10.7)$$

which to first order in the field amplitude yields the solution

$$\mathbf{r}_\omega = \frac{e}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}\mathbf{E}_\omega + \frac{be^3}{m^{*3}\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)^3\left(\omega_0^2 - \omega^2 - i\gamma\omega\right)}|\mathbf{E}_\omega|^2\mathbf{E}_\omega. \qquad (10.8a)$$

We can now use Eqn (10.8) to solve for the second of Eqn (10.5). Upon substitution we find

$$\mathbf{r}_{3\omega} = \frac{e}{m^*\left(\omega_0^2 - 9\omega^2 + 3i\gamma\omega\right)}\mathbf{E}_{3\omega}$$

$$+ \frac{be^3}{m^{*3}\left(\omega_0^2 - 9\omega^2 + 3i\gamma\omega\right)\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)^3}\left(\mathbf{E}_\omega\cdot\mathbf{E}_\omega\right)\mathbf{E}_\omega. \qquad (10.8b)$$

Recognizing that $\mathbf{p}_{\omega,3\omega} = e\mathbf{r}_{\omega,3\omega}$ is the dipole moment and that $\mathbf{P}_{\omega,3\omega} = Ne\mathbf{r}_{\omega,3\omega}$ is the total polarization density per unit volume (N is the number of oscillators in the volume of interest), as defined in Eqn (2.19), we may write

$$\mathbf{P}_\omega = \varepsilon_0\chi_\omega^{(1)}\mathbf{E}_\omega + \varepsilon_0\chi_\omega^{(3)}|\mathbf{E}_\omega|^2\mathbf{E}_\omega, \qquad (10.9a)$$

and

$$\mathbf{P}_{3\omega} = \varepsilon_0\chi_{3\omega}^{(1)}\mathbf{E}_\omega + \varepsilon_0\chi_{3\omega}^{(3)}\left(\mathbf{E}_\omega\cdot\mathbf{E}_\omega\right)\mathbf{E}_\omega, \qquad (10.9b)$$

where ε_0 is the permittivity of free space, and

$$\chi_\omega^{(1)} = \frac{Ne^2}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}; \quad \chi_\omega^{(3)} = \frac{Nbe^4}{m^{*3}\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)^3\left(\omega_0^2 - \omega^2 - i\gamma\omega\right)} \qquad (10.10)$$

and

$$\chi_{3\omega}^{(3)} = \frac{Nbe^4}{m^{*3}\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)^3\left(\omega_0^2 - 9\omega^2 + 3i\gamma\omega\right)} \qquad (10.11)$$

are the derived linear $(\chi_\omega^{(1)})$ and nonlinear $(\chi_{\omega,3\omega}^{(3)})$ medium susceptibilities experienced by the FF and TH fields. The second part of Eqn (10.10) reveals that this medium is endowed with nonlinear refraction, or self-phase modulation (SFM), via the $\mathrm{Re}\chi_\omega^{(3)}$ portion, and by so-called two-photon absorption via the term $\mathrm{Im}\chi_\omega^{(3)}$. By the same token, Eqn (10.11) shows that the generated TH signal experiences both nonlinear refraction and gain/loss via the terms $\mathrm{Re}\chi_{3\omega}^{(3)}$ and $\mathrm{Im}\chi_{3\omega}^{(3)}$, respectively. Finally, the linear dielectric function of this prototype, bulk medium may be written in terms of the linear susceptibility Eqn (10.10) as follows:

$$\varepsilon(\omega) = 1 + \chi_\omega^{(1)} = 1 + \frac{Ne^2}{m^*\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)}. \tag{10.12}$$

In Figure 10.2 we plot the resulting dielectric function (Figure 10.2(a)) and index of refraction (Figure 10.2(b)) of a fictional medium having an absorption resonance at $\lambda_0 = 200$ nm, damping coefficient $\gamma = 4.5 \times 10^{13}$ s^{-1} (relaxation time of order 1 ps), $m^* = m_e$ the rest mass of the electron, and number density $N = 10^{28}$ m^{-3}. At visible and longer wavelengths the value of the dielectric constant is $\varepsilon \approx 2$, which corresponds to an index of refraction $n = \sqrt{\varepsilon} \sim 1.41$.

Before we are able to quantify the magnitude of the respective nonlinear susceptibilities, $\chi_{\omega,3\omega}^{(3)}$, we must first estimate the value of the nonlinear coefficient b. Equation (10.2) contains two spring-related, internal forces acting on the charge. One may argue that the nonlinear term becomes important when linear and nonlinear internal spring-specific forces are of the same order of magnitude. That is to say, for a given spring deformation or stretch $\mathbf{r}_0$ one should have $k\mathbf{r}_0 \approx m^*b(\mathbf{r}_0 \cdot \mathbf{r}_0)\mathbf{r}_0$, which yields $b \approx \omega_0^2/|\mathbf{r}_0|^2$. Taking $\omega_0 = 2\pi c/\lambda_0 \sim 10^{16}$ s^{-1} and assuming that maximum spring
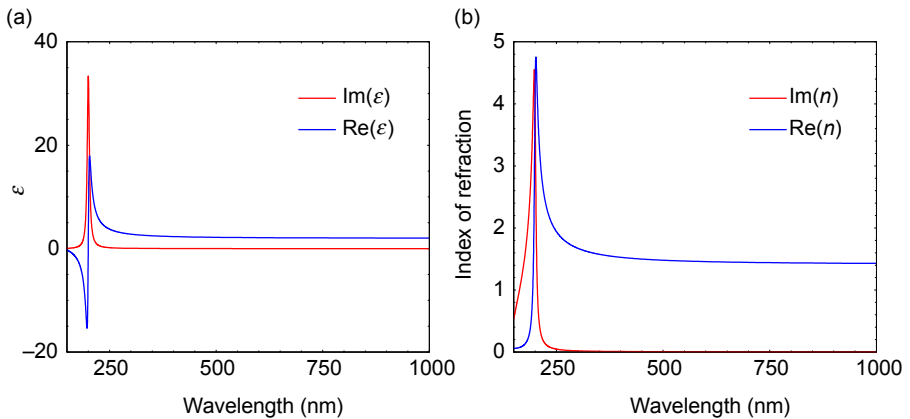
Figure 10.2 Dielectric constant (a) and index of refraction (b) for the hypothetical medium formulated in Eqn (10.12).
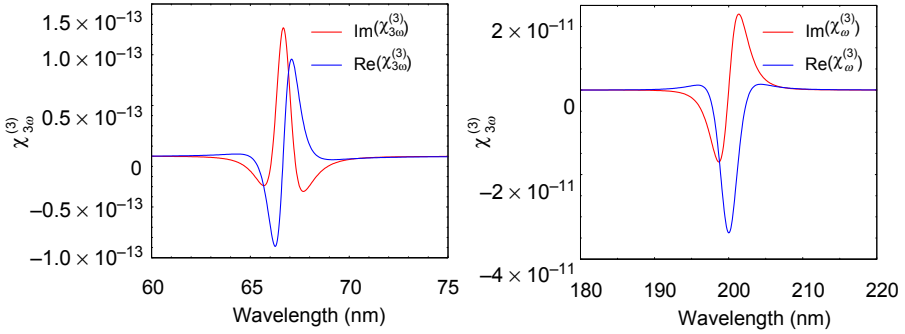
Figure 10.3 Nonlinear susceptibility for THG (left) and self-phase modulation and nonlinear absorption (right), obtained from Eqns (10.10) and (10.11). The nonlinear dispersion curves are very sensitive to the linear fit and the effective mass m^*. We have assumed the effective mass $m^* \approx m_e$, the electron mass.

deformation is comparable to the atomic diameter in question, we may assume that the value $|r_0| \sim 3$ Å (typical atomic diameter) represents a large, upper estimate of the oscillation amplitude. In Figure 10.3 we show the resulting nonlinear susceptibilities $\chi_{\omega,3\omega}^{(3)}$, as given by Eqns (10.10) and (10.11).

Given the simplicity of the model, the question that comes to mind is as follows: how do the predicted nonlinear susceptibilities compare with known values of $\chi_{\omega,3\omega}^{(3)}$? For example, it is known that *far from resonance* $\left|\chi_\omega^{(3)}\right|$ ranges from 10^{-21} (m/V)^2 for typical dielectric materials to 10^{-18} (m/V)^2 for semiconductors like GaAs or GaP and to 10^{-15} (m/V)^2 for metals like Cu and Au. While in Figure 10.3 $\left|\chi_\omega^{(3)}\right|$ is of order 10^{-11} (m/V)^2 near the resonance at 200 nm, far from resonance we find $\left|\chi_\omega^{(3)}\right| \sim 10^{-17}$ (m/V)^2, an estimate that is in line with most materials. More precise estimates may be obtained using the classical model by taking into account oscillator strengths (i.e., effective plasma frequency for the specific material), more accurate effective masses and atomic densities, damping coefficients, and maximum stretch or oscillation amplitude. Finally, we note that the same strategy may be followed to estimate the second order, nonlinear coefficients of noncentrosymmetric materials, $\chi_\omega^{(2)}$.

10.1 Linear and nonlinear refraction of ultrashort pulses: third harmonic generation

As we have seen above, the oscillator model described by Eqn (10.2) naturally contains both linear and nonlinear medium responses. In this section we will study the dynamics of an incident pump pulse as it refracts and propagates inside such a medium
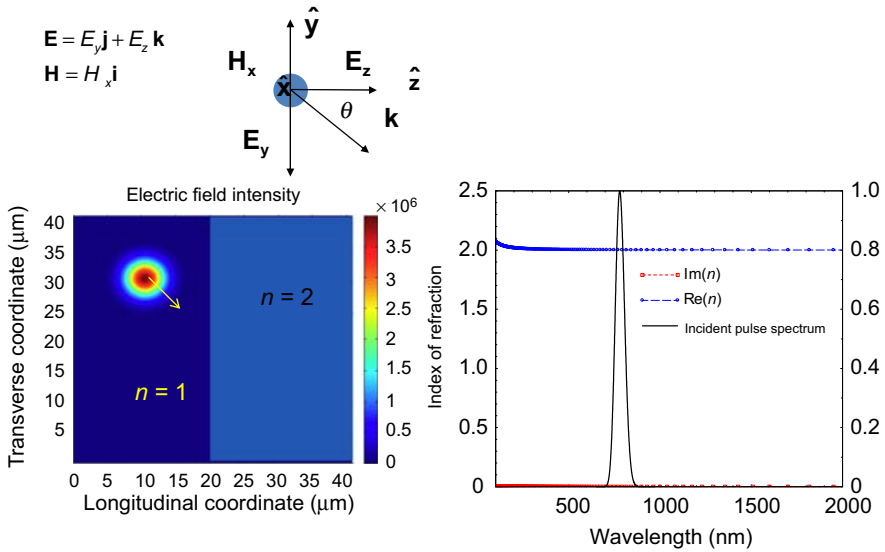
Figure 10.4 Left: an incident, 25 fs Gaussian pulse is initially located in vacuum ($n = 1$), just outside a material of index of refraction $n = 2$. Right: spectrum of a 25 fs pulse (solid line). The medium is nearly dispersionless and absorptionless across the entire visible spectrum.

and illustrate linear refraction of the pump accompanied by THG. Equation (10.2) must be coupled to Maxwell's equation in order for the wave to experience the medium. In Figure 10.4 we show a geometrical arrangement that contains a Gaussian pulse tuned to 780 nm (that is, the pulse's carrier or center wavelength), is approximately 25 fs in duration, and is incident at 45° with respect to the normal on a surface that separates vacuum ($n = 1$) from a medium whose index of refraction $n = 2$ (see Figure 10.4). The parameters of this hypothetical medium are chosen so that it displays little dispersion and practically no absorption [Im(n) $\sim$ 0] across the spectral range of interest, as depicted in Figure 10.4. Snell's law predicts that a light ray or beam will refract at an angle $\theta \approx 20.7°$. In Figure 10.5 we show several snapshots of the incident pulse as it traverses the surface and settles into the medium. Scattering from the smooth surface produces reflected (into vacuum) and transmitted (into the medium) pulses. The spatial compression of the transmitted pulse along the direction of propagation is due to the reduced wave velocity down to $\sim c/2$ and results in a corresponding broadening of the pulse in k-space. Snell's law is usually derived for incident monochromatic waves and smooth surfaces. The absence of extraneous scattering elements, apertures, or corners on the surface leads to the conservation of transverse momentum. For a wave incident on a surface at angle θ_{inc}, conservation of the k-vector component that points along the transverse coordinate yields $n_1\sin\theta_{inc} = n_2\sin\theta_{ref}$, where θ_{ref} is the angle of refraction. However, while this boundary condition is easily deduced from a conservation law, it must already be contained implicitly within Maxwell's equations. So, how does one go about extracting the angle of refraction for an incident pulse, using Maxwell's equations? We proceed to do so next.
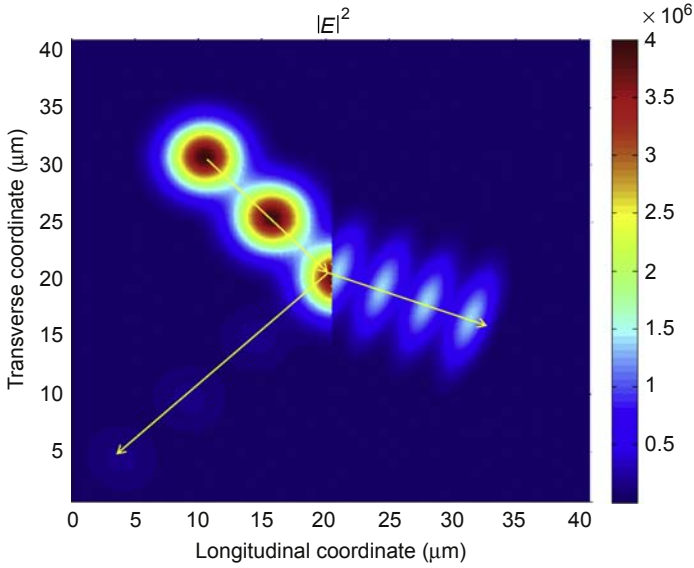
Figure 10.5 Sequential snapshots that depict the refraction of the pulse in Figure 10.4 as it happens. The transmitted pulse is flattened in the direction of propagation, as light slows down, while the refracted pulse is attenuated but does not change its shape.

The frequency makeup of a 25 fs pulse is several tens of nanometers (see Figure 10.4), but the flat dispersion curve ensures no significant departure from Snell's law's prediction for a single frequency component. Without loss of generality we may assume the magnetic field is linearly polarized along the x-direction (transverse magnetic, or TM) and points into the page, as illustrated on the sketch on Figure 10.4, and is incident at an arbitrary angle θ_{inc}. We thus expand the fields as follows:

$$
\mathbf{E} = E_y\mathbf{j} + E_z\mathbf{k} = \left(E_{\omega,y}(y,z,t)e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + E^*_{\omega,y}(y,z,t)e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \right)\widehat{\mathbf{y}}
$$

$$
+ \left(E_{\omega,z}(y,z,t)e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + E^*_{\omega,z}(y,z,t)e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \right)\widehat{\mathbf{z}},
$$

$$
\mathbf{H} = H_x\mathbf{i} = \left(H_{\omega,x}(y,z,t)e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} + H^*_{\omega,x}(y,z,t)e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \right)\widehat{\mathbf{x}}
$$

$$
(10.13)
$$

where $k_z = |\mathbf{k}|\cos\theta_i$ and $k_y = -|\mathbf{k}|\sin\theta_i$, $|\mathbf{k}| = k_0 = \omega/c$. This choice of carrier wave vector indicates the pulse is at first traveling along the $-y + z$ directions (down the page and to the right) and is consistent with the fact that the pulse is initially located in vacuum. We make no other assumptions about the makeup of the envelope functions, $E_{\omega,y}(y,z,t)$, $E_{\omega,z}(y,z,t)$, and $H_{\omega,x}(y,z,t)$, which are substituted into the vector Maxwell

equations and are allowed to evolve according to whatever physical boundaries and media are present within the spatial grid. Put another way, the fields' phase and amplitude are allowed to evolve free of preconditions, as functions of position and time, from an initial Gaussian centered at y_0 and z_0 such that $H_{\omega,x}(y, z, t = 0) = H_0 e^{-\frac{(z-z_0)^2}{w_z^2} - \frac{(y-y_0)^2}{w_y^2}}$, as depicted in Figure 10.4. w_z and w_y are the respective spatial pulse widths along the indicated directions. Using Eqn (10.13), we can explicitly write out the Poynting vector, which is equivalent to the *electromagnetic momentum density*, in terms of the field envelope functions. Neglecting components that oscillate at twice the carrier frequency, because they average out over an optical cycle, the Poynting vector has two components on the y-z plane:

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} = \mathbf{k}S_z(y, z, t) + \mathbf{j}S_y(y, z, t)$$
$$= -\mathbf{k}\left[E_y H_x^* + E_y^* H_x\right] + \mathbf{j}\left[E_z H_x^* + E_z^* H_x\right]. \qquad (10.14a)$$

The Poynting vector determines the direction of *energy flow*, which is usually the same as the direction of *phase refraction* as given by Snell's law for isotropic, homogeneous bulk materials, but may be different, as is the case for anisotropic or negative index materials. Maxwell's equations also provide a way to extract the phase refraction angle, which may be determined by performing a spectral decomposition of the fields and by calculating the wave vector's expectation value $\mathbf{K}(t)$ as a function of time, namely

$$\langle \mathbf{K}(t) \rangle = \int\limits_{k_y=-\infty}^{k_y=\infty} \int\limits_{k_z=-\infty}^{k_z=\infty} \mathbf{k}|H(\mathbf{k}, t)|^2 d\mathbf{k}. \qquad (10.14b)$$

One may also calculate the energy velocity as $\mathbf{V}_E = \frac{\langle \mathbf{S} \rangle}{\langle U \rangle}$, where U is the energy density [Eqn (2.58)]. The brackets denote spatial averages over the volume of interest. The solution to the problem is thus found by solving a set of coupled Maxwell-oscillator equations, namely

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t} \qquad (10.15)$$

$$\ddot{\mathbf{P}} + \gamma\dot{\mathbf{P}} + \omega_0^2 \mathbf{P} = \frac{Ne^2}{m}\mathbf{E}$$

For simplicity we temporarily neglect the nonlinear spring term and the magnetic portion of the Lorentz force. Substituting Eqn (10.13) into (10.15), we

obtain a set of coupled equations for the fields, currents, and polarization envelopes

$$\frac{\partial H_{\tilde{x}}}{\partial \tau} = i\beta\left(H_{\tilde{x}} + E_{\tilde{z}}\sin\theta_i + E_{\tilde{y}}\cos\theta_i\right) - \frac{\partial E_{\tilde{z}}}{\partial \tilde{y}} + \frac{\partial E_{\tilde{y}}}{\partial \tilde{z}}$$

$$\frac{\partial E_{\tilde{y}}}{\partial \tau} = i\beta\left(E_{\tilde{y}} + H_{\tilde{x}}\cos\theta_i\right) + \frac{\partial H_{\tilde{x}}}{\partial \tilde{z}} - 4\pi\left(J_{\tilde{y}} - i\beta P_{\tilde{y}}\right)$$

$$\frac{\partial E_{\tilde{z}}}{\partial \tau} = i\beta(E_{\tilde{z}} + H_{\tilde{x}}\sin\theta_i) - \frac{\partial H_{\tilde{x}}}{\partial \tilde{y}} - 4\pi(J_{\tilde{z}} - i\beta P_{\tilde{z}})$$

$$\frac{\partial J_{\tilde{y}}}{\partial \tau} = (2i\beta - \tilde{\gamma})J_{\tilde{y}} + \left(\beta^2 + i\tilde{\gamma}\beta - \beta_0^2\right)P_{\tilde{y}} + \frac{\pi\omega_p^2}{\omega_r^2}E_{\tilde{y}} \qquad (10.16)$$

$$\frac{\partial J_{\tilde{z}}}{\partial \tau} = (2i\beta - \tilde{\gamma})J_{\tilde{z}} + \left(\beta^2 + i\tilde{\gamma}\beta - \beta_0^2\right)P_{\tilde{z}} + \frac{\pi\omega_p^2}{\omega_r^2}E_{\tilde{z}}$$

$$\frac{\partial P_{\tilde{y}}}{\partial \tau} = J_{\tilde{y}}$$

$$\frac{\partial P_{\tilde{z}}}{\partial \tau} = J_{\tilde{z}}$$

We have used the scaled coordinates $\tilde{z} = z/\lambda_r$, $\tilde{y} = y/\lambda_r$, $\tilde{x} = x/\lambda_r$, and $\tau = ct/\lambda_r$, and scaled the frequencies $\beta = 2\pi\omega/\omega_r$, $\beta_0 = 2\pi\omega_0/\omega_r$, and damping coefficient $\tilde{\gamma} = \gamma(\lambda_r/c)$, where $\lambda_r = 1\,\mu m$ is a conveniently chosen reference wavelength such that $\omega_r = 2\pi c/\lambda_r$. We note that while the magnetic field is polarized in the x-direction, the fields are independent of x. Although scaling the equations of motion is not of crucial importance, it is nevertheless useful because it disentangles us from specific wavelength ranges. Equation (10.16) is then solved in the time domain using a modified fast Fourier transform pulse propagation method, which we discuss next. Close examination of the electric and magnetic field equations shows that they may be put into a Schrödinger-like form, namely

$$\frac{\partial H_{\tilde{x}}}{\partial \tau} = V_{H_{\tilde{x}}}H_{\tilde{x}} - \frac{\partial E_{\tilde{z}}}{\partial \tilde{y}} + \frac{\partial E_{\tilde{y}}}{\partial \tilde{z}}$$

$$\frac{\partial E_{\tilde{y}}}{\partial \tau} = V_{E_{\tilde{y}}}E_{\tilde{y}} + \frac{\partial H_{\tilde{x}}}{\partial \tilde{z}} - 4\pi\left(J_{\tilde{y}} - i\beta P_{\tilde{y}}\right), \qquad (10.17)$$

$$\frac{\partial E_{\tilde{z}}}{\partial \tau} = V_{E_{\tilde{z}}}E_{\tilde{z}} - \frac{\partial H_{\tilde{x}}}{\partial \tilde{y}} - 4\pi(J_{\tilde{z}} - i\beta P_{\tilde{z}}).$$

The potentials are $V_{H_{\tilde{x}}} = i\beta(H_{\tilde{x}} + E_{\tilde{z}}\sin\theta_i + E_{\tilde{y}}\cos\theta_i)/H_{\tilde{x}}$, $V_{E_{\tilde{y}}} = i\beta(E_{\tilde{y}} + H_{\tilde{x}}\cos\theta_i)/E_{\tilde{y}} - 4\pi(J_{\tilde{y}} - i\beta P_{\tilde{y}})/E_{\tilde{y}}$, and $V_{E_{\tilde{z}}} = i\beta(E_{\tilde{z}} + H_{\tilde{x}}\sin\theta_i)/E_{\tilde{z}} - 4\pi(J_{\tilde{z}} - i\beta P_{\tilde{z}})/E_{\tilde{z}}$. The equations are solvable using the classic, split-step, beam propagation method. The split-step algorithm usually calls for the separation of free-space and material equations and the integration of differential equations that are first order in time. The various parts of Eqn (10.17) are already first order in time, have no approximations, can be immediately separated into free-space and material equations, and are integrated in the time domain. The formal solutions of the free-space propagator may be derived from the free-space equations, obtained by setting the effective potentials equal to zero. Then, Eqn (10.17) is Fourier transformed in space, resulting in

$$\frac{\partial \tilde{H}_{\tilde{x}}}{\partial \tau} = -ik_{\tilde{y}}\tilde{E}_{\tilde{z}} + ik_{\tilde{z}}\tilde{E}_{\tilde{y}}$$

$$\frac{\partial \tilde{E}_{\tilde{x}}}{\partial \tau} = ik_{\tilde{z}}\tilde{H}_{\tilde{x}} \tag{10.18}$$

$$\frac{\partial \tilde{E}_{\tilde{z}}}{\partial \tau} = -ik_{\tilde{y}}\tilde{H}_{\tilde{x}}$$

Equation (10.18) may be integrated simultaneously using a midpoint trapezoidal method, so that

$$\tilde{H}_{\tilde{x}}(\delta\tau) = \tilde{H}_{\tilde{x}}(0) - \frac{ik_{\tilde{y}}\delta\tau}{2}\left(\tilde{E}_{\tilde{z}}(0) + \tilde{E}_{\tilde{z}}(\delta\tau)\right) + \frac{ik_{\tilde{z}}\delta\tau}{2}\left(\tilde{E}_y(0) + \tilde{E}_y(\delta\tau)\right)$$

$$\tilde{E}_{\tilde{y}}(\delta\tau) = \tilde{E}_{\tilde{y}}(0) + \frac{ik_{\tilde{z}}\delta\tau}{2}\left(\tilde{H}_{\tilde{x}}(0) + \tilde{H}_{\tilde{x}}(\delta\tau)\right) \tag{10.19}$$

$$\tilde{E}_{\tilde{z}}(\delta\tau) = \tilde{E}_{\tilde{z}}(0) - \frac{ik_{\tilde{y}}\delta\tau}{2}\left(\tilde{H}_{\tilde{x}}(0) + \tilde{H}_{\tilde{x}}(\delta\tau)\right).$$

The spatial dependence of the fields is implied. Solving for $\tilde{H}_x(\delta t)$ we find

$$\tilde{H}_{\tilde{x}}(\delta\tau) = \tilde{H}_{\tilde{x}}(0)\frac{\left(1 - \frac{(k_{\tilde{y}}^2 + k_{\tilde{z}}^2)\delta\tau^2}{4}\right)}{\left(1 + \frac{(k_{\tilde{y}}^2 + k_{\tilde{z}}^2)\delta\tau^2}{4}\right)} + \frac{(ik_{\tilde{z}}\tilde{E}_{\tilde{y}}(0) - ik_{\tilde{y}}\tilde{E}_{\tilde{z}}(0))\delta\tau}{\left(1 + \frac{(k_{\tilde{y}}^2 + k_{\tilde{z}}^2)\delta\tau^2}{4}\right)}, \tag{10.20}$$

Equation (10.20) is then substituted back into the second and third parts of Eqn (10.19) to calculate the electric fields. All fields are then inverse Fourier

transformed. The propagation step inside the medium is performed by integrating the material equations, also derived from Eqn (10.17), and written in terms of generic envelope functions as

$$\frac{\partial H_{\tilde{x}}}{\partial \tau} = i\beta\left(H_{\tilde{x}} + E_{\tilde{z}}\sin\theta_i + E_{\tilde{y}}\cos\theta_i\right)$$

$$\frac{\partial E_{\tilde{y}}}{\partial \tau} = i\beta\left(E_{\tilde{y}} + H_{\tilde{x}}\cos\theta_i\right) - 4\pi\left(J_{\tilde{y}} - i\beta P_{\tilde{y}}\right)$$

$$\frac{\partial E_{\tilde{z}}}{\partial \tau} = i\beta(E_{\tilde{z}} + H_{\tilde{x}}\sin\theta_i) - 4\pi(J_{\tilde{z}} - i\beta P_{\tilde{z}})$$

$$\frac{\partial J_{\tilde{y}}}{\partial \tau} = (2i\beta - \tilde{\gamma})J_{\tilde{y}} + \left(\beta^2 + i\tilde{\gamma}\beta - \beta_0^2\right)P_{\tilde{y}} + \frac{\pi\omega_p^2}{\omega_r^2}E_{\tilde{y}} \qquad (10.21)$$

$$\frac{\partial J_{\tilde{z}}}{\partial \tau} = (2i\beta - \tilde{\gamma})J_{\tilde{z}} + \left(\beta^2 + i\tilde{\gamma}\beta - \beta_0^2\right)P_{\tilde{z}} + \frac{\pi\omega_p^2}{\omega_r^2}E_{\tilde{z}}$$

$$\frac{\partial P_{\tilde{y}}}{\partial \tau} = J_{\tilde{y}}$$

$$\frac{\partial P_{\tilde{z}}}{\partial \tau} = J_{\tilde{z}}$$

Although we have neglected magnetic currents and polarizations, which typically characterize a magnetically active or negative index material, they may be reintroduced in a straightforward fashion. Then, an approach similar to the solution of Eqn (10.18) may be employed to solve Eqn (10.21). For instance, one may first obtain estimates of all fields, currents, and polarizations at $\tau = \delta\tau$ with a Euler method using only their initial values at $\tau = 0$. Using these estimates, the solutions for the currents are immediate and second order accurate, as follows

$$J_{\tilde{y},\tilde{z}}(\delta t) = J_{\tilde{y},\tilde{z}}(0)\frac{\left(1 + (2i\beta - \tilde{\gamma})\frac{\delta t}{2} + (\beta^2 + i\tilde{\gamma}\beta - \beta_0^2)\frac{\delta t^2}{4}\right)}{\left(1 - (2i\beta - \tilde{\gamma})\frac{\delta t}{2} - (\beta^2 + i\tilde{\gamma}\beta - \beta_0^2)\frac{\delta t^2}{4}\right)}$$

$$+ \frac{(\beta^2 + i\tilde{\gamma}\beta - \beta_0^2)P_{\tilde{y},\tilde{z}}(0)\delta t + \pi\frac{\omega_p^2}{\omega_r^2}\left(E_{\tilde{y},\tilde{z}}(0) + E_{P,\tilde{y},\tilde{z}}(\delta t)\right)\delta t \Big/ 2}{\left(1 - (2i\beta - \tilde{\gamma})\frac{\delta t}{2} - (\beta^2 + i\tilde{\gamma}\beta - \beta_0^2)\frac{\delta t^2}{4}\right)},$$

$$(10.22)$$

where $E_{P,\tilde{y},\tilde{z}}(\delta t)$ are first order accurate, predicted estimates of the fields at time $\tau = \delta\tau$. Once the currents are known, the polarizations may be found using the usual trapezoidal rule:

$$P_{\tilde{y},\tilde{z}}(\delta t) = P_{\tilde{y},\tilde{z}}(0) + \left(J_{\tilde{y},\tilde{z}}(\delta t) + J_{\tilde{y},\tilde{z}}(0)\right)\frac{\delta t}{2}. \tag{10.23}$$

In turn, knowledge of more accurate currents and polarization at time $\delta\tau$ allows second order accurate estimates of all electric and magnetic fields. The process is then repeated several times or until suitable convergence is achieved.

We now set out to establish a standard baseline of what is meant by ordinary refraction using the formalism developed above. In Figure 10.5 we illustrated the case where a 25 fs pulse crosses from vacuum into a material such that the pump experiences $\varepsilon \approx 4$ and $n \approx 2$ (see Figure 10.4). The pulse diameter (which is specified in space and then advanced in time) when the field amplitude is approximately 1/e of the on-axis value is roughly $10\lambda_r$, which translates into a pulse 10 wave cycles (λ_r/c) in duration. We thus avoid complications due to diffraction, at least for short propagation distances on the order of just a few pulse widths, because pulse width is more than several wavelengths long; the incident wave front is, and remains, nearly plane.

