Topics in Applied Physics 127

Kenneth Diest *Editor*

Numerical Methods for Metamaterial Design

Springer

Topics in Applied Physics
Volume 127

For further volumes:
www.springer.com/series/560

Topics in Applied Physics is a well-established series of review books, each of which presents a comprehensive survey of a selected topic within the broad area of applied physics. Edited and written by leading research scientists in the field concerned, each volume contains review contributions covering the various aspects of the topic. Together these provide an overview of the state of the art in the respective field, extending from an introduction to the subject right up to the frontiers of contemporary research.

Topics in Applied Physics is addressed to all scientists at universities and in industry who wish to obtain an overview and to keep abreast of advances in applied physics. The series also provides easy but comprehensive access to the fields for newcomers starting research.

Contributions are specially commissioned. The Managing Editors are open to any suggestions for topics coming from the community of applied physicists no matter what the field and encourage prospective editors to approach them with ideas.

Managing Editor

Dr. Claus E. Ascheron

Springer-Verlag GmbH
Tiergartenstr. 17
69121 Heidelberg
Germany
claus.ascheron@springer.com

Assistant Editor

Adelheid H. Duhm

Springer-Verlag GmbH
Tiergartenstr. 17
69121 Heidelberg
Germany
adelheid.duhm@springer.com

Kenneth Diest
Editor

Numerical Methods for Metamaterial Design

Springer

Editor
Kenneth Diest
MIT Lincoln Laboratory
Lexington, MA, USA

ISSN 0303-4216 ISSN 1437-0859 (electronic)
Topics in Applied Physics
ISBN 978-94-007-6663-1 ISBN 978-94-007-6664-8 (eBook)
DOI 10.1007/978-94-007-6664-8
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013945877

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my wife Merrielle, without whose support and patience this book would not have been possible

Preface

The idea of putting together this book is the result of the collective efforts of Daniel Marthaler, Luke Sweatlock, and myself while researching metamaterials and autonomous design tools within the Aerospace Research Labs at Northrop Grumman. We began looking into the area of metamaterial design using optimization methods combined with electromagnetic simulations in 2009 as an extension of our work on plasmonics and metamaterials. We quickly discovered that while there was decades worth of work on antenna and frequency selective surface design at microwave and radio frequencies, there was surprisingly little work done in the area of device design at infrared and optical frequencies using numerical methods. From a design standpoint, this also happened to be the frequency regime where materials' dispersion became most interesting. One of a few exceptions to this is the work done by Doug Werner's group at Penn State, and collaborations with Doug were an invaluable asset during our study of the field.

The design optimization methods discussed in this book are by no means a way to "hunt blindly" for solutions to design problems, and most instances of this yield extremely poor results. They are also not a replacement for a thorough understanding the underlying physics of how metamaterials respond; rather, they are a supplement to this knowledge and a valuable tool for those working in the field. In fact, the "objective functions" used to rank different candidate metamaterials' designs with respect to each other are a direct representation of the underlying physics involved in a given device design, and the success or failure of a given design optimization is critically dependent on this step. A former colleague was fond of saying "We must demand of ourselves understanding". To understand metamaterials, we must understand the underlying physics. To understand optimization, we must understand the underlying mathematics. To optimize the design of metamaterials, we must understand both.

While this text by no means encompasses all the work or methods that have been used to address the topic of optical metamaterial design, we hope that the topics covered here provide a fairly comprehensive overview of the main issues that arise when designing these structures, as well as which numerical methods are better suited for the task. This book is intended to provide a detailed enough treatment of the mathematical methods used, along with sufficient examples and additional references that senior level undergraduates or graduate students, who are new to the fields of plasmonics, metamaterials, or optimization methods, have an understanding of which approaches are best-suited for their work and how to implement these methods themselves.

The first chapter in the book is a brief overview of the major efforts within the field of metamaterials in the past and present. We touch on some of the key simulation and fabrication methods used in the field, as well as briefly review the physical mechanisms that contribute to metamaterial behavior at infrared and visible frequencies. This is done in an effort to frame the context of the book within the field as a whole.

The second chapter is an overview of the field of mathematical optimization, and describes where the methods covered in the book fit into the field as a whole. Chapters 3–5 describe the major optimization methods that are currently utilized in metamaterial design, and how differences in these methods make them more or less successful with increasing dimensionality. In Chap. 3, we discuss surrogate models that attempt to generate a maximally predictive, minimally complex model of the response of the entire design space. Chapter 4 discusses adaptive mesh optimization and examples of metamaterials designed using such an approach. Then in Chap. 5, we discuss fully stochastic methods that are based on techniques used in nature to efficiently optimize designs in very high dimensions, and numerous designs using these approaches are presented. The last two chapters of the book do not focus solely on the optimization method itself. These chapters integrate both optimization routines with novel methods for calculating and representing the shapes of the individual resonant structures within a metamaterial. These approaches are new to the field of metamaterial design; however, their applicability extends far beyond the focus of this book. This is clearly illustrated by the range of design examples that are covered throughout the last two chapters.

Finally, this book would never have been possible without a number of individuals who provided assistance, guidance, and support. A special thanks goes to Vladimir Liberman who provided thoughtful feedback regarding the individual chapters. In addition to all of the chapter authors, I would especially like to thank Harry Atwater, Erik Antonsson, Ray Bomeli, Ryan Briggs, Guy DeRose, Matthew Dicken, Bruce Esterline, James Ma, Doug Riley, Imogen Pryce, Mordechai Rothschild, Thom Schacher, Merrielle Spain, Robert Stewart, Seth Taylor, Ben Weiss, Northrop Grumman, MIT Lincoln Laboratory, and the staff at Springer.

Lexington, MA, USA Kenneth Diest

Contents

Contributors

Charles Audet GERAD and Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal, Québec, Canada

Zikri Bayraktar Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Jeremy A. Bossard Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Weitao Chen Department of Mathematics, The Ohio State University, Columbus, OH, USA

Kenneth Diest Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA, USA

Micah D. Gregory Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Zhi Hao Jiang Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Chiu-Yen Kao Department of Mathematical Sciences, Claremont McKenna College, Claremont, CA, USA

Sébastien Le Digabel GERAD and Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal, Québec, Canada

Jesse Lu Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA

Daniel E. Marthaler GE Global Research: Industrial Internet Analytics, San Ramon, CA, USA

Stanley Osher Department of Mathematics, University of California, Los Angeles, CA, USA

Tom Schaul Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Luke A. Sweatlock Northrop Grumman Aerospace Systems, Redondo Beach, CA, USA

Jelena Vuckovic Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA

Douglas H. Werner Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Pingjuan L. Werner Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Chapter 1
Introduction

Kenneth Diest

Abstract This chapter provides the context for the book in relation to the rest of the optical metamaterials community. First, a brief historical overview of optical metamaterial developments up to the start of the twentieth century is given. This is followed by a discussion of the field in relation to academic publications, nanofabrication, and electromagnetic simulations; and how developments in all three areas have contributed to the field as we know it today. The last section of the chapter presents the general framework for combining numerical optimization methods with full-field electromagnetic simulations for the design of metamaterials.

1.1 Introduction

The topic of electromagnetic metamaterials is a rich field that spans thousands of years and frequencies ranging from radio through the ultraviolet. These materials have a vast range of applications including art and jewelry, church decoration, frequency converters, electromagnetic cloaks, and sub-wavelength super lenses, just to name a few. Every day the field is expanding in new and different directions, and the range of new technologies being created seems limited only by the creativity of those involved. By combining our understanding of materials behavior with our unprecedented ability to model and fabricate structures at the nanoscale, researchers are bringing devices into the world that were previously only seen in the movies.

Here we define a metamaterial as "a man-made or otherwise artificially structured material with inclusions embedded in a host medium or patterned on a host surface, where the length scale of the inclusions is significantly smaller than the wavelength of interest." The macroscopic optical properties of the composite material are a result of the sub-wavelength unit structure, rather than the constituent materials; and by tuning the design of those inclusions, one can tune the overall electromagnetic properties of the metamaterial.

K. Diest (✉)
Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA 02420, USA
e-mail: diest@mit.edu

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_1,

1.2 Ancient Metamaterials

While there's no way of knowing for sure when and where the first optical metamaterials were produced, a strong candidate for this distinction is glass, specifically stained glass. The first evidence of glass making was around 3000 B.C.E. in an area called the Canaanite–Phoenician coast near the Mediterranean (just north of present-day Haifa). The sand from this region contained the right concentrations of lime and silica, so that the traders needed only to mix in natron (a mix of soda ash, baking soda, salt, and sodium sulfate) while the melt was placed into a hot fire [63]. The earliest glasses developed in this manner were opaque rather than transparent due to scattering from small air bubbles or particles trapped within the glass during formation. Later, during the first millennium B.C.E., hotter kilns were developed and artisans began to introduce metal oxides into the glass to control the color. By forming the glass either with or without charcoal, the glassworkers were able to either reduce or oxidize added copper, and as a result, produce glass that was either red or blue, respectively.

Since it seems that no discussion of plasmonics or metamaterials would be complete without referencing the Lycurgus Cup, we'll mention here that this artifact is one of the world's most famous examples of metal being introduced into glass. This cup was produced during the fourth century A.D. during the Roman Empire, Fig. 1.1. The cup depicts the death of King Lycurgus in Thrace at the hands of Dionysus. As can be seen from Fig. 1.1, the glass appears green (a) when seen with light reflected off the surface of the cup, and red (b) when seen with light transmitted through the cup. This remarkable behavior results from the introduction of colloidal gold and silver into the glass during formation. The resulting gold and silver nanoparticles within the glass reflect the green portion of the visible spectrum while transmitting the red portion of the visible spectrum [4].

It's truly remarkable that from this period in history all the way to present day, people have been studying how to produce new and different optical properties by simply combining nanoscale metallic inclusions within dielectrics, and that a theory for this type of scattering from metal spheres would not be formalized until 1908 by Gustav Mie [45], $\sim$1600 years after the cup was made. From the fourth century on, people began studying how to produce stained glass by annealing the material with the addition of metallic salts. One of the first books documenting these studies, and arguably the first book on metamaterials, was "*The Book of the Hidden Pearl*," written in the eight century A.D. by Jabir ibn Hayyan, discussing the manufacturing of colored glass as well as techniques for the coloring of gemstones [31].

Around the turn of the twentieth century, the theories required to explain the behavior of metamaterials began to take shape. The first attempt at developing modern metamaterials with sub-wavelength structures was by Jagadis Chunder Bose in 1898. Bose used pieces of twisted jute in an effort to develop an artificially structured chiral material [8]. In 1904, J.C.M. Garnett published his paper on "*Colours in Metal Glasses and in Metallic Films*" in the Philosophical Transactions of the Royal Society [24]. Here he used the Drude model for the optical properties of free electron metals to describe how colors arose and changed within glasses when

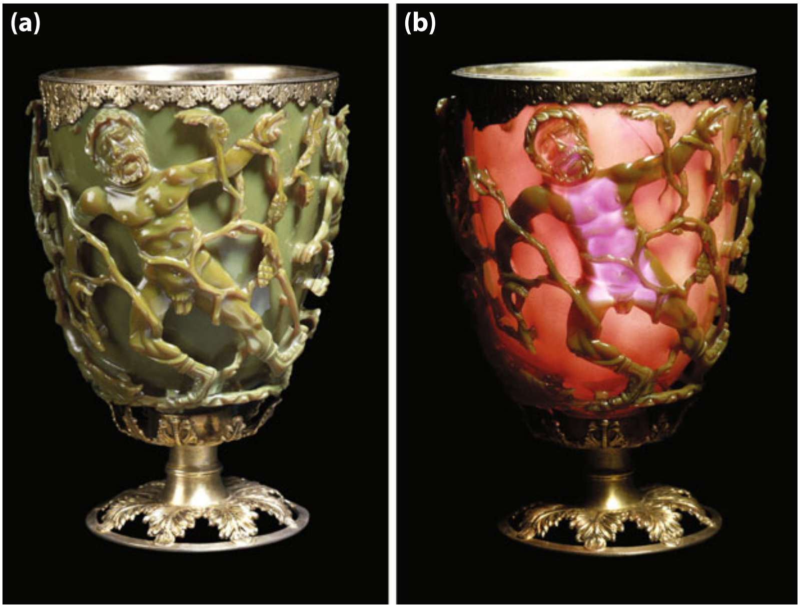


Fig. 1.1 The Lycurgus cup shown in both reflection (**a**) and transmission (**b**). Gold and silver nanoparticles are responsible for the strong reflection of green light and transmission of red light. ©Trustees of the British Museum

gold or silver films were annealed into nanoparticles that dispersed throughout soda glass. This work was followed in 1908 by the paper mentioned earlier from Gustav Mie that discussed the scattering of electromagnetic radiation by a sphere [45]. This work enabled the calculation of electric and magnetic fields inside and outside of a sphere which, in turn, can be used to calculate the scattering profile of incident light.

Taken together, these papers represent the first major effort using electromagnetic theory to analytically explain the behavior of nanoparticles, and more generally, optical metamaterials. This also represents a fundamental turning point in the history of metamaterials. Now, for the first time, the range of accessible metamaterials was not simply limited to those discovered by chance or passed down by word of mouth; rather, these theories could be applied in new and different ways to target specific applications and produce novel optical behavior.

1.3 Modern Metamaterials

In the late 1960s, the era of modern metamaterials began. Traditionally, modern developments in the field are attributed in large part to three seminal papers by Victor Veselago in 1967 [81], Sir John Pendry in 2000 [57], and David Smith in 2000 [71]. These papers certainly helped to inspire an entire field of researchers in the field of optical metamaterials and spark an enormous surge in related publications

(Fig. 1.2(b)); however, these publications and other like it represent one of three different communities that came together at the start of the twenty first century to bring about the field of metamaterials as we know it today. Current progress in the field was brought about by a combination of key papers as well as advances in, and the commercialization of, both nanofabrication capabilities and full-field electromagnetic modeling tools. These three areas all represent key components in the development and understanding of modern metamaterials, and the intersection of all three has brought the field to where it is today.