As expected, in Figure 10.5 the pulse appears to become compressed into an ellipse whose major axis is perpendicular to the direction of propagation. In the absence of any meaningful dispersion, the energy velocity quickly acquires a value of $c/2$. This slowdown largely reflects a loss in forward momentum, although part of the initial momentum and energy are back-reflected at the interface. The energy refraction angle (i.e., the direction of energy flow) may be calculated by monitoring the normalized, transverse, and longitudinal electromagnetic momenta inside the medium as functions of time:

$$M_{\tilde{z}}(\tau)_{\text{RHS}} = \frac{1}{c^2} \int_{\tilde{z}=0}^{\tilde{z}=\infty} \int_{\tilde{y}=-\infty}^{\tilde{y}=\infty} S_{\xi}(\tilde{y},\tilde{z},\tau)d\tilde{y}d\tilde{z};$$

$$M_{\tilde{y}}(\tau)_{\text{RHS}} = \frac{1}{c^2} \int_{\tilde{z}=0}^{\tilde{z}=\infty} \int_{\tilde{y}=-\infty}^{\tilde{y}=\infty} S_{\tilde{y}}(\tilde{y},\tilde{z},\tau)d\tilde{y}d\tilde{z}. \tag{10.24}$$

In Figure 10.6 we plot typical results for $M_{\tilde{z}}(\tau)_{\text{RHS}}$ and $M_{\tilde{y}}(\tau)_{\text{RHS}}$ versus time in units of $1/c^2$. The momenta are seen to build up until the entire pulse has entered the medium. At steady state, the energy refraction angle may be calculated as $\theta_{\text{ref}}^{\text{energy}} = \tan^{-1}(M_{\tilde{y}}/M_{\tilde{z}})$, which in our example yields the same results predicted by Snell's law to at least two parts in a 1000, i.e., $\theta_{\text{ref}} \approx 20.7°$. On the other hand, the phase refraction angle may be calculated using Eqn (10.15) as
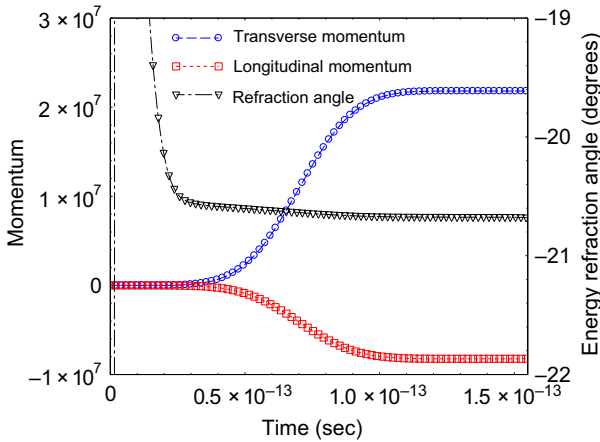
Figure 10.6 Total transverse and longitudinal momenta versus time of a pulse as it traverses into a medium, as in Figure 10.5. The total momentum vector points in a direction determined by $\theta_{ref}^{energy} = \tan^{-1}(M_{\tilde{y}}/M_{\tilde{z}})$, which is the direction of energy flow.

$\theta_{ref}^{phase} = \tan^{-1}(K_{\tilde{y}}/K_{\tilde{z}})$ and is identical to the energy refraction angle up to three decimal places. The differences are likely due mostly to numerical round-off and to a lesser extent to finite pulse bandwidth.

Having established a connection between phase and energy refraction, we are now ready to re-examine the dynamics by introducing a nonzero b-coefficient in Eqn (10.2). The analysis following Eqn (10.2) demonstrated that the presence of a third order nonlinear term leads to a combination of nonlinear refraction and absorption for the pump and the appearance of a TH signal, as represented by effective, complex nonlinear coefficients. In Figure 10.7 we display the results when $|b| \approx 10^{-9}$.
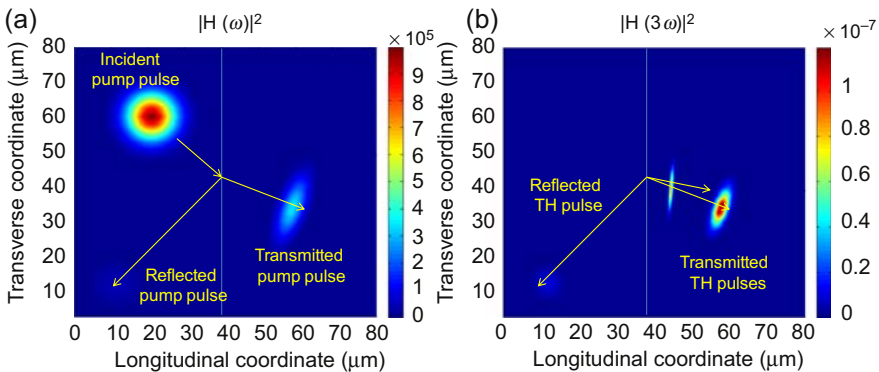
Figure 10.7 (a) Incident and scattered (transmitted and reflected) pump pulses. (b) Generated TH signal. The transmitted portion contains two signals: the inhomogeneous portion, which leads and is locked to the transmitted pump pulse, and the homogeneous signal, which lags and travels with the actual dispersion of the material at the TH wavelength. Absorption is neglected.

In Figure 10.7(a) we exhibit two snapshots of the pump's magnetic field intensity at $\tau = 0$ (incident pulse) and at some time after it has crossed the interface of our hypo-thetical medium (transmitted and reflected pulses). In Figure 10.7(b) we show the scat-tered TH signal corresponding to the final pump pulse snapshot. In addition to the reflected pulse, two separate, transmitted, TH pulses are visible, one precisely overlap-ping the pump pulse and the other lagging behind, traveling at a slower speed (as the enhanced pulse compression suggests) and in a slightly different direction.

The fact that for oblique incidence the TH signal presents two spatially separated solutions has been known for a long time. The presence of two separate pulses of the same frequency that appear to travel separately should not be surprising, as may be ascertained by performing a simple analysis of the differential equations that govern the dynamics. As we have seen above, the third order term in Eqn (10.2) leads directly to a TH field equation having the dominant, nonlinear driving term $\mathbf{P}_{3\omega}^{\mathrm{NL}} = \varepsilon_0 \chi_{3\omega}^{(3)} \mathbf{E}_{\omega}^3$. Let us now consider a particular scenario of Eqn (10.16) that corresponds to incident plane waves, which allows us to set terms that contain time derivatives equal to zero and thus convert the pulses into beams. Then, elimination of the magnetic field equation leads directly to the Helmholtz equation for the TH field amplitude inside the medium:

$$\nabla^2 \mathbf{E}_{3\omega} + \frac{9n_{3\omega}^2 \omega^2}{c^2} \mathbf{E}_{3\omega} = -\frac{9\omega^2}{c^2} \varepsilon_0 \chi_{3\omega}^{(3)} \mathbf{E}_{\omega}^3 e^{3i\mathbf{k}\cdot\mathbf{r}}, \tag{10.25}$$

where $\mathbf{k}$ is the pump wave vector inside the medium, and $n_{3\omega}$ is the index of refraction at the TH frequency. For simplicity we have assumed that the pump remains undepleted and of nearly constant amplitude throughout the medium. The general solution of Eqn (10.25) may be written as a superposition of two particular solutions: (1) the homo-geneous solution, found by setting the right-hand side of Eqn (10.25) equal to zero, characterized by a TH signal *propagating freely according to material dispersion*; and (2) the inhomogeneous or driven solution, with a TH field characterized by a wave vector equal to three times that of the pump, or $3\mathbf{k}$, and *propagating according to the same dispersion that characterizes the pump*. In short, the two signals that we observe in Figure 10.7(b) correspond to the homogeneous solution (the lagging pulse, which experiences the expected material dispersion with a higher index of refraction and smaller energy velocity) and the inhomogeneous pulse, which is trapped by the pump and thus appears to propagate and refract inside the medium with the same charac-teristics as the pump. Indeed, a study of the phase of the fields in Figure 10.7 reveals that the TH pulse is *phase-locked to* the pump. Put another way, the trapped TH signal is forced to refract in the direction of the pump, which for all intents and purposes means that the generated harmonic acquires the dispersive properties of the pump.

The dramatic significance of the pump coaxing the TH pulse to behave according to pump dispersion may be further augmented if the medium is assumed to be absorptive at the TH wavelength and continues to be transparent for the pump pulse. One may surmise that if indeed the phase-locked, TH signal propagates with the properties of the pump pulse, then it should follow that if absorption is present only at the TH wave-length the homogeneous pulse should be quickly absorbed, leaving the phase-locked component unscathed as it is intimately tied to the dynamics of the pump pulse.
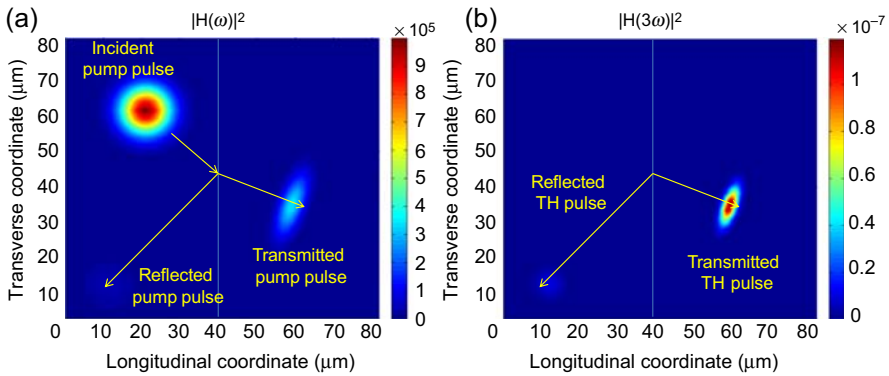
Figure 10.8 Same as Figure 10.7, except that the medium is now assumed to be highly absorptive at the TH wavelength. The transmitted SH signal is composed of only the inhomogeneous components, which continues to be locked to the pump and survives because the TH components travel with pump attributes. The homogeneous TH signal that lags in Figure 10.7 is now completely absorbed.

In Figure 10.8 we repropose the dynamic already shown in Figure 10.7, except that now dispersion has been augmented by strong absorption at the TH wavelength. In practice, this means reducing the oscillator's resonance closer to the TH wavelength. The material is still mostly transparent at the pump carrier wavelength. As predicted, Figure 10.8(b) shows that only the phase-locked signal clearly survives the presence of absorption at the TH wavelength for the reason that we have outlined, i.e., phase-locking induced transparency at the harmonic wavelength, regardless of material dispersion/absorption at the generated signal's wavelength.

10.2 Second and third harmonic generation

While centrosymmetric metals do not display a bulk, dipolar, second order nonlinearity, i.e., $\chi^{(2)} = 0$, they possess a large third order nonlinear susceptibility, $\chi^{(3)}$. SHG is most often examined as the direct product of the interaction of an incident field with free electrons only, with secondary contributions of bound electrons applied only to the linear dielectric constant. However, just as the linear dielectric constant is affected by both free electrons and interband transitions from electrons in the valence band, SHG can also arise from both conduction and inner-core electrons, due to a combination of spatial symmetry breaking (the mere presence of interfaces), the magnetic portion of the Lorentz force, to a lesser extent the interaction of TH and pump photons (down-conversion), and other effective nonlinearities induced by quantum tunneling mechanisms if metal components are in close proximity. By the same token, the third order nonlinearity, $\chi^{(3)}$, arising from bound charges generates most of the TH signal, subject to screening due to free-electron spill-out effect and geometrical considerations. To a much smaller degree, the interaction of pump and SH photons also

produces cascaded THG. Therefore, the classical, nonlinear oscillator model is pivotal in these systems, where a combination of free (Drude) and bound (Lorentz) electrons generally suffices to describe second and third order processes.

The influence of bound electrons to the linear dielectric constant and SHG is felt most in the visible and near infrared (IR) ranges. The uppermost filled valence level of silver, for example, is the $4\,d^{10}$ orbital, with 10 available bound electrons that may be modeled using Lorentz oscillators. All electrons are assumed to be under the influence of electric and magnetic forces, so that the dynamics that ensues in the metal and the dielectric contains surface and volume contributions simultaneously.

10.3   Free electrons

The description of free electrons inside the metal is mediated by an expanded form of the Drude model, which takes the following form:

$$m^*\frac{d\mathbf{v}}{dt} + \gamma m^*\mathbf{v} = e\mathbf{E} + \mathbf{v}\times\mathbf{H} - \frac{\nabla p}{n}, \tag{10.26}$$

where m^* is the effective mass of conduction electrons and n is their density; $\mathbf{v}$ is the electron velocity; $\mathbf{E}$ and $\mathbf{H}$ are electric and magnetic fields, respectively; and p is the electron gas pressure (the "gas" is confined by the metal walls). Since free electrons are not confined to any specific atomic site, the temporal derivative of the velocity in Eqn (10.26) is also position-dependent, as follows:

$$\frac{d\mathbf{v}}{dt} = \frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}, \tag{10.27}$$

so that Eqn (10.26) becomes

$$\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} + \gamma\mathbf{v} = \frac{e}{m^*}\mathbf{E} + \frac{e}{m^*}\mathbf{v}\times\mathbf{H} - \frac{\nabla p}{nm^*}. \tag{10.28}$$

Identifying the current density with $\mathbf{J} = ne\mathbf{v}$ makes it possible to rewrite Eqn (10.28) as

$$\frac{\partial\mathbf{J}}{\partial t} - \frac{\dot{n}}{n}\mathbf{J} + \mathbf{J}\cdot\nabla\left(\frac{\mathbf{J}}{ne}\right) + \gamma\mathbf{J} = \frac{ne^2}{m^*}\mathbf{E} + \frac{e}{m^*}\mathbf{J}\times\mathbf{H} - \frac{\nabla p}{nm^*}. \tag{10.29}$$

After defining $\dot{\mathbf{P}}_f = \mathbf{J}$, Eqn (10.29) becomes

$$\ddot{\mathbf{P}}_f - \frac{\dot{n}}{n}\dot{\mathbf{P}}_f + \left(\dot{\mathbf{P}}_f\cdot\nabla\right)\left(\frac{\dot{\mathbf{P}}_f}{ne}\right) + \gamma\dot{\mathbf{P}}_f = \frac{ne^2}{m^*}\mathbf{E} + \frac{e}{m^*}\dot{\mathbf{P}}_f\times\mathbf{H} - \frac{e\nabla p}{m^*}. \tag{10.30}$$

For free electrons the continuity equation $\dot{n}(\mathbf{r}, t) = -\frac{1}{e}\nabla \cdot \dot{\mathbf{P}}_f$ supplements the equations of motion and may be integrated directly to yield

$$n(\mathbf{r}, t) = n_0 - \frac{1}{e}\nabla \cdot \mathbf{P}_f, \tag{10.31}$$

where n_0 is the background, equilibrium charge density, and the initial value of the polarization is $\mathbf{P}_f(\mathbf{r}, t = 0) \equiv 0$ in the absence of any applied fields. Assuming that any local variations in the charge density are small as compared to the density itself, i.e., $|n_0| \gg |-\frac{1}{e}\nabla \cdot \mathbf{P}_f|$, then the ratio $\dot{n}/n$ may be expanded in powers of $1/(n_0 e)$ to obtain

$$\frac{\dot{n}}{n} = -\frac{1}{n_0 e}\nabla \cdot \dot{\mathbf{P}}_f \left(1 - \frac{1}{n_0 e}\nabla \cdot \mathbf{P}_f\right)^{-1} \sim -\frac{\nabla \cdot \dot{\mathbf{P}}_f}{e n_0} - \frac{1}{n_0^2 e^2}(\nabla \cdot \dot{\mathbf{P}}_f)(\nabla \cdot \mathbf{P}_f) + \vartheta\left(\frac{1}{n_0^3 e^3}\right) + \cdots \tag{10.32}$$

Substituting Eqn (10.32) back into Eqn (10.30) and neglecting terms of order $(1/n_0 e)^2$ and higher we have

$$\ddot{\mathbf{P}}_f + \gamma\dot{\mathbf{P}}_f = \frac{n_0 e^2}{m^*}\mathbf{E} - \frac{e}{m^*}\mathbf{E}(\nabla \cdot \mathbf{P}_f) + \frac{e}{m^*}\dot{\mathbf{P}}_f \times \mathbf{H}$$
$$- \frac{1}{n_0 e}\left[(\nabla \cdot \dot{\mathbf{P}}_f)\dot{\mathbf{P}}_f + (\dot{\mathbf{P}}_f \cdot \nabla)\dot{\mathbf{P}}_f\right] - \frac{e\nabla p}{m^*}. \tag{10.33}$$

The specific impact of pressure is seldom considered in the dynamics but may be treated either classically or quantum mechanically. If we assume that electrons form of a classical ideal gas, then the pressure equation may be written simply as $p = n K_B T$, where K_B is the Boltzmann constant and T is the temperature. The gradient of p in Eqn (10.33) becomes the gradient of n, which in turn may be related to the macroscopic polarization as follows:

$$-\frac{e\nabla p}{m^*} = -\frac{e}{m^*}K_B T\nabla\left(n_0 - \frac{1}{e}\nabla \cdot \mathbf{P}_f\right) = \frac{K_B T}{m^*}\nabla(\nabla \cdot \mathbf{P}_f). \tag{10.34}$$

In the quantum regime, the pressure takes the form $p = p_0(n/n_0)^\beta$, where $\beta = (D + 2)/D$ and D is the dimensionality of the problem. For $D = 3$, we have $p = p_0(n/n_0)^{5/3}$, where $p_0 = n_0 E_F$, E_F is the Fermi energy, and n_0 is the equilibrium charge density. The leading pressure terms are

$$-\frac{e\nabla p}{m^*} = -\frac{e p_0}{m^* n_0^{5/3}}\frac{5}{3}n^{2/3}\nabla n$$
$$= -\frac{5}{3}\frac{e n_0 E_F}{m^* n_0^{5/3}}n^{2/3}\nabla n \approx \frac{5}{3}\frac{E_F}{m^*}\nabla(\nabla \cdot \mathbf{P}_f) - \frac{10}{9}\frac{E_F}{m^*}\frac{1}{n_0 e}(\nabla \cdot \mathbf{P}_f)\nabla(\nabla \cdot \mathbf{P}_f). \tag{10.35}$$

A comparison between Eqns (10.34) and (10.35) shows that the quantum model intrinsically contains the classical, ideal electron gas contribution and a nonlinear quantum correction that leads to harmonic generation. If we once again scale the equation of motion as before, Eqn (10.33) becomes

$$
\begin{aligned}
\ddot{\mathbf{P}}_f + \tilde{\gamma}\dot{\mathbf{P}}_f ={} & r\frac{n_0 e^2}{m^*}\left(\frac{\lambda_0}{c}\right)^2 \mathbf{E} - \frac{e\lambda_0}{m^* c^2}\mathbf{E}(\nabla\cdot\mathbf{P}_f) + \frac{e\lambda_0}{m^*\, c}\dot{\mathbf{P}}_f \times \mathbf{H} \\
& - \frac{1}{n_0 e\lambda_0}\left[(\nabla\cdot\dot{\mathbf{P}}_f)\dot{\mathbf{P}}_f + (\dot{\mathbf{P}}_f\cdot\nabla)\dot{\mathbf{P}}_f\right] + \frac{5}{3}\frac{E_{\mathrm{F}}}{m^* c^2}\nabla(\nabla\cdot\mathbf{P}_f) \qquad (10.36) \\
& - \frac{10}{9}\frac{E_{\mathrm{F}}}{m^* c^2}\frac{1}{n_0 e\lambda_0}(\nabla\cdot\mathbf{P}_f)\nabla(\nabla\cdot\mathbf{P}_f).
\end{aligned}
$$

In addition to the magnetic Lorentz force $(e\lambda_0/m^* c)\dot{\mathbf{P}}_f \times \mathbf{H}$, we have an explicit quadrupole-like Coulomb term that arises from the continuity equation $-(e\lambda_0/m^* c^2)\mathbf{E}(\nabla\cdot\mathbf{P}_f)$, convective terms proportional to $[(\nabla\cdot\dot{\mathbf{P}}_f)\dot{\mathbf{P}}_f + (\dot{\mathbf{P}}_f\cdot\nabla)\dot{\mathbf{P}}_f]$, and linear and nonlinear pressure terms proportional to $\nabla(\nabla\cdot\mathbf{P}_f)$ and $(\nabla\cdot\mathbf{P}_f)\nabla(\nabla\cdot\mathbf{P}_f)$, respectively. For silver, the Fermi velocity $v_F \sim 10^6$ m/s so that $(E_F/m^* c^2) \sim 10^{-5}$. If for the moment we neglect all nonlinear contributions, Eqn (10.36) becomes

$$
\ddot{\mathbf{P}}_f + \tilde{\gamma}\dot{\mathbf{P}}_f = \frac{n_0 e^2}{m^*}\left(\frac{\lambda_0}{c}\right)^2 \mathbf{E} + \frac{5}{3}\frac{E_{\mathrm{F}}}{m^* c^2}\nabla(\nabla\cdot\mathbf{P}_f). \qquad (10.37)
$$

Expanding the terms on the right-hand side of Eqn (10.37) shows that the pressure couples orthogonal, free electron polarization states and introduces a dynamical anisotropy. Put another way, the effective dielectric constant depends not only on frequency but also on the spatial derivatives of the polarization field, i.e., $\varepsilon = \varepsilon(\omega, \mathbf{k})$, a situation that is usually referred to as nonlocality. Therefore, from a practical point of view one can easily see that electron gas pressure can dynamically modify the linear dielectric function of the metal, especially near its walls, should the fields become strongly confined and/or their derivatives be large enough (i.e., near sharp edges, corners, or in resonant, subwavelength cavities) to introduce large, evanescent k-vectors. The typical result is a blue-shift of transmission, reflection, and absorption spectra. The same is true for the nonlinear term: its magnitude could perturb Coulomb, Lorentz, or convective terms at high enough intensity and/or if large enough k-vectors are excited. Finally, we note that a decomposition of the fields into fundamental, second, and TH components causes Eqn (10.36) to split into three complex equations.

10.4   Bound electrons

Unlike free, conduction electrons, inner-core electrons respond to internal, linear, and nonlinear restoring forces, as described in Eqn (10.2), and they are not free to leave their atomic sites. The approach that we pursue differs slightly from the development
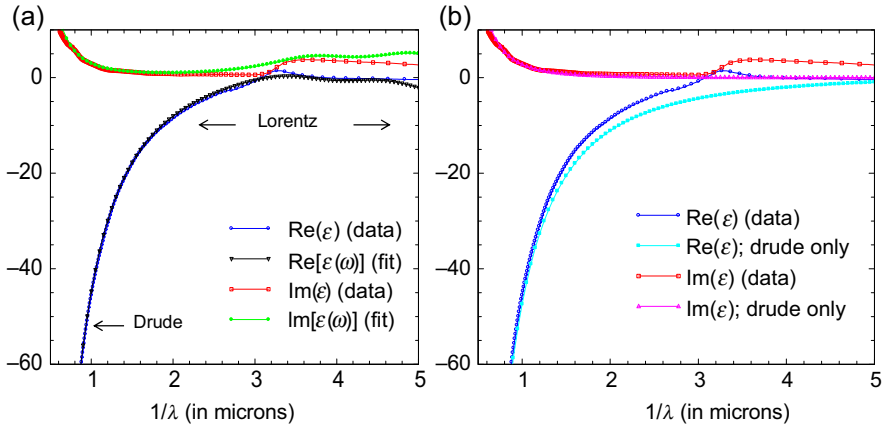
Figure 10.9 (a) Bulk silver data and the function in Eqn (10.38). (b) Palik's data as compared to a Drude-only contribution, i.e., Eqn (9.12) minus the Lorentz components. Both (a) and (b) suggest the importance of the Lorentz contributions even in the near IR region.

that follows Eqn (10.2) above, because we introduce harmonic components generated by the free electron gas. In Figure 10.9(a) we plot the dielectric response of silver reported in Palik's handbook, along with a possible fit using one Drude and two Lorentz oscillators, as follows:

$$\varepsilon(\tilde{\omega}) = 1 - \frac{6.965^2}{\tilde{\omega}^2 + i0.0573\tilde{\omega}} - \frac{4.25^2}{\tilde{\omega}^2 - 3.75^2 + i1.42\tilde{\omega}} - \frac{5.5^2}{\tilde{\omega}^2 - 5^2 + i1.42\tilde{\omega}}.$$

(10.38)

Here $\tilde{\omega} = 1/\lambda$, where λ is in microns. A comparison of the curves reveals reasonably good agreement in the range of interest (200–1200 nm). In Figure 10.9(b) we compare Palik's data only with the Drude term in Eqn (10.38). It is evident that purely Drude (free electron) behavior occurs at wavelengths approaching and exceeding 1 μm. In the near IR, visible, and UV portions of the spectrum inner-core contributions to the dielectric constant become dominant.

Following Eqn (10.2), Newton's second law for two separate bound species of core electrons leads to two equations of motion for bound charges that read as follows:

$$\ddot{\mathbf{P}}_1 + \tilde{\gamma}_{01}\dot{\mathbf{P}}_1 + \tilde{\omega}_{01}^2\mathbf{P}_1 - b_1(\mathbf{P}_1 \cdot \mathbf{P}_1)\mathbf{P}_1 = \frac{n_{01}e^2\lambda_0^2}{m_{b1}^*c^2}\mathbf{E} + \frac{e\lambda_0}{m_{b1}^*c}\dot{\mathbf{P}}_1 \times \mathbf{H},$$

(10.39)

$$\ddot{\mathbf{P}}_2 + \tilde{\gamma}_{02}\dot{\mathbf{P}}_2 + \tilde{\omega}_{02}^2\mathbf{P}_2 - b_2(\mathbf{P}_2 \cdot \mathbf{P}_2)\mathbf{P}_2 = \frac{n_{02}e^2\lambda_0^2}{m_{b2}^*c^2}\mathbf{E} + \frac{e\lambda_0}{m_{b2}^*c}\dot{\mathbf{P}}_2 \times \mathbf{H}.$$

(10.40)

Here $\mathbf{P}_{1,2} = n_{01,02}e\mathbf{r}_{1,2}$ are the polarizations; $\mathbf{r}_{1,2}$ are the bound electron's positions relative to an equilibrium origin; and $\dot{\mathbf{P}}_{1,2} = n_{01,02}e\dot{\mathbf{r}}_{1,2}$ are the bound current densities. To summarize, Eqn (10.36) is an augmented Drude model that describes free electrons inside the metal and Eqns (10.39) and (10.40) portray two separate

bound electron species, each subject to internal and external forces. Each electron species has its own damping rate $(\tilde{\gamma}_f, \tilde{\gamma}_{01}, \tilde{\gamma}_{02})$, effective mass $(m_0^*, m_{b1}^*, m_{b2}^*)$, density (n_{0f}, n_{01}, n_{02}), and, in the case of bound electrons, resonance frequency $(\tilde{\omega}_{01}^2, \tilde{\omega}_{02}^2)$ as well as associated nonlinear third order coefficient $\left(\tilde{b}_1 = b_1 \frac{\lambda_r^2}{n_{01}^2 e^2 c^2}, \tilde{b}_2 = b_2 \frac{\lambda_r^2}{n_{02}^2 e^2 c^2}\right)$. For simplicity, we have chosen to operate in two spatial dimensions plus time, so that the operator $\nabla \equiv \frac{\partial}{\partial \hat{z}} \widehat{\mathbf{k}} + \frac{\partial}{\partial \hat{y}} \widehat{\mathbf{j}}$, as the fields are assumed to be independent of $\hat{x}$.

Let us now consider one of the two bound oscillator species. In order to arrive at Eqns (10.39) and (10.40), let $\mathbf{r}_b$ represent one or the other bound electron position. Then, up to the TH frequency, the fields at the electron's position $\mathbf{r}_b$ may be written as

$$
\begin{aligned}
\mathbf{E} = \Big(& \mathbf{E}_\omega e^{i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{E}_\omega^* e^{-i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{E}_{2\omega} e^{2i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{E}_{2\omega}^* e^{-2i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} \\
& + \mathbf{E}_{3\omega} e^{3i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{E}_{3\omega}^* e^{-3i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} \Big) \\
\mathbf{H} = \Big(& \mathbf{H}_\omega e^{i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{H}_\omega^* e^{-i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{H}_{2\omega} e^{2i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{H}_{2\omega}^* e^{-2i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} \\
& + \mathbf{H}_{3\omega} e^{3i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} + \mathbf{H}_{3\omega}^* e^{-3i(\mathbf{k}\cdot\mathbf{r}_b - \omega t)} \Big).
\end{aligned}
$$

$$(10.41)$$

Expanding the fields in powers of $\mathbf{k}\cdot\mathbf{r}_b$ and retaining only first order terms, we have

$$
\mathbf{E} = \begin{pmatrix}
\mathbf{E}_\omega\, e^{-i\omega t}(1 + i\mathbf{k}\cdot\mathbf{r}_b + \ldots) + \mathbf{E}_\omega^*\, e^{i\omega t}(1 - i\mathbf{k}\cdot\mathbf{r}_b + \ldots) \\
+\mathbf{E}_{2\omega}\, e^{-2i\omega t}(1 + 2i\mathbf{k}\cdot\mathbf{r}_b + \ldots) + \mathbf{E}_{2\omega}^*\, e^{2i\omega t}(1 - 2i\mathbf{k}\cdot\mathbf{r}_b + \ldots) \\
+\mathbf{E}_{3\omega}\, e^{-3i\omega t}(1 + 3i\mathbf{k}\cdot\mathbf{r}_b + \ldots) + \mathbf{E}_{3\omega}^*\, e^{3i\omega t}(1 - 3i\mathbf{k}\cdot\mathbf{r}_b + \ldots)
\end{pmatrix},
$$

$$(10.42)$$

and similarly for the magnetic field. The solutions for the electron's position and its derivatives are

$$
\mathbf{r}_b = \mathbf{r}_\omega\, e^{-i\omega t} + \mathbf{r}_{2\omega}\, e^{-2i\omega t} + \mathbf{r}_{3\omega}\, e^{-3i\omega t} + c.c., \tag{10.43}
$$

$$
\dot{\mathbf{r}}_b = (\dot{\mathbf{r}}_\omega - i\omega\mathbf{r}_\omega)e^{-i\omega t} + (\dot{\mathbf{r}}_{2\omega} - 2i\omega\,\mathbf{r}_{2\omega})e^{-2i\omega t} + (\dot{\mathbf{r}}_{3\omega} - 3i\omega\mathbf{r}_{3\omega})e^{-3i\omega t} + c.c., \tag{10.44}
$$

$$
\begin{aligned}
\ddot{\mathbf{r}}_b = {} & \left(\ddot{\mathbf{r}}_\omega - 2i\omega\dot{\mathbf{r}}_\omega - \omega^2\mathbf{r}_\omega\right)e^{-i\omega t} + \left(\ddot{\mathbf{r}}_{2\omega} - 4i\omega\dot{\mathbf{r}}_{2\omega} - 4\omega^2\mathbf{r}_{2\omega}\right)e^{-2i\omega t} \\
& + \left(\ddot{\mathbf{r}}_{3\omega} - 6i\omega\dot{\mathbf{r}}_\omega - 9\omega^2\mathbf{r}_\omega\right)e^{-3i\omega t} + c.c.
\end{aligned}
$$

$$(10.45)$$

We do not assume slowly varying envelope functions, and all temporal and spatial derivatives will be retained. Neglecting for the moment the nonlinear restoring force, substitution of Eqns (10.41)−(10.45) into Eqn (10.2), for example, leads to

$$\ddot{\mathbf{r}}_\omega + (\gamma_b - 2i\omega)\dot{\mathbf{r}}_\omega + (\omega_0^2 - \omega^2 + i\gamma_b\omega)\mathbf{r}_\omega = \frac{e}{m_b^*}\begin{pmatrix} \mathbf{E}_\omega \\[8pt] -i\mathbf{k}\cdot\mathbf{r}_{2\omega}\ \mathbf{E}_\omega^* \\[8pt] -2i\mathbf{k}\cdot\mathbf{r}_{3\omega}\ \mathbf{E}_{2\omega}^* \\[8pt] +2i\mathbf{k}\cdot\mathbf{r}_\omega^*\ \mathbf{E}_{2\omega} \\[8pt] +3i\mathbf{k}\cdot\mathbf{r}_{2\omega}^*\ \mathbf{E}_{3\omega} \end{pmatrix}$$