1.3.1 Publications

In 1967, Victor Veselago published his seminal paper on "*The Electrodymanics of Substances with Simultaneously Negative Values of ε and μ* [81]." In this paper, Veselago describes such "*left-handed materials*" that would support electromagnetic waves where the phase velocity and pointing vector propagate in opposite directions. This work was not realized experimentally until 2000, when David Smith and colleagues fabricated the first ever negative-index material by structuring an array of copper strips and split ring resonators on printed circuit boards [71]. Smith and Schurig later used these techniques and the recently emerging field of transformation optics to design an electromagnetic invisibility cloak that operated over a band of microwave frequencies, Fig. 1.2(a) [39, 64, 65, 82]. Concurrently in 2000, Sir John Pendry published his paper entitled "*Negative Refraction Makes a Perfect Lens* [57]." Here Pendry took the concepts introduced in Veselago's paper and explored the possibility of producing a negative-index "*super lens*." In principle, this material could circumvent the normal diffraction limits of light and resolve structures "only a few nanometers across" at visible frequencies. The combination of these three efforts by Veselago, Smith, and Pendry helped kick-off an enormous amount of research within the field of negative index materials and metamaterials in general, and the field would not be where it is today without their work. Between the years 2000 and 2011, many areas of research have received tremendous attention including: optical metamaterials (∼2800 publications), negative index metamaterials (∼2450 publications) [59, 67, 72, 80], optical cloaking (∼650 publications) [11, 58, 64, 79], and nonlinear metamaterials (∼450 publications) [36, 37, 60, 87]. Representative references for each topic are listed above, and a compilation of all publications from the four representative research topics in the field as a function of year is plotted in Fig. 1.2(b).[1] While it should be noted that papers within the four topics are not mutually exclusive, the general trends clearly show enormous growth within the field starting around 2000 for optical, nonlinear, and negative index metamaterials, and 2006 for optical cloaking.

While the contributions of these researchers have certainly shaped the development of the field of optical metamaterials over the last half century, they are by

[1] Citations compiled through Web of Science, March 2012.

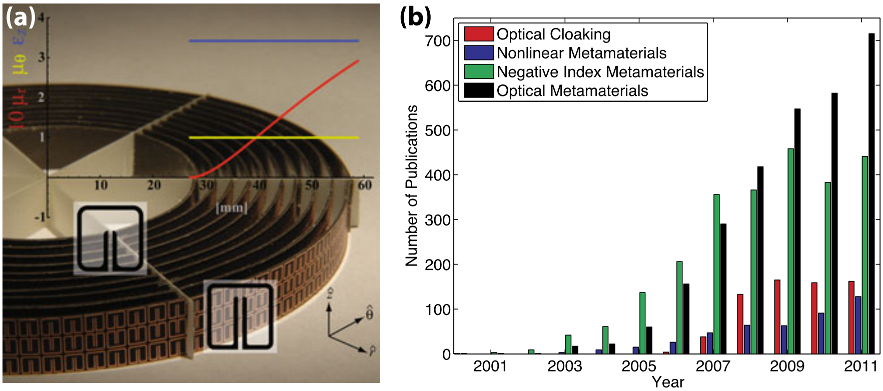


Fig. 1.2 In (**a**), the metamaterial-based, two-dimensional microwave cloak from Shurig et al. [64]. The structure consists of ten concentric cylinders of split ring resonators mounted on printed circuit board. The plot in the foreground shows the relevant materials parameters (μ_r, μ_θ, ε_z) as a function of distance from the center of the cloak. In (**b**), the number of papers published from 2000 to 2011 on the topics of: optical cloaking, nonlinear metamaterials, negative index metamaterials, and optical metamaterials

no means the only scientists to do so. From 1968–2009, Ben Munk became a pioneer within the field of frequency selective surfaces for their use in radar and other military applications [48–50]. Here, the military quickly realized the importance of being able to properly tune the design of antenna arrays for absorption and beam steering applications, and to this day, portions of his 1968 Ph.D. thesis are still classified. Research lead independently by Vladimir Shalaev and Xiang Zhang has taken the microwave cloaking and negative index of refraction concepts demonstrated by Smith and Shurig, and extended them to visible frequencies [9–11, 60, 67, 79, 80]. And finally, in a similar vein to the work done by Veselago, modeling and design work lead by Nader Engheta has predicted a wide range of exotic behavior from phase-shifters with novel medium to electromagnetic tunneling through waveguides of "*epsilon-near-zero*" materials, to the introduction of circuit nanoelement models in the optical domain using plasmonic and non-plasmonic nanoparticles [17–19, 62, 69].

While this list by no means encompasses the range of researchers who have made major contributions to the field, it does emphasize some of the major work done that has leveraged metamaterial theory (including Veselago, Mie, and Drude) to tailor the design of metamaterials for specific applications. These researchers have done a superb job of studying not only the fundamental resonances involved in these nanostructures, but also addressing the question that immediately follows: How to design these structures for specific applications?

1.3.2 Fabrication

As the field of metamaterials has moved to ever shorter wavelengths, the frequency ranges and resonances that we can study are limited by the modeling and fabrication capabilities at our disposal. During the 1960s, 1970s, and 1980s, the study of radio frequency (RF) metamaterials required structures with length scales on the order of centimeters, and these could be easily machined and assembled by hand. In the 1970s and 1980s, techniques for micro- and nano-lithography were developed to pattern structures with sizes spanning from tens of microns to tens of nanometers. For structures designed to operate at frequencies up to 1 THz, standard photolithography techniques allowed large arrays of these structures to be pattered very quickly by exposing ultraviolet light through a patterned glass photomask to transfer this pattern into a photosensitive resist; and subsequently transferring that pattern into the resonator material, either through etching or lift-off techniques. As the operating wavelength of the metamaterial grew shorter, standard lithography techniques ran up against the diffraction limit of the exposure light, and new methods were employed. While there are a number of techniques that have been developed to fabricate structures at the nanoscale, including X-ray lithography, interference lithography, extreme ultraviolet and immersion lithography, direct laser writing, and imprint lithography, over the past 20 years, there are two techniques that have played key roles in the fabrication work done in the metamaterials' community: electron beam lithography and focused ion beam patterning.

While both techniques are key components within modern, academic nanofabrication facilities, it is interesting to note that the two tools evolved along very different paths. Electron beam lithography was initially developed as far back as the 1960s and 1970s; however, for most of its history, this tool was largely used within the microelectronics industry and its price was such that it was prohibitively expensive for academic use. Even to this day, electron beam lithography tools in academia are mainly located within shared user facilities and have distributed ownership. In contrast, focused ion beams have been more of a research and development tool with many key developments coming from users of the tool. Only over the past few decades has the tool become mainstream enough to be commercialized by such companies as FEI Co. and Micrion Corp.

Scanning Electron Beam Lithography (SEBL) was first introduced as a commercial Gaussian beam system in 1962 by Philips, Eindhoven, and in 1974 as a commercially available shaped electron beam lithography system by Carl Zeiss, Jena. SEBL uses a set of electromagnetic lenses to focus a column of high energy (usually at 30, 50, or 100 keV) electrons onto a focal plane with typical spot sizes between 2 and 10 nm, Fig. 1.3(a). The lenses then raster the beam across the sample at predetermined positions to expose the resist where it should either remain or dissolve away during subsequent processing steps. Because the de Broglie wavelength [40] of these electrons is so much smaller than ultraviolet light used in standard photolithography, the minimum feature size is orders of magnitude smaller. Another benefit of electron beam lithography is in the flexibility of the tool when compared with photolithography. Besides the size limitations, photolithography requires the

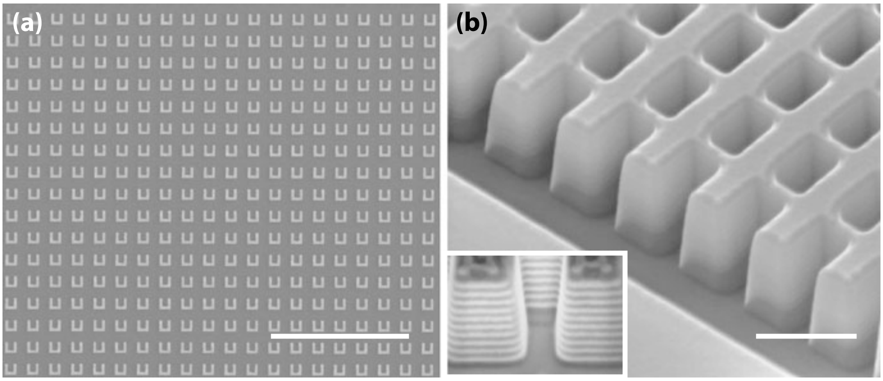


Fig. 1.3 An example of nanofabricated Split Ring Resonators fabricated on silicon substrates using Scanning Electron Beam Lithography (**a**). The scale bar corresponds to 5 μm. In (**b**), fishnet metamaterials fabricated using Focused Ion Beam milling were studied as negative index materials [80]. The scale bar corresponds to 1 μm

fabrication of a photomask which, once produced, is difficult to modify. In comparison, electron beam lithography is fed an electronic beam map before each run, which can be easily modified. While there are a number of benefits to using this technique, the primary limitation is throughput. Because the beam can only expose one spot at a time, the process is inherently serial, and as a result, patterning on 8″ or 12″ wafers would take orders of magnitude longer than with photolithography.

In comparison, the Focused Ion Beam (FIB) is a direct milling process and after patterning, does not require further etching or lift-off steps to fabricate nanostructures [38, 80], Fig. 1.3(b). Early developments in FIB milling came about in the 1970s when Levi-Setti [22], and Orloff & Swanson [56] independently introduced the first field emission Focused Ion Beam systems in 1975, and the first liquid metal ion source by Seliger in 1979 [66]; however, it wasn't until 1998 that FEI commercially produced the tool in its current form as a dual-beam Focused Ion Beam / Scanning Electron Microscope system.

Instead of transferring a pattern into a resist, which then requires further etching or lift-off steps to produce structures, the focused ion beam is a direct milling process. The FIB extracts gallium ions from a liquid metal source and accelerates these ions onto the sample in a focused beam with a radius of ∼10 nm. This allows the FIB to produce structures with critical dimensions equivalent to those with electron beam lithography; however, it has been shown that, for certain material sets and device designs, this process can be significantly quicker than electron beam lithography [20]. As with electron beam lithography, FIB is an inherently serial process, and as a result, is very time consuming for larger samples. Also, the materials selectivity of the etch process may be significantly reduced when compared with the variety of etches used in standard lithography. The etching must be timed properly or else significant erosion into the underlying substrate will occur. At the same time the sample is being milled, it's also being implanted with gallium ions. This effect will typically reduce the quality of the materials being etched, and depending on the material and structures being fabricated, can result in significantly modified material properties [29].

One method that address the issue of throughput with electron beam lithography is nanoimprint lithography [12, 27]. Using this technique, electron beam lithography is used to pattern a "master stamp" with the desired nanostructures. A sample is coated with a heat or ultraviolet curable resist and the stamp is directly pressed into the resist. The speed and effectiveness of this technique is then determined by the rate and extent to which the polymer conforms to the mask. Control of the adhesion between the stamp and the resist allows removal of the stamp while retaining the imprinted pattern. Once the stamp is fabricated, an entire 8″ wafer can be patterned in a matter of minutes. The resolution of this process is ∼10 nm, and there are no diffraction effects. The mold can be re-used many times and is only limited by the rate at which the stamping process erodes the stamp features. Not only does this process reduce patterning time by many orders of magnitude compared with electron beam lithography, but by stamping patterns onto the same substrate multiple times, three-dimensional structures can be fabricated layer by layer.

The last two benefits of imprint lithography cannot be emphasized enough. All of the methods discussed so far are inherently planar fabrication techniques, and fabricating fully three-dimensional metamaterials with these methods has two major problems. First, the accuracy with which subsequent layers can be positioned with respect to those already fabricated can be on the order of the resonator critical dimensions, which can significantly degrade the performance of the metamaterial. Second, the time required to fabricate a single, two-dimensional layer of metamaterials can be prohibitively long, and this essentially rules out repeating the process tens or hundreds of times to extend structures into the third dimension.

While there are promising alternative approaches to address these issues such as self-assembly techniques [25, 68] and direct laser writing [15, 21], these methods are still under development and are not yet integrated in standard fabrication facilities. Examples of both methods are shown in Fig. 1.4(a)–(b).

1.3.3 Modeling

Before the 1960s, electromagnetic modeling was mainly limited to closed-form and infinite series analytical solutions to the problems of interest. During the 1960s, both the Finite Element Method (FEM) and the Finite-Difference Time-Domain method (FDTD) were reported for the first time in the field of computational electromagnetics, and since then have become two of the main methods for analyzing complex optical metamaterials. The following sections give a brief overview and comparison of the two methods. For a rigorous treatment of these methods, the reader is referred to [33, 34, 76].