$$+\frac{e}{m_b^* c}\begin{pmatrix} (\dot{\mathbf{r}}_\omega^* + i\omega\mathbf{r}_\omega^*)\times\mathbf{H}_{2\omega} + (\dot{\mathbf{r}}_{2\omega} - 2i\omega\ \mathbf{r}_{2\omega})\times\mathbf{H}_\omega^* \\[8pt] (\dot{\mathbf{r}}_{2\omega}^* + 2i\omega\ \mathbf{r}_{2\omega}^*)\times\mathbf{H}_{3\omega} + (\dot{\mathbf{r}}_{3\omega} - 3i\omega\mathbf{r}_{3\omega})\times\mathbf{H}_{2\omega}^* \\[8pt] -(i\dot{\mathbf{r}}_\omega + \omega\mathbf{r}_\omega)\times\mathbf{H}_\omega^*\mathbf{k}\cdot\mathbf{r}_\omega - (-i\dot{\mathbf{r}}_\omega^* + \omega\mathbf{r}_\omega^*)\times\mathbf{H}_\omega\mathbf{k}\cdot\mathbf{r}_\omega \\[8pt] +(i\dot{\mathbf{r}}_\omega + \omega\mathbf{r}_\omega)\times\mathbf{H}_\omega\mathbf{k}\cdot\mathbf{r}_\omega^* \end{pmatrix}$$

$$\ddot{\mathbf{r}}_{2\omega} + (\gamma_b - 4i\omega)\dot{\mathbf{r}}_{2\omega} + (\omega_0^2 - 4\omega^2 + 2i\gamma_b\omega)\mathbf{r}_{2\omega}$$

$$= \frac{e}{m_b^*}\begin{pmatrix} +\mathbf{E}_{2\omega} \\[8pt] +i\mathbf{k}\cdot\mathbf{r}_\omega\ \mathbf{E}_\omega \\[8pt] -i\mathbf{k}\cdot\mathbf{r}_{3\omega}\ \mathbf{E}_\omega^* \\[8pt] +3i\mathbf{k}\cdot\mathbf{r}_\omega^*\ \mathbf{E}_{3\omega} \end{pmatrix} + \frac{e}{m_b^* c}\begin{pmatrix} (\dot{\mathbf{r}}_\omega - i\omega\mathbf{r}_\omega)\times\mathbf{H}_\omega \\[8pt] +(\dot{\mathbf{r}}_{3\omega} - 3i\omega\mathbf{r}_{3\omega})\times\mathbf{H}_\omega^* + (\dot{\mathbf{r}}_\omega^* + i\omega\mathbf{r}_\omega^*)\times\mathbf{H}_{3\omega} \\[8pt] -(i\dot{\mathbf{r}}_{2\omega} + 2\omega\ \mathbf{r}_{2\omega})\times\mathbf{H}_\omega^*\mathbf{k}\cdot\mathbf{r}_\omega - 2(i\dot{\mathbf{r}}_\omega + \omega\mathbf{r}_\omega)\times\mathbf{H}_\omega^*\mathbf{k}\cdot\mathbf{r}_{2\omega} \\[8pt] -(-i\dot{\mathbf{r}}_\omega^* + \omega\mathbf{r}_\omega^*)\times\mathbf{H}_\omega\mathbf{k}\cdot\mathbf{r}_{2\omega} - 2(-i\dot{\mathbf{r}}_\omega^* + \omega\mathbf{r}_\omega^*)\times\mathbf{H}_{2\omega}\mathbf{k}\cdot\mathbf{r}_\omega \\[8pt] +(i\dot{\mathbf{r}}_{2\omega} + 2\omega\ \mathbf{r}_{2\omega})\times\mathbf{H}_\omega\mathbf{k}\cdot\mathbf{r}_\omega^* + 2(i\dot{\mathbf{r}}_\omega + \omega\ \mathbf{r}_\omega)\times\mathbf{H}_{2\omega}\mathbf{k}\cdot\mathbf{r}_\omega^* \end{pmatrix}$$

$$\ddot{\mathbf{r}}_{3\omega} + (\gamma_b - 6i\omega)\dot{\mathbf{r}}_{3\omega} + (\omega_0^2 - 9\omega^2 + 3i\gamma_b\omega)\mathbf{r}_{3\omega}$$

$$= \frac{e}{m_b^*}\begin{pmatrix} +\mathbf{E}_{3\omega} \\[8pt] +i\mathbf{k}\cdot\mathbf{r}_{2\omega}\ \mathbf{E}_\omega \\[8pt] +2i\mathbf{k}\cdot\mathbf{r}_\omega\mathbf{E}_{2\omega} \end{pmatrix} + \frac{e}{m_b^* c}\begin{pmatrix} (\dot{\mathbf{r}}_{2\omega} - 2i\omega\mathbf{r}_{2\omega})\times\mathbf{H}_\omega + (\dot{\mathbf{r}}_\omega - i\omega\mathbf{r}_\omega)\times\mathbf{H}_{2\omega} \\[8pt] +(i\dot{\mathbf{r}}_\omega + \omega\ \mathbf{r}_\omega)\times\mathbf{H}_\omega\mathbf{k}\cdot\mathbf{r}_\omega \end{pmatrix}$$

$$(10.46)$$

Finally, we simplify Eqn (10.46) if we identify $\mathbf{P}_{b,\omega} = n_{0b}e\mathbf{r}_\omega$, $\mathbf{P}_{b,2\omega} = n_{0b}e\mathbf{r}_{2\omega}$, and $\mathbf{P}_{b,3\omega} = n_{0b}e\mathbf{r}_{3\omega}$ and make the following assumptions:

$$i\mathbf{k}\cdot n_{0b}e\mathbf{r}_\omega \approx \nabla\cdot\mathbf{P}_{b,\omega} \quad 2i\mathbf{k}\cdot n_{0b}e\mathbf{r}_{2\omega} \approx \nabla\cdot\mathbf{P}_{b,2\omega} \quad 3i\mathbf{k}\cdot n_{0b}e\mathbf{r}_{3\omega} \approx \nabla\cdot\mathbf{P}_{b,3\omega}.$$

$$(10.47)$$

Then, Eqn (10.46) becomes

$$\ddot{\mathbf{P}}_{b,\omega} + \tilde{\gamma}_{b,\omega}\dot{\mathbf{P}}_{b,\omega} + \tilde{\omega}_{0,b,\omega}^2 \mathbf{P}_{b,\omega} \approx \frac{n_{0,b}e^2\lambda_0^2}{m_b^* c^2}\mathbf{E}_\omega$$

$$+ \frac{e\lambda_0}{m_b^* c^2}\left(\begin{array}{c} -\dfrac{1}{2}\mathbf{E}_\omega^* \nabla \cdot \mathbf{P}_{b,2\omega} \\[2mm] +2\mathbf{E}_{2\omega}\nabla \cdot \mathbf{P}_{b,\omega}^* \\[2mm] -\dfrac{2}{3}\mathbf{E}_{2\omega}^* \nabla \cdot \mathbf{P}_{b,3\omega} \\[2mm] -\dfrac{3}{2}\mathbf{E}_{3\omega}\nabla \cdot \mathbf{P}_{b,2\omega}^* \end{array} \right) + \frac{e\lambda_0}{m_b^* c}\left(\begin{array}{c} \left(\dot{\mathbf{P}}_{b,\omega}^* + i\omega\mathbf{P}_{b,\omega}^*\right) \times \mathbf{H}_{2\omega} \\[2mm] +\left(\dot{\mathbf{P}}_{b,2\omega} - 2i\omega\mathbf{P}_{b,2\omega}\right) \times \mathbf{H}_\omega^* \\[2mm] +\left(\dot{\mathbf{P}}_{b,2\omega}^* + 2i\omega\mathbf{P}_{b,2\omega}^*\right) \times \mathbf{H}_{3\omega} \\[2mm] +\left(\dot{\mathbf{P}}_{b,3\omega} - 3i\omega\mathbf{P}_{b,3\omega}\right) \times \mathbf{H}_{2\omega}^* \end{array} \right)$$

$$\ddot{\mathbf{P}}_{b,2\omega} + \tilde{\gamma}_{b,2\omega}\dot{\mathbf{P}}_{b,2\omega} + \tilde{\omega}_{0,b,2\omega}^2 \mathbf{P}_{b,2\omega} \approx \frac{n_{0,b}e^2\lambda_0^2}{m_b^* c^2}\mathbf{E}_{2\omega}$$

$$+ \frac{e\lambda_0}{m_b^* c^2}\left(\begin{array}{c} \mathbf{E}_\omega\nabla \cdot \mathbf{P}_{b,\omega} \\[2mm] -\dfrac{1}{3}\mathbf{E}_\omega^* \nabla \cdot \mathbf{P}_{b,3\omega} \\[2mm] -3\mathbf{E}_{3\omega}\nabla \cdot \mathbf{P}_{b,\omega}^* \end{array} \right) + \frac{e\lambda_0}{m_b^* c}\left(\begin{array}{c} \left(\dot{\mathbf{P}}_{b,\omega} - i\omega\mathbf{P}_{b,\omega}\right) \times \mathbf{H}_\omega \\[2mm] +\left(\dot{\mathbf{P}}_{b,\omega}^* + i\omega\mathbf{P}_{b,\omega}^*\right) \times \mathbf{H}_{3\omega} \\[2mm] +\left(\dot{\mathbf{P}}_{b,3\omega} - 3i\omega\mathbf{P}_{b,3\omega}\right) \times \mathbf{H}_\omega^* \end{array} \right)$$

$$\ddot{\mathbf{P}}_{b,3\omega} + \tilde{\gamma}_{b,3\omega}\dot{\mathbf{P}}_{b,3\omega} + \tilde{\omega}_{0,b,3\omega}^2 \mathbf{P}_{b,3\omega} \approx \frac{n_{0,b}e^2\lambda_0^2}{m_b^* c^2}\mathbf{E}_{3\omega}$$

$$+ \frac{e\lambda_0}{m_b^* c^2}\left(\begin{array}{c} \dfrac{1}{2}\mathbf{E}_\omega\nabla \cdot \mathbf{P}_{b,2\omega} \\[2mm] +2\mathbf{E}_{2\omega}\nabla \cdot \mathbf{P}_{b,\omega} \end{array} \right) + \frac{e\lambda_0}{m_b^* c}\left(\begin{array}{c} \left(\dot{\mathbf{P}}_{b,2\omega} - 2i\omega\mathbf{P}_{b,2\omega}\right) \times \mathbf{H}_\omega \\[2mm] +\left(\dot{\mathbf{P}}_{b,\omega} - i\omega\mathbf{P}_{b,\omega}\right) \times \mathbf{H}_{2\omega} \end{array} \right)$$

$$(10.48)$$

The scaled coefficients are $\tilde{\gamma}_{b,N\omega} = (\gamma_b - Ni\omega)\frac{\lambda_r}{c}$, $\tilde{\omega}_{0,b,N\omega}^2 = (\omega_0^2 - (N\omega)^2 + i\gamma_b N\omega)\frac{\lambda_r^2}{c^2}$, where N is an integer that denotes the given harmonic. We emphasize that all envelope functions in Eqn (10.48) are allowed to vary rapidly in space and time, as demonstrated by the presence of spatial and temporal derivatives up to all orders of dispersion.

The description of nonlinear $\chi^{(3)}$ contributions is usually done beginning with the general expansion of the third order polarization as follows:

$$P_{\mathrm{NL},i}^{(3)} = \sum_{j=1,3} \sum_{k=1,3} \sum_{l=1,3} \chi_{i,j,k,l} E_j E_k E_l \tag{10.49}$$

However, having foregone the direct implementation of $\chi^{(3)}$, and having opted instead for the use of dynamic terms, Eqn (10.49) becomes

$$P_{\mathrm{NL},i}^{(3)} = -\sum_{j=1,3} \sum_{k=1,3} \sum_{l=1,3} b_{i,j,k,l} P_j P_k P_l. \tag{10.50}$$

For example, for a material like GaAs, having cubic symmetry of the type $\overline{4}3m$, Eqn (10.50) reduces to

$$P_{\mathrm{NL},x}^{(3)} = b_{xxxx}^{(3)} P_x^3 + 3b_{xxyy}^{(3)} P_y^2 P_x + 3b_{xxzz}^{(3)} P_z^2 P_x$$

$$P_{\mathrm{NL},y}^{(3)} = b_{yyyy}^{(3)} P_y^3 + 3b_{xxyy}^{(3)} P_x^2 P_y + 3b_{yyzz}^{(3)} P_z^2 P_y \tag{10.51}$$

$$P_{\mathrm{NL},z}^{(3)} = b_{zzzz}^{(3)} P_z^3 + 3b_{zzxx}^{(3)} P_x^2 P_z + 3b_{zzyy}^{(3)} P_y^2 P_z$$

For metals the situation is comparable, except that for isotropic crystal symmetry the relations between the tensor components allow one to more simply write

$$P_{\mathrm{NL},x}^{(3)} = b_{\mathrm{Ag}}^{(3)} \left(P_x^3 + P_y^2 P_x + P_z^2 P_x \right)$$

$$P_{\mathrm{NL},y}^{(3)} = b_{\mathrm{Ag}}^{(3)} \left(P_y^3 + P_x^2 P_y + P_z^2 P_y \right) \tag{10.52}$$

$$P_{\mathrm{NL},z}^{(3)} = b_{\mathrm{Ag}}^{(3)} \left(P_z^3 + P_x^2 P_z + P_y^2 P_z \right)$$

Adding each component in Eqn (10.52) to the respective components of Eqn (10.48) leads to nonlinear contributions to all harmonic components, with self- and cross-phase modulation and up- and down-conversion, along with terms that taken together with Eqn (10.36) can couple orthogonal polarization states. Phase modulation of the pump field is always important, especially in situations where the local field intensity can be amplified hundreds or perhaps thousands of times relative to its incident value. Equations (10.36) and (10.48) thus form a set of coupled equations that
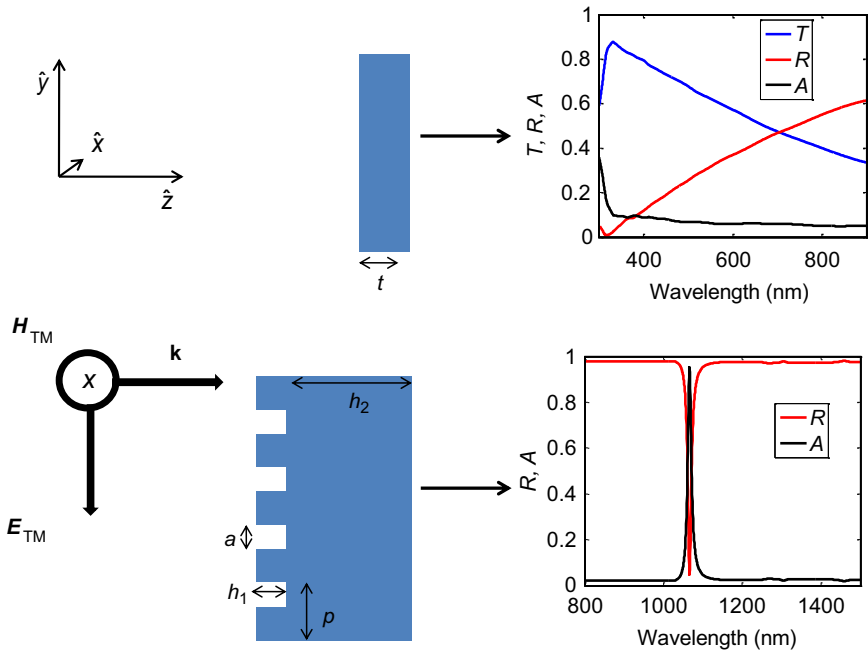
Figure 10.10 Top row: a single silver layer, $t = 10$ nm; Bottom row: silver grating with $p = 1030$ nm, $a = 300$ nm, $h_1 = 45$ nm, and $h_2 = 200$ nm. The respective transmission, reflection, and absorption spectra at normal incidence are depicted to the right of each object. The grating displays zero transmission and is not shown. The field is incident from the left and polarized as shown.

describe free and bound charges that give rise to second and THG in metallic (centro-symmetric) structures of arbitrary geometry and are valid in the pulsed regime.

In Figure 10.10 we show the two examples that we have investigated: a bare silver layer (top) and an opaque plasmonic grating (bottom panel). For each structure in Figure 10.10 we also show the corresponding linear transmission, reflection, and absorption spectra obtained at normal incidence. Worthy of note is the narrow, plasmonic reflection resonance displayed by the grating.

At resonance, the plasmonic grating is able to confine the field in close proximity to the surface, producing an intense field near internal and external corners. A bound wave travels along the grating, with an evanescent tail that spills inside the metal, which in turn absorbs all the incident energy, resulting in near-zero reflections.

In Figure 10.11 we show the results of vectorial, nonlinear calculations performed using incident Gaussian pulses approximately 50 fs in duration for metal layers and approximately 700 fs for the grating in order to resolve the narrow resonance; peak field intensity is approximately 2 GW/cm^2; and $\tilde{b}_1 = \tilde{b}_2 = \frac{\omega_0^2 \lambda_r^2}{r_0^2 n_{01}^2 \mathrm{e}^2 c^2} = 10^{-6}$. We compare reflected, TH conversion efficiencies at normal incidence for the two structures depicted in Figure 10.10. The increased conversion efficiency versus decreasing
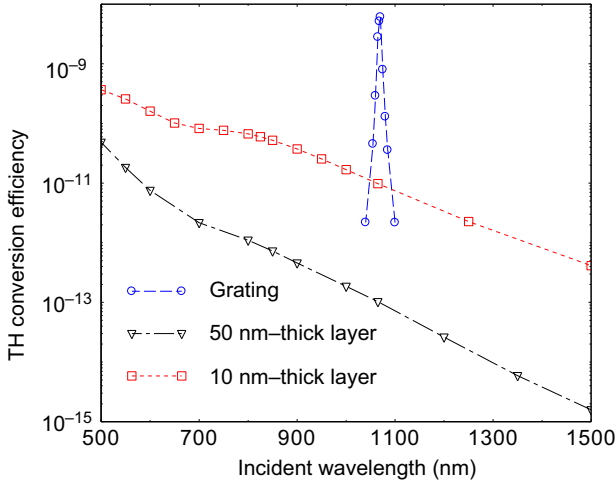
Figure 10.11 Reflected THG conversion efficiency versus incident pump wavelength for the structures depicted in Figure 10.10. The calculations are carried out for two different silver layer thicknesses in order to highlight the impact of thickness on harmonic generation and for the grating. At resonance, the plasmonic grating outperforms both metal layers by several orders of magnitude. Off-resonance, the efficiency of the 10 nm-thick layer overtakes the grating.

wavelength for the smooth 10 nm- and 50 nm-thick metallayers may be understood in terms of increased penetration depth, which is more pronounced for thinner layers and shorter wavelengths, or a combination of both. The results in Figure 10.11 suggest that plasmon-assisted THG mediated by the grating yields the largest predicted conversion efficiencies for the two types of structures we have examined, even though conversion efficiency is large over a relatively narrow band of wavelengths. Figure 10.10 also suggests that off-resonance, even the 10 nm-thick metal layer, can outperform the plasmonic grating by several orders of magnitudes.

The angular dependence of THG from 100 nm- to 10 nm-thick metal layers is reported in Figure 10.12. The results for the 100 nm-thick layer are very similar to
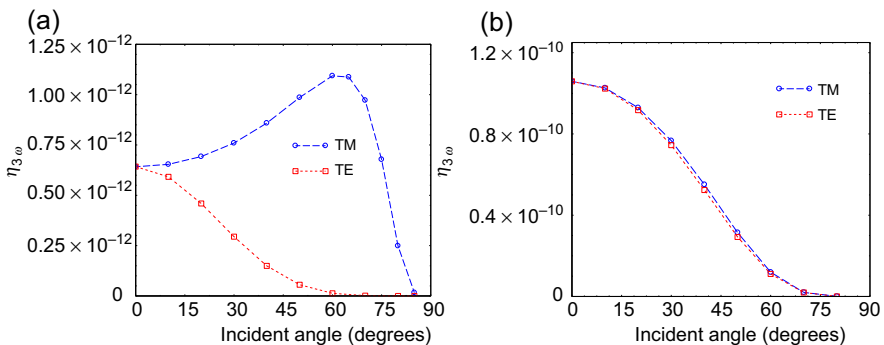
Figure 10.12 TM- and TE-polarized reflected THG conversion efficiency versus incident angle for (a) 100 nm-thick and (b) 10 nm-thick silver layers.

previously reported results. Our calculations yield similar results at 1064 and 850 nm for both TE- and TM-polarized light. The peak of maximum TH conversion efficiency shifts gradually to smaller angles as incident pump wavelength decreases, reaching 56° at 632 nm.

The qualitative differences that we predict between thick and thin metal layers are apparent: the 10 nm-thick layer—Figure 10.12(b)—displays maximum conversion efficiency at normal incidence, along with a concurrent increase of two orders of magnitude in conversion efficiency as compared to the 100 nm-thick layer. The 10 nm-thick layer displays improved conversion efficiencies thanks to simultaneously improved field penetration and larger local fields that tend to concentrate inside the metal and to exploit its bulk nonlinearity.

In Figure 10.13 we show the results for SHG for the smooth metal layers and the grating. Once again the grating yields the largest narrow-band conversion efficiencies near 1070 nm, the plasmonic resonance, and it is about two orders of magnitude larger than the reflected SHG arising from the maximum SH peak displayed by the 100 nm-thick silver mirror. The silver mirror's maximum conversion efficiency occurs near 68° and is predicted to be larger than the conversion efficiency peaks generated by either metal layer, notwithstanding improved field penetration in the thin layer,
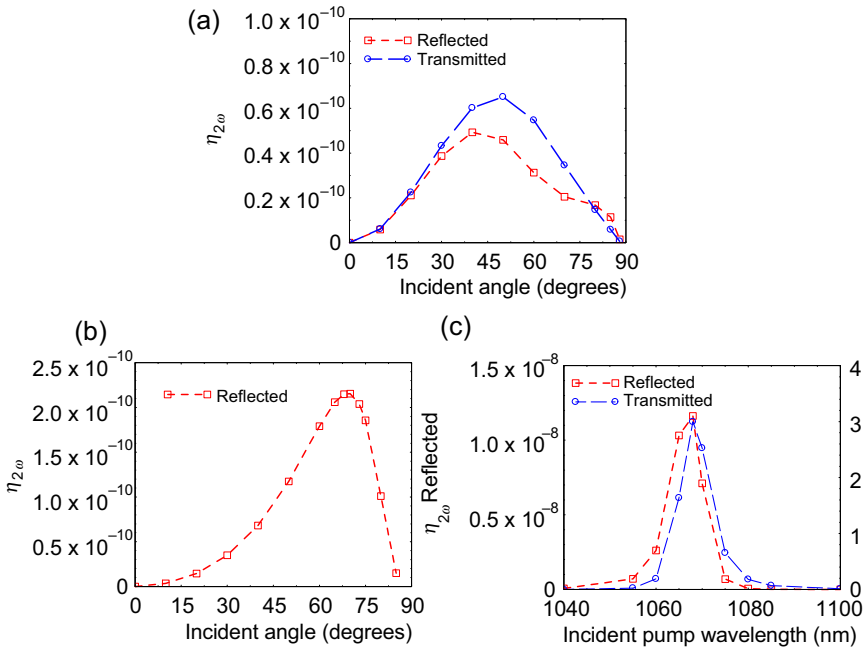
Figure 10.13 (a) SHG versus incident angle for 10 nm-thick silver and (b) a 100 nm-thick silver mirror and incident wavelength of 850 nm. The transmitted SHG for the 100 nm-thick layer may be neglected. (c) Transmitted and reflected SHG spectra for the grating described in Figure 10.2. The overall enhancement the grating provides with respect to the flat-layered structure is nearly two orders of magnitude with respect to the maximum SH peak in (b).

testimony to the fact that SHG remains mostly a surface phenomenon. The fact that maximum conversion efficiency of the arrangements that contain only flat metal layers occurs off-axis is due to the nature of the intrinsic second order nonlinearities and brings up the question of what is a reasonable benchmark for the purpose of comparing conversion efficiencies and assessing performance.

A reasonable benchmark for SHG arising from the grating is the peak produced by the simplest, most efficient structure, which in this case appears to be the peak of the silver mirror at $68°$. Therefore, realistic comparisons of the type we have discussed clearly suggest that resonant plasmonic gratings and metasurfaces may indeed provide a few orders of magnitude of enhanced performance, if compared to reasonably and appropriately chosen benchmarks.

Problems

1. Rewrite Eqn (10.2) for a second order nonlinearity, assuming fundamental and SH fields are present, and derive expressions equivalent to Eqns (10.10) and (10.11) for linear and second order nonlinear dispersion.
2. The dielectric function of silver may be expressed as follows in the range $200-1200$ nm:
$$\varepsilon(\tilde{\omega}) = 1 - \frac{\tilde{\omega}_{pf}^2}{\tilde{\omega}^2 + i\tilde{\gamma}_f\tilde{\omega}} - \frac{\tilde{\omega}_{p1}^2}{\tilde{\omega}^2 - \tilde{\omega}_{01}^2 + i\tilde{\gamma}_{01}\tilde{\omega}} - \frac{\tilde{\omega}_{p2}^2}{\tilde{\omega}^2 - \tilde{\omega}_{02}^2 + i\tilde{\gamma}_{02}\tilde{\omega}}, \text{where } \tilde{\omega} = 1/\lambda, \text{where } \lambda \text{ is in microns,}$$
   with $\tilde{\gamma}_f = 0.0573$, $\tilde{\gamma}_{01} = \tilde{\gamma}_{02} = 1.42\tilde{\omega}_{pf} = 6.965$, $\tilde{\omega}_{p1} = 4.25$, $\tilde{\omega}_{p2} = 5.5$, $\tilde{\omega}_{01} = 3.75$, $\tilde{\omega}_{02} = 5$. Using the definitions of the scaled plasma frequency as $\tilde{\omega}_{pj}^2 = 4\pi n_{0j}e^2\lambda_r^2/(m_{bj}^*c^2)$, show that the effective bound electron masses for each of the Lorentz oscillators are $m_{b1}^* \approx 98m_e$ and $m_{b2}^* \approx 58m_e$. Assume $\lambda_r = 1$ μm and $n_{01} = n_{02} = 5 \times 10^{22}$ cm^{-3}.
3. Beginning with the field Eqn (10.13), derive Eqn (10.14) and expand the Poynting vector in its various harmonic components.
4. Derive Eqn (10.16) for the fundamental field by substituting Eqn (10.13) into Eqn (10.15).
5. Derive the solutions (10.19) and (10.20) from Eqn (10.18) using the midpoint trapezoidal method.
6. Use the definition of quantum pressure to derive Eqn (10.35).
7. Substitute the assumptions in Eqn (10.47) into Eqn (10.46) to derive the set of Eqn (10.48).

References

[1] R.W. Boyd, Nonlinear Optics, Academic, 2003.
[2] E. Adler, Nonlinear optical frequency polarization in a dielectric, Phys. Rev. 134 (1964) A728.
[3] J.W. Haus, D. de Ceglia, M.A. Vincenti, M. Scalora, Quantum conductivity for metal-insulator-metal nanostructures, J. Opt. Soc. Am. B 31 (2014) 259.
[4] J.W. Haus, D. de Ceglia, M.A. Vincenti, M. Scalora, Nonlinear quantum tunneling effects in nano-plasmonic environments: two-photon absorption and harmonic generation, J. Opt. Soc. Am. B 31 (2014) A13-A19.

[5] M. Scalora, D. de Ceglia, M.A. Vincenti, J.W. Haus, Nonlocal and quantum tunneling contributions to harmonic generation in nanostructures: electron cloud screening effects, Phys. Rev. A 90 (2014) 013831.

[6] C. Ciracì, M. Scalora, D.R. Smith, Third-harmonic generation in the presence of classical nonlocal effects in gap-plasmon nanostructures, Phys. Rev. B 90 (2015) 205403.

[7] O.L. de Lange, R.E. Raab, Surprises in the multipole description of macroscopic electrodynamics, Am. J. Phys. 74 (2006) 301−312.

[8] J.D. Jackson, The Classical Electromagnetic Field, Wiley, 1999.

[9] J.A. Goldstone, E. Garmire, Intrinsic optical bistability in nonlinear media, Phys. Rev. Lett. 53 (1984) 910−913.

[10] R.C. Miller, Optical second harmonic generation in piezoelectric crystals, Appl. Phys. Lett. 5 (1964) 17−18.

[11] C.G. Garrett, F.N.H. Robinson, Miller's phenomenological rule for computing nonlinear susceptibilities, IEEE J. Quantum Elect. 2 (1966) 328−329.

[12] L. Allen, J.H. Eberly, Optical Resonance and Two-level Atoms, Dover, 1987.

[13] J. Koga, Simulation model for the effects of nonlinear polarization on the propagation of intense pulse lasers, Opt. Lett. 24 (1999) 408−410.

[14] C. Conti, A. Di Falco, G. Assanto, Frequency generation within the forbidden band gap: all optical rabi-like splitting in photonic crystals and microcavities, Phys. Rev. E 70 (2004) 066614.

[15] A.M.A. Ibrahim, P.K. Choudury, On the Maxwell-Duffing approach to model photonic deflection sensor, IEEE Photon. J. 5 (2013) 6800812.

[16] H. Ehrenreich, H.R. Philipp, Optical properties of Ag and Cu, Phys. Rev. 128 (1962) 1622.

[17] E.D. Palik, Handbook of Optical Constants of Solids, Academic Press, London-New York, 1985.

[18] R.W. Boyd, Z. Shi, I. De Leon, The third order nonlinear optical susceptibility of gold, Opt. Comm. 326 (2014) 74−79.

[19] M. Scalora, M.E. Crenshaw, A beam propagation method that handles reflections, Opt. Comm. 108 (1994) 191.

[20] M. Scalora, M.A. Vincenti, D. de Ceglia, V. Roppo, M. Centini, N. Akozbek, M.J. Bloemer, Second- and third-harmonic generation in metal-based structures, Phys. Rev. A 82 (2010) 043828.

[21] M.A. Vincenti, D. De Ceglia, V. Roppo, M. Scalora, Harmonic generation in metallic, GaAs-filled nanocavities in the enhanced transmission regime at visible and UV wavelengths, Opt. Express 19 (2011) 2064−2078.

[22] M. Scalora, G. D'Aguanno, N. Mattiucci, M.J. Bloemer, D. de Ceglia, M. Centini, A. Mandatori, C. Sibilia, N. Akozbek, M.G. Cappeddu, M. Fowler, J.W. Haus, Negative refraction and sub-wavelength focusing in the visible range using transparent metallo-dielectric stacks, Opt. Express 15 (2007) 508−523.

[23] D. de Ceglia, M. Vincenti, M. Cappeddu, M. Centini, N. Akozbek, A. D'Orazio, J. Haus, M. Bloemer, M. Scalora, Tailoring metallodielectric structures for super resolution and superguiding applications in the visible and near-IR ranges, Phys. Rev. A 77 (2008) 033848.

[24] R.S. Bennink, Y.K. Yoon, R.W. Boyd, J.E. Sipe, Accessing the optical nonlinearity of metals with metal-dielectric photonic bandgap structures, Opt. Lett. 24 (1999) 1416.

[25] N.N. Lepeshkin, A. Schweinsberg, G. Piredda, R.S. Bennink, R.W. Boyd, Enhanced nonlinear optical response of one-dimensional metal-dielectric photonic crystals, Phys. Rev. Lett. 93 (2004) 123902.

[26] K. Li, X. Li, D.Y. Lei, S. Wu, Y. Zhan, Plasmon gap mode-assisted third-harmonic generation from metal film-coupled nanowires, Appl. Phys. Lett. 104 (2014) 261105.

[27] H. Aouani, M. Rahmani, M. Navarro-Cía, S.A. Maier, Third-harmonic-upconversion enhancement from a single semiconductor nanoparticle coupled to a plasmonic antenna, Nat. Nanotechnology 9 (2014) 290−294.

[28] W.K. Burns, N. Bloembergen, Third-harmonic generation in absorbing media of cubic or isotropic symmetry, Phys. Rev. B 4 (1971) 3437−3450.

[29] J.E. Sipe, V.C.Y. So, M. Fukui, G.I. Stegeman, Analysis of second-harmonic generation at metal surfaces, Phys. Rev. B 21 (1980) 4389.

[30] D. Krause, C.W. Teplin, C.T. Rogers, Optical surface second harmonic measurements of isotropic thin-film metals: gold, silver, copper, aluminum, and tantalum, J. Appl. Phys. 96 (2004) 3626.

[31] N. Bloembergen, R.K. Chang, S.S. Jha, C.H. Lee, Optical harmonic generation in reflection from media with inversion symmetry, Phys. Rev. 174 (3) (1968) 813−822.

[32] H.W.K. Tom, T.F. Heinz, Y.R. Shen, Second-harmonic reflection from silicon surfaces and its relation to structural symmetry, Phys. Rev. Lett. 51 (1983) 1983−1986.

[33] M. Kauranen, T. Verbiest, A. Persoons, Second-order nonlinear optical signatures of surface chirality, J. Mod. Opt. 45 (1998) 403−423.

[34] P. Guyot-Sionnest, Y.R. Shen, Local and nonlocal surface nonlinearities for surface optical second harmonic generation, Phys. Rev. B 35 (1987) 4420−4426.

[35] P. Guyot-Sionnest, Y.R. Shen, Bulk contributions in surface second harmonic generation, Phys. Rev. B 38 (1988) 7985−7989.

[36] M. Galli, D. Gerace, K. Welna, T.F. Krauss, L. O'Faolain, G. Guizzetti, L.C. Andreani, Low-power continuous-wave generation of visible harmonics in silicon photonic crystal nanocavities, Opt. Exp. 18 (2010) 26613−26624.

[37] V. Roppo, M. Centini, C. Sibilia, M. Bertolotti, D. de Ceglia, M. Scalora, N. Akozbek, M.J. Bloemer, J.W. Haus, O.G. Kosareva, V.P. Kandidov, Role of phase matching in pulsed second-harmonic generation: walk-off and phase-locked twin pulses in negative-index media, Phys. Rev. A 76 (2007) 033829.

[38] V. Roppo, J.V. Foreman, N. Akozbek, M.A. Vincenti, M. Scalora, Third harmonic generation at 223 nm in the metallic regime of GaP, Appl. Phys. Lett. 98 (2011) 11105.

[39] D.J. Moss, E. Ghahramani, J.E. Sipe, H.M. van Driel, Band-structure calculation of dispersion and anisotropy in for third-harmonic generation in Si, Ge, and GaAs, Phys. Rev. B 41 (1990) 1542−1560.

[40] D.E. Aspnes, A.A. Studna, Anisotropies in the above-band-gap optical spectra of cubic semiconductors, Phys. Rev. Lett. 54 (1985) 1956−1959.

[41] R.W.J. Hollering, Angular dependence of optical second-harmonic generation at a Ge (111) surface, J. Opt. Soc. Am. B 8 (1991) 374−377.

[42] N. Bloembergen, Y.R. Shen, "Optical nonlinearity of a plasma, Phys. Rev. 141 (1966) 298−305.

[43] E.J. Adles, D.E. Aspnes, Application of the anisotropic bond model to second-harmonic generation from amorphous media, Phys. Rev. B 77 (2008) 165102.

[44] P.S. Pershan, Nonlinear optical properties of solids: energy considerations, Phys. Rev. 130 (1963) 919−928.

[45] T. Verbiest, K. Clays, V. Rodriguez, Second-Order Nonlinear Optical Characterization Techniques: An Introduction, Taylor and Francis, 2009.

[46] J.P. Gordon, Radiation forces and momenta in dielectric media, Phys. Rev. A 8 (1973) 14.

[47] D. Epperlein, B. Dick, G. Marowsky, G.A. Reider, Second harmonic generation in centro-symmetric media, Appl. Phys. B 44 (1987) 5−10.

[48] F. Xiang Wang, F.J. Rodríguez, W.M. Albers, R. Ahorinta, J.E. Sipe, M. Kauranen, Surface and bulk contributions to the second-order nonlinear optical response of a gold film, Phys. Rev. B 80 (2009) 233402.

[49] V. Roppo, F. Raineri, R. Raj, I. Sagnes, J. Trull, R. Vilaseca, M. Scalora, C. Cojocaru, Enhanced efficiency of the second harmonic inhomogeneous component in an opaque in a cavity, Opt. Lett. 36 (2011) 1809−1811.

This page intentionally left blank

Nanophotonic devices

11

J.W. Haus
University of Dayton, Dayton, OH, USA

> $\varepsilon \breve{v} \rho \eta \kappa \alpha$ *(Eureka).* **Archimedes** *famous remark after working out the connection between water displacement by weight and volume measurements to determine the purity of gold coins.*
>
> *The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny…'*
>
> **Isaac Asimov**
>
> *The science of today is the technology of tomorrow.*
>
> **Edward Teller**

11.1   Introduction

Photonic devices, sometimes referred to as optoelectronic devices, are essential in modern, high-performance applications, such as optical communications, biomolecule sensing, and data storage and retrieval. Devices, such as light sources, detectors, modulators, and amplifiers, are photonic elements commonly found in fiber communication and data storage systems. Photonic devices are the workhorses of our optical systems, and their operational requirements are as varied as the number of system applications. They perform the tasks necessary to keep our information economy running. Photonic devices are designed to encode information by gigahertz modulation of the optical field using electro-optic nonlinear effects, while other devices route the information stream from one point to its destination through an optical fiber communications system.

Another class of photonic devices are those developed to sense the presence of biomolecules. An aqueous sample with low concentrations of impurities requires a device with high sensitivity that is reliable for use outside of the laboratory. The application of photonic sensors to medicine demands both high-sensitivity concomitant with a low false alarm probability. Many photonic device strategies have been proposed to detect biomolecules at low concentrations. Those based on surface plasmon resonances (SPRs) have become ubiquitous for laboratory studies. Since the fields are confined relatively close to the surface, the angle of coupling light to the SPR is highly sensitive to the refractive index near the metal's surface. This has enabled the application of *label-free detection*, which is a technique of covalently attaching a layer of biomolecules, such as antibodies, to the surface. The biomolecules bind specific molecules, e.g., antigens that specifically attach to an antibody, in the surrounding environment and change the local properties, such as refractive index, absorption, or fluorescence yield.

Other sensors are based on resonator concepts and fabricated using nanotechnology tools. The high-Q characteristics of the resonator concentrate the field and are expected to increase the sensitivity of the measurements. They are expected to have a wide range of field applications, including monitoring the food supply for pathogens and environmental monitoring of water and air for pollutants. As improvements in device design are conceived, nanofabrication will be challenged to meet the specifications. In this chapter we highlight a few selected devices that are outcomes of nanophotonics research; by restricting the scope of the chapter, it will be kept to a manageable length.

11.2   Semiconductor optoelectronic devices

Semiconductor materials form the basis of the remarkable evolution of electronic devices whose development accelerated with the invention of integrated circuitry. As discussed in earlier chapters, the tools for epitaxial deposition of thin films are available with exquisite, atomic layer control of thin film thickness. The design of complex optoelectronic devices includes both the electronic and photonic properties of the materials. The diode is a common structure used in many optoelectronic device designs, whether it is a light-emitting diode, semiconductor laser, photovoltaic device, or a photodetector.

All sources of photons are designed to operate by well-known physical principles. The most common radiative effect is spontaneous emission, where an electron in an excited quantum state emits a photon while making a transition to a lower energy state, as illustrated in Figure 11.1(a). The energy of the ground and excited states are E_1 and
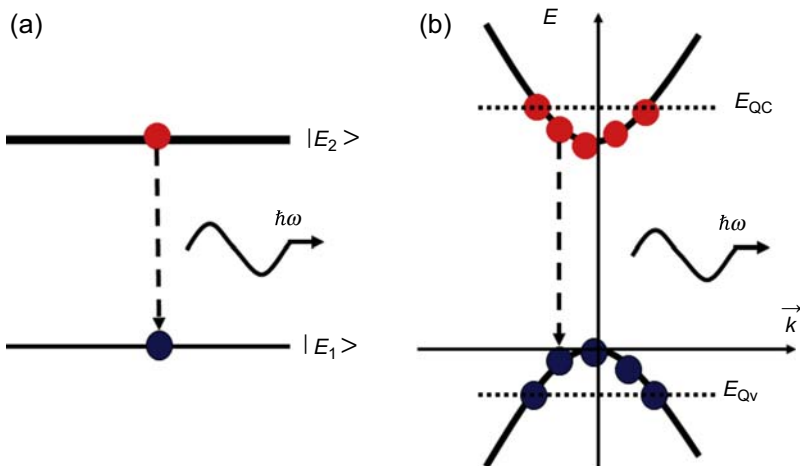
Figure 11.1 Illustration of spontaneous emission. (a) An electron initially in an excited state of energy E_2 transitions to a lower energy state of energy E_1 and at the same time emits a photon to conserve energy. (b) Electrons in the conduction band can recombine with holes in the valence band and emit a photon during process. The quasi-Fermi levels for the conduction and valence bands are denoted by E_{Qc} and E_{Qv}, respectively.

E_2, respectively. The emitted photon has energy $\hbar\omega = E_2 - E_1$, and energy is conserved in the process. The electronic wave functions for the atomic or molecular states are denoted by the "ket" notation $|E_a\rangle, a = 1, 2$, where the argument simply denotes the energy of the state. Spontaneous emission is a ubiquitous process found in all materials: atoms, molecules, and semiconductors. In the following section we will explore emission and absorption in semiconductors and discuss the changes in the optical processes due to nanoscale structures.

11.2.1 Photon emission and absorption processes

An electron and hole that occupy states in the conduction and valence bands can recombine, specifically in semiconductors, and in the process emit a photon. Two electronic bands (conduction and valence) with momentum states $|\vec{k}\rangle$ occupied by electrons and holes can recombine, as shown in Figure 11.1(b), and emit a photon. The momentum of the photon is much smaller than the electron/hole momenta, and the transition occurs without changing the particle momenta. The conduction and valence bands are separated by at least the band gap energy E_g. The kinetic energy of the conduction electrons of mass m_c is $\frac{\hbar^2 k^2}{2m_c}$, and the kinetic energy of the holes of mass m_v is $-\frac{\hbar^2 k^2}{2m_v}$ in the valence band.

The electron energy in the conduction and valence bands is

$$E_c\left(\vec{k}\right) = E_g + \frac{\hbar^2 k^2}{2m_c} \quad \text{and} \quad E_v\left(\vec{k}\right) = -\frac{\hbar^2 k^2}{2m_v}. \tag{11.1}$$

Conservation of energy for the electron−photon interaction dictates that the electronic transition satisfy

$$\hbar\omega = E_c\left(\vec{k}\right) - E_v\left(\vec{k}\right) = E_g + \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_c} + \frac{1}{m_v}\right) = E_g + \frac{\hbar^2 k^2}{2m_r}. \tag{11.2}$$

The mass m_r is dubbed the reduced mass. The minimum energy of the photons is the band gap energy, and most of the spontaneously emitted photons have energies close to this value. Spontaneous emission depends on the occupation probability of the electronic states. In equilibrium the occupation probability is dictated by the Fermi distribution function

$$f(E, E_F) = \frac{1}{e^{\beta(E - E_F)} + 1}, \tag{11.3}$$

where E_F is the Fermi level and $\beta = 1/k_B T$, T is the temperature, and k_B is Boltzmann's constant. The Fermi level marks the energy value where the occupation probability is a half. The Fermi level may lie between bands as it does in undoped semiconductors where it lies in the middle of the band gap. In doped semiconductors, the Fermi level is

shifted toward one of the band edges; for n-doped semiconductors, the Fermi level lies close to but below the conduction band edge, and in p-doped semiconductors it lies near to and above the valence band edge.

The equilibrium the density of carriers in the conduction and in valence band are denoted as n_0 and p_0, respectively. However, for optoelectronic devices excess carriers $\Delta n = \Delta p$ are injected into the active region, increasing the carrier concentration of electrons ($n = n_0 + \Delta n$) and holes ($p = p_0 + \Delta n$) and driving the system to a nonequilibrium state. To operate optoelectronic devices, the injected carrier density is large, i.e., $\Delta n \gg n_0, p_0$. However, even with a system that is far from equilibrium, the carrier distribution is expressed using the Fermi distribution but modified by replacing the equilibrium Fermi level with two quasi-Fermi levels, E_{Qc} and E_{Qv}, for the conduction and valence bands, respectively.

The quasi-equilibrium approximation is used for materials where the intraband relaxation times are much faster than the interband transitions. For instance, hot electrons within the conduction band will rapidly exchange energy with the lattice ($\sim 10^{-12}$ s) and relax to the bottom of the band, while the interband transitions depend on radiative and nonradiative processes, which are relatively slow ($\sim 10^{-9}$ s). In three-dimensional systems with a high density of injected carriers, the quasi-Fermi levels lie inside the bands. In these cases they are defined by the relation

$$\Delta n = \frac{2}{(2\pi)^3} \iiint \mathrm{d}^3 \vec{k} \, f(E, E_{Qa}), \quad a = v \ \text{ or } \ c. \tag{11.4}$$

At zero temperature the quasi-Fermi levels are expressed as functions of the excess carriers; using the parabolic approximation for the energy bands, the result is

$$E_{Qc} = E_g + \frac{\hbar^2}{2m_c}\left(3\pi^2 \Delta n\right)^{2/3}, \tag{11.5}$$

$$E_{Qv} = -\frac{\hbar^2}{2m_v}\left(3\pi^2 \Delta n\right)^{2/3}. \tag{11.6}$$

The energy at the top of the valence band is set to zero, and the bottom of the conduction band is shifted above zero by the band gap energy E_g. The quasi-Fermi levels appearing in Figure 11.1(b) illustrate that their position can lie well within the bands.

There are two additional physical processes that play important roles in semiconductor optoelectronics: absorption and stimulated emission. Both processes are depicted in Figure 11.2. In the absorption process in Figure 11.2(a), a photon is absorbed, and the energy and momentum are transferred to an electron−hole pair created in different bands. Stimulated emission in Figure 11.2(b) is the creation of a second photon that is catalyzed by the presence of the first photon. The result is the clone of the photon with the same wavelength and phase. Again, the dotted lines in Figure 11.2(b) denote the energies of the quasi-Fermi levels, as in Figure 11.1(b).

Emission and absorption rates are calculated by applying the perturbative quantum mechanical result called Fermi's golden rule, which is the basis for decay rate
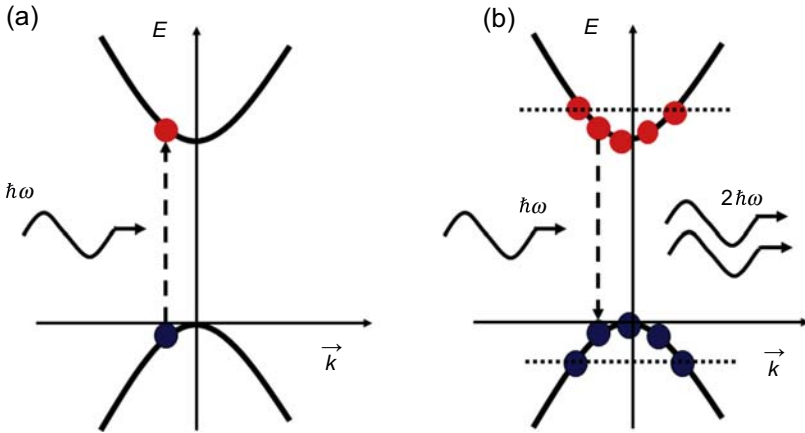
Figure 11.2 (a) Absorption of a photon creates occupied electron and hole states in the
conduction and valence bands. (b) An incident photon induces an electron−hole recombination
event with the emission of a photon in phase with the incident photon. The dotted lines denote
the energies of the quasi-Fermi levels.

calculations in quantum mechanics. For a quantum system initially in a state $|i\rangle$, the
transition rate (per second) to the final state $|f\rangle$ is

$$\Gamma_{if} = \frac{1}{V} \sum_f \frac{2\pi}{\hbar} |\langle f|H_{int}|i\rangle|^2, \tag{11.7}$$

where H_{int} is the interaction Hamiltonian, and the sum extends over the final states. The
interaction Hamiltonian captures the connection between the electrons and the photons.
The Hamiltonian for the interaction between an electromagnetic field and the electrons
(setting the static field contribution to zero) is a generalization of Schrödinger's equation
in Chapter 3 (the momentum operator is $\vec{p} = \frac{\hbar}{i}\vec{\nabla}$):

$$H = \frac{1}{2m_0}(\vec{p} - e\vec{A})^2 + V(\vec{r}) = \frac{1}{2m_0}p^2 + V(\vec{r}) + H_{int}. \tag{11.8}$$

The free mass of the electron is m_0; the vector potential $\vec{A}$ is related to the electro-
magnetic fields by the relations

$$\vec{B} = \vec{\nabla} \times \vec{A}, \text{ and } \vec{E} = -\frac{\partial \vec{A}}{\partial t}. \tag{11.9}$$

The vector potential satisfies the relation $\vec{\nabla} \cdot \vec{A} = 0$, which is called the Coulomb or
radiation gauge. The nonlinear vector potential contribution in the interaction Hamil-
tonian is small even for concentrated laser fields and neglected, yielding

$$H_{int} = -\frac{e\vec{A} \cdot \vec{p}}{m_0}. \tag{11.10}$$

The quantized version of the vector potential is written in a plane-wave expansion as

$$\vec{A} = \sum_{\vec{K}} \hat{e}_{\vec{K}} \frac{A_{\vec{K}}}{2} \left(a_{\vec{K}} e^{i\vec{K}\cdot\vec{r}} + a_{\vec{K}}^{+} e^{-i\vec{K}\cdot\vec{r}} \right). \tag{11.11}$$

$\hat{e}_{\vec{K}}$ is the polarization of the mode, and $A_{\vec{K}}$ is its amplitude. The operators $(a_{\vec{K}}^{+}, a_{\vec{K}})$ are photon creation and annihilation operators in the plane-wave basis that satisfy the commutation relations

$$\left[a_{\vec{K}'}, a_{\vec{K}}^{+} \right] = a_{\vec{K}'} a_{\vec{K}}^{+} - a_{\vec{K}}^{+} a_{\vec{K}'} = \delta_{\vec{K}',\vec{K}}. \tag{11.12}$$

The last equality contains the Kronecker delta function, which is zero unless the wave vectors for both photonic modes are equal when it is equal to unity. The quantum states for photons can be expanded in the set of photon number states $\left\{ \left| n_{\text{ph},\vec{K}} \right\rangle, \text{for all } \vec{k}, n_{\text{ph},\vec{K}} = 0, 1, 2... \right\}$, also called Fock states. The property of the creation and annihilation operators acting on the Fock states is

$$a_{\vec{K}} \left| n_{\text{ph},\vec{K}} \right\rangle = \sqrt{n_{\text{ph},\vec{K}}} \left| n_{\text{ph},\vec{K}} - 1 \right\rangle,$$

$$a_{\vec{K}}^{+} \left| n_{\text{ph},\vec{k}} \right\rangle = \sqrt{n_{\text{ph},\vec{K}} + 1} \left| n_{\text{ph},\vec{K}} + 1 \right\rangle. \tag{11.13}$$

11.2.2 Density of states

The density of states (DOS) plays an important role in determining the transition rates; the sum over final states in Eqn (11.7) can be related to the DOS, which has an important role in the determination of the rates. The DOS is not immutable, in other words it is not a quantity that is fixed by nature, but the DOS can be manipulated to improve device performance. The electronic or photonic DOS in momentum space is derived by counting up the states to a given wave number value is determined from the number of waves confined to a box with a volume in D dimensions give as $V(D) = L^{D}$, where each side of the box has a length L. The smallest component of the wave vector for periodic boundary conditions is $\frac{2\pi}{L}$. The number of modes, $N(k)$, is calculated in Eqn (4.24) in three dimensions by integration over the normalized volume in k-space the number of modes is generalized to D dimensions. The DOS, as a function of wave number, is defined by

$$\rho(k) = \frac{1}{V(D)} \frac{dN(k)}{dk} = s \left(\frac{1}{(2\pi)^{D}} \right) \Omega_{D} k^{D-1}, \tag{11.14}$$

where Ω_{D} is expression for angular integrals $\Omega_{1} = 2$, $\Omega_{2} = 2\pi$, $\Omega_{3} = 4\pi$; s adds spin/polarization degrees of freedom for the waves; and $s = 2$ for both electrons and

photons. The DOS has units that depend on the dimensionality —for example, states per cubic meter per inverse meter. The DOS is expressed in terms of the energy (e.g., states per cubic meter per Joule) using the relation

$$\rho(E)dE = \rho(k)dk. \tag{11.15}$$

The DOS in three dimensions for the carriers in the conduction band or the valence band within the parabolic approximation for the band energies is

$$\rho_c(E) = \frac{\sqrt{2}m_c^{3/2}}{\pi^2\hbar^3}\sqrt{E - E_g}, \text{ and } \rho_v(E) = \frac{\sqrt{2}m_v^{3/2}}{\pi^2\hbar^3}\sqrt{-E}. \tag{11.16}$$

The electronic DOS are used to express the carrier densities as

$$n + \Delta n = \int dE\rho_c(E)f(E, E_{Qc}), \tag{11.17}$$

$$p + \Delta n = \int dE\rho_v(E)f(E, E_{Qv}). \tag{11.18}$$

For completeness, since the energy−momentum relation for photons is a linear function, the photonic DOS in three dimensions in a medium of (real) dielectric permittivity is written below. The dielectric function is expressed in terms of the real part of the refractive index r_r and the permittivity of free space ε_0, as $\varepsilon = n_r^2\varepsilon_0$.

$$\rho_{ph}(E) = \frac{n_r^3}{\pi^2\hbar c}\left(\frac{E}{\hbar c}\right)^2. \tag{11.19}$$

The photonic DOS defined in Eqn (11.19) is number of photon states in an energy decrement, dE, per unit volume per unit energy.

11.2.3 Optical properties of semiconductors

We calculate the absorption coefficient using a single plane-wave mode (say wave vector $\vec{K}_0$) for the electromagnetic field occupied by a photon number n_{ph} and the initial wave function of the electronic system. The initial state denoting the photon mode and the multielectronic state is

$$|i\rangle = \left|n_{ph}, \psi_{i,\vec{k}}\right\rangle. \tag{11.20}$$

The final state is composed of an electromagnetic field with one less photon and an excited electronic state:

$$|f\rangle = \left|n_{ph} - 1, \psi_{f,\vec{k}}\right\rangle. \tag{11.21}$$

The wave functions $\psi_{i,\vec{k}}$ and $\psi_{f,\vec{k}}$ are the Bloch functions for the semiconductor lattice (refer to Chapter 4, Supplement B). For interband transitions the semiconductor wave function for the initial state of the electron is in the conduction band, and in the final state it is in the valence band. They are written in the form

$$\psi_{i,\vec{k}}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u_{\mathrm{c}}(\vec{r}) \text{ and } \psi_{f,\vec{k}}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u_{\mathrm{v}}(\vec{r}). \tag{11.22}$$

The amplitude of the quantized vector potential is defined in Eqn (11.11). Using the interaction matrix element for photon absorption processes and the perturbative Hamiltonian operator the matrix element between the initial and final state is

$$\left\langle \psi_{f,\vec{k}_f}, n_{\mathrm{ph}} - 1 \left| -\frac{e}{m_0}\vec{A}\cdot\vec{p} \right| n_{\mathrm{ph}}, \psi_{i,\vec{k}_i} \right\rangle =$$
$$-\frac{e}{m_0}\sqrt{n_{\mathrm{ph}}}A_{\vec{K}}\left\langle \psi_{f,\vec{k}_f} \left| \hat{e}\cdot\vec{p} \right| \psi_{i,\vec{k}_i} \right\rangle. \tag{11.23}$$

$\hat{e}$ is the polarization unit vector of the electromagnetic wave. The last expression can be written as

$$\left\langle \psi_{f,\vec{k}_f} \left| \hat{e}\cdot\vec{p} \right| \psi_{i,\vec{k}_i} \right\rangle = \hat{e}\cdot\iiint \mathrm{d}^3 r\, e^{-i\vec{k}_f\cdot\vec{r}}u_{\mathrm{c}}^*(\vec{r})e^{i\vec{K}\cdot\vec{r}}$$
$$\times\left(\vec{k}_i + \vec{p}\right)e^{i\vec{k}_i\cdot\vec{r}}u_{\mathrm{v}}(\vec{r}). \tag{11.24}$$

The functions $\{u_{\mathrm{c}}(\vec{r}), u_{\mathrm{v}}(\vec{r})\}$ are periodic functions over the lattice period of the unit cell, and $\vec{k}_i$ is a much more slowly varying function. The integral over the volume can be reduced to a sum over the lattice with integrals over the primitive unit cell:

$$\iiint \mathrm{d}^3 r\, e^{-i\vec{k}_f\cdot\vec{r}}u_{\mathrm{c}}^*(\vec{r})e^{i\vec{K}\cdot\vec{r}}\left(\vec{k}_i + \vec{p}\right)e^{i\vec{k}_i\cdot\vec{r}}u_{\mathrm{v}}(\vec{r})$$
$$= \sum_{\text{lattice }\{\vec{R}\}} e^{-i(\vec{k}_i+\vec{K}-\vec{k}_f)\cdot\vec{R}}\iiint_{\text{unit cell}} \mathrm{d}^3 r\, u_{\mathrm{c}}^*(\vec{r})\left(\vec{k}_i + \vec{p}\right)u_{\mathrm{v}}(\vec{r}). \tag{11.25}$$

In Chapter 3 the valence bands are identified as states with p-state symmetry, and the conduction band has s-state symmetry; therefore, the first integral vanishes, i.e.,

$$\vec{k}_i\iiint \mathrm{d}^3 r\, u_{\mathrm{c}}^*(\vec{r})u_{\mathrm{v}}(\vec{r}) = 0. \tag{11.26}$$

The second integral has a momentum matrix element denoted as

$$\vec{p}_{cv} = N \iiint_{\text{unit cell}} d^3r \, u_c^*(\vec{r})\vec{p}u_v(\vec{r}). \tag{11.27}$$

N is the number of lattice sites. The sum over the lattice imposes momentum conservation for each component

$$\sum_{\text{lattice } \{\vec{R}\}} e^{-i(\vec{k}_i + \vec{K} - \vec{k}_f)\cdot\vec{R}} = N\delta_{\vec{k}_i + \vec{K} - \vec{k}_f}. \tag{11.28}$$

$\delta_{\vec{k}_i + \vec{K} - \vec{k}_f}$ is a product of Kronecker deltas for each momentum component. The photon momentum ($\vec{K}$) is small so that electron momenta are the same for the initial and final states.

Applying Eqn (11.7) for an electronic transition from the ground state to an excited state, the absorption rate (per second) is

$$\Gamma_{\text{abs}} = \frac{2\pi}{\hbar}\left(\frac{e}{m_0}\right)^2 n_{\text{ph}}\left|\frac{A_{\vec{K}}}{2}\right|^2 \left|\hat{e}\cdot\vec{p}_{cv}\right|^2 \delta\left(E_v(\vec{k}_i) - E_c(\vec{k}_i) - \hbar\omega\right). \tag{11.29}$$

The delta function expresses the requirement that the transitions conserve energy. Following the same reasoning, the stimulated and spontaneous emission rate is expressed as

$$\Gamma_{\text{st}} + \Gamma_{\text{sp}} = \frac{2\pi}{\hbar}\left(\frac{e}{m_0}\right)^2 (n_{\text{ph}} + 1)\left|\frac{A_{\vec{K}}}{2}\right|^2 \left|\hat{e}\cdot\vec{p}_{cv}\right|^2 \delta\left(E_v(\vec{k}_i) - E_c(\vec{k}_i) - \hbar\omega\right). \tag{11.30}$$

The spontaneous emission contribution is the rate calculated when the electromagnetic field is in the vacuum state $n_{\text{ph}} = 0$, and the stimulated emission rate is distinguished by the presence of initial photons in the state.

For many optoelectronic devices, electrons and holes are injected into the semiconductor conduction band. Valence bands and the transition rates determined from Fermi's golden rule are incomplete, since the initial electronic state is not externally controlled. The initial electronic energy is summed over, and the occupation of the initial and final states is accounted for.

Also, the emission or absorption rates are subject to the occupation of the initial and final states. To calculate the emission rate, the initial state in the conduction band should be occupied, and the final state in the valence band should be empty. The probability that the initial electronic state is occupied is determined according to the Fermi distribution in the conduction band

$$f_c(E_i, E_{Qc}), \tag{11.31}$$

and the probability that the final state is unoccupied by an electron is related to the Fermi distribution in the valence band

$$1 - f_v(E_f, E_{Qv}).$$ (11.32)

The subscripts (v, c) are added to identify the band for the Fermi functions. Equation (11.32) is interpreted as the occupation probability for a hole in the final energy state E_f. The quasi-Fermi levels were previously kept in the argument of the Fermi functions to remind the reader that charges are injected to operate the devices. In the following, we will suppress writing the Fermi functions explicitly in the arguments to streamline the notation.

The product of these two probabilities determines the emission probability:

$$f_e(E_i, E_f) = f_c(E_i, E_{Qc})(1 - f_v(E_f, E_{Qv})).$$ (11.33)

The quasi-Fermi energies are implicit in the emission probability arguments to keep the notation compact. To calculate the total emission rate, the transition rate from Fermi's golden rule is summed over initial states. The sum is transformed into an integral over the plane-wave momenta

$$\frac{2}{V} \sum_i \ldots \rightarrow 2 \iiint \frac{d^3 k_i}{(2\pi)^3} \ldots .$$ (11.34)

The factor of 2 is for the electronic spin states. The frequency-dependent stimulated emission rate for an electronic transition from the conduction band to the valence band is

$$\overline{R}_{st}(\omega) = \frac{2\pi}{\hbar} \left(\frac{e}{m_0}\right)^2 n_p \left|\frac{A_{\vec{K}}}{2}\right|^2 \left|\hat{e} \cdot \vec{p}_{cv}\right|^2$$

$$\times \iiint \frac{d^3 k}{(2\pi)^3} \delta\left(E_c\left(\vec{k}\right) - E_v\left(\vec{k}\right) - \hbar\omega\right) f_e\left(E_c\left(\vec{k}\right), E_v\left(\vec{k}\right)\right).$$ (11.35)

The stimulated emission rate has units of photons per unit volume per unit time. The delta function reduces the integral to an expression for the electronic DOS; it is convenient to replace the electronic DOS by the *optical joint* DOS using the relation

$$\rho_j(E)d(\hbar\omega) = \rho_v(E)dE,$$ (11.36)

with the optical joint DOS result for $D = 3$ as

$$\rho_{j3}(\hbar\omega) = \frac{\sqrt{2} m_r^{3/2}}{\pi^2 \hbar^3} \sqrt{\hbar\omega - E_g}.$$ (11.37)

The function vanishes for photon energies less than the band gap energy. The stimulated emission rate is

$$\overline{R}_{st}(\omega) = \frac{2\pi}{\hbar}\left(\frac{e}{m_0}\right)^2 n_p \left|\frac{A_{\vec{K}}}{2}\right|^2 \left|\hat{e}\cdot\vec{p}_{cv}\right|^2 \rho_{j3}(\hbar\omega)$$
$$\times \left(\left[1-f_v\left(-(\hbar\omega - E_g)\frac{m_r}{m_h}\right)\right]f_c\left(E_g + (\hbar\omega - E_g)\frac{m_r}{m_c}\right)\right). \tag{11.38}$$

The vector potential amplitude is determined from the Poynting vector for a single frequency field. The electric field amplitude is related to the vector potential amplitude by $\left|E_{\vec{K}}\right| = \omega\left|A_{\vec{K}}\right|$, and the time averaged Poynting vector is the power per unit area

$$I = \frac{n_r}{2\mu_0 c}n_p\left|E_{\vec{K}}\right|^2 = \frac{n_r\omega^2}{2\mu_0 c}n_p\left|A_{\vec{K}}\right|^2. \tag{11.39}$$

The number of photons per unit volume is extracted from the intensity as

$$N_p = \frac{n_p}{V} = \frac{I}{\hbar\omega v_g} = \frac{n_r\omega}{2\hbar\mu_0 c v_g}n_p\left|A_{\vec{K}}\right|^2. \tag{11.40}$$

The group velocity ($v_g = c/n_g$) of the wave defined in Chapter 2 takes into account dispersion in the medium. The vector potential amplitude is

$$\left|A_{\vec{K}}\right|^2 = \frac{2\hbar}{\omega n_r n_g \varepsilon_0 V}. \tag{11.41}$$

The stimulated emission rate can now be determined by replacing the vector potential amplitude in Eqn (11.41). The coefficient of the emission occupation probability is

$$\overline{R}_{st,0}(\omega) = \frac{2\pi}{\hbar}\left(\frac{e}{m_0}\right)^2 N_p \frac{\hbar}{2\omega n_r n_g \varepsilon_0}\left|\hat{e}\cdot\vec{p}_{cv}\right|^2 \rho_{j3}(\hbar\omega). \tag{11.42}$$

The stimulated emission rate is

$$\overline{R}_{st}(\omega) = \overline{R}_{st,0}(\omega)\left(\left[1-f_v\left(-(\hbar\omega - E_g)\frac{m_r}{m_h}\right)\right]f_c\left(E_g + (\hbar\omega - E_g)\frac{m_r}{m_c}\right)\right). \tag{11.43}$$

The spontaneous emission rate is determined starting from Eqn (11.30) in the absence of a photon density. It depends on the emission probability, as used to calculate the stimulate emission rate, and the photons are emitted with random polarization.