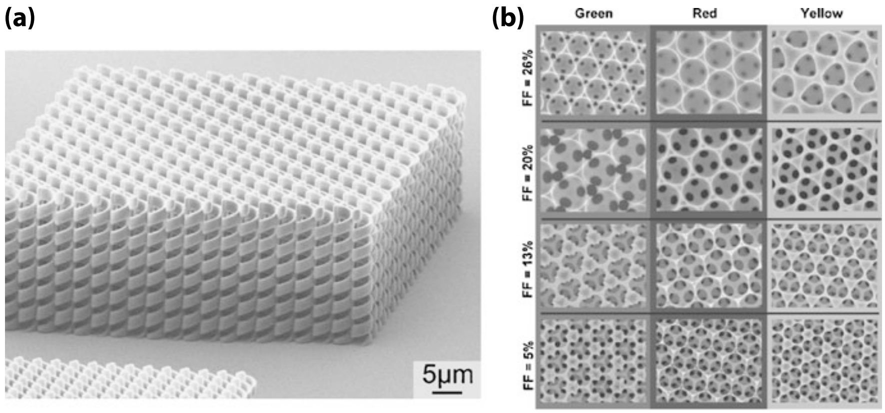


Fig. 1.4 In (**a**), an example of chiral metamaterials fabricated using Direct Laser Writing [78]. The cubic lattice is written in SU-8 negative tone polymer on a glass substrate. In (**b**), an example of metamaterial arrays fabricated using self assembly [86]. Nickel is patterned in an inverse-opal structure. The *four rows* correspond to different filling fractions of nickel, and the *three columns* correspond to observed far-field colors of the structure as a result of differing surface topographies

1.3.3.1 Finite-Difference Time-Domain Method

The Finite-Difference Time-Domain (FDTD) method is a time-stepping approach that models how electromagnetic waves actually move through a structure. The FDTD method has a number of different implementations; however, the most well-structured is the highly accurate algorithm introduced by Kane Yee in 1966 [85], and was first made commercially available by Panoramic Technology in 1999. Using the Yee algorithm, the structure to be simulated is first broken up into a rectangular grid, with the corresponding ε and μ calculated at each spatial position. The method starts with the time-dependent, differential form of Maxwell's equations:

$$\nabla \cdot \mathbf{D} = 0, \tag{1.1a}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{1.1b}$$

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} - \mathbf{M}, \tag{1.1c}$$

$$\frac{\partial \mathbf{D}}{\partial t} = \nabla \times \mathbf{H} - \mathbf{J}. \tag{1.1d}$$

Here, Faraday's law (Eq. (1.1c)) relates the magnetic flux density **B**, the electric field **E**, and the magnetic current density **M**; while Ampere's law (Eq. (1.1d)) relates the electric flux density **D**, the magnetic field **H**, and the electric current density **J**. Equations (1.1a)–(1.1d) are then discretized using a central-difference approximation which is accurate to second-order. The Yee algorithm solves for the electric and magnetic fields in space and time by utilizing Maxwell's curl equations. By combining Eqs. (1.1c) and (1.1d) with the constitutive equations for the electric flux density

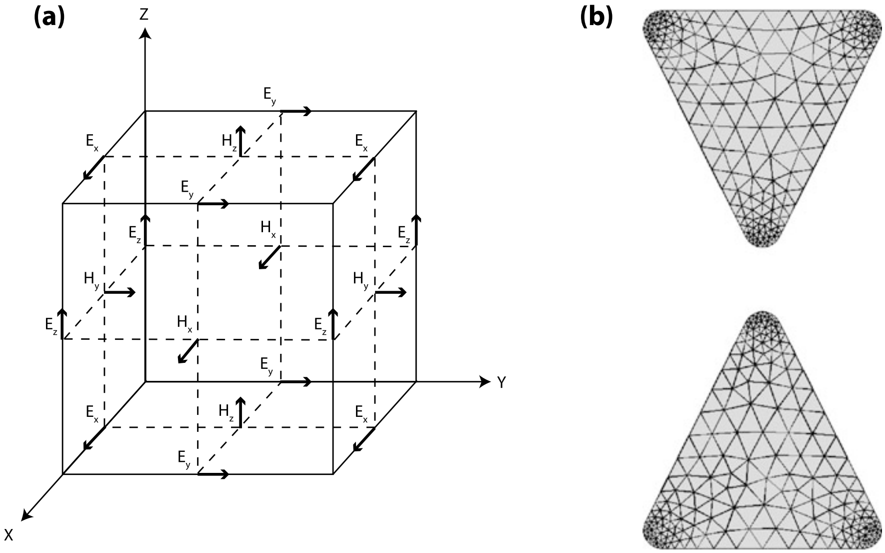


Fig. 1.5 The rectangular Yee cell is shown in (**a**). This visualization shows the distribution of electric and magnetic field vector components. Using Finite-Difference Time-Domain simulations, the three-dimensional volume of structures and spaces to be simulated consists of an array of these cells. An example of a bow-tie antenna discretized using the triangular, conformal meshing used in Finite Element Methods is shown in (**b**). The benefits of a conformal mesh over the rectangular grid used in FDTD can be seen near the corners of the antenna

(Eq. (1.2)), magnetic flux density (Eq. (1.3)):

$$\mathbf{D} = \varepsilon_0 \varepsilon \mathbf{E}, \tag{1.2}$$

$$\mathbf{B} = 2\mu_0 \mu \mathbf{H}, \tag{1.3}$$

along with the fact that the electric and magnetic current densities can serve as additional sources of electric ($\mathbf{J}_{\text{source}}$) and magnetic ($\mathbf{M}_{\text{source}}$) energy:

$$\mathbf{J} = \mathbf{J}_{\text{source}} + \sigma_E \mathbf{E}, \tag{1.4}$$

$$\mathbf{M} = \mathbf{M}_{\text{source}} + \sigma_H \mathbf{H}, \tag{1.5}$$

where σ_E is the electrical conductivity and σ_H is the magnetic loss, we arrive at Maxwell's curl equations for linear, isotropic, non-dispersive materials:

$$\frac{\partial \mathbf{H}}{\partial t} = -\frac{1}{\mu} \nabla \times \mathbf{E} - \frac{1}{\mu} (\mathbf{M}_{\text{source}} + \sigma_H \mathbf{H}), \tag{1.6}$$

$$\frac{\partial \mathbf{E}}{\partial t} = -\frac{1}{\varepsilon} \nabla \times \mathbf{H} - \frac{1}{\varepsilon} (\mathbf{J}_{\text{source}} + \sigma_E \mathbf{E}). \tag{1.7}$$

This produces a set of six coupled scalar equations for $\frac{\partial H_{x,y,z}}{\partial t}$ and $\frac{\partial E_{x,y,z}}{\partial t}$ that represent a "Yee cell," Fig. 1.5(a), where every electric-field component is surrounded by four circulating magnetic-field components and every magnetic-field component is surrounded by four circulating electric-field components. The simulation volume is then spanned by an array of Faraday's law and Ampere's law contours. Thus, the method accurately simulates both the differential and integral forms of Maxwell's equations at every point in the simulation volume.

To obtain each $\mathbf{E}$ and $\mathbf{H}$ component at time t and position (x, y, z), the curl equations are discretized in time, and the electromagnetic pulse propagates through the simulation volume by leapfrogging from $E_{x,y,z}(x, y, z, t = 0)$ to $H_{x,y,z}(x + 1/2, y + 1/2, z + 1/2, t = 1/2\Delta t)$ to $E_{x,y,z}(x + 1, y + 1, z + 1, t = \Delta t)$, and so on. Finally, a Fourier transform of these results yields the field magnitudes and phases at every point and every frequency.

1.3.3.2 Finite Element Method

Compared with the FDTD method, the Finite Element Method (FEM) is an inherently more complex and universal method. FEM is a numerical procedure to find stable solutions to boundary-value partial differential equations. This approach was first reported in 1943 by Richard Courant in his study of elasticity and structural analysis [14], where the concept of mesh discretization of a simulation was introduced. It was not until 1969 that the method was introduced to the field of electromagnetic engineering, when Silvester used this approach in the field of microwave engineering [70].

The Finite Element Method for electromagnetics is a frequency-domain method, which is again based on Maxwell's equations (1.1a)–(1.1d). When reformulated to include the anisotropic materials permittivity (ε_{ij}) and permeability (μ_{ij}) of the structure under consideration, we start with:

$$\nabla \cdot (\varepsilon_{ij} \cdot \mathbf{E}) = -\frac{1}{i\omega}\nabla \cdot \mathbf{J}, \tag{1.8a}$$

$$\nabla \cdot (\mu_{ij} \cdot \mathbf{H}) = -\frac{1}{i\omega}\nabla \cdot \mathbf{M}, \tag{1.8b}$$

$$i\omega\mu_{ij} \cdot \mathbf{H} = -\nabla \times \mathbf{E} - \mathbf{M}, \tag{1.8c}$$

$$-i\omega\varepsilon_{ij} \cdot \mathbf{E} = -\nabla \times \mathbf{H} - \mathbf{J}. \tag{1.8d}$$

In this context, FEM assumes that the resonant structure under consideration is surrounded by an artificial absorbing boundary condition which approximates the electric and magnetic fields approaching zero at infinity:

$$\hat{n} \times \nabla \times \mathbf{E} + ik_0\hat{n} \times \hat{n} \times \mathbf{E} \approx 0, \tag{1.9a}$$

$$\hat{n} \times \nabla \times \mathbf{H} + ik_0\hat{n} \times \hat{n} \times \mathbf{H} \approx 0. \tag{1.9b}$$

Here, $\hat{n}$ represents the vector normal to the boundary surface and k_0 is the incident free-space wave vector. Following the treatment by Jin and Riley [34], when the boundary conditions in Eqs. (1.9a) and (1.9b) are combined with the vector wave equation for the electric field of Maxwell's equations:

$$\nabla \times (\mu_0/\mu_{ij} \cdot \nabla \times \mathbf{E}) - k_0^2 \varepsilon_{ij} \cdot \mathbf{E} = -ik_0\sqrt{\mu_0/\varepsilon_0}\mathbf{J} - \nabla \times (\mu_0/\mu_{ij} \cdot \mathbf{M}), \quad (1.10)$$

we arrive at

$$\begin{aligned} &\iiint_V \left[(\nabla \times \mathbf{T}) \cdot \mu_0/\mu_{ij} \cdot (\nabla \times \mathbf{E}) - k_0^2 \mathbf{T} \cdot \varepsilon_{ij} \cdot \mathbf{E}\right] dV \\ &\quad = \oiint_{S_B \cup S_{\text{surf}}} \hat{n} \cdot \left[\mathbf{T} \times (\mu_0/\mu_{ij} \cdot \nabla \times \mathbf{E})\right] dS \\ &\qquad - \iiint_V \mathbf{T} \cdot \left[ik_0\sqrt{\mu_0/\varepsilon_0}\mathbf{J} + \nabla \times (\mu_0/\mu_{ij} \cdot \mathbf{M})\right] dV, \end{aligned}$$

where V represents the entire volume of integration, S_B is the artificial absorbing boundary surface, S_{surf} is the surface of the structure being simulated, and $\mathbf{T}$ is an appropriate test function used for integration. Combining this with the artificial absorbing boundary condition in Eqs. (1.9a) and (1.9b) gives

$$\begin{aligned} &\iiint_V \left[(\nabla \times \mathbf{T}) \cdot \mu_0/\mu_{ij} \cdot (\nabla \times \mathbf{E}) - k_0^2 \mathbf{T} \cdot \varepsilon_{ij} \cdot \mathbf{E}\right] dV \\ &\quad = \oiint_{S_{\text{surf}}} (\hat{n} \times \mathbf{T}) \cdot \mu_0/\mu_{ij} \cdot (\nabla \times \mathbf{E})\, dS - ik_0 \oiint_{S_B} (\hat{n} \times \mathbf{T}) \cdot (\hat{n} \times \mathbf{E})\, dS \\ &\qquad - \iiint_V \mathbf{T} \cdot \left[ik_0\sqrt{\mu_0/\varepsilon_0}\mathbf{J} + \nabla \times (\mu_0/\mu_{ij} \cdot \mathbf{M})\right] dV. \end{aligned} \quad (1.11)$$

The volume of integration "V" is then meshed into subregions using trapezoidal, tetrahedral, or other types of meshing schemes. One example of this is shown in Fig. 1.5(b), where the conformal nature of these meshes is especially useful with curved surfaces such as the corners of a bow-tie antenna.[2] To find a solution to the electromagnetics problem posed in Eq. (1.11), the **E**-field tangent to each edge of an individual meshing cell is calculated, and a set of basis vectors are used to extrapolate the resulting fields throughout the remaining simulation volume. The **E**-field within the entire structure is then given by

$$\mathbf{E} = \sum_{k=1}^{R_{\max}} \mathbf{R_k} E_k, \quad (1.12)$$

where **R** is the vector field component along a given meshing cell edge, $R_{\max}$ is the total number of edges within the simulation with the exception of S_{surf}, and E_k is the tangential electric field component along the same edge. When the basis vectors, **R**,

[2] Figure 1.5(b) was produced using the resources of MIT Lincoln Laboratory.

are the same as those in the test function, **T**, we can combine Eqs. (1.11) and (1.12) to obtain the discretized, Galerkin formulation of the electromagnetics problem for a single frequency:

$$\sum_{h,k=1}^{R_{\max}} \mathcal{M}_{hk} E_k = -\iiint\limits_V \mathbf{R_h} \cdot \left[i k_0 \sqrt{\mu_0/\varepsilon_0} \mathbf{J} + \nabla \times (\mu_0/\mu_{ij} \cdot \mathbf{M}) \right] dV, \qquad (1.13)$$

where $\mathcal{M}_{hk}$ is given by

$$\mathcal{M}_{hk} = \iiint\limits_V \left[(\nabla \times \mathbf{R_h}) \cdot \mu_0/\mu_{ij} \cdot (\nabla \times \mathbf{R_k}) - k_0^2 \mathbf{R_h} \cdot \varepsilon_{ij} \cdot \mathbf{R_k} \right] dV$$
$$- i k_0 \oiint\limits_{S_B} (\hat{n} \times \mathbf{R_h}) \cdot (\hat{n} \times \mathbf{R_k}) \, dS. \qquad (1.14)$$

1.3.3.3 FDTD and FEM

While both FDTD and FEM accurately model the three-dimensional response of structures to incident electromagnetic radiation, a number of similarities and differences can be seen. FDTD is based on the relatively straightforward implementation of the Yee algorithm. The structure is excited using a broad-band pulse and, as a result, Fourier transforming the fields provides broad-band information with a single simulation. As a result, the hardware limitations of FDTD are based more on the speed and number of processors, rather than the amount of RAM available. The amount of RAM required is determined by the size of the simulation, density of the mesh, and amount of data being stored throughout the simulation. The method is highly efficient, and there is no large matrix to invert, as with FEM. The method is highly parallelizable, and can easily handle both anisotropic and inhomogeneous structures, including nonlinear and dispersive media. However, the method is naturally based on a rectangular grid. As a result, accommodating curved or highly dynamic surfaces is a limitation, even with recent advances in conformal meshing techniques. Further, a new broadband simulation is required every time the excitation conditions are changed.

In comparison, FEM is a more complicated approach, in terms of formulation, meshing, and computation. Proper mesh generation and boundary truncation can be a significant challenge. Also, to arrive at a solution for a given frequency, a system of linear equations need to be solved. The large matrix inversions required can result in substantial computational requirements; however, recent advances in sparse matrix solvers can be incorporated to significantly reduce these issues. As a result, the hardware limitations of FEM are based more on the amount of RAM needed to complete the matrix inversion. While techniques exist to mitigate this problem, the matrices involved scale with the number of degrees of freedom and the meshing density and can quickly grow to the point where sizable amounts of RAM are required to obtain any solution. The resulting solution is independent of excitation, and once the matrix is inverted, it is fairly easy to find other solutions.

Like FDTD, this technique can be highly parallelizable by simply running simulations at different frequencies on different processors. Also, while it is more complex to generate the triangular or tetrahedral meshing, the resulting mesh is conformal. Hence, the method excels with curved or highly dynamic surfaces. From a design and optimization standpoint, when the simulated structure requires only one illumination condition over a wide range of frequencies, FDTD is usually the stronger method. When the frequency range is narrow, or even limited to a single frequency, but a wide range of illumination angles and conditions are required, FEM is usually the stronger method.

Finally, it should also be mentioned that while FEM and FDTD are two of the main methods used for this type of electromagnetic analysis, they are by no means the only methods. In specific situations, other techniques such as: Rigorous Coupled Waveguide Analysis, the Method of Moments, the Boundary Element Method, and others are also utilized to solve design problems; and when studying radio frequency or radar designs, may actually be more applicable than FDTD or FEM. In the end, the design optimization methods discussed throughout the rest of this book should be applicable to any of these design problems and able to be combined with any of these simulation methods.

1.3.4 The Union of Fields

The key developments within the fields of Simulation (FDTD and FEM), Fabrication (SEBL and FIB), and Publications that have been discussed throughout Sect. 1.3 of this chapter, are listed in Table 1.1 along with other relevant developments within these fields. While developments within all three areas date back to the 1960s, it wasn't until the turn of the century that the field started growing into what we know today. This coincides with the first commercial dual beam SEM/FIB, the first releases of commercial FDTD and method-of-moments-based FEM solvers, and publications by Pendry and Smith et al. [57, 71]. The union of these fields allowed a rigorous study of the electromagnetic resonances that occur within a wide range of today's metamaterials that operate at infrared and visible frequencies.

Just as the focus of metamaterials/Frequency Selective Surfaces that operate at microwave and radio frequencies has largely been on device applications, efforts within the field of metamaterials that operate at terahertz, infrared, and visible frequencies will become increasingly applications driven. To that end, advances in the field of metamaterial design will come about by manipulating fabrication and simulation capabilities in new and different ways.

Table 1.1 Key developments within the areas of simulation, fabrication, and publication as they relate to the study of optical metamaterials

Year	FEM	FDTD	SEBL	FIB	Pubs	Development
1943	x					Courant introduces FEM for the study of elasticity and structural mechanics [14]
1960	x					Clough coins the term "Finite Element Method"
1960			x			Möllenstedt and Speidel report writing "fine lines" with a 20 nm electron beam in collodium film [46]
1962			x			Philips, Eindhoven, introduces the first commercial Gaussian beam system
1964	x					Control Data Systems releases the first commercial FEM software for the study of mechanics
1966		x				Yee introduces the FDTD method for solving Maxwell's curl equations on a discretized grid [85]
1967					x	Victor Veselago publishes his manuscript on substances with negative values of ε and μ [81]
1969			x			Hatzakis introduces PMMA as an electron beam resist [28]
1969	x					Silvester publishes the first article using the Finite Element Method to study electromagnetics [70]
1974			x			Carl Zeiss, Jena, launches the first commercial shaped electron beam lithography system
1975				x		Levi-Setti, Orloff, and Swanson introduce the first field emission Focused Ion Beam systems [22, 56]
1979				x		Seliger introduces the first Focused Ion Beam based on the liquid metal ion source [66]
1980		x				Taflove coins the term "FDTD" acronym [75]
1980			x			Gaussian systems first introduced for low-resolution pattern generation in large scale manufacturing [3]
1987	x					Greengard and Rokhlin introduce the Fast Multipole Method [26]
1988				x		Sudraud et al. first use dual-beam SEM/FIB for microcircuit repair [73]
1994		x				Berenger and Katz introduce perfectly matched boundary layers (PMLs) for FDTD [5, 35]
1994		x				Reuter introduces modeling of dispersive materials within FDTD [61]
1995	x	x				Wu and Itoh introduce the first combination of FDTD and FEM for objects with curved boundaries [84]
1998				x		FEI releases its first commercial dual-beam Scanning Electron Microscope / Focused Ion Beam
1999		x				Panoramic Technology releases the first commercial rigorous FDTD electromagnetics solver

Table 1.1 (Continued)

Year	FEM	FDTD	SEBL	FIB	Pubs	Development
2000					x	Pendry publishes "*Negative Refraction Makes a Perfect Lens*" [57]
2000					x	Smith and colleagues fabricate the first negative index material [71]
2004	x					FEKO releases the first commercial FEM software utilizing the Method of Moments solver
2006					x	Shurig et al. demonstrate the first electromagnetic invisibility cloak [64]

1.4 Design

As with any design problem, to understand the observable extrinsic properties, you must first understand the intrinsic properties of the constituent components of a system. In the case of optical metamaterials, we will combine dielectrics and metals. For the designs studied in this book, the dielectric acts as the host medium that supports either structured or unstructured arrays of metallic resonators. In the case of bulk metamaterials, where the resonant arrays are distributed throughout the volume of the structure, the dielectric needs to be transparent to the wavelengths of interest. Any significant amount of absorption would result in higher metamaterial losses and decreased device performance. In contrast, this requirement is relaxed when working with frequency selective surfaces, where the resonant array is patterned above the dielectric substrate. Here, substrates such as semiconductors are sometimes utilized to tune the local dielectric environment in unique ways that would otherwise not be available to bulk metamaterials.

As we will see in the following sections, materials behavior at these frequencies is dominated by free electrons. To provide both strong optical contrast between the metals and dielectrics, and minimize the contribution of the dielectric to the overall metamaterial performance, the electrons within the dielectric are tightly bound to the atomic lattice. Additionally, we will see that the surface plasmon resonances that play a major factor in the device performance, are only supported at interfaces between negative (metals) and positive dielectric constants.

1.4.1 Optical Properties of Metals

As the operation range of metamaterials has moved from radio and microwave to infrared, visible, and ultraviolet frequencies, the materials used to fabricate these structures have also changed. Structures such as frequency selective surfaces that operate at radio and microwave frequencies are oftentimes constructed using printed copper wires on top of printed circuit board material. At these frequencies, the metal can be treated as a perfect electrical conductor. The materials can be thought of as having infinite electrical conductivity and are without losses. The dielectric properties of the metal are fixed, and the penetration of the electric field below the surface of the metal is assumed to be negligible.

As we move to infrared and visible frequencies, this assumption breaks down and things become more complex. Here, the electromagnetic fields penetrate tens of nanometers into the metallic resonators, and materials' response is dominated by the behavior of the free electrons within a material. As a result, gold, silver, copper, and aluminum become the dominant materials used because their electron density and configuration are such that bulk plasma oscillations and surface plasmons can be supported. These materials' resonances are the result of significant dispersion throughout the frequency range of interest.

The resulting shift in size of the individual resonant structure is tens to hundreds of nanometers. At these dimensions, the field penetration depth into the metal resonators becomes a significant fraction of the overall thickness. Additionally, when the geometry of the meta-atoms is tuned to have resonances that coincide with the natural resonances of the metals, we observe significant electromagnetic field enhancement around the resonators, and striking bulk optical properties.

As with any material, we can describe the optical properties with a frequency-dependent, complex dielectric function. For the plasmonic materials mentioned above, we express the dielectric function in terms of both free-electron effects (ε_D) using the Drude–Sommerfeld model, and interband transitions ($\varepsilon_{\mathrm{IB}}$). Each of these effects will be discussed, and we then finish the section by discussing the surface plasmon effects that arise within these metals.

1.4.1.1 Drude Metals

Under illumination by a time-harmonic external electric field $\mathbf{E}_0\mathrm{e}^{-i\omega t}$, the equation of motion for free electrons in a metal is given by

$$m_D^* \frac{\partial^2 \mathbf{r}(t)}{\partial t^2} + m_D^* \frac{1}{\tau} \frac{\partial \mathbf{r}(t)}{\partial t} = e\mathbf{E}_0\mathrm{e}^{-i\omega t}, \tag{1.15}$$

where e and m_D^* are the charge and effective mass of the free electrons, and $\mathbf{r}$ is the displacement of an electron under an external field. τ is the average relaxation time of the free electrons. τ is proportional to $\tau = \frac{\ell}{v_F}$, where v_F is the Fermi velocity and ℓ is the electron mean free path. These values for aluminum, copper, silver, and gold are listed in Table 1.2. Solving for $\mathbf{r}$ gives

$$\mathbf{r} = \frac{e}{m} \frac{\mathbf{E}_0\mathrm{e}^{-i\omega t}}{(\omega^2 + i\omega/\tau)}. \tag{1.16}$$

Table 1.2 Drude and Drude–Sommerfeld values for plasmonic metals including the: plasma frequency ω_p [10], Fermi velocity v_F (cm/s) [2], Drude relaxation time τ_D [2], frequency of interband transitions ω_{IB} [13], and the electron configuration

	Relevant physical constants for plasmonic metals				
	ω_p (eV)	v_F (cm/s)	τ_D (10^{-14} s)	ω_{IB} (eV)	e^- configuration
Aluminum	15.1	2.03	0.80	1.41	[Ne]$3s^2 3p^1$
Copper	8.8	1.57	2.7	2.1	[Ar]$3d^{10} 4s^1$
Silver	9.2	1.39	4.0	3.9	[Kr]$4d^{10} 5s^1$
Gold	9.1	1.40	3.0	2.3	[Xe]$4f^{14} 5d^{10} 6s^1$

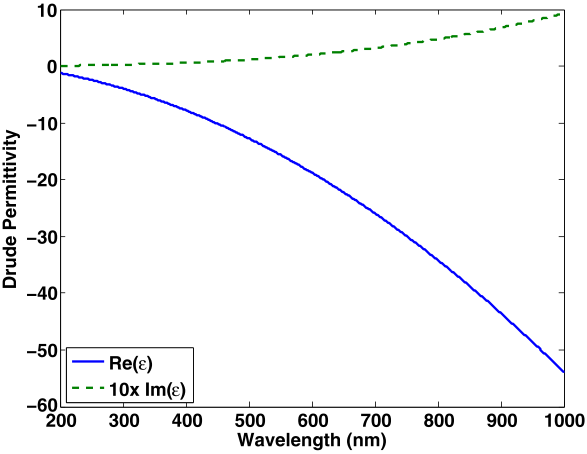


Fig. 1.6 Real (*solid blue*) and imaginary (*dashed green*) Drude components of the dielectric function of silver. In this plot, the imaginary permittivity has been scaled up by a factor of ten

Combining this result with Eq. (1.2), we obtain the complex Drude model for frequency-dependent permittivity:

$$\varepsilon_D(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\omega/\tau}. \tag{1.17}$$

Here, the term ω_p is the bulk plasma frequency given by $\omega_p = \sqrt{(ne^2)/(m_D^* \varepsilon_0)}$. Finally, we can separate Eq. (1.17) into its real and imaginary components:

$$\varepsilon_D(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + 1/\tau^2} + i\frac{\omega_p^2}{\omega\tau(\omega^2 + 1/\tau^2)}. \tag{1.18}$$

A plot of the real and imaginary Drude components of the dielectric function are shown in Fig. 1.6 for silver.[3] In this figure, the real part of the dielectric constant is shown as the solid blue line and the imaginary part is shown as the dashed green line.

[3] Figure 1.6 was produced using the resources of MIT Lincoln Laboratory.

Also, to plot both constants on the same y-axis, the imaginary part of the dielectric constant has been plotted as ten times its actual value. Here we see that the real part of the dielectric constant is negative across visible and infrared frequencies. This indicates that under external illumination, the electrons are driven 180° out of phase with the incident light. This results in the high reflectivity that is typically associated with metals. We also see a significant contribution from the imaginary part of the dielectric constant. The optical losses associated with these metals are an inherent limitation for certain types of metamaterial designs.

1.4.1.2 Interband Transitions

While the Drude–Sommerfeld model for metals provides a nice starting point for their understanding, it is by no means a complete explanation of their optical behavior. The fact that gold, silver, and copper refer to colors as well as metals clearly indicates that there's more going on than the model in the previous section can explain. The explanation for such effects lies with interband transitions.

Gold, silver, and copper are all monovalent, Face-Centered Cubic metals. For these noble metals, the Fermi surface of the metal strongly resembles a free electron sphere with the exception of the $\langle 111 \rangle$ direction, where the surface intersects the Brillouin zone face. From Table 1.2 we see that the electron configuration of all three have 10 electrons occupying the d-bands and 1 electron occupying the s-band. Additionally, all three metals have the fully occupied d-bands 2–4 eV below the s-band. As a result, absorption can occur when light above this interband transition energy is incident upon the surface of the metal. This explains why copper has a somewhat reddish appearance, gold appears to be yellow, and silver strongly reflects across the entire visible spectrum.