Their radiation covers all solid angles. The spontaneous emission rate (photons per unit volume per unit time) is

$$
\overline{R}_{\text{sp}} = \frac{4}{3} \int d(\hbar\omega) \left(\frac{e}{m_0}\right)^2 \frac{n_{\text{r}}\omega}{\hbar} |p_{\text{cv}}|^2
$$
$$
\times \iiint \frac{d^3k}{(2\pi)^3} \delta\left(E_{\text{c}}\left(\vec{k}\right) - E_{\text{v}}\left(\vec{k}\right) - \hbar\omega\right) f_{\text{e}}\left(E_{\text{c}}\left(\vec{k}\right), E_{\text{v}}\left(\vec{k}\right)\right). \tag{11.44}
$$

The expression can be simplified using the optical joint DOS and

$$
\overline{R}_{\text{sp}} = \Gamma_{\text{sp}} \int d(\hbar\omega) \rho_{j3}(\hbar\omega) f_{\text{e}}\left(E_{\text{c}}\left(\vec{k}\right), E_{\text{v}}\left(\vec{k}\right)\right). \tag{11.45}
$$

The photon absorption rate follows a derivation that is similar to the stimulated emission rate with the result

$$
\overline{R}_{\text{abs}}(\omega) = \overline{R}_{\text{abs},0}(\omega)\left(f_{\text{v}}\left(-(\hbar\omega - E_{\text{g}})\frac{m_{\text{r}}}{m_{\text{h}}}\right)\left[1 - f_{\text{c}}\left(E_{\text{g}} + (\hbar\omega - E_{\text{g}})\frac{m_{\text{r}}}{m_{\text{c}}}\right)\right]\right). \tag{11.46}
$$

The absorption Fermi occupation probability in parentheses represents a process where the valence band state is occupied and the conduction band state is empty. The coefficient is defined as

$$
\overline{R}_{\text{abs},0}(\omega) = \frac{2\pi}{\hbar}\left(\frac{e}{m_0}\right)^2 N_{\text{p}} \frac{\hbar}{2\omega n_{\text{r}} n_{\text{g}} \varepsilon_0}\left|\hat{e}\cdot\vec{p}_{\text{cv}}\right|^2 \rho_{j3}(\hbar\omega). \tag{11.47}
$$

The absorption coefficient $\alpha(\omega)$ is defined by setting the power loss per unit volume to the absorption rate

$$
\alpha(\omega)I = \hbar\omega \overline{R}_{\text{abs}}(\omega). \tag{11.48}
$$

The absorption coefficient is

$$
\alpha(\omega) = \frac{\pi}{2\omega n_{\text{r}} c\varepsilon_0}\left(\frac{e}{m_0}\right)^2 \left|\hat{e}\cdot\vec{p}_{\text{cv}}\right|^2 \rho_{j3}(\hbar\omega)
$$
$$
\times \left(f_{\text{v}}\left(-(\hbar\omega - E_{\text{g}})\frac{m_{\text{r}}}{m_{\text{h}}}\right)\left[1 - f_{\text{c}}\left(E_{\text{g}} + (\hbar\omega - E_{\text{g}})\frac{m_{\text{r}}}{m_{\text{c}}}\right)\right]\right). \tag{11.49}
$$

The absorption coefficient ignores stimulated emission, which can dominate for strong carrier injection. The net gain coefficient is defined, similar to the absorption coefficient, by using both gain and loss contributions:

$$
g(\omega)I = \hbar\omega\left(\overline{R}_{\text{st}}(\omega) - \overline{R}_{\text{abs}}(\omega)\right). \tag{11.50}
$$

The net gain spectrum is extracted from Eqn (11.50) after dividing by the intensity in Eqn (11.39):

$$
\begin{aligned}
g(\omega) = {} & \frac{\pi}{\omega n_{\mathrm{r}} c \varepsilon_0} \left(\frac{e}{m_0}\right)^2 \left|\widehat{e}\cdot\vec{p}_{\mathrm{cv}}\right|^2 \rho_{j3}(\hbar\omega) \\
& \times \left(f_{\mathrm{c}}\left(E_{\mathrm{g}} + (\hbar\omega - E_{\mathrm{g}})\frac{m_{\mathrm{r}}}{m_{\mathrm{c}}} \right) - f_{\mathrm{v}}\left(-(\hbar\omega - E_{\mathrm{g}})\frac{m_{\mathrm{r}}}{m_{\mathrm{h}}} \right) \right).
\end{aligned}
\tag{11.51}
$$

For $g(\omega) > 0$ the signal is amplified in the semiconductor. The gain energy bandwidth can be determined from Eqn (11.51), since the sign is determined by the last term. From the condition

$$
f_{\mathrm{v}}(E_{\mathrm{v}}) - f_{\mathrm{c}}(E_{\mathrm{c}}) < 0.
\tag{11.52}
$$

the following expression can be derived:

$$
E_{\mathrm{c}} - E_{\mathrm{v}} = \hbar\omega < E_{\mathrm{Qc}} - E_{\mathrm{Qv}}.
\tag{11.53}
$$

The gain lies in the energy interval between the band gap energy and the difference between the quasi-Fermi levels (E_{g}, $E_{\mathrm{Qc}} - E_{\mathrm{Qv}}$).

Momentum interband matrix elements are related to a single parameter by the relation

$$
\left|\widehat{e}\cdot\vec{p}_{\mathrm{cv}}\right|^2 = \frac{m_0}{2} E_{\mathrm{p}}.
\tag{11.54}
$$

The semiempirical E_{p} values for several diamond and zinc-blende lattice materials is given in Table 11.1. The numbers do not vary their values a great deal for the semiconductor materials. The value for diamond is large in correspondence with the large band gap for this material; this property was noted and discussed in Chapter 4.

The E_{p} values in Table 11.1 have an interesting interpretation when they are related to the free electron kinetic energy, i.e., $E_{\mathrm{p}} = \frac{\hbar^2}{2m_0}\left(\frac{2\pi}{\Lambda}\right)^2$. The wavelength corresponds to values in the range of lattice constants $\Lambda = 0.24 - 0.27$ nm with the exception of diamond, where the wavelength is $\Lambda = 0.17$ nm. In fact, these results loosely correlate with the lattice constants of the materials, since diamond's lattice constant is smaller than that for the semiconductor materials.

Table 11.1 E_{p} values for several diamond/zinc-blende semiconductors

Material	C	Si	GaAs	AlAs	InP	InAs	GaP	CdTe
E_{p} (eV)	49.8	21.6	25.7	21.1	20.9	22.2	22.2	20.7

11.2.4 Optical properties in quantum confined semiconductors

Nanometer thickness, semiconductor films over a substrate, restrict the kinetic energy of carriers and quantize the energy. The growth of precise films has improved the wall-plug efficiency of semiconductor lasers as well as their wavelength tunability. Quantum well technology has been successfully transitioned to commercial devices, such as laser diodes, light-emitting diodes, and photodetectors.

The effect of quantum confinement on the electronic states is illustrated in Figure 11.3. In quantum wells one degree of motion is quantized. The two unaffected momenta are parallel to the plane of the film and denoted as the vector $\vec{k}_{\parallel}$. For a film of thickness L_z and infinite well height, the conduction and valence band energies are

$$E_c(n) = E_g + \frac{\hbar^2}{2m_c}\left(\frac{n_c \pi}{L_z}\right)^2 + \frac{\hbar^2}{2m_c}k_{\parallel}^2, \tag{11.55}$$

$$E_v(n) = -\frac{\hbar^2}{2m_v}\left(\frac{n_v \pi}{L_z}\right)^2 - \frac{\hbar^2}{2m_v}k_{\parallel}^2. \tag{11.56}$$
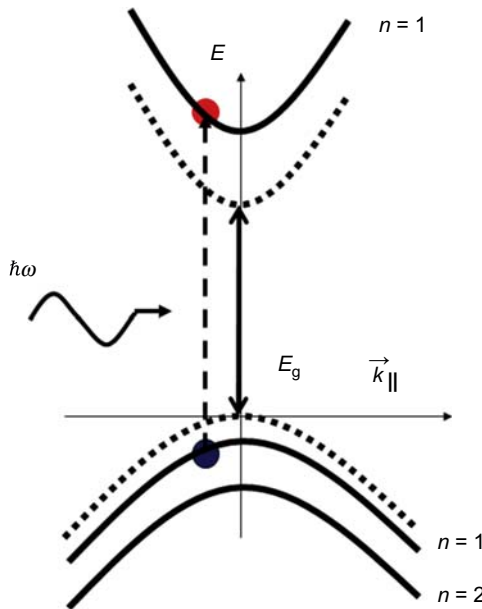
Figure 11.3 Schematic of quantum confined electronic states due to a thin film of thickness L_z and quantum well structure using the parabolic approximation for the bands. The band gap energy E_g with no quantum confinement effect (i.e., when $L_z \to \infty$) is drawn as dotted lines. The integer n is the level carrier state in the quantum well.

The integer indices (n_c, n_v) take integer values $\{1, 2, ...\}$. Including the quantum confinement effect, the transition energy between two states is modified to

$$E(n_v, n_c) = E_g + \frac{\hbar^2 n_c^2}{2m_c}\left(\frac{\pi}{L_z}\right)^2 + \frac{\hbar^2 n_v^2}{2m_v}\left(\frac{\pi}{L_z}\right)^2 + \frac{\hbar^2 k_\parallel^2}{2m_r}. \tag{11.57}$$

In comparison with Eqn (11.2), the energy bands are shifted, yielding a larger transition energy due to the quantum confinement. The integers (n_c, n_v) are independent of one another, and the energy shift due to confinement in the quantum well is inversely proportional to the carrier mass. However, the transitions between states with equal index values can be dominant for low-lying states in the absence of an applied electric field. The optical joint DOS is modified by the reduced dimensionality from Eqn (11.36) to $\left(\text{assuming } n_c = n_v,\ E_n = E(n, n) = \frac{\hbar^2}{2m_r}\left(\frac{n\pi}{L_z}\right)^2\right)$

$$\rho_{j2}(\hbar\omega) = \frac{m_r}{\pi\hbar^2 L_z} \sum_{n=1,2,} \Theta(\hbar\omega - (E_g + E_n)), \tag{11.58}$$

where the Heaviside function is

$$\Theta(\hbar\omega - (E_g + E_n)) = \begin{cases} 1, & \hbar\omega > E_g + E_n \\ 0, & \hbar\omega < E_g + E_n \end{cases}. \tag{11.59}$$

The interesting and useful feature of the two-dimensional optical joint DOS is that the discrete steps as transitions are energetically allowed for higher energies. For comparison, the DOS for $D = 3$ and 2 is plotted in Figure 11.4.

In quantum wells, the Bloch wave function is modified by the confining potential. The functions that are periodic over the unit cell are relatively insensitive to the boundary. The absorption is affected due to the shift of the band edge to higher values and the change of the DOS. The wave functions are modified by the introduction of an envelope function for the electron wave. The position vector is separated into components
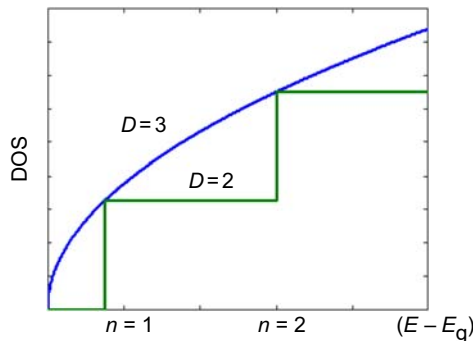
Figure 11.4 Optical joint DOS for $D = 3$ and $D = 2$. The latter shows two steps: $n = 1$ and $n = 2$.

parallel $(\vec{\rho})$ and perpendicular $(\hat{e}_z z)$ to the walls $(\vec{r} = \vec{\rho} + \hat{e}_z z)$. The quantum well wave functions for the conduction and valence bands, analogous to the two-band expression in Eqn (11.22), are

$$
\psi_{c,\vec{k}n'}(\vec{r}) = e^{i\vec{\kappa}\cdot\vec{\rho}}F_{cn'}(z)u_{cn'}(\vec{r}) \quad \text{and}
$$

$$
\psi_{v,\vec{k}n}(\vec{r}) = e^{i\vec{\kappa}\cdot\vec{\rho}}F_{vn}(z)u_{vn}(\vec{r}).
$$

(11.60)

The wave vector $\vec{\kappa}$ is the Bloch vector for the two unrestricted dimensions. The envelope functions $\{F_{cn'}(z), F_{vn}(z)\}$ are Schrödinger equation solutions—bound walls of the quantum well separated by a distance L_z. The momentum matrix elements are

$$
\vec{p}_{cv,n} = N \iiint_{\text{unit cell}} \mathrm{d}^3 r F_{cn'}^*(z)u_{cn'}^*(\vec{r})\vec{p}F_{vn}(z)u_{vn}(\vec{r}).
$$

(11.61)

For deep states in a potential well in the absence of an applied field, the envelope functions are normalized and approximately orthogonal

$$
\int \mathrm{d}z F_{cn'}^*(z)F_{vn}(z) = \delta_{n,n'}.
$$

(11.62)

The momentum matrix elements for interband transitions are largely unaffected by the quantum confinement potential. The emission and absorption spectra are modified by the shift of the energy band edge and the DOS. The gain spectrum is

$$
g(\omega) = \frac{\pi}{\omega n_r c \varepsilon_0}\left(\frac{e}{m_0}\right)^2 \left|\hat{e}\cdot\vec{p}_{cv}\right|^2 \sum_n \frac{m_r}{\pi\hbar^2 L_z}\Theta(\hbar\omega - (E_g + E_n))(f_v(E_v) - f_c(E_c)).
$$

(11.63)

where the argument of the Fermi functions are

$$
E_v = -E_{vn} - (\hbar\omega - E_g)\frac{m_r}{m_h} \quad \text{and} \quad E_c = E_g + E_{cn} + (\hbar\omega - E_g)\frac{m_r}{m_c}.
$$

(11.64)

with

$$
E_{cn} = \frac{\hbar^2 n_c^2}{2m_c}\left(\frac{\pi}{L_z}\right)^2 \quad \text{and} \quad E_{vn} = \frac{\hbar^2 n_v^2}{2m_v}\left(\frac{\pi}{L_z}\right)^2.
$$

(11.65)

Figure 11.5 illustrates a distinction between the gain in a bulk semiconductor system from that in a quantum well system. The amplitude of the gains is scaled in these figures, and a finite temperature is used. The discontinuity in the quantum well gain is due to the DOS. At low temperatures, the gain spectrums sharpen around the crossover from gain to loss.
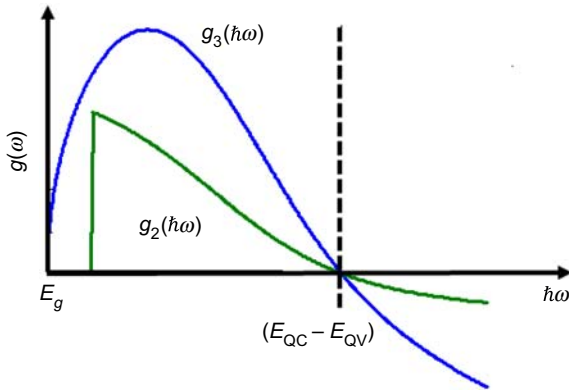
Figure 11.5 Illustration of the gain spectrums in two and three dimensions. The vertical dotted line is the difference in quasi-Fermi levels separating the gain from loss regions of the spectrum.

The design of quantum well lasers provides control of the emission wavelength over a wide range of values by choosing the quantum well width. The details of quantum well designs in the presence of applied fields were discussed in Chapter 3. The confinement dimensions are smaller than the de Broglie thermal wavelength ($\lambda_T = h/\sqrt{m_0 k_B T}$), which at room temperature is about 15 nm for one degree of freedom of a free electron, improving the device performance by freezing out one degree of freedom for the kinetic energy. Thus, quantum confinement has the effect of reducing the device's temperature sensitivity. Multiple quantum wells increase the gain via the additional confinement regions; a material with N quantum wells correspondingly increases by N times the gain of a single well.

Other important factors that play a role in the design of optoelectronic devices are the injection of electrons and holes in the active region and the overlap of the active region with the electromagnetic mode. The photon emission and absorption calculations presented above provide useful guidance for designing light-emitting diodes and semiconductor lasers. The quantum well devices have superior performance in terms of higher gains that lower the current threshold for lasing and a narrower bandwidth, which narrows the line width of the laser line.

Further dimensional confinement such as quantum wires and quantum dots will, in principle, improve performance by creating a higher gain with a narrower profile and further reduced temperature dependence. However, in practice, the fabrication of quantum wires has proven to be especially difficult. There has been serious device research on designing and fabricating quantum dot lasers. They can be grown using epitaxial methods by mismatching the lattice constants between the two materials. The induced strain on a two-dimensional layer as a few monolayers grow epitaxially across the surface causes additional atoms to migrate and form nanometer-size islands. This process to grow nanoparticles is called Stranski–Krastanow growth. The nanoparticle size fluctuations provide additional inhomogeneous broadening mechanisms that degrade potential optoelectronic device performance. Quantum dots have been used to design lasers, light-emitting diodes, and photovoltaic devices. The narrow spectral emission, reduced temperature dependence, controlled wavelength, and high-electronic DOS are contributing factors that promote quantum dots applications.

11.2.5   *Quantum cascade lasers*

The discussion of single and multiple coupled quantum wells in Chapter 3 included a description of several classes of semiconductor lasers, including quantum well lasers and double heterostructure lasers. Quantum confinement effects are used to adjust the emission wavelength over a small range by engineering the width of the quantum well. There is another relatively new class of lasers called quantum cascade lasers (QCLs) introduced in Chapter 3 that can be designed to emit photons over a wide range of infrared to near terahertz wavelengths. When the barrier layer is thinned, the electronic wave functions extend through the barrier and into the neighboring wells. In other words, the structures form mini-bands within the electronic conduction band. The photon emission occurs via electron transitions from one mini-subband state to another. The electronic transitions occur entirely in the conduction band; on one level such transitions are called intraband transitions or intersubband transitions.

The operational characteristic of the QCL is a spatially repeated unit of multiple quantum wells, and each layer of the unit is designed to optimize each function. The main function is electron injection into an excited, conduction-band electronic state $|3\rangle$. The wave function of state $|1\rangle$ is repeated at the front and back ends of a single unit. The electron transitions to state $|3\rangle$, and designing the wave function overlaps with state $|2\rangle$. Photon emission is enabled between an excited state and a lower energy state (the shaded region in Figure 11.6(a)). Finally, the electron relaxes to the ground
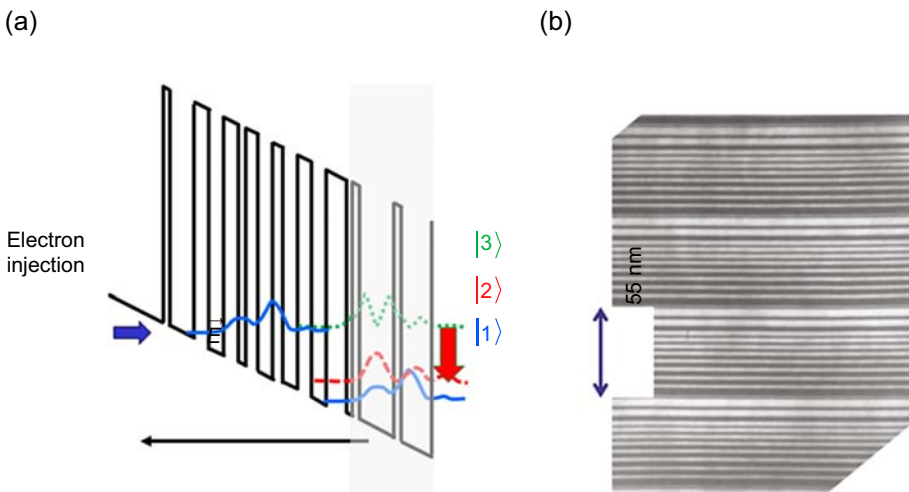
Figure 11.6 Quantum cascade laser. (a) The conduction band with three states and the highest energy state forming the widest mini-band is the injector band; the shaded block is the active region where the lasing transition occurs between states $|3\rangle$ and $|2\rangle$. Relaxation to state $|1\rangle$ provides a pathway to the next stage. (b) An scanning electron microscope picture of several QCL stages.

state $|1\rangle$, and due to the applied field the electron is injected into the next unit to repeat the emission process. Figure 11.6(a) illustrates the essential features of a QCL; a conduction band with several quantum wells and barrier height is chosen by alloying concentrations. A typical barrier height would be on the order of 0.5 eV. The design requirements of the three states result in a complex set of quantum well and quantum barrier widths, as shown in the figure. The tilt of the wells is the result of an applied field to promote wave function overlap and electron injection. The SEM picture of a QCL in Figure 11.6(b) shows the multiple-layered structure and repetitions of the basic unit.

The device consists of several identical stages; the number of repetitions of the basic unit is variable between 20 and 100. Hence, one injected electron can generate between 20 and 100 photons as it travels across the device. Typical QCL devices are fabricated using lattice matched GaAs/AlGaAs materials. By engineering the mini-subband energies the wavelength operation can be tuned from 2 to 100 μm. However, each QCL device operates over a narrow wavelength band. QCL designs have wavelengths that are not accessible using interband electronic transitions, and they can deliver several watts of peak power or continuous wave output power.

11.3   Photonic crystal phenomena

A photonic crystal is a periodic lattice whose constituents are materials with different dielectric constants. Due to the periodicity photonic crystals have a close analogy to crystalline electronic materials. However, the photonic lattice alters the physical properties of light, i.e., photonic dispersion and scattering, instead of modifying the electronic states. Therefore, designing a photonic crystal is conceptually similar to engineering the physical properties of electrons. The control over photonic properties offered by photonic crystal fabrication includes the DOS, spatial and temporal dispersion, and mode symmetries.

A simple example of a photonic crystal is a structure fabricated using multiple layers of two dielectric materials to impose periodicity in one dimension. Although thin film fabrication has been available for more than a century, the recent research has proven that there are always new properties that are uncovered when studying it from a different perspective. One-dimensional photonic crystals can be fabricated over large area planar substrates to make conventional optical devices, such as optical filters, beam splitters, and antireflection optical devices. More unconventional devices include phase-matching waves for frequency conversion, slow light for optical delays, and increased absorption or scattering using nonlinear processes.

Large area, one-dimensional photonic crystal devices can be fabricated by using physical and chemical thin film deposition techniques. However, fabricating thick, volume photonic crystals with periodicity in two- and three-dimensions remains a serious barrier to transitioning the technology into practical photonic devices. The explosion of research activities in the field can be traced to Eli Yablonovitch's 1987 publication that described how a three-dimensional periodic dielectric structure could modify the photonic DOS, which as discussed above will control light−matter

interactions, such as spontaneous and stimulated emission. The photonic DOS is the electromagnetic analogue to the electronic DOS discussed above. Edward Purcell is credited with the notion that the photonic DOS is not an immutable property of space but can be altered by controlling the environment about an atom or molecule. In a homogeneous, three-dimensional space the DOS is a smooth and monotonic growing function of energy or wave number. However, in periodic dielectric materials the DOS is highly structured, and band gaps in the spectrum appear where there are no electromagnetic modes available. It was the region of the band gaps that Yablonovitch exploited to suppress spontaneous emission in lasers.

The photonic DOS in D dimensions can be extracted from Eqn (11.14). In a homogeneous medium or composite medium in the long-wavelength limit, the dispersion relation is described by a linear function $\omega(\vec{k}) = vk$, where v is the phase velocity of the wave. In homogenized materials, the photonic DOS is a smooth function given by

$$\rho(\omega) = \Gamma(D)(\omega/v)^{D-1}. \tag{11.66}$$

The mathematical treatment of photonic crystals is based, of course, on Maxwell's equations. The coefficient is $\Gamma(D)$ extracted from Eqn (11.14). There are many computational tools available to solve Maxwell's equations, including transfer matrix methods (TMMs), finite difference methods, and plane-wave band structure methods. It is instructive to introduce the plane-wave method here in the context of electromagnetism to gain a deeper insight into the methodologies.

The optical properties of photonic crystals can employ the plane-wave method to determine the photonic band structure; the conceptual foundation of the calculations is similar to previous discussions of electronic band structures. However, in photonic crystals the properties of waves are guided by Maxwell's equations, presented in Chapter 2. A vector wave equation is derived from Maxwell's equations and can be expressed in terms of magnetic or electric fields. Consider materials that have isotropic dielectric functions. For periodic dielectric permittivity the wave equation is

$$\vec{\nabla} \times \left(\vec{\nabla} \times \vec{E}(\vec{r}, \omega) \right) = -\frac{\omega^2}{c^2} \varepsilon(\vec{r}, \omega) \vec{E}(\vec{r}, \omega), \tag{11.67}$$

for the electric field; the magnetic field satisfies the equation

$$\vec{\nabla} \times \left(\eta(\vec{r}, \omega) \vec{\nabla} \times \vec{H}(\vec{r}, \omega) \right) = -\frac{\omega^2}{c^2} \vec{H}(\vec{r}, \omega), \tag{11.68}$$

where $\eta(\vec{r}, \omega) = 1/\varepsilon(\vec{r}, \omega)$. The periodicity of the dielectric function is imposed by the relation $\varepsilon(\vec{r}, \omega) = \varepsilon(\vec{r} + \vec{a}, \omega)$ with $\vec{a}$ a lattice translation vector. We note that the above formalism can be expanded to incorporate a periodic magnetic permeability function. Magnetic effects are absent at optical frequencies in homogeneous systems, and the requirements to make a single layer homogenized metamaterial with magnetic properties at optical frequencies is a daunting challenge. Making multiple layers of such material is beyond the present fabrication capabilities.

The electromagnetic field is expanded in a set of plane waves in the format specified by Bloch's theorem, discussed in the context of electronic problems in Supplement B of Chapter 4:

$$\vec{E}(\vec{r}, \omega) = e^{i\vec{k}\cdot\vec{r}} \sum_{\vec{G}} \vec{E}(\vec{G}) e^{i\vec{G}\cdot\vec{r}} \text{ and}$$

$$\vec{H}(\vec{r}, \omega) = e^{i\vec{k}\cdot\vec{r}} \sum_{\vec{G}} \vec{H}(\vec{G}) e^{i\vec{G}\cdot\vec{r}}.$$

(11.69)

The wave vector $\vec{k}$ is the Bloch wave vector and is restricted to the first Brillouin zone. The summation extends over the reciprocal lattice wave vectors. The field amplitudes are implicitly dependent on the Bloch wave vector. The wave equations are transformed to

$$\left(\vec{G} + \vec{k}\right) \times \left(\left(\vec{G} + \vec{k}\right) \times \vec{E}(\vec{G})\right) = \frac{\omega^2}{c^2} \sum_{\vec{G}'} \varepsilon\left(\vec{G} - \vec{G}'\right) \vec{E}\left(\vec{G}'\right),$$

(11.70)

$$\left(\vec{G} + \vec{k}\right) \times \left(\sum_{\vec{G}'} \eta\left(\vec{G} - \vec{G}'\right)\left(\vec{G}' + \vec{k}\right) \times \vec{H}\left(\vec{G}'\right)\right) = \frac{\omega^2}{c^2} \vec{H}(\vec{G}).$$

(11.71)

The material functions $\varepsilon(\vec{G} - \vec{G}')$ are Fourier transforms of the real-space functions. Equations (11.70) and (11.71) are vector equations that are solved to determine the photonic band structure of photonic crystals in any dimension; they are called the E-method (E for Electric field) and H-method (H for the magnetic field), respectively. The matrix equations are solved by determining the eigenvalues represented by the variable ω^2/c^2.

Truncating the summations introduces error in the band structure that can be significant. The convergence of the bands depends on several factors, including the frequency of the band (i.e., lower bands converge faster than higher bands), the ratio of the dielectric constants of the constituent materials, and the geometry of the photonic crystal. The difference in convergence for the same lattice using the E- or H-method in three dimensions is illustrated in Figure 11.7. The frequencies of photonic bands are calculated at the M-point of the Brillouin zone for the simple cubic lattice. The lattice is composed of air-spheres with dielectric constant 1 embedded in a host material with dielectric constant 13. The scaled lattice constant is unity, and the spheres having a scaled radius of 0.495 are nearly touching. The number of plane waves used is successively increased to $N > 1000$, which requires determining the eigenvalues of a matrix with more than one million elements. In Figure 11.7, the H-method converges more slowly than the E-method for this case. One should not conclude from this that the E-method is always preferred over the H-method; other numerical examples show cases where the H-method converges more rapidly than the E-method. Asymptotically, the two methods converge to the same values.

Air hole, *r* = 0.495, lowest energies at M point
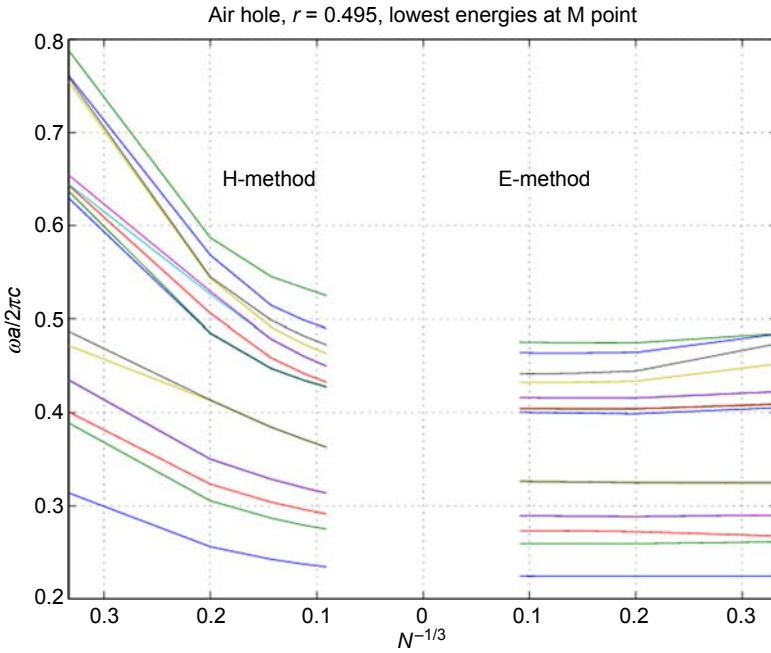
Figure 11.7 Convergence of the two forms of the vector wave equation.
From H. S. Sozuer et al. Phys. Rev. B (1992).