To model the contribution of interband transitions to the overall dielectric function, we modify Eq. (1.15) to include damping from bound electrons γ, and the electron restoring force α:

$$m_B \frac{\partial^2 \mathbf{r}(t)}{\partial t^2} + m_B \gamma \frac{\partial \mathbf{r}(t)}{\partial t} + \alpha \mathbf{r} = e\mathbf{E}_0 \mathrm{e}^{-i\omega t}, \tag{1.19}$$

where m_B is the mass of bound electrons. Solving Eq. (1.19) following the same method as in Sect. 1.4.1.2, we arrive at

$$\varepsilon_{\mathrm{IB}}(\omega) = 1 - \frac{\tilde{\omega}_p^2}{(\omega_0^2 - \omega^2) - i\gamma\omega}. \tag{1.20}$$

Here, the term $\tilde{\omega}_p$ is the Drude–Sommerfeld plasma frequency given by $\tilde{\omega}_p = \sqrt{(\tilde{n}e^2)/(m_B\varepsilon_0)}$, $\tilde{n}$ is the concentration of bound electrons, and, $\omega_0 = \sqrt{\alpha/m_B}$. In a similar manner to Eq. (1.18), we can separate Eq. (1.20) into its real and imaginary components:

$$\varepsilon_{\mathrm{IB}}(\omega) = 1 - \frac{\tilde{\omega}_p^2(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2} + i\frac{\tilde{\omega}_p^2\omega\gamma}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2}. \tag{1.21}$$

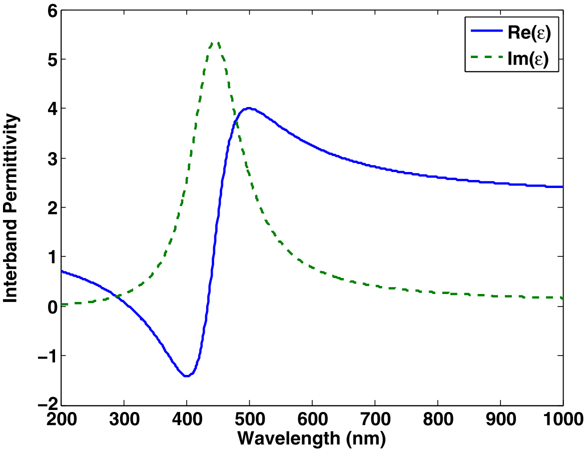


Fig. 1.7 Real (*solid blue*) and imaginary (*dashed green*) interband components of the dielectric function of gold

Plots of the real and imaginary contributions to the dielectric constant of gold are shown in Fig. 1.7. In this figure, the real part of the dielectric constant is shown as the solid blue line and the imaginary part is shown as the dashed green line.[4] Here the interband transitions can clearly be seen as spikes in ε_2. Finally, at frequencies far from where interband transitions occur, these effects continue to have an influence on the overall dielectric function of the material. This manifests itself as a constant offset term in the overall dielectric function. Typical values of this offset ε_∞ for gold are between 6.5 and 9 and for silver are between 4.5 and 5.

1.4.1.3 Dispersion and Surface Plasmons

Separate from bulk plasmons within the metals mentioned above are a type of electron density oscillation at the interface between a metal and a dielectric. These resonances are known as surface plasmons, and play a significant role on the overall behavior of optical metamaterials that operate at infrared, visible, and ultraviolet frequencies. In addition, when these oscillations propagate along the metal surface in the form of a guided wave, they are referred to as surface plasmon polaritons (SPPs).

Even though ε and $\tilde{n}$ are referred to as constants, we know that at optical frequencies these properties can vary significantly, depending on the configurations in which they are used as well as the frequency of the light involved. This property of materials is known as dispersion. To calculate the dispersion of these structures, we start with an incident electromagnetic wave of the form [16]:

$$\mathbf{E}(x, y, z) = E_0 \mathrm{e}^{i(k_x x - k_z |z| - \omega t)} \tag{1.22}$$

whose electric field has a perpendicular component to the waveguide (transverse-magnetic polarization). Here the components of the electric field within the metal

[4] Figure 1.7 was produced using the resources of MIT Lincoln Laboratory.

are given by:

$$E_x^{\text{metal}} = E_0 e^{i(k_x x - k_{z1}|z| - \omega t)}, \tag{1.23a}$$

$$E_y^{\text{metal}} = 0, \tag{1.23b}$$

$$E_z^{\text{metal}} = E_0 \left(\frac{-k_x}{k_{z1}} \right) e^{i(k_x x - k_{z1}|z| - \omega t)}, \tag{1.23c}$$

and the components of the electric field within the dielectric are given by:

$$E_x^{\text{dielectric}} = E_0 e^{i(k_x x - k_{z2}|z| - \omega t)}, \tag{1.24a}$$

$$E_y^{\text{dielectric}} = 0, \tag{1.24b}$$

$$E_z^{\text{dielectric}} = E_0 \left(\frac{-\varepsilon_1 k_x}{\varepsilon_2 k_{z1}} \right) e^{i(k_x x - k_{z2}|z| - \omega t)}, \tag{1.24c}$$

where k_{z1} and ε_1 represent the wave vector and dielectric constant within the metal layer, and k_{z2} and ε_2 represent the wavevector and dielectric constant within the dielectric layer, respectively. For both sets of equations, k_x represents the component of the wave vector in the direction of propagation along the metal-dielectric interface. Similarly, k_z represents the component of the wave vector perpendicular to the metal-dielectric interface, and from this we obtain the decay length of the electro-magnetic field into the layers, or the "skin depth":

$$\hat{z} = \frac{1}{|k_z|}. \tag{1.25}$$

Note that for metamaterial structures with thicknesses on the order of twice the skin depth, interactions between both surfaces can occur and further modify the behavior of the individual resonant structure. By requiring continuity of the **E** and **B** fields at the interface between the two layers, we obtain the dispersion relation for a single metal–dielectric interface [42, 74]:

$$k_x = \frac{\omega}{c} n_{\text{spp}}, \tag{1.26a}$$

$$k_{z1,2}^2 = \varepsilon_{1,2} \left(\frac{\omega}{c} \right)^2 - k_x^2, \tag{1.26b}$$

where the effective surface plasmon index is given by

$$n_{\text{spp}} = \sqrt{\frac{\varepsilon_1 \varepsilon_2}{\varepsilon_1 + \varepsilon_2}}. \tag{1.27}$$

These relations show an exponential decay into both the metal and dielectric, although the decay is much shorter into the metal. Additionally, these relations are for ideal metals with no defects. As the size of the individual resonant element within the metamaterial is decreased, grain boundary and surface roughness scattering will play an increasing role in the performance of the device. This effect, along with the decreased size of the total structure, manifests itself in the form of a modified scattering time [10].

1.4.2 Current Designs

While advances in fabrication and simulation capabilities have allowed the operating frequency of optical metamaterials to increase over the past few decades, it is interesting to note that many of the most prominently studied designs within the field continue to be variations on structures adopted from radio and microwave frequency antenna design. Such structures include bow tie antennas [38, 44], dipole antennas [41, 47, 55], fishnet structures [43, 80], and perhaps the best example of this, the split-ring resonator (SRR).[5] These four structures are shown in Fig. 1.8. With structures such as dipole antenna, the individual resonators are basic enough that the resonance can be calculated using either full-field electromagnetic simulations, or obtained analytically using a basic LC circuit model; however, we see from the literature that variations in the constituent materials, geometrical parameters, host medium, and three-dimensional array layout quickly increase the complexity of the design to the point where full-field electromagnetic simulations are required.

As is often the case with these structures and studies, a combination of fabricated samples and full-field electromagnetic simulations sweep through a few of the critical design parameters and analyze how the resulting resonator response is affected. This may then be followed by highlighting the optimized structure to take advantage of the resonance under consideration. When the number of parameters under consideration is small, and the question is "how does each design variant change the overall metamaterial response," this is certainly a valuable and viable approach; however, as the primary focus shifts to optimizing resonances for a given application and the number of parameters increases, it quickly becomes apparent that from a time standpoint, this exhaustive approach is no longer feasible. At this point in the design process, we arrive at the central question of this book:

> What is the most accurate and efficient way to tailor the broadband optical properties of a metamaterial to have predetermined responses at predetermined wavelengths?

Throughout the rest of the book, we address one answer to this question. By combining numerical optimization methods with full-field electromagnetic simulations, we are able to explore high-dimensional design spaces, orders of magnitude faster than performing traditional parameter sweeps. Using this approach, the researcher determines the design parameters to be varied, along with the range of interest for

[5] Figure 1.8 was produced using the resources of MIT Lincoln Laboratory.

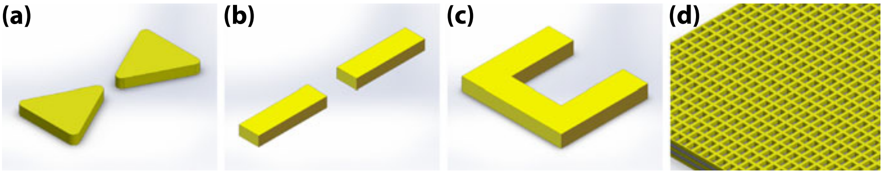


Fig. 1.8 Some of the most common individual metamaterial resonators including a bow tie antenna (**a**), a dipole antenna (**b**), a split ring resonator (**c**), and fishnet metamaterials (**d**)

each parameter. The optimization routine then steps through a simplex of test points. For each point, the program executes a function call by sending the metamaterial design to an electromagnetic solver, and then extracts the relevant figure(s) of merit. The figure(s) of merit are then combined based on a user defined "*cost function*" or "*objective function*" to rank the metamaterial design with respect to all other designs.

The range of optimization routines that can be used for this approach span the entire spectrum. Surrogate optimization methods, such as Curiosity Driven Optimization, choose test points in an effort to generate a maximally predictive, minimally complex model of the response of every possible geometrical variation within the specified design space (see Chap. 3). Gradient-free optimization techniques, such as Mesh Adaptive Search algorithms, are extremely robust in terms of their ability to survey non-smooth "parameter space," and based on specified criteria of convergence, can do a remarkable job of finding global optimum designs (see Chap. 4). Evolutionary algorithms, such as Particle Swarm Optimization and Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES), rely on evaluating sets of test points and based on the results, permuting the sets to generate different, and hopefully better, sets of geometrical solutions (see Chap. 5). Finally, in much the same way genetic algorithms mutate their solution set to develop new, better solution sets; new optimization methods are always being developed. Conjugate gradient methods are being combined with objective-first designs, which start with the desired electromagnetic fields, and work backwards to calculate the required dielectric distribution (see Chap. 6). Level Set Methods, which are computational techniques traditionally applied to fluid dynamics, have already shown promise for designing photonic crystals and are now being explored for metamaterial applications (see Chap. 7). Finally, when all else fails, the Black Box Optimization Benchmarking has established a yearly workshop to assess the performance of newly developed optimizers to understand their strengths and weaknesses, and this organization is a constant source of new and different ideas [1].

Finally, while the techniques mentioned in the previous paragraph summarize the extent to which avenues of metamaterial design optimization are covered in this text; everything here, as well as most work in the literature, has focused on selecting a specific material for the resonator design and then using geometrical permutations to obtain optimized or novel device performance. While this is certainly a rich field of study, one can imagine other avenues by which new metamaterial designs can be achieved. One such avenue that is receiving increased attention is described in Sect. 1.4.3.

1.4.3 Future Designs

Throughout the history of optical metamaterials, gold, silver, and copper have been the dominant materials used. This is in large part because in these metals, the free electrons necessary to support plasmon resonances are in high enough concentrations to resonate at near-infrared, visible, and ultraviolet frequencies. Unfortunately, the same resonances that give these exotic optical properties introduce high losses and limit the overall performance of devices. This limitation with traditional plasmonic materials has provided an opportunity for both alternative plasmonic materials, as well as additional design degrees of freedom, by tuning their resonant frequencies [7].

In recent years, a variety of material sets have been proposed as alternative plasmonic materials including doped semiconductors [30, 52, 77], intermetallics [6], transparent conducting oxides [23, 53, 83], transition metal nitrides [53], and graphene [32]. One material set in particular, Transparent Conducting Oxides (TCOs), have shown significant tunability across the near-infrared spectrum by varying the concentration of oxygen vacancies and interstitial metal dopants introduced into the films during deposition. These materials, including aluminum zinc oxide, indium zinc oxide, and indium tin oxide have primarily been used as components in touch screen displays; however, their low losses (five times smaller than silver) [51, 54], tunability, and compatibility with standard fabrication processes have resulted in increasing attention from the plasmonics and metamaterials communities. From a design and optimization standpoint, they offer another interesting benefit. From Sect. 1.4.1.1 we know that the Drude dielectric constant is given by:

$$\varepsilon = 1 - \frac{\omega_p^2}{\omega^2 + i\omega/\tau},$$

$$\omega_p^2 = \frac{ne^2}{\varepsilon_\infty m^*}.$$

TCOs, such as indium tin oxide or indium zinc oxide, can typically be doped to have carrier concentrations between 10^{19}–10^{21} cm^{-3}. Based on this model, Fig. 1.9 shows that by adjusting the carrier concentration within the material during deposition, we can tune the plasma frequency ($\varepsilon = 0$) across the near-infrared spectrum.

To date, virtually all optimized metamaterial design has focused on parametrically tuning the topology of the metamaterial unit cell, and a given material with preset electronic and optical properties is chosen in a binary manor. With the introduction of TCOs as alternative plasmonic materials for metamaterial design, we can now include the resonant frequencies of the material itself as another design parameter to be optimized. This can be taken one step further, by considering metamaterial

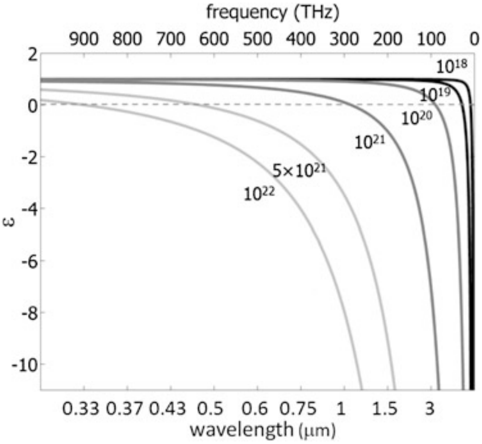


Fig. 1.9 Permittivity dispersion modified by a change in the carrier concentration. As the carrier density (per cubic centimeter) increases, the plasma frequency ($\varepsilon = 0$) shifts toward visible frequencies, and the dispersion becomes substantially different in that regime. Reprinted with permission from E. Feigenbaum et al.,"Unity-Order Index Change in Transparent Conducting Oxides at Visible Frequencies," *Nano Letters* **10**, 2111–2116 (2010). Copyright 2010 American Chemical Society

designs where the doping concentration and resulting plasma frequency are shifted as a function of resonator thickness. These additional design degrees of freedom present an interesting opportunity for future metamaterial designs, and are left as an exercise for the reader.

References

1. Comparing continuous optimisers: Coco. http://coco.gforge.inria.fr/doku.php?id=start
2. N.W. Ashcroft, N.D. Mermin, *Solid State Physics* (Brooks/Cole, Pacific Grove, 1976)
3. J.P. Ballantyne, Mask fabrication by electron-beam lithography, in *Electron-Beam Technology in Microelectronic Fabrication*, ed. by G.R. Brewer (Academic Press, New York, 1980), pp. 259–307
4. D.J. Barber, I.C. Freestone, An investigation of the origin of the color of the Lycurgus cup by analytical transmission electron-microscopy. Archaeometry **32**, 33–45 (1990)
5. J.P. Berenger, A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**, 185–200 (1994)
6. M.G. Blaber, M.D. Arnold, M.J. Ford, Optical properties of intermetallic compounds from first principles calculations: a search for the ideal plasmonic material. J. Phys. Condens. Matter **21**, 144211 (2009)
7. A. Boltasseva, H.A. Atwater, Low-loss plasmonic metamaterials. Science **331**(6015), 290–291 (2011)
8. J.C. Bose, On the rotation of plane polarization of electric waves by a twisted structure. Proc. R. Soc. Lond. **63**, 146–152 (1898)
9. W. Cai, U.K. Chettiar, H.K. Yuan, V.C. de Silva, A.K. Sarychev, A.V. Kildishev, V.P. Drachev, V.M. Shalaev, Metamagnetics with rainbow colors. Opt. Express **15**(6), 3333–3341 (2007)

10. W. Cai, V. Shalaev, *Optical Metamaterials: Fundamentals and Applications* (Springer, Berlin, 2010)
11. W.S. Cai, U.K. Chettiar, A.V. Kildishev, V.M. Shalaev, Optical cloaking with metamaterials. Nat. Photonics **1**(4), 224–227 (2007)
12. Y.F. Chen, J.R. Tao, X.Z. Zhao, Z. Cui, A.S. Schwanecke, N.I. Zheludev, Nanoimprint lithography for planar chiral photonics meta-materials. Microelectron. Eng. **78–79**, 612–617 (2005)
13. B.R. Cooper, H. Ehrenreich, H.R. Philipp, Optical properties of Nobel metals 2. Phys. Rev. **138**, 494–507 (1965)
14. R.L. Courant, Variational methods for the solution of problems of equilibrium and vibration. Bull. Am. Math. Soc. **49**, 1–23 (1943)
15. M. Deubel, G. Von Freymann, M. Wegener, S. Pereira, K. Busch, C.M. Soukoulis, Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. Nat. Mater. **3**(7), 444–447 (2004)
16. K. Diest, Active metal–insulator–metal plasmonic devices. Ph.D. thesis, California Institute of Technology, September 2012
17. N. Engheta, Circuits with light at nanoscales: optical nanocircuits inspired by metamaterials. Science **317**(5845), 1698–1702 (2007)
18. N. Engheta, A. Salandrino, A. Alu, Circuit elements at optical frequencies: nanoinductors, nanocapacitors, and nanoresistors. Phys. Rev. Lett. **95**, 095504 (2005)
19. N. Engheta, R.W. Ziolkowski, *Metamaterials: Physica and Engineering Explorations* (IEEE Press, New York, 2006)
20. C. Enkrich, R. Perez-Willard, D. Gerthsen, J.F. Zhou, T. Koschny, C.M. Soukoulis, M. Wegener, Focused-ion-beam nanofabrication of near-infrared magnetic metamaterials. Adv. Mater. **17**, 2547–2549 (2005)
21. T. Ergin, N. Stenger, P. Brenner, J.B. Pendry, M. Wegener, Three-dimensional invisibility cloak at optical wavelengths. Science **328**(5976), 337–339 (2010)
22. W.H. Escovitz, T.R. Fox, R. Levi-Setti, Scanning-transmission ion-microscope with a field-ion source. Proc. Natl. Acad. Sci. USA **72**(5), 1826–1828 (1975)
23. E. Feigenbaum, K. Diest, H.A. Atwater, Unity-order index change in transparent conducting oxides at visible frequencies. Nano Lett. **10**, 2111–2116 (2010)
24. J.C.M. Garnett, Colours in metal glasses and in metallic films. Philos. Trans. R. Soc. Lond. A **203**, 385–420 (1904)
25. D.H. Gracias, J. Tien, T.L. Breen, C. Hsu, G.M. Whitesides, Forming electrical networks in three dimensions by self-assembly. Science **289**(5482), 1170–1172 (2000)
26. L. Greengard, V. Rokhlin, A fast algorithm for particle simulations. J. Comput. Phys. **73**, 325–348 (1987)
27. L.J. Guo, Nanoimprint lithography: methods and material requirements. Adv. Mater. **19**, 495–513 (2007)
28. M. Hatzakis, Electron resists for microcircuit and mask production. J. Electrochem. Soc. **116**, 1033–1037 (1969)
29. M.D. Henry, M.J. Shearn, B. Chhim, A. Scherer, Ga^+ beam lithography for nanoscale silicon reactive ion etching. Nanotechnology **21**, 245303 (2010)
30. A.J. Hoffman, L. Alekseyev, S.S. Howard, K.J. Franz, D. Wasserman, V.A. Podolskiy, E.E. Narimanov, D.L. Sivco, C. Gmachl, Negative refraction in semiconductor metamaterials. Nat. Mater. **6**(12), 946–950 (2007)
31. J. ibn Hayyan, *The Book of the Hidden Pearl*
32. M. Jablan, H. Buljan, M. Soljacic, Plasmonics in graphene at infrared frequencies. Phys. Rev. B **80**, 245435 (2009)
33. J.M. Jin, *The Finite Element Method in Electromagnetics*, 2nd edn. (Wiley, Hoboken, 2002)
34. J.M. Jin, D.J. Riley, *Finite Element Analysis of Antennas and Arrays* (Wiley-IEEE Press, Hoboken, 2009)
35. D.S. Katz, E.T. Thiele, A. Taflove, Validation and extension to three dimensions of the Berenger PML absorbing boundary condition for FD–TD meshes. IEEE Microw. Guided Wave Lett. **4**(8), 268–270 (1994)

36. M.W. Klein, C. Enkrich, M. Wegener, S. Linden, Second-harmonic generation from magnetic metamaterials. Science **313**(5786), 502–504 (2006)
37. M.W. Klein, M. Wegener, N. Feth, S. Linden, Experiments on second- and third-harmonic generation from magnetic metamaterials. Opt. Express **15**(8), 5238–5247 (2007)
38. A.L. Koh, A.I. Fernandez-Dominguez, D.W. McComb, S.A. Maier, J.K.W. Yang, High-resolution mapping of electron-beam-excited plasmon modes in lithographically defined gold nanostructures. Nano Lett. **11**(3), 1323–1330 (2011)
39. U. Leonhardt, T.G. Philbin, General relativity in electrical engineering. New J. Phys. **8**, 247 (2006)
40. R. Liboff, *Introductory Quantum Mechanics*, 4th edn. (Addison-Wesley, Reading, 2002)
41. N. Liu, L. Langguth, T. Weiss, J. Kastel, M. Fleischhauer, T. Pfau, H. Giessen, Plasmonic analogue of electromagnetically induced transparency at the Drude damping limit. Nat. Mater. **8**(9), 758–762 (2009)
42. S. Maier, *Plasmonics: Fundamentals and Applications* (Springer, Berlin, 2007)
43. A. Mary, S.G. Rodrigo, F.J. Garcia-Vidal, L. Martin-Moreno, Theory of negative-refractive-index response of double-fishnet structures. Phys. Rev. Lett. **101**, 103902 (2008)
44. J. Merlein, M. Kahl, A. Zuschlag, A. Sell, A. Halm, J. Boneberg, P. Leiderer, A. Leitenstorfer, R. Bratschitsch, Nanomechanical control of an optical antenna. Nat. Photonics **2**(4), 230–233 (2008)
45. G. Mie, Articles on the optical characteristics of turbid tubes, especially colloidal metal solutions. Ann. Phys. **25**(3), 377–445 (1908)
46. G. Moellenstedt, R. Speidel, Elektronenoptischer Mikroschreiber unter Elektronenmikroskopischer Arbeitskontrolle. Phys. Bl. **16**, 192 (1960)
47. P. Muhlschlegel, H.J. Eisler, O.J.F. Martin, B. Hecht, D.W. Pohl, Resonant optical antennas. Science **308**(5728), 1607–1609 (2005)
48. B.A. Munk, *Frequency Selective Surfaces: Theory and Design* (Wiley-Interscience, New York, 2000)
49. B.A. Munk, *Finite Antenna Arrays and FSS* (Wiley/IEEE Press, New York, 2003)
50. B.A. Munk, G.A. Burrell, Plane-wave expansion for arrays of arbitrarily oriented piecewise linear elements and its application in determining the impedance of a single linear antenna in a lossy half-space. IEEE Trans. Antennas Propag. **27**(3), 331–343 (1979)
51. G. Naik, A. Boltasseva, A comparative study of semiconductor-based plasmonic metamaterials. Metamaterials **5**, 1–7 (2011)
52. G.V. Naik, A. Boltasseva, Semiconductors for plasmonics and metamaterials. Phys. Status Solidi RRL **4**(10), 295–297 (2010)
53. G.V. Naik, J. Kim, A. Boltasseva, Oxides and nitrides as alternative plasmonic materials in the optical range. Opt. Mater. Express **1**(6), 1090–1099 (2011)
54. M. Noginov, L. Gu, J. Livenere, G. Zhu, A. Pradhan, R. Mundle, M. Bahoura, Y. Barnakov, V. Podolskiy, Transparent conductive oxides: plasmonic materials for Telecom wavelengths. Appl. Phys. Lett. **99**, 021101 (2011)
55. D.M. O'Carroll, C.E. Hofmann, H.A. Atwater, Conjugated polymer/metal nanowire heterostructure plasmonic antennas. Adv. Mater. **22**(11), 1223 (2010)
56. J.H. Orloff, L.W. Swanson, Study of a field-ionization source for microprobe applications. J. Vac. Sci. Technol. **12**(6), 1209–1213 (1975)
57. J.B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**(18), 3966–3969 (2000)
58. J.B. Pendry, D. Schurig, D.R. Smith, Controlling electromagnetic fields. Science **312**(5781), 1780–1782 (2006)
59. E. Plum, J. Zhou, J. Dong, V.A. Fedotov, T. Koschny, C.M. Soukoulis, N.I. Zheludev, Metamaterial with negative index due to chirality. Phys. Rev. B **79**, 035407 (2009)
60. A.K. Popov, V.M. Shalaev, Compensating losses in negative-index metamaterials by optical parametric amplification. Opt. Lett. **31**(14), 2169–2171 (2006)

61. C.E. Reuter, R.M. Joseph, E.T. Thiele, D.S. Katz, A. Taflove, Ultrawideband absorbing boundary condition for termination of waveguide structures in FD–TD simulations. IEEE Microw. Guided Wave Lett. **4**(10), 344–346 (1994)
62. M.M.I. Saadoun, N. Engheta, A reciprocal phase-shifter using novel pseudochiral or omega-medium. Microw. Opt. Technol. Lett. **5**(4), 184–188 (1992)
63. S.L. Sass, *The Substance of Civilization* (Arcade, New York, 1998)
64. 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**, 977–980 (2006)
65. D. Schurig, J.B. Pendry, D.R. Smith, Calculation of material properties and ray tracing in transformation media. Opt. Express **14**(21), 9794–9804 (2006)
66. R. Seliger, J.W. Ward, V. Wang, R.L. Kubena, A high-intensity scanning ion probe with submicrometer spot size. Appl. Phys. Lett. **34**(5), 310–312 (1979)
67. V.M. Shalaev, W.S. Cai, U.K. Chettiar, H.K. Yuan, A.K. Sarychev, V.P. Drachev, A.V. Kildishev, Negative index of refraction in optical metamaterials. Opt. Lett. **30**(24), 3356–3358 (2005)
68. E.V. Shevchenko, D.V. Talapin, N.A. Kotov, S. O'Brien, C.B. Murray, Structural diversity in binary nanoparticle superlattices. Nature **439**(7072), 55–59 (2006)
69. M. Silveirinha, N. Engheta, Tunneling of electromagnetic energy through subwavelength channels and bends using epsilon-near-zero materials. Phys. Rev. Lett. **97**, 157403 (2006)
70. P.P. Silvester, Finite element solution of homogeneous waveguide problems. Alta Freq. **38**, 313–317 (1969)
71. D.R. Smith, W.J. Padilla, D.C. Vier, S.C. Nemat-Nasser, S. Schultz, Composite medium with simultaneously negative permeability and permittivity. Phys. Rev. Lett. **84**, 4184–4187 (2000)
72. C.M. Soukoulis, S. Linden, M. Wegener, Negative refractive index at optical wavelengths. Science **315**(5808), 47–49 (2007)
73. P. Sudraud, G. Assayag, M. Bon, Focused ion beam milling, scanning electron microscopy, and focused droplet deposition in a single microsurgery tool. J. Vac. Sci. Technol. B **6**, 234–238 (1988)
74. L.A. Sweatlock, Plasmonics: numerical methods and device applications. Ph.D. thesis, California Institute of Technology, 2008
75. A. Taflove, Application of the finite-difference time-domain method to sinusoidal steady-state electromagnetic penetration problems. IEEE Trans. Electromagn. Compat. **22**, 191–202 (1980)
76. A. Taflove, S.C. Hagness, *Computational Electromagnetics: The Finite-Difference Time-Domain Method*, 3rd edn. (Artech House, Norwood, 2005)
77. T. Taubner, D. Korobkin, Y. Urzhumov, G. Shvets, R. Hillenbrand, Near-field microscopy through a sic superlens. Science **313**(5793), 1595 (2006)
78. M. Thiel, H. Fischer, G.V. Freymann, M. Wegener, Three-dimensional chiral photonic superlatticies. Opt. Lett. **35**(2), 166 (2010)
79. J. Valentine, J.S. Li, T. Zentgraf, G. Bartal, X. Zhang, An optical cloak made of dielectrics. Nat. Mater. **8**(7), 568–571 (2009)
80. J. Valentine, S. Zhang, T. Zentgraf, E. Ulin-Avila, D.A. Genov, G. Bartal, X. Zhang, Three-dimensional optical metamaterial with a negative refractive index. Nature **455**, 376–379 (2008)
81. V.G. Veselago, The electrodynamics of substances with simultaneously negative values of epsilon and mu. Sov. Phys. Usp. **10**(4), 509–514 (1968)
82. A.J. Ward, J.B. Pendry, Refraction and geometry in Maxwell's equations. J. Mod. Opt. **43**(4), 773–793 (1996)
83. P. West, S. Ishii, G. Naik, N. Emani, V.M. Shalaev, A. Boltasseva, Searching for better plasmonic materials. Laser Photonics Rev. **4**(6), 795–808 (2010)
84. R.B. Wu, T. Itoh, Hybridizing FDTD analysis with unconditionally stable FEM for objects of curved boundary, in *IEEE Microwave Theory and Techniques Society Symposium Digest*, vol. 2 (1995), pp. 833–836