11.3.1   Two-dimensional photonic crystals

The wave equations can be simplified to scalar equations in one and two dimensions. Specifically in two dimensions when the dielectric function is a periodic function over the entire (x, y) plane, the scalar equation is obtained for the electric field polarized along the z-axis $\vec{E} = E(x,y)\hat{z}$ using Eqn (11.69) the wave equation as a function of the Fourier amplitudes is

$$\left|\vec{G}+\vec{k}\right|^2 E(\vec{G}) = \frac{\omega^2}{c^2}\sum_{\vec{G}'}\varepsilon\left(\vec{G}-\vec{G}'\right)E\left(\vec{G}'\right). \tag{11.72}$$

In the same way, the scalar equation for the magnetic field polarized along the z-axis $\vec{H} = H(x,y)\hat{z}$ is written as a function of its Fourier amplitudes in the algebraic form

$$\left(\sum_{\vec{G}'}\eta\left(\vec{G}-\vec{G}'\right)\left(\vec{G}+\vec{k}\right)\cdot\left(\vec{G}'+\vec{k}\right)H\left(\vec{G}'\right)\right) = \frac{\omega^2}{c^2}H(\vec{G}). \tag{11.73}$$
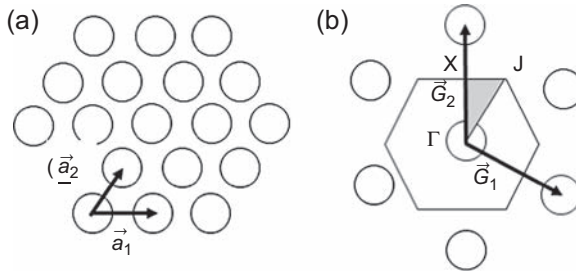
Figure 11.8 Triangular lattice of circular holes and the corresponding reciprocal lattice. (a) The triangular lattice with lattice points at the center of the circles; the real-space basis vectors are $(\vec{a}_1, \vec{a}_2)$. (b) The reciprocal lattice with basis vectors are $(\vec{G}_1, \vec{G}_2)$ and the hexagon-shaped first Brillouin zone. Three high-symmetry points (Γ, X, J) of the Brillouin zone are shown.

Numerical codes in Matlab for solving Eqns (11.72) and (11.73) are listed and described in Supplement A. The triangular lattice is illustrated in Figure 11.8. Figure 11.8(a) is a cross-section of the photonic crystal perpendicular to the axis of the rods, which is fabricated with dielectric circular cylinders embedded in a dielectric of the host medium. The basis vectors for the triangular lattice are labeled as $(\vec{a}_1, \vec{a}_2)$, and they are not orthogonal, as discussed in Chapter 4. The reciprocal lattice in Figure 11.8(b) has the Γ-point and the six nearest-neighbor reciprocal lattice points. The first Brillouin zone construction around the Γ-point has two high-symmetry points on the surface of the Brillouin zone labeled (X, J). The shaded triangle is the reduced area of the Brillouin that is repeated throughout the rest of the area using operations of the group (rotations, mirror, etc.)—elements that leave the lattice invariant.

Figure 11.9 shows the photonic band structure for the lowest bands of a triangular lattice with air-cylinders ($n = 1$) embedded in a dielectric with relative dielectric
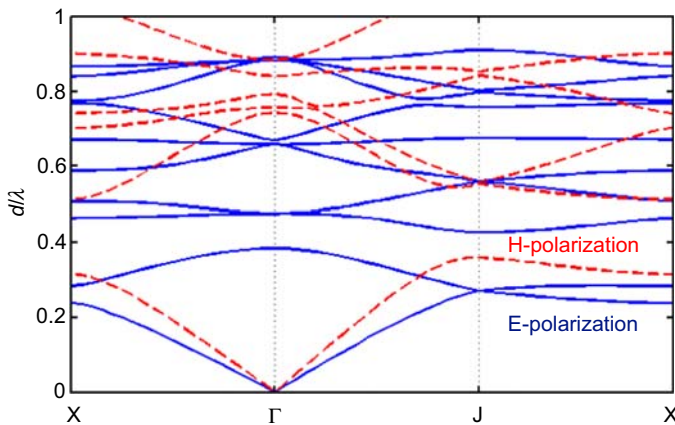
Figure 11.9 Triangular lattice band structure for air-rods ($n = 1$) embedded in a high-dielectric material ($n = \sqrt{13}$). The rod radius is 0.45 in units scaled relative to the lattice constant.

constant $\varepsilon_r = n^2 = 13$. The cylinders are nearly touching with a radius $r = 0.45\, d$, where d is the lattice constant. The E-polarization is the solution of Eqn (11.72) with the E-field parallel to the rods, and the H-polarization is the solution of Eqn (11.73). There is a band gap between the second and third bands for the E-polarization and a large gap between the first and second bands for the H-polarization. The band structure displays a complete band gap around $\frac{d}{\lambda} = 0.4$. Nearing the γ-point the wavelength approaches zero where the effective medium theory is valid. The sample is birefringent in this limit.

Full photonic band gaps in all directions are required to suppress spontaneous emission. Two-dimensional structures have propagating modes along the symmetry axis and therefore do not possess full photonic band gaps; to have that property three-dimensional photonic crystal structures are required. The difficulties in fabricating photonic crystals are discussed below. Despite the shortcomings of two-dimensional photonic crystals, they have many of the physical properties that are found in their three-dimensional cousins.

11.3.2   Mode symmetry

The missing element from the previous discussion is the symmetry of the modes corresponding to each band. The modes have been classified as coupled and uncoupled modes, according to whether they couple to plane-waves incident on the crystal surface. Put another way, the modes are symmetric or antisymmetric when an applicable mirror transformation is applied to its eigenfunction.

Photonic band structures for both polarizations in a two-dimensional triangular lattice are plotted in Figure 11.10; the lowest bands are shown. The symmetry of the bands is labeled by a letter A when the field amplitude is symmetric with respect to translation and by a letter B when the field amplitude is antisymmetric for that translation operation and therefore represents an uncoupled mode. The shaded region for the E-polarized band diagram is a region of interest. Light propagating along the ΓJ direction encounters a band of frequencies where coupling is forbidden by symmetry. Correspondingly, the experimental transmission spectra on the right in Figure 11.10 have a dip. The result could be misconstrued as a band gap while band dispersion calculations reveal that it is due to the antisymmetry of the mode. The experiment uses a sample with a triangular array of circular air-rods fabricated in a block of material with a relative dielectric constant of 2.6. The lattice constant is 1.17 μm, and the radius of the rods is 0.5 μm. The transmission spectra cover a range from 1.4 to 5.0 μm. Other dips seen in the transmission spectra for both polarizations correspond to band gaps identified in the band structure.

Uncoupled modes are ubiquitous and can be used as a transmission filter in cases where the refractive index contrast is too small to form a complete band gap. Due to their lack of coupling to external modes, it was suggested that they could form a good cavity inside a photonic crystal. By adding a gain medium to the photonic crystal the cavity can support a lasing mode. Experiments were performed based on a photonic crystal's uncoupled mode behavior to confirm this unusual cavity characteristic.
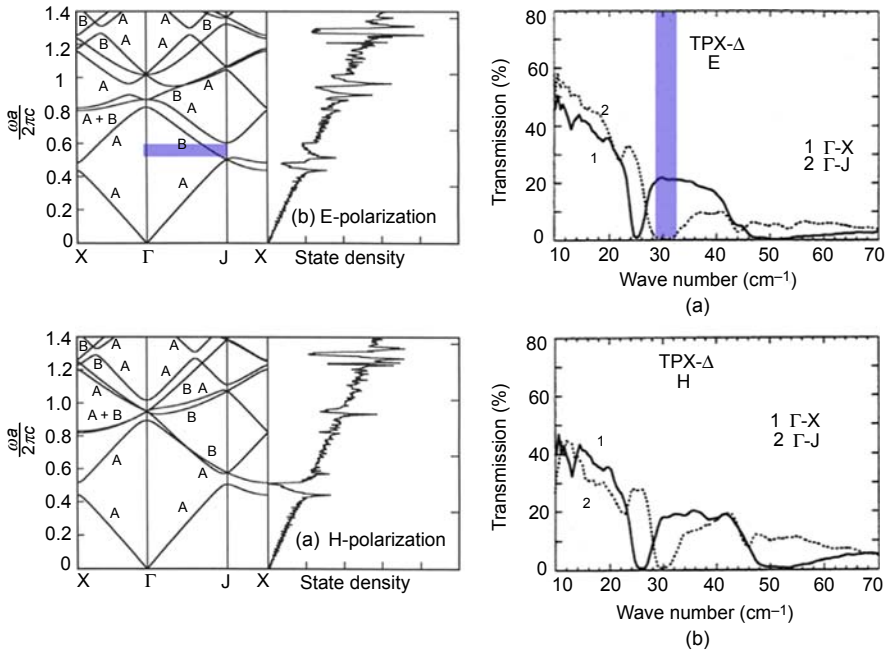
Figure 11.10 Two-dimensional band structures compared with experimental results. (Left) Theoretical calculations. *(Copyright (1994) The Japan Society of Applied Physics.)* (Right) Experimental results. *(From Phy. Rev. B, 55, 10443 (1997).)*

11.3.3 Photonic crystal fibers

Photonic crystal fibers are an independent technology that was pioneered by Phillip Russell. It has the form of a finite, two-dimensional photonic crystal, usually with a defect in the center. There are several variations of photonic crystal fibers that are used in applications. Figure 11.11 classifies photonic crystal fibers into solid (a) and
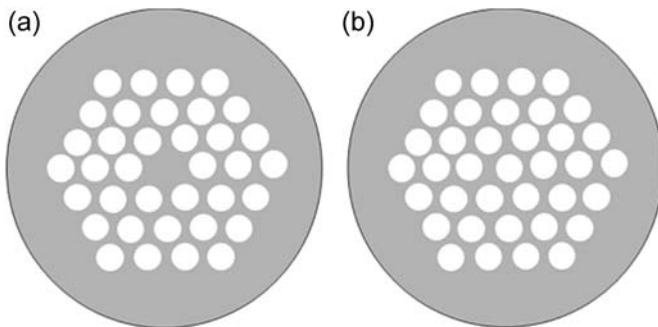
Figure 11.11 Two broad classifications of photonic crystal fibers. (a) Solid core fiber cross-section. (b) Hollow core fiber cross-section.

hollow (b) core types. Under these two types, there are many different types of photonic crystal fibers. The lattice symmetry, pitch, and hole size can be changed to suit the design of a specific fiber type. The solid core class of fibers has a high-index solid core with a lower effective index surrounding medium. The hollow core fibers have a lower index than the surrounding medium, and confinement to the center core hole is made possible by scattering and interference, combining to form an evanescent field in the cladding region.

We discuss here just a few of the many types of specialty photonic crystal fibers available. Endlessly single mode fibers are uniquely suited to applications where widely different wavelengths are used. In common core/cladding fibers, the single mode regime is defined by the V parameter, which is the scaled frequency, i.e., $V = 2\pi a \sqrt{n_{\mathrm{core}}^2 - n_{\mathrm{clad}}^2}/\lambda$, where a is the core radius, and the indices of the core and cladding are n_{core} and n_{clad}. For $V < 2.4048$ ordinary fiber has a single mode, and this is desirable to avoid modal pulse dispersion effects in the multimode regime. In forever single mode fibers, the mode is designed to change the effective index, so that the indices are replaced by effective wavelength-dependent functions $n_{\mathrm{core}}(\lambda)$ and $n_{\mathrm{clad}}(\lambda)$, and the core radius is replaced by the hole pitch Λ. The effective index changes with wavelength to keep the effective V parameter for the endless single mode fiber below a threshold value. To maintain a single mode photonic crystal fiber with a single central hole missing, one needs to keep the ratio of the hole diameter to the hole pitch small, i.e., $d/\Lambda < 0.45$. For fibers with more missing holes the ratio is smaller, e.g., for seven missing holes $d/\Lambda < 0.15$. An example of a forever single mode fiber is shown in Figure 11.12(a). Note that the holes are small as compared to the lattice pitch so that the fill factor is small for this fiber; the fiber designs with large mode areas can suppress nonlinear effects when they are undesirable. Endless single mode fibers are used to align several laser wavelengths before insertion into crystals for nonlinear parametric frequency conversion.

The spiderweb fiber shown in Figure 11.12(b) has a small core that concentrates light and a very high-index contrast with the surrounding medium, which increases the local light intensity. This fiber is designed to enhance nonlinear effects that include stimulated Raman scattering and self-phase modulation. Applications include super-continuum generation where the frequency bandwidth is more than doubled (i.e., an octave or
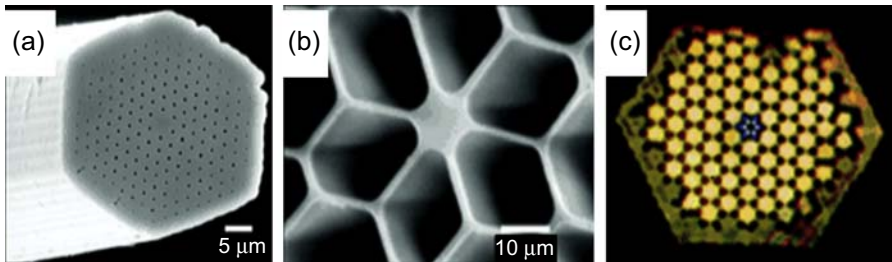
Figure 11.12 Three examples of photonic crystal fibers. (a) Endless single mode fiber. (b) Spiderweb fiber. (c) Hollow core fiber.
With permission from source: P. Russell, Science (2003).

more of light frequencies) due to nonlinear propagation in the fiber. Super-continuum generation has a bright output. The applications include spectroscopy and optical coherence tomography. The fibers can be designed to enhance selected nonlinear effects, such as Raman signal generation to amplify signals at the Raman shifted frequency, four-wave mixing signals, and nonlinear parametric processes.

The hollow core fiber confines a band of light wavelengths within the air-core region, such as seen in Figure 11.12(c). In the figure blue wavelengths are confined in the core while other wavelengths leak throughout the cladding region. These fibers can transport intense laser fields that might otherwise damage a fiber by confining most of the energy in the hollow core, and they can serve as sensors that have a long path length to determine the presence of trace amounts of gases or biomolecules contained in aerosols flowing through the hollow core.

11.3.4 Three-dimensional photonic crystals

Since Yablonovitch's seminal article, published in 1987 there have been many attempts to fabricate three-dimensional photonic crystals that have *full* photonic band gaps in visible or near-infrared wavelength regimes. A full photonic band gap structure has no propagating mode in any direction. Researchers have applied different design approaches, and yet it remains a difficult challenge. Scaling down fabrication processes that work for Radio or Terahertz frequencies requires very delicate processing tools with nanometer precision. There is a spectrum of innovative technologies that have been developed to make photonic crystals with nanoscale-critical dimensions. On one end of the fabrication spectrum, so-called top-down approaches, they are fabricated using combinations of lithographic patterning materials; deposition and etching technologies are costly to purchase and maintain, making it a high barrier to access this direction of research. On the other end of the fabrication spectrum, there is the lab chemist or materials scientist who is harnessing self-assembly techniques to build a volume crystal using bottom-up approaches.

Band structure simulations are important in the search to identify candidates for full photonic band gaps. They identified structures that could be fabricated as candidates for experimental investigations. The appearance of photonic band gaps depends on the symmetry and the dielectric materials used. There are two rules of thumb to help determine candidate structures that may have full photonic band gaps. The first observation is that the contrast between the dielectric constants should be much larger than unity $\frac{\varepsilon_1}{\varepsilon_2} \gg 1$. Physically this means that band gaps occur in the strong scattering regime. As a rule ratios larger than 10 open full band gaps, which has driven research using semiconductor materials with air-holes sculpted into them. The second observation is the opening of a band gap occurs when the structure is mostly filled with low dielectric material; in other words, the structure would have most of its material cut away, leaving a delicate scaffolding behind.

Initially, the diamond lattice of spheres was identified to open a full photonic band gap between the second and third bands; similarly, the related face-centered cubic (FCC) lattice could also have a full band gap between the eighth and ninth bands. An example is shown in Figure 11.13(a). The high-symmetry points labeled on the

(a)

(b)

(c)
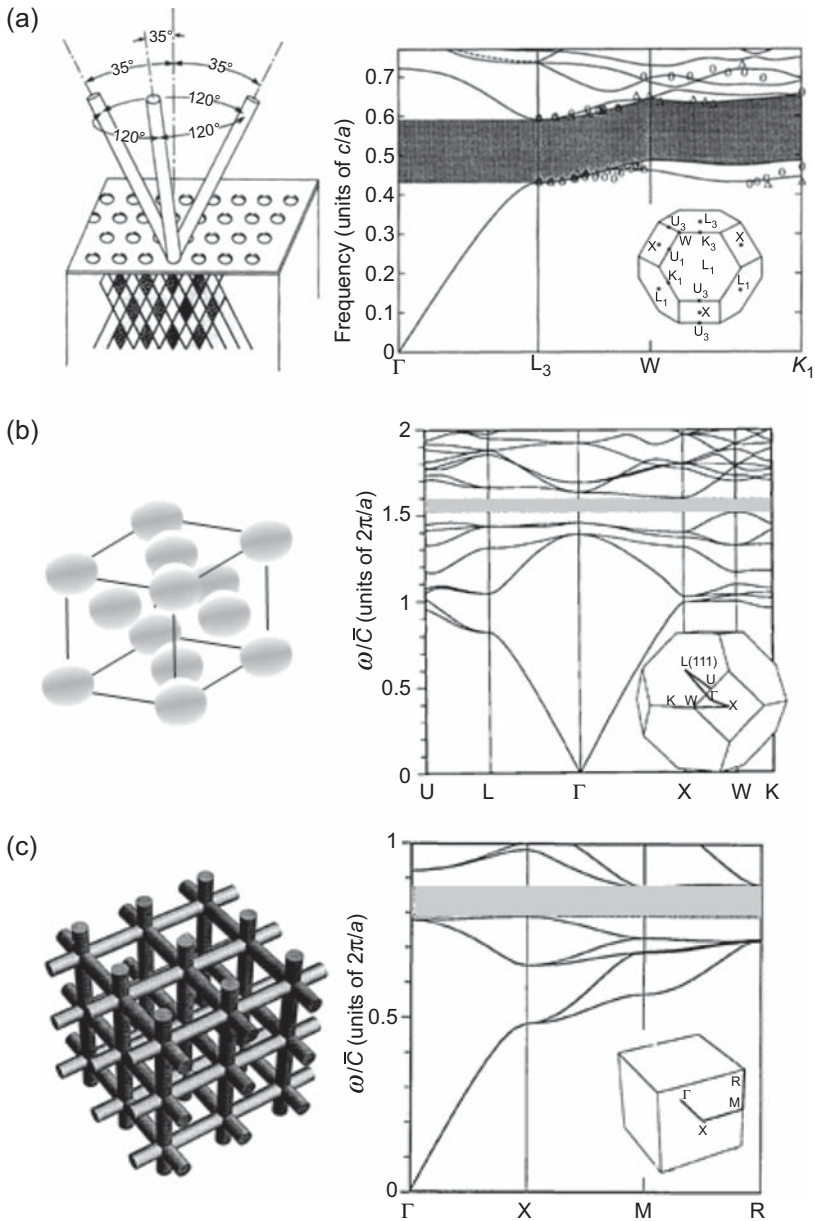
Figure 11.13 Lattice structures and the corresponding photonic band diagrams for selected three-dimensional lattices. The relative dielectric constant for the host medium in (a) is 12.3; for (b) and (c) it is 13. (a) Yablonovite $\varepsilon_h = 12.3$. (b) FCC lattice of air-spheres $\varepsilon_h = 16$; the volume fraction is 0.74 (close-packed spheres). (c) SC lattice $\varepsilon_h = 13$ of circular air-cylinders with volume fraction 0.81.

Figures reproduced with permission from APS and from OSA: (a) E. Yablonovitch et al. Phys. Rev. Lett. (1991); (b) H. S. Sozuer et al., Phys. Rev. B (1992). (c) H. S. Sozuer, J. Opt. Soc. Am. B (1993).

ordinate axis are labeled in the inset, showing the first Brillouin zone. The parameters for the band structure are found in the figure caption. A small full band gap was reported for a simple cubic lattice between the fifth and sixth bands. Since it is desirable to form a band gap between lower order bands, many of the fabricated three-dimensional photonic crystals are based on the diamond lattice symmetry. The first demonstration of a diamond symmetry lattice in the RF regime was reported by Yablonovitch's group, fabricating a structure now called Yablonovite, which can be made by drilling three holes through a surface with a two-dimensional lattice. The diagram showing how Yablonovite can be fabricated by drilling three holes from a patterned top plane is indicated on the left in Figure 11.13(b). On the right is the corresponding photonic band diagram for Yablonovite.

Full photonic band gaps have been realized in a number of structures, including scaffold structures of the ilk shown on the left in Figure 11.13(c), which is based on the simple cubic (SC) lattice. The SC lattice can have a full photonic band gap for large enough air fraction and high enough dielectric contrast.

Full photonic band gaps were also found for woodpile arrangements, layer by layer placements, and inverse opal structures. Inverse opals are fabricated using self-assembly combined with a process to infuse material in the interstices between the spheres and finally an acid etch process to dissolve and remove the spheres. The process of removing the spheres uses an acid that attaches the opal material while leaving the scaffold material around it unaffected. For instance, silica spheres are dissolved in a buffered HF solution while a semiconductor material, such as silicon or CdS, remains behind in the interstitial space.

11.3.5   Transfer matrix methods: application to one dimension

Band structure calculations performed by plane wave generally provide limited information about the wave propagation applicable to infinite lattices and real dielectric functions. Other numerical techniques such as finite difference methods and finite element methods calculate properties of finite systems, including complex dielectric functions, material anisotropy, and finite lattice size effects. A description of these calculations is beyond the scope of this book and the interested reader can follow references at the end of this chapter to find more details on numerical methods. This section ends with a description of a one-dimensional transfer matrix technique. It provides details that can be compared with experiments or used to design multilayer samples for specific applications.

Optical transmission and reflection spectra calculations are obtained using the TMM. They are numerically efficient, allow complex dielectric coefficients to be studied, and the samples are finite in one direction. The lateral size of the structure is infinite for TMM. The technique can be extended to two- and three-dimensional photonic crystals using finite difference or plane-wave decompositions; the results provide direct correspondence with experimental data. Here we refer to the one-dimensional TMM to illustrate its usefulness. Supplement B has a transfer matrix program that the reader can use to explore the properties of samples with finite thicknesses.

The transmission and reflection of the light wave though isotropic dielectric multilayers is calculated using 2×2 matrices called the transfer matrix. A multilayer
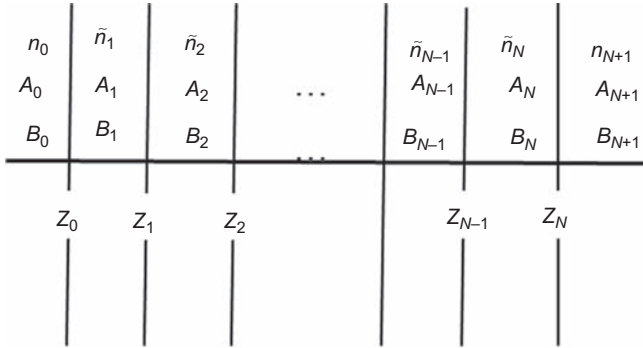
Figure 11.14 Multilayered dielectric structure to illustrate the one-dimensional TMM.

sample is illustrated in Figure 11.14. The indices are shown in the top row, and the next two rows indicate the amplitudes of the forward and backward waves, respectively. The position of each interface is denoted by z_l, where $l = 0$ to N.

In Figure 11.14 $\tilde{n}_1$ is the index of the first dielectric layer and $\tilde{n}_2$ is the index of the second dielectric layer, etc.; all functions may be complex, i.e., $\tilde{n}_a = (n_a - i\kappa_a)$ $a = 1, 2, \ldots N$. n_0 is the index of the incident medium, and n_{N+1} is the index of the substrate.

The incident electric field of a plane wave polarized in the y-direction (the polarization perpendicular to the plane in Figure 11.14 and is also called s-polarization) is expressed as a function of z as

$$\vec{E}(z) = \hat{y}(E_0 \exp(i(\omega t - k_z z))). \tag{11.74}$$

Note that this definition of the plane-wave function differs from Chapter 2 by a minus sign in the exponent. In each region, the scalar component of the field is described by forward $\{A_l\}$ and backward $\{B_l\}$ wave amplitudes:

$$E(z) = \begin{cases} A_0 \exp(-ik_{0z}(z - z_0)) + B_0 \exp(ik_{0z}(z - z_0)) & z < z_0 \\ A_l \exp(-ik_{lz}(z - z_l)) + B_l \exp(ik_{lz}(z - z_l)) & z_{l-1} < z < z_l \\ A_{N+1} \exp(-ik_{0z}(z - z_N)) + B_{N+1} \exp(ik_{0z}(z - z_N)) & z_N < z \end{cases} \tag{11.75}$$

where A_l is the amplitude at the interface $z = z_l$. $k_{lz} = n_l \frac{\omega}{c} \cos \theta_l$, θ_l is the incident ray angle. Using the continuity boundary condition for the tangential component of the electric and magnetic field at each interface, the relation between amplitude from the previous medium is connected to the amplitudes in the following medium:

$$\begin{pmatrix} A_0 \\ B_0 \end{pmatrix} = M_0^{-1} M_1 \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \tag{11.76}$$

$$\begin{pmatrix} A_l \\ B_l \end{pmatrix} = P_l M_l^{-1} M_{l+1} \begin{pmatrix} A_{l+1} \\ B_{l+1} \end{pmatrix} \tag{11.77}$$

where $l = 1, 2, 3...N$.

$$\begin{pmatrix} A_0 \\ B_0 \end{pmatrix} = M_0^{-1} \left[M_l P_l M_l^{-1} \right]^N M_{l+1} \begin{pmatrix} A_{l+1} \\ B_{l+1} \end{pmatrix}. \tag{11.78}$$

The result of taking a product of all 2×2 matrices is the matrix

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = M_0^{-1} \left[M_l P_l M_l^{-1} \right]^N M_{l+1}. \tag{11.79}$$

For s-polarization the matrix for the lth layer is

$$M_{sl} = \begin{pmatrix} 1 & 1 \\ k_{lz} & -k_{lz} \end{pmatrix}. \tag{11.80}$$

For p-polarization it is

$$M_{pl} = \begin{pmatrix} 1 & 1 \\ \dfrac{k_{lz}}{\tilde{n}_l^2} & -\dfrac{k_{lz}}{\tilde{n}_l^2} \end{pmatrix}. \tag{11.81}$$

Defining d_l as the layer thickness, the phase change of the forward or backward waves across the medium is described by the matrix

$$P_l = \begin{pmatrix} \exp(ik_{lz}d_l) & 0 \\ 0 & \exp(-ik_{lz}d_l) \end{pmatrix}. \tag{11.82}$$

The matrix relationship between the field amplitudes and the input surface and at the output surface is

$$\begin{pmatrix} A_0 \\ B_0 \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} A_{N+1} \\ B_{N+1} \end{pmatrix}. \tag{11.83}$$

Setting backward wave amplitude from the substrate to zero, i.e., $B_{N+1} = 0$, the complex reflection amplitude is

$$r = \frac{B_0}{A_0} = \frac{M_{21}}{M_{11}}, \tag{11.84}$$

and the complex transmission amplitude is

$$t = \frac{A_{N+1}}{A_0} = \frac{1}{M_{11}}. \tag{11.85}$$

The reflection and transmission coefficients at oblique angles of incidence are

$$R = |r|^2 \text{ and } T = \frac{n_{N+1}\cos\theta_{N+1}}{n_0\cos\theta_0}|t|^2. \tag{11.86}$$

Transmission spectra for 10 pairs of dielectric films are shown in Figure 11.15. The indices are modest with $n_1 = 1.5$ and $n_2 = 2$, and the thicknesses are chosen as quarter-wave thickness, i.e., $d_a = \lambda/(4n_a)$ where $a = 1, 2$ is the material layer label. The superstrate and substrate are assumed to be air. At normal incidence, the transmission has a wide band gap centered on 750 nm; a change of the angle of incidence to 45° shifts the band gap center by a half width of the band gap.

Substituting a metal for one of the dielectric materials yields an interesting system that is called a metallodielectric, which has surprising and unique properties. Two results are illustrated in Figure 11.16. In Figure 11.16(a) the transmittance of two material configurations is illustrated, called the Periodic and Symmetric stacks. The Periodic stack has six periods of an Ag metal film and a fictitious dielectric with constant index 4. The Symmetric stack has the same amount of material but with half the thickness of the last dielectric layer moved in front of the first metal layer. The difference in the transmittance between the two cases is significant with more than a factor of 2 increase for the Symmetric stack.

Figure 11.16(b), a silver dielectric Symmetric stack at normal incidence and at a 45° angle of incidence, illustrates another stark difference between dielectric and
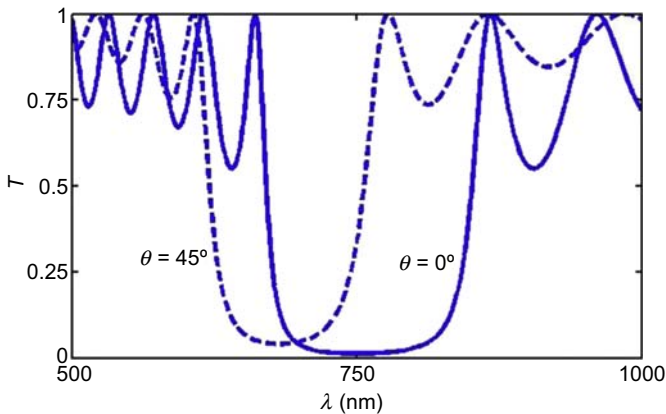
Figure 11.15 p-polarization transmission spectra at two angles of incidence for 10 periods of two dielectric films with indices $n_1 = 1.5$ and $n_2 = 2$. The layers are a quarter-wave thick, $d_1 = 125$ nm and $d_2 = 93.75$ nm at a wavelength of 750 nm. The superstrate and substrate have index 1.
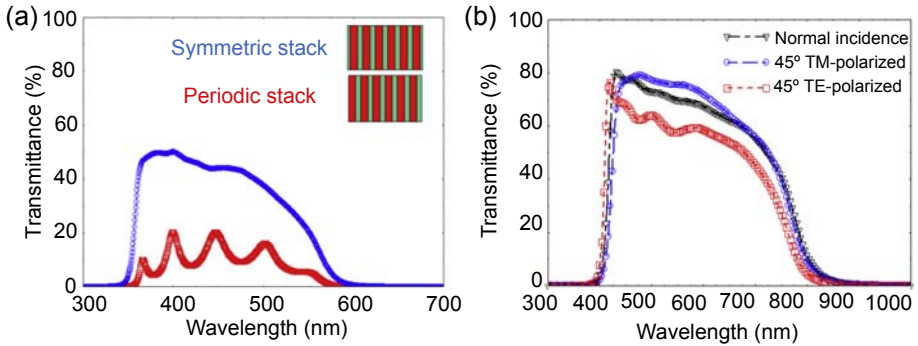
Figure 11.16 (a) Normal incidence transmittance of two metallodielectric stacks with six silver metal layers, each 32 nm thick. The dielectric material has refractive index 4 and layer thickness 22 nm. The Periodic stack has six periods, and the Symmetric stack has a half dielectric layer removed from one end and placed on the opposite end. (b) Angle dependence of the transmittance for a Symmetric dielectric stack using four periods; both polarizations are shown. There are 10 layers; the materials and thicknesses are as follows: TiO_2 (25 nm) (Ag (16 nm)/ TiO_2 (50 nm) $\times$ 4)/Ag (16 nm)/TiO_2 (25 nm).
Figure (a) with permission from OSA, M. Scalora et al. Opt. Express (2007); Figure (b) from M. Scalora.

metallodielectric stacks by comparison with Figure 11.15. The transmittance of the metallodielectric structure shows a modest shift, while the dielectric structure has a shift of the band gap center by more than 50 nm as well as a change in the width and depth of the gap.