85. S.K. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 302–307 (1966)
86. X. Yu, Y.J. Lee, R. Furstenberg, J.O. White, P.V. Braun, Filling fraction dependent properties of inverse opal metallic photonic crystals. Adv. Mater. **19**, 1689–1692 (2007)
87. A.A. Zharov, I.V. Shadrivov, Y.S. Kivshar, Nonlinear properties of left-handed metamaterials. Phys. Rev. Lett. **91**, 037401 (2003)

Chapter 2
An Overview of Mathematical Methods for Numerical Optimization

Daniel E. Marthaler

Abstract This chapter serves as a basic overview of mathematical optimization problems and reviews how certain classes of these problems are solved. For the general category of nonlinear problems, both smooth and nonsmooth "*Derivative Free*" topics are discussed with and without constraints.

2.1 Introduction

This book is concerned with finding the "best" solution to particular metamaterial design problems. Best is put in quotations because the idea of what represents a good design is defined by the user, and very much depends on the application. The best design for some problems may be the one that reflects the most light transmitted at a given wavelength. Others might be those that absorb the most light throughout a range of wavelengths. Whatever the definition used to define what the "best" design implies, once it is established, we actually want to determine the structure that will yield this best solution. Mathematical optimization is the process we will use to select an optimal choice from a set of alternatives for this determination.

In this chapter, we give an overview of mathematical optimization and introduce the general (nonlinear) problem. The concepts introduced informally here will be covered in more detail in later chapters as specific applications and instantiations are discussed. We attempt to give a summary of the major work that has been done in this field, structuring it around different classes of the general problem. For a chapter of this type, brevity is a must, as the shear amount of material covered would (and does) fill entire textbooks.

Furthermore, when discussing mathematical optimization, we implicitly assume that we have a problem to optimize. For the scope of this book, we focus on metamaterial design problems. In general though, the problem we seek to optimize has an *objective function* and in most cases, actually determining the correct form of this function is one of the most difficult aspects to conduct.

D.E. Marthaler (✉)
GE Global Research: Industrial Internet Analytics, San Ramon, CA 94583, USA
e-mail: marthaler@ge.com

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_2,

When modeling mathematical optimization problems, we separate them into different classes according to the type of problem they are attempting to solve. The problems may have models that are linear or nonlinear and may or may not be constrained. The objective and constraint functions might be differentiable or non-differentiable, convex or non-convex. In some cases, the problems may only be given via a *black box*, that is, we only know the outputs of the objective function given certain inputs, but not any actual analytical form. Nice references on fundamental theory, methods, algorithm analysis and advice on how to obtain and implement good algorithms for different classes of optimization are provided in [1, 2, 7, 8, 12, 29, 30, 37, 57] among others. We give only a cursory overview of various types of solution techniques. Interested readers are encouraged to refer to the references for more detail.

The rest of the chapter is organized as follows: Sect. 2.2 lays out the general optimization problem and includes a high level discussion on constructing viable objective functions. Section 2.3 discuses linear and convex models and solutions, in particular, the least squares method and different regularizers. Section 2.4 discusses optimization problems that utilize derivatives of the objective function, with subsections focusing on those with and without constraints. Finally, Sect. 2.5 looks at algorithms for optimization problems where derivative information is not available, either because the objective function is not differentiable, the derivative is not available, or the derivative is just too expensive to compute. We conclude with a short summary.

2.2 Mathematical Optimization

The present work considers general multi-objective optimization problems that may be written in the following form:

$$
\begin{aligned}
&\min_{\mathbf{x}} \mathbf{F}(\mathbf{x}) = \left[f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\right]^T \\
&\text{subject to} \\
&\quad g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \ldots, m_{\text{ieq}}, \\
&\quad h_i(\mathbf{x}) = 0, \quad i = 1, 2, \ldots, m_{\text{eq}}.
\end{aligned}
\tag{2.1}
$$

Here $\mathbf{x} = (x_1, \ldots, x_n)$ is the variable to be minimized, $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^k$ is a multi-valued objective function, the functions $g_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, m_{\text{ieq}}$, are the *inequality constraint functions*, and the functions $h_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m_{\text{eq}}$, are the *equality constraint functions*.

We define the space of *feasible solutions* or the *feasible set* as the set of all points that satisfy the constraints:

$$\Omega = \left\{\mathbf{y} \in \mathbb{R}^n : g_i(\mathbf{y}) \leq 0,\ i = 1, \ldots, m_{\text{ieq}} \text{ and } h_j(\mathbf{y}) = 0,\ j = 1 \ldots, m_{\text{eq}}\right\}.$$

The *attainable set* is the range of the feasible set under the objective function:

$$\mathbb{A} = \{\mathbf{F}(\mathbf{x}) : \mathbf{x} \in \Omega\}.$$

Typically in multi-objective optimization, there is no single global solution. It is often necessary to instead seek solutions satisfying Pareto optimality. A point $\mathbf{x}^* \in \Omega$ is *Pareto optimal* if and only if there is no other point $\mathbf{x} \in \Omega$ such that $\mathbf{F}(\mathbf{x}) \leq \mathbf{F}(\mathbf{x}^*)$ and $F_i(\mathbf{x}) < F_i(\mathbf{x}^*)$ for at least one i. That is, no element of $\mathbf{F}$ can be made better without (at least) one other element being made worse [32].

The concept of Pareto optimality invariably leads practitioners to decide which elements of $\mathbf{F}$ are "more important" than others. Having such a ranking of the elements of the objective function, the theory of preferences [38, 43, 44] allows for the construction of a *utility function*. This allows us to convert the general multi-objective function into a single scalar-valued objective function.

One of the most general utility functions is the weighted exponential sum:

$$U = \sum_{i=1}^{k} w_i \left[F_i(\mathbf{x})\right]^p \tag{2.2}$$

for some $p > 0$. Generally, p is proportional to the amount of emphasis placed on minimizing the function with the largest difference between $F_i(\mathbf{x})$ and the minimizer of $F_i(\mathbf{x})$ [28]. Without loss of generality, we can assume $F_i(\mathbf{x}) > 0$, for all i, otherwise we can rescale the objective function to make it so. Here, $\mathbf{w} = \{w_1, \ldots, w_k\}$ is a vector of weights, typically set by the practitioner, such that $\sum_{i=1}^{k} w_i = 1,\ w_i > 0$. Generally, the relative ordering of the weights reflects the relative importance of the objectives.

The most common implementation of Eq. (2.2) is to set $p = 1$, i.e.,

$$U = \sum_{i=1}^{k} w_i F_i(\mathbf{x}), \tag{2.3}$$

which is commonly referred to as the *weighted sum method*. If all of the weights are positive, then the minimum of Eq. (2.3) is Pareto optimal [56], that is, a minimizer of Eq. (2.3) is a Pareto solution of Eq. (2.1).

Selecting non-arbitrary weights is a difficult undertaking. Many approaches exist in selecting weights, surveys of which are provided by [16, 19, 23, 55]. Unfortunately, a satisfactory method to select appropriate weights does not guarantee that the final solution will be acceptable, that is, aligned with predefined preferences. In fact, it is known that weights must be functions of the original objectives in order for a weighted sum to mimic a list of preferences accurately [34]. They cannot be constants. Nevertheless, we proceed in assuming that our multi-objective function in Eq. (2.1) will be converted into a scalar objective, leading to our general problem for the remainder of the chapter:

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) \\ &\text{subject to} \\ &\quad g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \ldots, m_{\text{ieq}}, \\ &\quad h_i(\mathbf{x}) = 0, \quad i = 1, 2, \ldots, m_{\text{eq}}, \end{aligned} \tag{2.4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, and the other functions are as in Eq. (2.1).

2.3 Finding Solutions

In attempting to solve all but the most trivial of problems in the form of Eq. (2.4), a numerical algorithm is used to find a solution $\mathbf{x}^*$. Different objective functions f and constraint functions g, h are more efficiently solved with different types of algorithms. To deduce which algorithm would best assist in finding optimal solutions, we first determine the class of problem characterized by particular forms of the objective and constraint functions.

The simplest form of Eq. (2.4) is in fitting a regression line $y = mx + b$ through a pair of points (x_i, y_i), $i = 1, 2$. We choose the objective function $f(\mathbf{x}) = (\mathbf{y} - m\mathbf{x} - b)^2$ and there are no constraints. Here, $\mathbf{x} = (x_1, x_2)$, and $\mathbf{y} = (y_1, y_2)$. The optimal solution to this problem is given by

$$\begin{aligned} m &= \frac{y_2 - y_1}{x_2 - x_1}, \\ b &= y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1. \end{aligned}$$

When there are more than two points, it is usually impossible to fit a line through all of the points, so instead, we find the line that minimizes the total squared distance to the points:

$$\min \sum_{i=1}^{N} (ax_i + b - y_i)^2.$$

In higher dimensions, the analog to this line fitting problem is to find constants $(a_1, a_2, \ldots, a_n)$ that solve

$$\min \sum_{i=1}^{N} \left(a_i x_i^{(j)} - y_i^{(j)}\right)^2$$

for each $\mathbf{x}^{(j)}, \mathbf{y}^{(j)}$ pair (we omit b for clarity). In matrix notation, this is equivalent to finding the minimum of the function

$$f(\mathbf{a}) = |X\mathbf{a} - \mathbf{y}|_2^2$$

where X is the matrix whose ith row is $\mathbf{x}^{(i)}$ and $\mathbf{y} = (y_1, \ldots, y_N)^T$. A more common designation to this problem is writing X as A, $\mathbf{a}$ as $\mathbf{x}$ and $\mathbf{y}$ as $\mathbf{b}$. We then solve the problem $A\mathbf{x} = \mathbf{b}$. Problems of this type are referred to as *Least Squares* problems and formulating them as minimization problems

$$\min_{\mathbf{x}} |A\mathbf{x} - \mathbf{b}|_2^2 \tag{2.5}$$

leads to a *residual least squares* (RSS) problem. There are many algorithms that solve RSS problems. For a list and introduction, see, for example, [17].

It is well known that attempting to minimize an RSS problem via a numerical method can lead to instabilities. This occurs when the matrix A is not of full rank or when the matrix $A^T A$ is not invertible. In such situations, Eq. (2.5) is stabilized by including a *regularization* term:

$$|A\mathbf{x} - \mathbf{b}|_2^2 + |\Gamma \mathbf{x}|_2^2 \tag{2.6}$$

where Γ is a suitably chosen matrix called a Tikhonov matrix [50]. Usually, Γ is taken to be the identity $\Gamma = I$. An explicit solution to Eq. (2.6) is

$$\mathbf{x}^* = \left(A^T A + \Gamma^T \Gamma\right)^{-1} A^T \mathbf{b}, \tag{2.7}$$

and with $\Gamma = I$ the problem is usually formulated with a *regularization parameter* λ:

$$|A\mathbf{x} - \mathbf{b}|_2^2 + \lambda |\mathbf{x}|_2^2 \tag{2.8}$$

which is commonly known as *Ridge regression* since the parameter λ makes a “ridge” along the diagonal of $A^T A$.

Other regularizations are possible. In particular, we can take a different p-norm in the regularization term. A common choice is the 1-norm, producing the *Least Absolute Selection and Shrinkage Operator* (LASSO) formulation [49]:

$$\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1. \tag{2.9}$$

The multitude of methods that can be used to solve problems of type Eq. (2.9) and its constrained formulation

$$\begin{aligned} &\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ &\text{s.t. } |\mathbf{x}|_1 \le t \end{aligned} \tag{2.10}$$

are discussed within [8], but we mention here that there are many solvers that can be proved to solve the problem to a specified accuracy with a number of operations that does not exceed a polynomial of the problem dimensions.

Although the RSS and LASSO formulations described above were for linear formulations of the objective function, nonlinear formulations exist, one such can be seen in Chap. 6. In general, these problems belong to a class of problems known as *Convex* optimization. We classify a *convex* optimization problem as one in which the objective and constraint functions are convex, i.e., they satisfy the inequalities

$$\begin{aligned} f(\alpha\mathbf{x}+\beta\mathbf{y}) &\le \alpha f(\mathbf{x})+\beta f(\mathbf{y}) \quad \text{and} \\ g_i(\alpha\mathbf{x}+\beta\mathbf{y}) &\le \alpha g_i(\mathbf{x})+\beta g_i(\mathbf{y}), \quad i=1,\ldots,m_{\text{leq}}, \\ h_i(\alpha\mathbf{x}+\beta\mathbf{y}) &\le \alpha h_i(\mathbf{x})+\beta h_i(\mathbf{y}), \quad i=1,\ldots,m_{\text{eq}} \end{aligned} \tag{2.11}$$

for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ and all $\alpha,\beta\in\mathbb{R}$ with $\alpha+\beta=1$, $\alpha\ge 0$, $\beta\ge 0$.