The TMM is an efficient numerical method that can be extended in several ways to extract more information about multilayer structures. The fields throughout the material can be calculated, and the complex transmission coefficient phase can determine an effective phase and group velocity of a wave. By a plane-wave expansion a beam of arbitrary shape can be studied, even including evanescent waves in the description of the optical beams to examine propagation of subwavelength diameter beams leading to super-resolution effects. Other extensions include using more than two materials, fabrication with arbitrary layer thicknesses, and anisotropic component materials in each layer.

The TMM enables system designs to improve nonlinear interactions by exploiting local field confinement and enhancement at transmission resonances. This analysis has aided the study of enhanced second- and third-harmonic generation where quantitative and accurate simulations are required. A number of publications on this topic can be found in the references at the end of this chapter.

The TMM generalized for higher dimensions is a powerful method for studying photonic crystal transmission properties. On the one hand, structure calculations give important information on the dispersion and mode symmetry properties of an infinite photonic crystal; on the other hand, TMM yields complementary information on the transmission amplitude and phase. Together they are applied to identify and quantify uncoupled mode transmissivity. A thorough discussion of the topic of mode symmetry can be found in Sakoda's book and in selected references at the end of this chapter.

11.4   Metasurfaces: nanoantennas

The word "metasurface" is a relatively new name coined to focus attention on the structured thin film characteristics on the surface. Like the topic of metamaterials in Chapter 8, metasurfaces transcend the usual properties of smooth, flat surfaces to produce new physical effects using structures that are smaller than a wavelength. Similar to the effective medium concept, the surface should have an effective optical characteristic. Metasurfaces are fabricated with nanoscale elements that are designed to transform the incident optical beam into an output beam with desired characteristics. Ideally, surface relief features are subwavelength in size to treat the structure as a homogeneous film but with spatially variant properties. Because they are a surface modification, metasurfaces are much simpler to fabricate than bulk metamaterials, and therefore they have a better chance of making a technological impact, especially at shorter wavelengths. Metamaterials, on the other hand, are difficult to fabricate as bulk materials.

A plasmonic nanostructure placed on a surface has properties of an antenna, which is why one calls it a nanoantenna. As the carriers in the structure are excited and oscillated in an external field, electromagnetic waves are reradiated into the surrounding environment with a characteristic amplitude and phase. By carefully designing a set of different nanoantennas, they can be placed in different positions across a surface to form phase array antennas.

To illustrate how properties of nanoantennas can be useful in designing metasurfaces, we consider the simple case of a single prolate spheroid, i.e., a nanorod. The local field near a prolate spheroid is derived in the electrostatic regime, that is, in the limit where the antenna size is small as compared to the wavelength, in Supplement C.

Chapter 7 introduced the concept of the induced dipole moment with components defined by

$$p_i = \alpha_i E_i. \tag{11.87}$$

The polarizability tensor in principal axes coordinates is simplified to the expression

$$\alpha_i = V \frac{(\varepsilon_m - \varepsilon_h)}{(L_i \varepsilon_m + (1 - L_i)\varepsilon_h)}. \tag{11.88}$$

The shape of the nanoparticle is characterized by the variable L_i, which takes values between 0 and 1, previously discussed in Chapter 7. By keeping the wavelength constant and modifying the shape we can illustrate the amplitude and phase changes of the nanorod polarizability normalized by the volume that are ultimately observable in the scattered light from the nanorods. The polarizability for a gold nanorod for different shapes is shown in Figure 11.17. The amplitude changes by more than a factor of 10, which can be managed by changing the volume of each nanorod. The phase varies by π over the entire range, which means it falls short of controlling the phase by 2π. Another problem with this nanorod model is that the polarization is restricted to fields applied along a single axis.
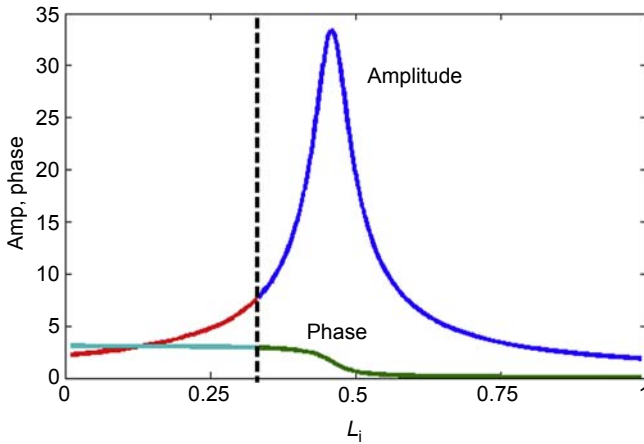
Figure 11.17 The amplitude and phase of a gold nanorod. The wavelength is 633 nm, $\varepsilon_m = 10.56 + i1.28$ and $\varepsilon_h = 9$. The vertical dotted line separates the values of L_i where the field is parallel to the rod ($L_i > 1/3$) from those that are perpendicular to it.

The group of Federico Capasso solved the phase restriction by modifying the shape of the nanorod to V-shaped antennas that incorporate two oscillating modes. Their design provided amplitude, phase, and polarization controls for many beam-forming applications. A set of V-shaped antennas was used as elements of a metasurface that together could act as phase arrays. The symmetry axis passes through the vertex; the vertex angle Δ and the arm lengths are two variables that are adjusted to find the desired operational characteristics.

Figure 11.18 illustrates the geometry of a V-shaped antenna and the applied field. The amplitude of the incident field is denoted by $\vec{E}_\parallel$; from the two arms of the antenna
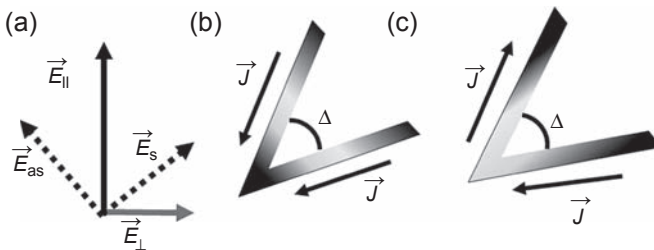
Figure 11.18 Illustration of the V-shaped antenna. The vertex angle Δ is variable, as are the lengths of the two arms. (a) An applied field $\vec{E}_\parallel$ is illustrated by a vertical solid arrow. The field is decomposed into two orthogonal components: one excites a symmetric mode $\vec{E}_s$, and the second excites an antisymmetric mode $\vec{E}_{as}$. The output scattered field of interest is the perpendicular field $\vec{E}_\perp$. (b) The symmetric mode is defined by the current density in both arms driven in the same direction about the vertex. (c) The antisymmetric mode is defined by the current density in both arms driven in the opposite directions through the vertex. High-charge density is denoted by the dark, solid coloring filling the antenna volume.
Redrawn from figure in N. Yu et al. Science (2011).

the induced currents, denoted by $\vec{J}$, are decomposed into two cases symmetric in both arms, or they are antisymmetric. The symmetric case in Figure 11.18(b) is driven by the field $\vec{E}_s$, that is, at a 45° angle with respect to the horizontal line. The V-shaped antenna is oriented so that $\vec{E}_s$ points along the symmetry axis, which bisects the vertex angle Δ. The antisymmetric case in Figure 11.18(c) is driven by the field $\vec{E}_{as}$, which is perpendicular to the symmetry axis. The currents separate the changes differently as well. The charge densities for the symmetric case are high at the vertex and at the opposite ends of the antenna, whereas the charge densities are high only at the ends of the antenna, and the vertex is charge neutral. The scattered light has an orthogonal polarization from the incident light, which makes it easier to separate from the unscattered and scattered incident light.

Numerical simulations are ultimately required to determine a set of antenna designs to construct a metasurface. Eight V-shaped antenna elements have been used to approximate a smooth phase gradient function. The set of V-shaped antennas used by Capasso's group is illustrated in Figure 11.19(a). First we note that there are two subsets of four V-shaped antennas. The second subset of four is derived from the first set by a mirror symmetry through the horizontal axis.

Each V-shaped antenna has a scattered field amplitude that can be represented for the nth element as

$$\vec{E}_{\perp}(n) = \hat{e}_{\perp}\left(A_s(n)e^{i\varphi_s(n)} + A_{as}(n)e^{i\varphi_{as}(n)}\right). \tag{11.89}$$

The amplitude and phase of the symmetric and antisymmetric scattered amplitudes combine to form the scattered beam. The scattered amplitudes of each nanoantenna are adjusted to be approximately equal; more importantly, the phases of the scattered waves are designed to have approximately an equal distribution of phase differences. Figure 11.20 shows the results of a design by Capasso's group with nearly equal phase
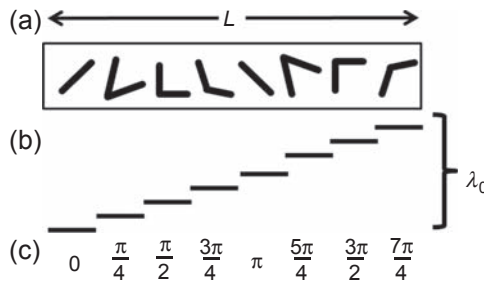
Figure 11.19 (a) A set of eight V-shaped antenna elements that are designed to span 2π phase change with nearly equal phase jumps of $\pi/4$. The period of the eight elements is the length L. (b) An illustration of the phase discontinuities across the set of elements, and the distance is equivalently the wavelength of light in vacuum λ_0. (c) A count of the phase steps; the first element is designated as the reference phase 0.
Redrawn from figure in N. Yu et al. Science (2011).
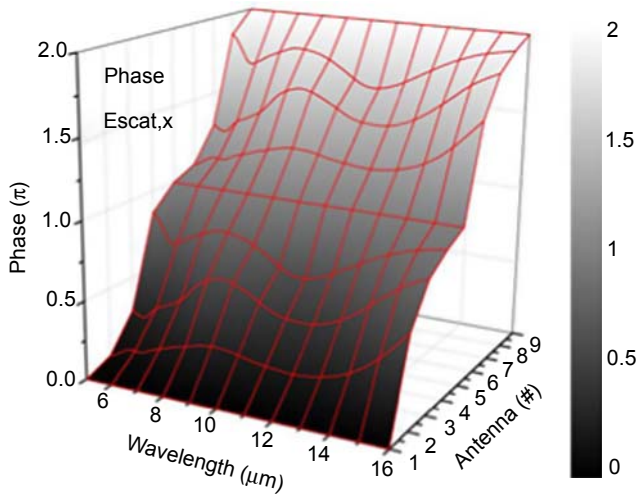
Figure 11.20 Phase difference between the eight nanoantenna elements for wavelengths ranging from 5 to 16 μm.
Figure used with permission from N. Yu et al. IEEE Journal of Selected Topics in Quantum Electronics (2013).

π/4 differences with antenna #1 used for the reference phase for the eight antennas across a wide range of wavelengths.

The basic set of nanoantennas are mosaicked on the surface of a substrate in an array to form a large phased array, as illustrated in Figure 11.21(a). The outlined set is repeated throughout the surface to form the beam deflection device. Figure 11.21(b) is a schematic of the beam interaction with the surface. The beam is incident on the metasurface from below; its polarization is denoted by an arrow on the phase lines. After the metasurface the beam separates into a diffracted part with the output polarization (also called the anomalous or extraordinary wave polarization) perpendicular to the input beam and the undiffracted part, which passes through as an ordinary wave. The angle of the diffracted beam is determined by the basic set of elements.
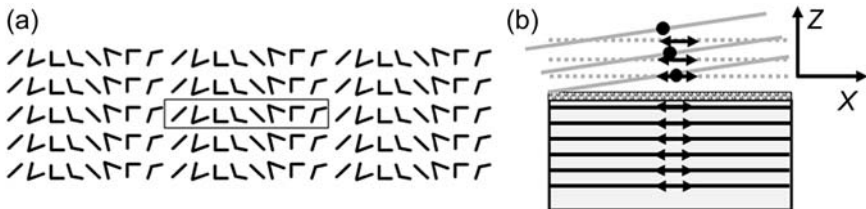
Figure 11.21 (a) The basic building block is the set of nanoantennas in the box. (b) The schematic of an incident beam interacting with the metasurface on the top that is placed on top of the substrate. For clarity, the reflected waves are not shown.
Redrawn from image in N. Yu et al. Science (2011).

On the macroscopic scale the results can be summarized by a generalization of Snell's law, referring to the coordinate axes in Figure 11.21(b) as

$$k_{x2} - k_{x1'} = \frac{\partial \Phi}{\partial x} = \frac{2\pi}{L}. \tag{11.90}$$

The last equality is the gradient from the example of a 2π phase shift over a length L, discussed above. The tangential component of the wave vector has a discontinuity imposed by the phase gradient on the surface. The more traditional results are expressed in terms of angles of incidence (θ_i) and refraction (θ_t):

$$n_2 \sin \theta_t - n_1 \sin \theta_i = \frac{\lambda_0}{2\pi} \frac{\partial \Phi}{\partial x} = \frac{\lambda_0}{L}. \tag{11.91}$$

Again, the last equality refers to the special situation previously discussed. The generalization of Snell's law can be extended to the second direction and the result in vectorial form:

$$\vec{k}_{\perp 2} - \vec{k}_{\perp 1'} = \vec{\nabla}\Phi. \tag{11.92}$$

The gradient operator and the wave vectors lie in the (x, y) plane perpendicular to the surface normal.

By rotating the center axis of the V-shaped antennas relative to the incident wave polarization, the scattered polarization is rotated out of the plane. If the symmetry axis of the V-shaped antenna has an angle β with respect to a laboratory-based axis and the incident field has an angle α with that axis, then the scattered wave polarization is at an angle $2\beta - \alpha$. In this way, for a fixed $\alpha = 0$ incident polarization and two sets of V-shaped antennas with $\beta = 0$ and $\pi/4$, light is scattered with orthogonal polarizations in the same direction. Using a neighboring pair of nanoantennas with orthogonal scattered polarizations in the same direction and a $\pi/2$ phase shift between the two sets of nanoantennas, a linear polarized beam is transformed into a transmitted circular polarized beam. A wide band circular polarizer design was fabricated and characterized by Capasso's group.

The concept of metasurfaces fabricated with distributions of nanoantennas is extended in many other ways. Using variations of the antenna shapes and distributions on surfaces, beams can be shaped into formats such as nondiffracting Bessel or Airy beams. Orbital angular momentum of the optical beams can be impressed or detected by introducing holographic-type structures on the surface, and a linear array of antenna structures cut to a metal film that scatters incident light into surface plasmon polariton modes on a metal surface. The V-shaped antenna array is distributed to radiate the incident light with a phase shift that radially changes in a circular pattern to produce a phase front that focuses light on-axis at a defined focal distance. The flat lens design in principle can have a large numerical aperture for better light collection efficiency.

11.5   The future of photonic devices

An extensive, international community of researchers is devoted to transforming interesting new nanophotonics concepts into practical devices. Integrating electronic and photonic properties is an ultimate direction of the present studies. Researchers studying photovoltaic devices have integrated plasmonic structures, quantum dots, or dye molecules in an effort to improve their energy-harvesting efficiency. New perovskite materials have made rapid progress and achieved a conversion efficiency around 20%.

An interesting, new, and as yet nascent photonic device concept is the so-called rectenna structure to transform electromagnetic energy into direct current. Rectenna is a contracted name for "rectifying antenna," which is fabricated using a thin nanometer/thick dielectric film sandwiched between two different metal films. The design asymmetry electromagnetic radiation incident on the rectenna generates a direct current (DC) from the quantum mechanical process of electron tunneling between the two metals. The energy-harvesting process is purely quantum mechanical. In particular, it does not involve interband transitions that are prevalent in photovoltaics nor does it rely on excited carrier dynamics.

A simplified rectenna geometry is illustrated in Figure 11.22(a). Incident electromagnetic radiation drives electrons in the arms of the antenna, which are fabricated using two different metals. In the region where the two metals overlap, there is a thin film insulator separating the two metals, shown as the inset in the figure; this is also called an MIM to denote its metal−insulator−metal structure. Its thickness is on the order of nanometers, which enables electrons to tunnel through the barrier. The quantum mechanical in Figure 11.22(b) shows the essential structure and function of the MIM diode, which involves the metal work functions and the insulator electron affinity. The work functions
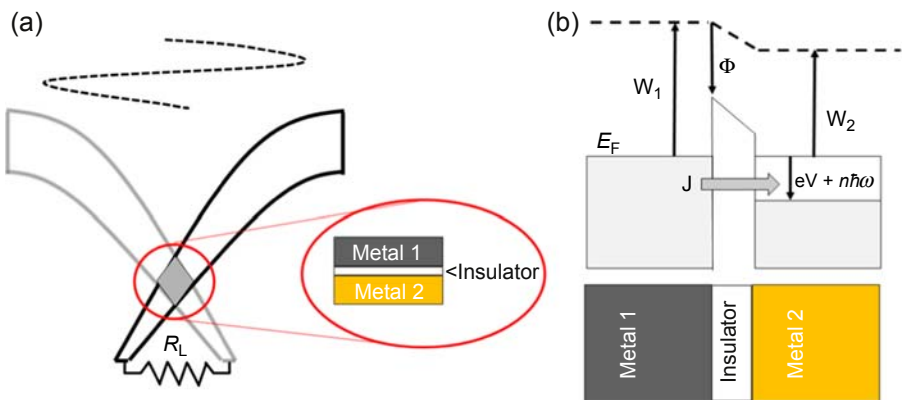
Figure 11.22 (a) Schematic of a rectenna design with the metal−insulator−metal (MIM) diode in the overlap region. The load resistor is denoted by R_L. The inset is a side view of the MIM diode. (b) A view of the MIM with essential physical parameters to determine the barrier height and shape; the metal work functions W_1 and W_2 and electron affinity of the insulator Φ.

of the two metals (W_1, W_2) are the minimum energy required to free the electron from the metal, i.e., the energy gain for the electron at the Fermi level to raise it to the vacuum energy level, shown by a dashed line in the figure. The electron affinity in solid state materials is the energy released by an electron that is captured to typically occupy an energy level at the bottom of the conduction band.

The current flow denoted as J is due to the electron tunneling from filled electronic states below the Fermi level in Metal 1 to empty states above the depressed Fermi level in Metal 2. The Fermi level depression in Metal 2 is a combination of an applied DC voltage (V) and an oscillating energy shift due to the energy of n photons ($n\hbar\omega$). The asymmetric shape of the barrier region creates an asymmetry in the tunneling current, leading to a rectified direct current flowing through the load R_L. Due to the fact that the current is not the result of a band-to-band carrier excitation, the current induced by the electromagnetic wave has a very broad band response. Ultimately, it is the scattering relaxation time of electrons in the metal that limits the response of the device. The time scale for many metals is of order femtoseconds, which means that frequencies in the infrared can be rectified. At the radio frequency end of the electromagnetic spectrum, an antenna is the best choice for coupling electromagnetic waves to a diode rectifier. The antenna strategy has been scaled to shorter wavelengths, but fabricating an antenna becomes increasingly difficult at near-infrared and optical frequencies. A nanoantenna concept would be better. Potential technology applications of rectenna designs include wide bandwidth and ultrafast photodetection. They could be studied for solar energy harvesting as an alternative approach to photovoltaics.

Advances in nanosensors will have future benefits in diagnosing medical conditions, detecting food-borne pathogens, and environmental monitoring for pollution and toxins in air or water. The sensitivity can be at unprecedented low concentrations to parts per billion. Nanoparticles and carbon nanotubes have been studied for molecular imaging and sensing and drug delivery. While fluorescent molecules lose their efficacy by photobleaching during irradiation, semiconductor nanoparticles continue to radiate. Nanoparticles have access to cells and may be tailored to target diseased cells, such as malignant growth. Gold nanoparticles are biocompatible, which has made them useful for drug agent delivery. Their small size, around 10 nm, enables them to be carried through small capillaries, and they can pass through cell membranes to reach the target cells. Their shape-dependent surface plasmon response can be tailored for photothermal therapy in wavelength regions where light penetrates deep into tissue. The uptake and accumulation of gold nanoparticles in cancerous tumors is important for subsequent laser irradiation near the specific nanoparticle absorption resonance to kill the cells by thermal ablation.

Surface enhanced Raman scattering (SERS) is a widely used nonlinear optical technique with extreme sensitivity for detecting even single molecules. The electromagnetic scattered intensity near a nanostructured metal surface, usually gold or silver, is enhanced by 10−12 orders of magnitude. Initially, surfaces were randomly roughened with results that were difficult to reproduce. Electromagnetic simulations have revealed that the field enhancement is the result of extreme local fields from cavity-type SPRs and nanoscale separation between two metal protrusions due to the presence of complex structures on the metal surface. Using this knowledge

SERS surfaces could be designed by optimizing performance using nanofabrication techniques and known high sensitivity positions.

Knowledge of the field has resulted in the top-down design and fabrication of metal surfaces with nanostructures that promote local field confinement. Local field enhancement for these nanostructures is typically in the range of $100-1000$ times the applied field. The Raman scattered intensity being proportional to the fourth power of the incident field accounts for the extreme signal sensitivity. Since the Raman shifts provide a unique molecular spectrum, they can be used to identify molecules at very low concentrations. SERS is being developed as a useful tool for the detection of food-borne disease agents, chemical explosives, biomolecules, and environmental contaminants.

These examples represent a few of the many interesting applications of nanophotonics. The choices of topics mentioned in this book may serve to elucidate the potential of nanophotonics concepts and phenomena. We predict that the speed of new innovations will continue to accelerate and that future developments will be as interesting and surprising as those that have already been discovered. The rest will be left for the reader to explore by reading the literature, discussions with colleagues, and surfing the web.

Problems

1. Use the E and H programs in Supplement A to study the band structure diagrams for different cases. Plot the transmission and reflection spectra for a range of wavelengths scaled to the lattice constant a.
   a. Keeping the cylinder radius fixed at $r = 0.45a$, where a is the lattice constant, and cylinder dielectric constant at 1, change the dielectric constant of the host medium from 1 to 25 and plot the reflection and transmission spectra. Identify the lowest band gaps in the ΓX and ΓJ directions when they appear, and plot their width in units of (a/λ) as a function of the host dielectric constant.
   b. For the dielectric constants of the host and cylinder given by 15 and 1, respectively, plot the transmission and reflection spectra as the cylinder radius is changed from 0 to 0.5.
2. Formulate a band structure program for the rectangular lattice with lattice constants (a, b). It will be analogous to the E and H programs in Supplement A for the triangular lattice. Plot the band structure for the following cases: the cylinder radius $r = 0.45a$, where $a = b$ is the lattice constant, and cylinder dielectric constant at 1 and dielectric constant of the host medium varied from 1 to 15.
3. Using the transfer matrix in Supplement B, plot the transmission and reflection coefficient for wavelengths from 400 to 1000 nm, consisting of two materials with dielectric constants 1.5 and 2.2, respectively. Neglect dispersion of the material properties. The thickness of the periods is a quarter wavelength at 750 nm.
   a. Change the number of periods from 2 to 20, and plot the transmission coefficient at normal incidence around the band gap.
      - Plot the minimum transmission as a function of the number of periods.
      - Plot the width of the band gap (measured from the transmission maxima closest to the band gap) versus the number of periods.
   b. For a stack with 10 periods, plot the E-polarization and H-polarization gap minimum transmission and its wavelength versus an angle of incidence from $0°$ to $75°$.

Supplement A: Numerical solution of the scalar wave equations in two-dimensional periodic media

The two-dimensional solution of photonic crystal band structures illustrates the essential complexity found in three-dimensional photonic crystals with the simplification of using scalar equations. The wave equations are solved using a plane-wave expansion to determine the complex dispersion properties of periodic dielectric materials. The photonic band structures can be used to determine different physical properties, such as spatial and temporal dispersion.

The programs solve the wave equations for a circular cylinder with dielectric constant EPSA in a background dielectric constant EPSB. The lattice constant, d, is scaled to unity, and the scaled cylinder radius, a, is less than 0.5 so that the cylinders do not overlap. The fractional volume of the cylinder relative to the unit cell volume is defined by the variable beta. The vector components for the first Brillouin zone symmetry points are $\Gamma = (gxg, gyg)$, $X = (gxx, gyx)$, and $J = (gxi, gyj)$. To plot the band structure, an itinerary around the Brillouin zone starts at point X, follows a line to the Γ point and then to the J point, and finishes at the X point.

The programs for E- and H-polarization are run using a set of plane waves that are symmetrically constructed in hexagonal rings around the Γ point. The basis vectors to generate the points on the reciprocal lattice are $G1 = (gx1, gy1)$ and $G2 = (gx2, gy2)$. The user can choose the variable $N1$ to determine the number of reciprocal lattice plane waves used to calculate the band structure. Figure 11.23 shows the plane-wave reciprocal
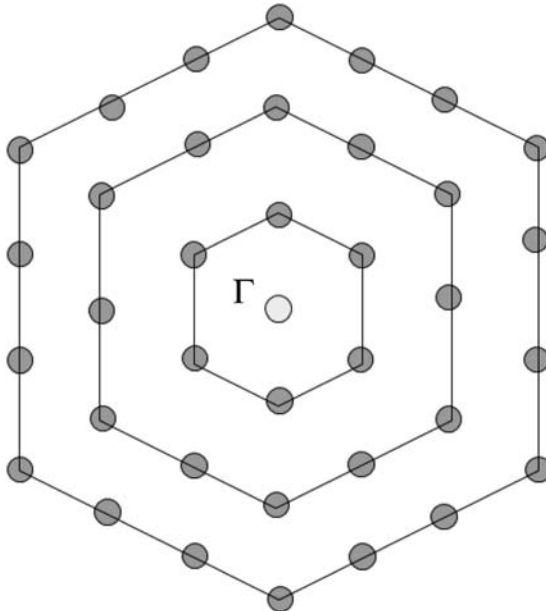
Figure 11.23 Construction of the number of plane waves in hexagonal rings.

lattice points in the set with $N1 = 1$, 2, and 3, which has 7, 19, and 37 elements, respectively.

Once the set of plane waves are chosen, the matrices for the dielectric function $\varepsilon(\vec{G}, \vec{G}')$ (E-method) or its reciprocal function $\eta(\vec{G}, \vec{G}')$ (H-method) are constructed. The diagonal elements of the material matrices are the average background value of the function, and the off-diagonal values are the Fourier transform of a circular cylinder minus the background. For the E-method the diagonal and off-diagonal terms are as follows:

EPS(ii, ii) = bbeta*EPSA + (1.D0-bbeta)*EPSB;

EPS(ii, j) = 2.d0*pi*(EPSA-EPSB)*a*a/(vcel) * besselj(1, x)/x;

The off-diagonal dielectric function is related to the modulation dielectric constant (EPSA-EPSB), i.e., the average background is subtracted. Similar consideration holds for the H-method.

With the above preparation, the equations are solved by matrix methods. The E-method solves the generalized eigenvalue problem defined by

$$Ax = \Omega Bx.$$

A and B are matrices, and $\Omega = \left(\frac{\omega}{c}\right)^2$ is the corresponding eigenvalue. On the other hand, for the H-method, the matrix B is replaced by the unit matrix, and the eigenvalues are found using common matrix methods.

Programs

E-Polarization Program

```
% E-polarization solution of the Wave equation in 2D periodic media
clear all;
pi  = acos(-1.0);% constant
N1  = 7; % Defines the size of the reciprocal space.
d=1.;% lattice constant scaled to unity.
a=.45;% Radius of the circular cylinder
bbeta = 3.14159*a*a/d/d*2.0/sqrt(3.0);% fractional area of the cylinder
%  relative to the unit cell
EPSA=1.; % Circular cylinder dielectric constant
EPSB=13;% Matrix dielectric constant
% The symmetry points in the Brillouin zone are defined
gxg  =  0.0; gyg  =  0.0;% Gamma point
gxx  =  0.0; gyx  =  2.0*pi/sqrt(3.0);% X point
gxj  =  2.0*pi/3.0; gyj  =  2.0*pi/sqrt(3.0);% J point
% The itinerary around the brillouin zone is laid out. X->G->J->X
gxi(1)  =  gxx; gyi(1)  =  gyx;
gxi(2)  =  gxg;gyi(2)  =  gyg;
gxi(3)  =  gxj;gyi(3)  =  gyj;
gxi(4)  =  gxx;gyi(4)  =  gyx;
  % Construct the set of reciprocal lattice wave vectors for the plane
wave calculation
      L=0;
```

```
        gx1  =  2.0*pi;
        gy1  =  -2.0*pi/sqrt(3.0);
        gx2  =  0.0;
        gy2  =  2.0*pi*2.0/sqrt(3.0);
%        generate  all  the  wave  vectors  within  a  hexagon  of
%        Maximum  width  2*n1.
for  ii=-N1:N1;
   for  j=-N1:N1;
        if  (  (-(N1+1)<(ii+j))  &  ((ii+j)<(N1+1))),
            L=L+1;
            gx(L)  =  gx1*ii  +  gx2*j;
            gy(L)  =  gy1*ii  +  gy2*j;
         end;
    end;
end;
% Construct the matrix of Fourier components of the periodic dielectric
%  function
     vcel = sqrt(3.0)/2.d0*d*d;% Wigner-Seitz unit cell volume
     % Average  dielectric  constant
    for  ii  =1:L;
        EPS(ii,ii)  =  bbeta*EPSA  +  (1.D0-bbeta)*EPSB;
    end;
    % off-diagonal  matrix  elements
     for  ii=1:L;
          i1=  ii  +  1;
          for  j=i1:L;
             x1  =  gx(ii)  -  gx(j);
             x2  =  gy(ii)  -  gy(j);
             x  =  sqrt(x1*x1+x2*x2)*a/d;
             EPS(ii,j)  =  2.d0*pi*(EPSA-EPSB)*a*a/(vcel)  *
besselj(1,x)/x;
% The  matrix  is  symmetric
             EPS(j,ii)  =  EPS(ii,j);
         end;
     end
  % Use the itinerary defined above to constructed above to move around
  %  the  Brillouin  zone
      n=L;
  % Counter  for  the  number  of  points  in  the  itinerary
      nknt  =  0;
for  nk  =  1:3;
          nk1  =  nk+1;
       nmax=  20;
     for  l=1:nmax;
           dkx  =(l-1)*(gxi(nk1)-gxi(nk))/(nmax)  +  gxi(nk);
           dky  =(l-1)*(gyi(nk1)-gyi(nk))/(nmax)  +  gyi(nk);
       for  ii=1:n;
           a(ii,ii) = ((dkx+gx(ii))^2 + (dky+gy(ii))^2)/4.d0/pi/pi;
           b(ii,ii)  =  EPS(ii,ii);
```

```
              for  j=ii+1:n;
                  a(ii,j)  =  0.d0;
                  a(j,ii)  =  0.d0;
                  b(ii,j)  =  EPS(ii,j);
                  b(j,ii)  =  EPS(ii,j);
              end
         end
  %  Find  the  eignevalues
           v=eig(a,b);
  %  Store  the  lowest  ten  eigenvalues  for  plotting
         nknt  =  nknt  +  1;
         for  li=1:10;
              rr(nknt,li)=sqrt(v(li));
              kxx(nknt)=nknt-1;
         end
       end
  end

  %  Plot  lowest  ten  eigenvalues
  figure(1)
  hold  on;
  for  ii=1:10;
     plot(kxx,rr(:,ii));
  end;
  ylabel('a/\lambda');
  xlabel('wave  vector  itinerary');
  legend('E  polarization')
  grid  on;
```

H-Polarization Program

```
  %   H-polarization solution of the Wave equation in 2D periodic media
  clear  all;
  pi  =  acos(-1.0);% constant
  N1  =  7;  %  size  of  the  reciprocal  space.
  d=1.;% lattice  constant  scaled  to  unity.
  a=.45;% Scaled radius of the circular cylinder relative to the lattice
        %  constant
  bbeta   =   3.14159*a*a/d/d*2.0/sqrt(3.0);%  fractional   area   of   the
  cylinder relative to the unit cell
  EPSA=1.;  %  Circular  cylinder  dielectric  constant
  EPSB=13;%  Matrix  dielectric  constant
  %  The  coordinates  of  symmetry  points  for  the  Brillouin  zone
  gxg  =  0.0;  gyg  =  0.0;%  Gamma  point
  gxx  =  0.0;  gyx  =  2.0*pi/sqrt(3.0);%  X  point
  gxj  =  2.0*pi/3.0;  gyj  =  2.0*pi/sqrt(3.0);%  J  point
  %  The  itinerary around the brillouin zone is laid out. X->G->J->X
  gxi(1)  =  gxx;  gyi(1)  =  gyx;
  gxi(2)  =  gxg;
  gyi(2)  =  gyg;
```

```
gxi(3)  =  gxj;
gyi(3)  =  gyj;
gxi(4)  =  gxx;
gyi(4)  =  gyx;
  %   Construct  the  set  of  reciprocal  lattice  wave  vectors  for  the
plane wave calculation
        L=0;
        gx1  =  2.0*pi;
        gy1  =  -2.0*pi/sqrt(3.0);
        gx2  =  0.0;
        gy2  =  2.0*pi*2.0/sqrt(3.0);
%        generate  all  the  wave  vectors  within  a  hexagon  of
%         Maximum  width  2*n1.
for  ii=-N1:N1;
    for  j=-N1:N1;
        if  (  (-(N1+1)<(ii+j))  &  ((ii+j)<(N1+1))),
            L=L+1;
            gx(L)  =  gx1*ii  +  gx2*j;
            gy(L)  =  gy1*ii  +  gy2*j;
        end;
    end;
end;
% Construct the matrix of Fourier components of the periodic dielectric
% function
        vcel  =  sqrt(3.0)/2.d0*d*d;%  Wigner-Seitz  unit  cell  volume
%  Average  of  the  inverse  dielectric  constant
      for  ii  =1:L;
        ETA(ii,ii)  =  bbeta/EPSA  +  (1.D0-bbeta)/EPSB;
    end;
%  off-diagonal  matrix  elements
    for  ii=1:L;
          i1=  ii  +  1;
          for  j=i1:L;
             x1  =  gx(ii)  -  gx(j);
             x2  =  gy(ii)  -  gy(j);
             x  =  sqrt(x1*x1+x2*x2)*a/d;
             ETA(ii,j)  =  2.d0*pi*(1/EPSA-1/EPSB)*a*a/(vcel)
*  besselj(1,x)/x;
%  The  matrix  is  symmetric
            ETA(j,ii)  =  ETA(ii,j);
          end;
      end
%  Use the itinerary constructed above to move around the Brillouin zone
      n=L;
%  Counter  for  the  number  of  points  in  the  itinerary
      nknt  =  0;
for  nk  =  1:3;
        nk1  =  nk+1
      nmax=  20;
```

```
      for  l=1:nmax;
           dkx  =(l-1)*(gxi(nk1)-gxi(nk))/(nmax)  +  gxi(nk);
           dky  =(l-1)*(gyi(nk1)-gyi(nk))/(nmax)  +  gyi(nk);
      for  ii=1:n;
           a(ii,ii)  =  ETA(ii,ii)*((dkx+gx(ii))^2  +  (dky+gy(ii))^2)/
4.d0/pi/pi;
           for  j=ii+1:n;
               a(ii,j)  =  ETA  (ii,j)*((dkx+gx(ii))*(dkx+gx(j))  +
(dky+gy(ii))*(dky+gy(j)))/4.d0/pi/pi;
               a(j,ii)  =  a(ii,j);
          end
      end
%  Find  the  eignevalues
          v=eig(a);
%  Store  the  lowest  ten  eigenvalues  for  plotting
          nknt  =  nknt  +  1;
          kxx(nknt)=nknt-1;
          for  li=1:10;
           rr(nknt,li)=sqrt(v(li));
          end

     end

end
%  Plot  lowest  ten  eigenvalues
figure(1)
hold  on;
for  ii=1:10;
   plot(kxx,rr(:,ii),'r');
end;
ylabel('a/\lambda');
xlabel('wave  vector  itinerary');
legend('H  polarization')
grid  on;
```

Supplement B: TMM for one-dimensional periodic media

The one-dimensional TMM illustrates the power of the method. This version contains fixed refractive indices, which can be replaced by experimental data or functions to cover absorption and dispersion over the wavelength range of interest.

The main program

```
clear  all;
lambda_min=300;  %  Minimum  wavelength  for  the  spectrum
lambda_max=800;  %  Maximum  wavelength  for  the  spectrum
%  data  for  refractive  indices.  This  can  be  replaced  by  experimental  data
ld0=[180  500  600  800  6100];  %  wavelengths  for  index  data
```

```
n00=[1. 1. 1. 1.   1.]; % superstrate medium has a real index
n10=[1.3 1.3 1.3 1.3  1.3]; % Medium 1 real part of the index
k10=[0. 0. 0. 0.   0.]; % Medium 1 imaginary part of the index
n20=[2. 2. 2. 2.   2.]; % Medium 2 real part of the index
k20=[0. 0. 0. 0.   0.]; % Medium 2 imaginary part of the index
n30=[1. 1. 1. 1.   1.]; % substrate medium has a real index
% number of periods
m=7;
% Thicknesses of each layer
d1(1:m)=400/4/1.3; % quarter-wave thickness center wavelength 400 nm
d2(1:m)=400/4/2;  % quarter-wave thickness center wavelength 400 nm
%   Total sample thickness calculated
L=sum(d1+d2);
%   angle of incidence
phi0=50/180*pi;
kx=2*pi*sin(phi0);% Transverse wave vector times wavelength (scaling)
lambda=lambda_min:1:lambda_max; % vector that spans a set of wavelengths
% refractive indices are interpolated from the data
n0=interp1(ld0,n00,lambda);
n1= interp1(ld0,n10,lambda);
k1= interp1(ld0,k10,lambda);
n2= interp1(ld0,n20,lambda);
k2= interp1(ld0,k20,lambda);
n3=interp1(ld0,n30,lambda);
% Calculate reflection and transmission coefficient
[Tpp,Rpp,Tss,Rss]=f_2mlyr(n0,n1,k1,n2,k2,n3,d1,d2,m,kx,lambda);
Trans=Tpp;
Refl=Rpp;
Tp=Tpp;Rp=Rpp;Ts=Tss;Rs=Rss;
figure(1);
plot(lambda,Tp,lambda,Rp)
xlabel('\lambda(nm)','fontsize',18,'FontName','Arial','fontweight','b')
ylabel('T_p,R_p','fontsize',18,'FontName','Arial','fontweight','b')
legend('T_p','R_p');
figure(2);
plot(lambda,Ts,lambda,Rs)
xlabel('\lambda(nm)','fontsize',18,'FontName','Arial','fontweight','b')
ylabel('T_s,R_s','fontsize',18,'FontName','Arial','fontweight','b')
legend('T_s','R_s');
```

Function to calculate the transmission and reflection coefficients

```
function [Tp,Rp,Ts,Rs]=f_2mlyr(n0,n1,k1,n2,k2,n3,d1,d2,m,kx,lambda)
% The output from this version are the p and s transmission and
reflection
% coeficients.  It may be modified to output the complex amplitudes.
for n=1:length(lambda)
%  Complex longitudinal wave vector component (scaled by the wavelength)
```

```
      kz0=conj(sqrt((2*pi*n0(n))^2-kx^2));
      kz1=conj(sqrt((2*pi*(n1(n)+1i*k1(n)))^2-kx^2));
      kz2=conj(sqrt((2*pi*(n2(n)+1i*k2(n)))^2-kx^2));
      kz3=conj(sqrt((2*pi*(n3(n)))^2-kx^2));
%   Reflection and transmission amplitudes for s- and p-polarization
%   in medium.
      [r01p,t01p,r01s,t01s]=f_rtampx(n0(n),0,n1(n),k1(n),kz0,kz1);
      [r12p,t12p,r12s,t12s]=f_rtampx(n1(n),k1(n),n2(n),k2(n),kz1,kz2);
      [r21p,t21p,r21s,t21s]=f_rtampx(n2(n),k2(n),n1(n),k1(n),kz2,kz1);
      [r23p,t23p,r23s,t23s]=f_rtampx(n2(n),k2(n),n3(n),0,kz2,kz3);
%   Superstrate interface matrices
        C01p=[1  r01p;  r01p  1];
        C01s=[1  r01s;  r01s  1];
        Ap=C01p;
        tp=t01p;
        As=C01s;
        ts=t01s;
%   Phase change due to path lengths in each medium for m periods.
    for  jj=1:m-1
     delta12=kz1*d1(jj)/lambda(n);
     delta21=kz2*d2(jj)/lambda(n);
%   Matrices from medium 1 to 2 and 2 to 1 for p and s poalrizations.
  C12p=[exp(1i*delta12)  r12p*exp(1i*delta12); r12p*exp(-1i*delta12)
exp(-1i*delta12)];
  C21p=[exp(1i*delta21)  r21p*exp(1i*delta21); r21p*exp(-1i*delta21)
exp(-1i*delta21)];
  C12s=[exp(1i*delta12)  r12s*exp(1i*delta12); r12s*exp(-1i*delta12)
exp(-1i*delta12)];
  C21s=[exp(1i*delta21)  r21s*exp(1i*delta21); r21s*exp(-1i*delta21)
exp(-1i*delta21)];
        Ap=Ap*C12p*C21p;
        tp=tp*t21p*t12p;
        As=As*C12s*C21s;
      ts=ts*t21s*t12s;
    end;
%   The last layer and interface with the substrate
      delta12=kz1*d1(m)/lambda(n);
      delta23=kz2*d2(m)/lambda(n);
      C12p=[exp(1i*delta12) r12p*exp(1i*delta12); r12p*exp(-1i*delta12)
exp(-1i*delta12)];
      C12s=[exp(1i*delta12) r12s*exp(1i*delta12); r12s*exp(-1i*delta12)
exp(-1i*delta12)];
      C23p=[exp(1i*delta23) r23p*exp(1i*delta23); r23p*exp(-1i*delta23)
exp(-1i*delta23)];
      C23s=[exp(1i*delta23) r23s*exp(1i*delta23); r23s*exp(-1i*delta23)
exp(-1i*delta23)];
       Ap=Ap*C12p*C23p;
      tp=tp*t12p*t23p;
%   Calculations of the reflection and transmission amplitudes and
```

```
%    coefficients  for  p  and  s  polarizations.
     rpp(n)=Ap(2,1)/Ap(1,1);
     Rp(n)=abs(Ap(2,1)/Ap(1,1))^2;
     tpp(n)=tp/Ap(1,1);
     Tp(n)=kz3/kz0*abs(tp/Ap(1,1))^2;
      As=As*C12s*C23s;
      ts=ts*t12s*t23s;
      rss(n)=As(2,1)/As(1,1);
      Rs(n)=abs(As(2,1)/As(1,1))^2;
      tss(n)=ts/As(1,1);
      Ts(n)=kz3/kz0*abs(ts/As(1,1))^2;
      end;
```

Function to compute the Fresnel Coefficients

```
function[rp,tp,rs,ts]=f_rtampx(n1,k1,n2,k2,kz1,kz2)
%calculate amplitude of transmission and reflection ith Gaussian beam
 nc1=n1-1i*k1;nc2=n2-1i*k2;
rp=(nc2^2*kz1-nc1^2*kz2)/(nc1^2*kz2+nc2^2*kz1+eps);
rs=(kz1-kz2)/(kz1+kz2+eps);
tp=2*nc2^2*kz1/(nc1^2*kz2+nc2^2*kz1+eps);
ts=2*kz1/(kz1+kz2+eps);
```

Supplement C: Spheroidal nanorods

As a first treatment of nanoantennas consider the electrostatic treatment that was previously used for LRC circuits in Chapter 2. For the nanoantenna in the form of a needle there is an exact solution for a prolate spheroid that can describe any aspect ratio of the needle. The coordinate system is are called prolate spheroidal coordinates, which have rotational symmetry about the z-axis. The coordinates are constructed with two foci at $z_\pm = \pm a/2$. The distance from the foci to a point (x, y, z) is

$$r_\pm = \sqrt{\rho^2 + \left(z \mp \frac{a}{2}\right)^2}, \tag{11.C1}$$

where the (ρ, z) plane is used with $\rho = \sqrt{x^2 + y^2}$.

Using this definition spheroidal coordinates are

$$\xi = \frac{r_+ + r_-}{a}, \ \eta = \frac{r_+ - r_-}{a}, \ \text{and} \ \phi = \tan^{-1}\left(\frac{x}{y}\right). \tag{11.C2}$$

The coordinates form an orthogonal triple with ξ, the radial variable over domain $\xi \in (1, \infty)$, and η, the "angular" variable, which has the domain $\eta \in (-1, 1)$. The coordinate ϕ is the usual azimuthal coordinate (longitude angle) with domain

$\phi \in (0, 2\pi)$. The Cartesian coordinates are expressed in terms of the spheroidal coordinates:

$$x = \frac{a}{2}\sqrt{(\xi^2 - 1)(1 - \eta^2)}\cos\phi, \ y = \frac{a}{2}\sqrt{(\xi^2 - 1)(1 - \eta^2)}\sin\phi, z = \frac{a}{2}\xi\eta.$$

$$(11.C3)$$

The (ρ, z) planes obey the following relations:

$$\frac{\rho^2}{(\xi^2 - 1)} + \frac{z^2}{\xi^2} = \frac{a^2}{4} \quad \text{and} \quad \frac{z^2}{\eta^2} - \frac{\rho^2}{(1 - \eta^2)} = \frac{a^2}{4}.$$

$$(11.C4)$$

The first equation is the expression for a family ellipsoids parameterized by ξ, and the second is a family of hyperboloids parameterized by η. A plot of the ellipses and hyperbolas in the (x, z) plane are given in Figure 11.24. The ratio of the major and minor axes (r) for the ellipse is related to the coordinate by $\xi = 1/\sqrt{1 - r^2}$. In other words, the variable ξ is the inverse of the particle's eccentricity, as discussed in Chapter 7.

The Laplacian is separable in spheroidal coordinates, and the solutions of Laplace's equation for the angular variable ϕ are the familiar trigonometric functions. The solutions of Laplace's equation are the associated Legendre polynomial, which is a function of the angular variable, $P_n^m(\eta)$, and also the radial variable solutions are $(P_n^m(\xi), Q_n^m(\xi))$, where $Q_n^m(\xi)$ is the associated Legendre function of the second kind.
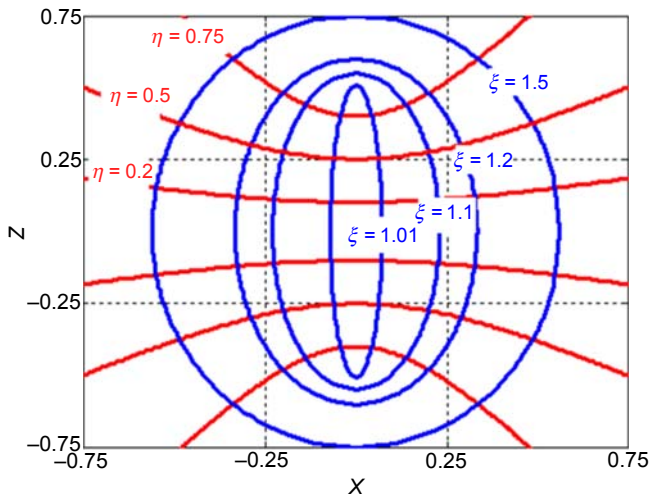
Figure 11.24 Spheroidal coordinate cross-sectional surfaces. Those curves parameterized by constant ξ form a family of ellipses. The curves parameterized by constant η form a family of hyberbolas.

The form of the electrostatic potential for a dielectric prolate spheroid when a constant field is applied along the z-axis is

$$\Phi_> = -E_0 z + B_> P_1(\eta) Q_1(\xi), \tag{11.C5}$$

$$\Phi_< = A_< P_1(\eta)[P_1(\xi)]. \tag{11.C6}$$

The Legendre functions are $P_1(\xi) = \xi$, $Q_1(\xi) = \frac{\xi}{2}\ln\left(\frac{\xi+1}{\xi-1}\right) - 1$. Applying the electrostatic boundary conditions at the surface ξ_1 for a prolate spheroid of dielectric constant ε_m embedded in a dielectric medium with ε_d yields two equations:

$$-E_0 \frac{a}{2} P_1(\xi_1) + B_> Q_1(\xi_1) = A_< P_1(\xi_1), \tag{11.C7}$$

$$\varepsilon_d \left(-E_0 \frac{a}{2} + B_> \frac{\partial Q_1(\xi_1)}{\partial \xi_1}\right) = \varepsilon_m A_<. \tag{11.C8}$$

To solve for the unknown coefficients $(A_<, B_>)$:

$$A_< = \frac{-E_0 \frac{a}{2} \varepsilon_d}{(\varepsilon_d - \varepsilon_m) 2 L_z + \varepsilon_m}, \tag{11.C9}$$

$$B_> = \frac{-E_0 \frac{a}{2} (\varepsilon_m - \varepsilon_d) \xi_1}{(\varepsilon_d - \varepsilon_m) 2 L_z + \varepsilon_m}. \tag{11.C10}$$

The coefficient in the denominator related to the result discussed in Chapter 7 is

$$L_z = \frac{\xi_1(\xi_1^2 - 1)}{2} \left(\frac{\xi_1}{(\xi_1^2 - 1)} - \frac{1}{2}\ln\left[\frac{\xi_1 + 1}{\xi_1 - 1}\right]\right). \tag{11.C11}$$

The derivation of the potential when the field is applied along the minor axis direction is left to the reader. The electric field inside the prolate spheroid is constant and points in the direction of the applied field when it is aligned with a principal axis. The electric field outside the spheroid has a constant applied value and a dipole contribution. In the limit of a spherical particle ($\xi_1 \rightarrow \infty$) $L_z \rightarrow \frac{1}{3}$, and the results of a spherical particle are recovered.

Further reading

Books
[1] B.A.A. Saleh, M. Teich, Fundamental of Photonics, second ed., Wiley, New York, 2013.
[2] J. Singh, Semiconductor Devices: Basic Principles, Wiley, New York, 2001.
[3] S.L. Chuang, Physics of Photonic Devices, second ed., Wiley, New York, 2009.

[4]   K. Sakoda, Optical Properties of Photonic Crystals, second ed., Springer-Verlag, Berlin, 2004.

[5]   A. Yariv, P. Yeh, Optical Waves in Crystals, Wiley, New York, 1984.

[6]   A. Taflove, Computational Electrodynamics, Artech House, Boston, 1995.

[7]   S.A. Maier, Plasmonics: Fundamentals and Applications, Springer, New York, 2007.

[8]   G. Moddel, S. Grover (Eds.), Rectenna Solar Cells, Springer, New York, 2013.

[9]   S. Logothetidis (Ed.), Nanomedicine and Nanobiotechnology, Nanoscience and Technology, Springer-Verlag, Berlin, 2012.

Reviews

Semiconductor lasers

[10] C. Gmachl, F. Capasso, D.L. Sivco, A.Y. Cho, Recent progress in quantum cascade lasers and applications, Rep. Prog. Phys. 64 (2001) 1533.

Photonic crystals

[11] H. Benisty, C. Weisbuch, Photonic crystals, Prog. Opt. 49 (2006) 177−315.

[12] K. Busch, G. von Freymann, S. Linden, S.F. Mingaleev, L. Tkeshelashvili, M. Wegener, Periodic nanostructures for photonics, Phys. Rep. 444 (2007) 101−202.

[13] K. Sakoda, J.W. Haus, Science and engineering of photonic crystals, Prog. Opt. 54 (2009) 271−317.

Photonic crystal fibers

[14] P.St.J. Russell, Photonic crystal fibers, Science 299 (2003) 358−362.

[15] P.St.J. Russell, Photonic crystal fibers, J. Lightwave. Technol. 24 (2006) 4729−4749.

Metasurfaces

[16] P. Genevet, F. Capasso, Holographic optical metasurfaces: a review of current progress, Rep. Prog. Phys. 78 (2015) 024401 (19 pp.).

[17] N. Yu, F. Capasso, Flat optics with designer metasurfaces, Nat. Mater. 13 (2014) 139−150.

[18] N. Yu, P. Genevet, F. Aieta, M.A. Kats, R. Blanchard, G. Aoust, J.P. Tetienne, Z. Gaburro, F. Capasso, Flat optics: controlling wavefronts with optical antenna metasurfaces, IEEE J. Sel. Top. Quantum Electron. 19 (2013) 4700423.

Selected papers

Photonic crystals

[19] E. Yablonovitch, Inhibited spontaneous emission in solid-state physics and electronics, Phys. Rev. Lett. 58 (1987) 2059.

[20] E. Yablonovitch, T.J. Gmitter, Photonic band structure: the face-centered-cubic case, Phys. Rev. Lett. 63 (1989) 1950.

[21] E. Yablonovitch, T.J. Gmitter, K.M. Leung, Photonic band structure: the face-centered-cubic case employing nonspherical atoms, Phys. Rev. Lett. 67 (1991) 2295.

[22] T. Kawashima, K. Miura, T. Sato, S. Kawakami, Self-healing effects in the fabrication process of photonic crystals, Appl. Phys. Lett. 77 (2000) 2613.

[23] H.S. Sozuer, J.W. Haus, R. Inguva, Photonic bands: convergence problems with the plane-wave method, Phys. Rev. B 45 (1992) 13962.

[24] H.S. Sozuer, J.W. Haus, Photonic bands: simple-cubic lattice, J. Opt. Soc. Am. B 10 (1993) 296.

[25] K. Busch, S. John, Photonic band gap formation in certain self-organizing systems, Phys. Rev. E 58 (1998) 3896.

[26] D.J. Norris, Y.A. Vlasov, Chemical approaches to three dimensional semiconductor photonic crystals, Adv. Mater. 13 (2001) 371.

[27] W.M. Robertson, G. Arjavalingam, R.D. Meade, K.D. Brommer, A.M. Rappe, J.D. Joannopoulos, Measurement of photonic band structure in a two-dimensional periodic dielectric array, Phys. Rev. Lett. 68 (1992) 2023.

[28] E.M. Purcell, Spontaneous emission probabilities at radio frequencies, Phys. Rev. 69 (1946) 681.

[29] W.M. Robertson, G. Arjavalingam, R.D. Meade, K.D. Brommer, A.M. Rappe, J.D. Joannopoulos, Measurement of the photon dispersion relation in two-dimensional ordered dielectric arrays, J. Opt. Soc. Am. B 10 (1993) 322.

[30] K. Sakoda, Symmetry, degeneracy, and uncoupled modes in two-dimensional photonic lattices, Phys. Rev. B 52 (1995) 7982.

[31] K. Sakoda, Group-theoretical classification of eigenmodes in three-dimensional photonic lattices, Phys. Rev. B 55 (1997) 15345.

[32] Z. Yuan, J.W. Haus, K. Sakoda, Eigenmode symmetry for simple cubic lattices and the transmission spectra, Opt. Express 3 (1998) 19.

Metallodielectrics

[33] M. Scalora, M.J. Bloemer, A.S. Manka, S.D. Pethel, J.P. Dowling, C.M. Bowden, Transparent, metallo-dielectric one dimensional photonic band gap structures, J. Appl. Phys. 83 (1998) 2377.

[34] M.J. Bloemer, M. Scalora, Transmissive properties of Ag/MgF$_2$ photonic band gaps, Appl. Phys. Lett. 72 (1998) 1676−1678.

[35] M. Scalora, M.J. Bloemer, C.M. Bowden, Laminated photonic band structures with high conductivity and high transparency: metals under a new light, Opt. Photon. News 10 (1999) 23.

[36] M. Scalora, G. D'Aguanno, N. Mattiucci, M.J. Bloemer, D. de Ceglia, M. Centini, A. Mandatori, C. Sibilia, N. Akozbek, M.G. Cappeddu, M. Fowler, J.W. Haus, Negative refraction and sub-wavelength focusing in the visible range using transparent metallo-dielectric stacks, Opt. Express 15 (2007) 508−523.

Nonlinear optics in photonic crystals

[37] Y. Dumeige, P. Vidakovic, S. Sauvage, I. Sagnes, J.A. Levenson, C. Sibilia, M. Centini, G. D'Aguanno, M. Scalora, Enhancement of second-harmonic generation in a one-dimensional semiconductor photonic band gap, Appl. Phys. Lett. 78 (2001) 3021−3023.

[38] D. Pezzetta, C. Sibilia, M. Bertolotti, R. Ramponi, R. Osellame, M. Marangoni, J.W. Haus, M. Scalora, M.J. Bloemer, C.M. Bowden, Enhanced Čerenkov second-harmonic generation in a planar nonlinear waveguide that reproduces a one-dimensional photonic bandgap structure, J. Opt. Soc. Am. B 19 (2002) 2102.

[39] C. Deng, J.W. Haus, A. Sarangan, A.M. Zheltikov, M. Scalora, M. Bloemer, C. Sibilia, Photonic band gap enhanced second-harmonic generation in planar lithium niobate waveguide, Laser Phys. 16 (2006) 927.

[40] P.P. Markowicz, H. Tiryaki, P.N. Prasad, V.P. Tondiglia, L.V. Natarajan, T.J. Bunning, J.W. Haus, Electrically switchable third-harmonic generation in photonic crystals, J. Appl. Phys. 97 (2005) 083512.

Metasurfaces

[41]   N. Yu, P. Genevet, M.A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, Z. Gaburro, Light propagation with phase discontinuities: generalized laws of reflection and refraction, Science 334 (2011) 333.

[42]   P. Genevet, N. Yu, F. Aieta, J. Lin, M.A. Kats, R. Blanchard, Z. Gaburro, F. Capasso, Ultra-thin plasmonic optical vortex plate based on phase discontinuities, Appl. Phys. Lett. 100 (2012) 13101.

[43]   F. Aieta, P. Genevet, N. Yu, M.A. Kats, Z. Gaburro, F. Capasso, Out-of-plane reflection and refraction of light by anisotropic optical antenna metasurfaces with phase discontinuities, Nano Lett. 12 (2012) 1702.

Rectennas: metal−insulator−metal structures

[44]   M. Dagenais, K. Choi, F. Yesilkoy, A.N. Chryssis, M.C. Peckerar, Solar spectrum rectification using nano-antennas and tunneling diodes, Proc. SPIE 7605 (2010) 76050E.

[45]   S. Bhansali, S. Krishnan, E. Stefanakos, D.Y. Goswami, Tunneling junction based rectenna - a key to ultrahigh efficiency solar/thermal energy conversion, AIP Conf. Proc. 1313 (2010) 79.

[46]   S. Grover, G. Moddel, Engineering the current−voltage characteristics of metal−insulator−metal diodes using double-insulator tunnel barriers, Solid-State Electron. 67 (2012) 94.

[47]   S. Grover, G. Moddel, Applicability of metal/insulator/metal (MIM) diodes to solar rectennas, IEEE J. Photovoltaics 1 (2011) 78.

[48]   J.W. Haus, L. Li, N. Katte, C. Deng, M. Scalora, D. de Ceglia, M.A. Vincenti, Nanowire metal-insulator-metal plasmonic devices, in: P. Buranasiri, S. Sumriddetchkajorn (Eds.), ICPS 2013: International Conference on Photonics Solutions, Proc. of SPIE, 8883, 2013, p. 888303.

[49]   J.W. Haus, D. de Ceglia, M.A. Vincenti, M. Scalora, Quantum conductivity for metal-insulator-metal nanostructures, J. Opt. Soc. Am. B 31 (2014) 259.

Nanosensors

[50]   P.L. Stiles, J.A. Dieringer, N.C. Shah, R.P. Van Duyne, Surface-enhanced raman spectroscopy, Annu. Rev. Anal. Chem. 1 (2008) 601−626.

[51]   J.H. Grossman, S.F. McNeil, Nanotechnology in cancer medicine, Phys. Today 65 (2012) 38−42.

This page intentionally left blank

Index

Chemical vapor deposition (CVD), 115–116, 157
Chiral vector, 120
Chirality parameter, 230–231
Classical oscillator model, 309
Clausius–Mossotti formula, 214
CMOS. *See* Complementary metal-oxide semiconductor (CMOS)
CNTs. *See* Carbon nanotubes (CNTs)
Complementary metal-oxide semiconductor (CMOS), 10, 157–158
Compound microscope, 186–187, 186f
Compound semiconductors, 123–127
Conductivity, 40
Constituent materials, 220–221
Constitutive relations, 16–19
Contact photolithography, 164–165, 164f
Continuity equation, 326
Copper, 111
Core electrons, 108
Correlation energy, 108
Coulomb gauge, 345
Coulomb interaction energy, 106
Coulomb's law, 14–15
Coupled wells with applied electric field, 60
Covalent bond, 3–4, 100–102, 113–115, 124
Critical dimension (CD), 164
Crystal structure, 91
  periodic lattices, 91–93
  reciprocal lattice, 93–103
  two-dimensional rectangular lattice, 92f
Crystalline films, 159
Crystals, 91
CVD. *See* Chemical vapor deposition (CVD)
Czochralski process, 123, 159–160

D
deBroglie's principle, 172
Decomposition, 96
Density functional theory (DFT), 103
Density of states (DOS), 110–111, 346–347
Depth of field, 195
Depth of focus (DOF), 167
Devices, 8–11
DFT. *See* Density functional theory (DFT)
Diamond, 115–116

Diatomic molecules, 102
Diazonaphthoquinone (DNQ), 163
Diode, 342
Dipole moment, 312–313
Dirac points, 117–118
Direct bandgap, 113–114, 123
Dispersive medium, dissipation and energy density in, 24–26
Dissipation in dispersive medium, 24–26
DNQ. *See* Diazonaphthoquinone (DNQ)
DOF. *See* Depth of focus (DOF)
Doping, 114–115
DOS. *See* Density of states (DOS)
Double barrier, resonant tunneling across, 68–70
Double patterning, 168–169
"Double-fishnet" structure, 264
Double-negative metamaterials, 254–265
Drude model, 20, 40, 233–235, 325
Drude–Lorentz dispersion model, 235
Dry etching. *See* Plasma etching
Dual tone photoresists, 169

E
e-beam lithography (EBL). *See* Electron-beam lithography (EBL)
EBL. *See* Electron-beam lithography (EBL)
EFCs. *See* Equi-frequency contours (EFCs)
Effective mass, 71–73
Effective medium approximations, 211
Effective medium theories, 211
  anisotropic mixtures, 225–227
  Bruggeman theory, 218–220
  Maxwell Garnett theory, 212–218
  microstructure investigation, 212f
  Nicolson–Ross–Weir method, 223–225
  quasi-static numerical approaches, 220–223
  spatial dispersion effects, 227–231
  two-phase inhomogeneous material, 218f
Effective permittivity, 212–213
Electric field distribution, 238f
Electric ring resonator (ERR), 290
Electrical conductivity, 112
Electrodynamics, 13
  Maxwell's equations, 13–19
  microscopic dynamical models, 19–20
  quasistatic limits, 32–40
  wave equations, 21–32

This page intentionally left blank

The field of nanophotonics covers several key enabling technologies that have the potential to drive economic growth and impact numerous application domains, such as integrated circuits technology, the environment, healthcare, military, transport, manufacturing and energy. This book gives readers the theoretical underpinnings needed to understand the latest advances in the field.

After an introduction to the area, chapters 2 and 3 cover the essential topics of electrodynamics, quantum mechanics and computation as these relate to nanophotonics. Subsequent chapters explore materials for nanophotonics including nanoparticles, photonic crystals, nanosilicon, nanocarbon, III-V and II-VI semiconductors. Fabrication and characterization techniques are addressed, and the important field of plasmonics is covered. The final chapter describes applications of nanophotonics in devices such as lasers, LEDs and photodetectors.

Fundamentals and Applications of Nanophotonics provides a systematic review of this important area for academics and post-graduate students in photonics, optics and nanoscience as well as those working in industry in the areas of nanotechnology, photonics, microelectronics, nanoelectronics, semiconductors, sensors, lasers, optical materials and devices, biology and nanomaterials.

Joseph Haus is a professor and past director of the Electro-Optics Program at the University of Dayton, USA.