Most, if not all, metamaterial design problems will have nonlinear objective functions, and, when applicable, nonlinear constraints that unfortunately do not satisfy Eq. (2.11) everywhere in their domains. Fortunately though, many problems will have the property that Eq. (2.11) will be satisfied *locally* everywhere. That is, for any point $\mathbf{x}$ in the domain of f, there is a hypersphere about $\mathbf{x}$ where Eq. (2.11) is satisfied (although the α and β will be dependent upon the point $\mathbf{x}$). Such functions are called *locally convex*.

Unfortunately, the absence of global convexity limits the capability of most algorithms to guarantee finding the global minimum of Eq. (2.4). The best most algorithms can achieve is to find a local solution to the problem.

Techniques for solving Eq. (2.4) comprise two types: those that utilize gradient information and those that do not. Recall that a function has C^k smoothness if it is differentiable and its derivative is C^{k-1} smooth. This recursive definition starts with the class C^0, the continuous functions.

2.4 Algorithms Utilizing Gradient Information

We first discuss methods utilizing gradient information that are targeted for optimization problems with no constraints.

2.4.1 Unconstrained Nonlinear Optimization

To find the solution, $\mathbf{x}^*$, to Eq. (2.4) in the case where $\Omega=\mathbb{R}^n$ (i.e., an unconstrained problem), we must satisfy the *second order optimality conditions* [12]:

1. (necessity) If $\mathbf{x}^*$ is a local solution to Eq. (2.4), then $\nabla f(\mathbf{x}^*)=0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite.
2. (sufficiency) If $\nabla f(\mathbf{x}^*)=0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite, then there exists an $\alpha>0$ such that $f(\mathbf{x})\ge +\alpha\|\mathbf{x}-\mathbf{x}^*\|$ for all $\mathbf{x}$ near $\mathbf{x}^*$.

Satisfying these conditions only guarantees a *local* optimum for the general case. Most algorithms used to find solutions are iterative and take the form of Algorithm 2.1:

Algorithm 2.1: General iterative algorithm

```
input: Objective function $f$, initial point $x_0$
repeat
    Determine a descent direction $\mathbf{d}_k$
    Determine a step length $\alpha_k$
    Update Candidate $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$.
until $\nabla f(\mathbf{x}) \approx 0$
```

This is a consistent meme in solving mathematical optimization problems: from your current solution estimate, choose a better candidate and continue until the optimality conditions are satisfied. Algorithms for computing solutions to Eq. (2.4) differ in how they select descent directions $\mathbf{d}_k$ and step sizes α_k. We now discuss some possibilities for both.

2.4.1.1 Descent

Two methods for selection of a descent direction are:

1. Steepest Descent
2. Conjugate Gradient

The *steepest descent*, or gradient descent, algorithms choose descent directions $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ based on the idea that f decreases fastest in the direction of its negative gradient. Unfortunately, due to the iterative nature of Algorithm 2.1, gradient descent's subsequent iterations may undo some minimization progress made on previous descents. To combat this, the *conjugate gradient* algorithm selects successive descent directions in a conjugate direction to previous descent directions. At iteration k, one evaluates the current negative gradient vector $-\nabla f(\mathbf{x}_k)$ and adds to it a linear combination of the previous descent iterates to obtain a new conjugate direction along which to descend. Initially, the descent is in the direction of the negative gradient, but each subsequent step moves in a direction that modifies the negative of the current gradient by a factor of the previous direction. The CG algorithm is shown in Algorithm 2.2.

Different Conjugate Gradient methods correspond to different choices for the scalar β_k. Three of the best known versions are:

- Fletcher–Reeves: $\beta_k^{\mathrm{FR}} = \frac{\mathbf{s}_k^T \mathbf{s}_k}{\mathbf{s}_{k-1}^T \mathbf{s}_{k-1}}$
- Polak–Ribiére: $\beta_k^{\mathrm{PR}} = \frac{\mathbf{s}_k^T (\mathbf{s}_k - \mathbf{s}_{k-1})}{\mathbf{s}_{k-1}^T \mathbf{s}_{k-1}}$
- Hestenes–Stiefel: $\beta_k^{\mathrm{HS}} = \frac{\mathbf{s}_k^T (\mathbf{s}_k - \mathbf{s}_{k-1})}{\mathbf{d}_{k-1}^T (\mathbf{s}_k - \mathbf{s}_{k-1})}$

for a full list, consult [18].

Algorithm 2.2: Nonlinear conjugate gradient

```
input: Objective function $f$, initial point $x_0$
$\mathbf{d}_0 = -\nabla f(\mathbf{x}_0)$
(Line Search) $\alpha_0 = \arg\min_\alpha f(\mathbf{x}_0 + \alpha \mathbf{d}_0)$
$\mathbf{x}_1 = \mathbf{x}_0 + \alpha \mathbf{x}_0$ repeat
    Determine steepest direction $\mathbf{s}_k = -\nabla f(\mathbf{x}_k)$
    Determine the scalar $\beta_k$ (see below)
    Update the conjugate direction $\mathbf{d}_k = \mathbf{s}_k + \beta_k \mathbf{d}_{k-1}$
    Determine a step length (Line search) $\alpha_k = \arg\min_\alpha f(\mathbf{x}_k + \alpha \mathbf{d}_k)$
    Update Candidate $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$.
until $\nabla f(\mathbf{x}) \approx 0$
```

2.4.1.2 Step Length

Having a descent direction, we must now determine how far along that direction to move for the next iterate. Ideally, we would move a length α along the line where α solves

$$\min_\alpha f(\mathbf{x}_k + \alpha \mathbf{d}_k), \tag{2.12}$$

i.e., the distance that minimizes the objective function in the direction $\mathbf{d}_k$. Notice that this is a one dimensional optimization problem in α. Finding an optimal solution to this problem would imply a method of solving the original nonlinear optimization problem! Therefore, instead of solving (2.12), we seek an efficient way of computing an acceptable α that guarantees that Algorithm 2.1 will converge to a $\mathbf{x}^*$.

To do this, we must find an α satisfying the following two conditions:

$$\begin{aligned} f(\mathbf{x}_k + \alpha \mathbf{d}_k) &\le f(\mathbf{x}_k) + c_1 \alpha \mathbf{d}_k^T \nabla f(\mathbf{x}_k), \\ \mathbf{d}_k^T \nabla f(\mathbf{x}_k + \alpha \mathbf{d}_k) &\ge c_2 \mathbf{d}_k^T \nabla f(\mathbf{x}_k) \end{aligned} \tag{2.13}$$

with $0 < c_1 < c_2 < 1$. The first condition is known as the *Armijo rule*. It ensures that the step length decreases f sufficiently for this iteration. The second condition is known as the *curvature condition*. It ensures that the slope of f has been reduced sufficiently for this iteration. Unfortunately, these two conditions may result in an α that is not close to an actual minimum of (2.12). Therefore, we modify the curvature condition to include

$$\left| \mathbf{d}_k^T \nabla f(\mathbf{x}_k + \alpha \mathbf{d}_k) \right| \le c_2 \left| \mathbf{d}_k^T \nabla f(\mathbf{x}_k) \right|, \tag{2.14}$$

and this ensures that α will lie close to a minimum critical point of Eq. (2.12). These three conditions taken together form the *Strong Wolfe conditions* [12] and are a prerequisite to any step length determination algorithm. Many methods exist for solving the general unconstrained problem, but they all utilize an algorithm similar to Algorithm 2.1 in their strategy.

2.4.1.3 Quasi-Newton Methods

In general, methods that utilize gradient information seek to find a stationary point of f by finding a zero of the gradient ∇f. A general class of methods, *quasi-Newton methods,* seek to do this by using Newton's method to find a root of ∇f. The underlying assumption in these methods is that the function f can locally be approximated by a quadratic.

Regular Newton's method updates candidate solutions at each iteration via

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left[\nabla^2 f(\mathbf{x}_k)\right]^{-1} \nabla f(\mathbf{x}_k)$$

where $\nabla^2 f(\mathbf{x})$ denotes the Hessian, or the second derivative of f. Updates can be very expensive since we must find the inverse of an $n \times n$ matrix at every iteration. To ease computational cost, approximations to the Hessian and its inverse are used. There are multiple ways the Hessian can be approximated, one method that is extensively employed is from the Broyden family which uses a convex combination of Daviodon–Fletcher–Powell [14] and BFGS [45] updates. An extensive survey of Quasi-Newton methods may be found in [40].

2.4.2 Constrained Nonlinear Optimization

When dealing with the general form of Eq. (2.4), i.e., when the constraints exist, the first question to answer is how to ascertain if a candidate $\mathbf{x}^*$ is indeed a solution.

First, we define a constraint g_i to be *active* (resp., *inactive*) at a point $\mathbf{x}$ if $g_i(\mathbf{x}) = 0$ (resp., $g_i(\mathbf{x}) < 0$). (Note, equality constraints are always active.) We define the *active set* at $\mathbf{x}$, $\mathcal{A}(\mathbf{x})$, as the indices of those constraints $g_i(\mathbf{x})$ that are active at the given point. For a given candidate solution, $\mathbf{x}_k$, if no constraints are active, then the necessary and sufficient conditions are the same as for the unconstrained case. In the case where the candidate lies on the boundary of the feasible set (i.e., at least one constraint is active), the second order optimality conditions for the unconstrained case do not apply because the direction of the negative gradient (or even a descent direction in a conjugate direction) will push the next iterate into the infeasible set.

We will specify the optimality conditions for a solution $\mathbf{x}^*$ to solve Eq. (2.4) through the use of a Lagrangian function:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m_{\text{leq}}} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \lambda_i h_i(\mathbf{x})$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m_{\text{leq}}})$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{m_{\text{eq}}})$ are vectors called *KKT multipliers*. Now, if $\mathbf{x}^*$ is an optimal solution to Eq. (2.4), then there exist KKT multipliers

$\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ such that

$$\begin{aligned}
&\nabla f\left(\mathbf{x}^*\right) + \sum_{i=1}^{m_{\text{leq}}} \lambda_i^* \nabla g_i\left(\mathbf{x}^*\right) + \sum_{i=1}^{m_{\text{eq}}} \mu_i^* \nabla h_i\left(\mathbf{x}^*\right) = 0,\\
&g_i\left(\mathbf{x}^*\right) \le 0 \quad \text{for } i = 1, \ldots, m_{\text{leq}},\\
&h_i\left(\mathbf{x}^*\right) = 0 \quad \text{for } i = 1, \ldots, m_{\text{eq}},\\
&\lambda_i^* \ge 0 \quad \text{for } i = 1, \ldots, m_{\text{leq}},\\
&\mu_i^* \ge 0 \quad \text{for } i = 1, \ldots, m_{\text{eq}},\\
&\lambda_i^* g_i\left(\mathbf{x}^*\right) = 0 \quad \text{for } i = 1, \ldots, m_{\text{leq}}.
\end{aligned} \tag{2.15}$$

The above conditions are known as the *Karush–Kuhn–Tucker* conditions (KKT conditions) [8]. Points that satisfy them are critical points of the original problem. To determine if these critical points are indeed solutions of Eq. (2.4), we impose second order conditions on the points (for they could be a maximizer or a saddle point).

Before stating the second order sufficient and necessary conditions, we first define the *tangent space* for feasible points $\bar{\mathbf{x}}$

$$T = \left\{\mathbf{v} : \nabla g_j(\bar{\mathbf{x}})\mathbf{v} = 0 \; \forall j \in \mathcal{A}(\bar{\mathbf{x}}),\; \nabla h(\bar{\mathbf{x}})\mathbf{v} = 0\right\}$$

where $\mathcal{A}(\bar{\mathbf{x}})$ denotes the active set.

For a KKT point, we also define the *relaxed tangent space*

$$T' = \left\{\mathbf{v} : \nabla g_j(\bar{\mathbf{x}})\mathbf{v} = 0 \; \forall j \in \{j : \lambda_j > 0\},\; \nabla h(\bar{\mathbf{x}})\mathbf{v} = 0\right\}.$$

Having these definitions, we now state the second order necessary and sufficient conditions for a feasible candidate $\mathbf{x}^*$ with KKT multipliers $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ satisfying Eq. (2.15) to be a solution to Eq. (2.4):

$$\mathbf{w}^T \nabla_x L^2\left(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*\right)\mathbf{w} > 0 \quad \forall \mathbf{w} \in T',\; \mathbf{w} \ne \mathbf{0}. \tag{2.16}$$

Methods for finding a suitable optimum satisfying Eqs. (2.15) and (2.16) for constrained optimization problems are ubiquitous. We focus on two categories:

1. Primal methods
2. Penalty and Barrier Methods

We will briefly describe each type below.

2.4.2.1 Primal Methods

Primal methods are those that solve Eq. (2.4) by starting with a candidate in the feasible set Ω and searching only the feasible set for an optimal solution. The main characteristics of these algorithms is that they find new candidates that simultaneously decrease the objective function at each step, while remaining feasible. To update a given candidate $\mathbf{x}_k$, a vector $\mathbf{d}_k$ is chosen such that it is both descending and feasible. The following must hold for $\mathbf{d}_k$ to be a feasible direction:

$$\nabla f(\mathbf{x})^T \mathbf{d}_k < 0, \tag{2.17}$$

$$\nabla g_i(\mathbf{x})^T \mathbf{d}_k < 0, \tag{2.18}$$

$$\nabla h_i(\mathbf{x})^T \mathbf{d}_k = 0. \tag{2.19}$$

Equation (2.17) implies that we are descending, and Eqs. (2.18) and (2.19) imply that we are increasing feasibility (by moving in the direction tangential to the active set for the inequality constraints and parallel for the equality constraints).

Feasible direction methods suffer from requiring a feasible initial candidate, from situations where no feasible descent direction exists, and may be subject to jamming, or oscillations that prevent convergence of the algorithm [12].

Gradient projection methods are motivated from steepest descent algorithms in unconstrained optimization. Their basic idea is to take the negative of the gradient of the objective function and project it onto the *working surface* in order to determine a feasible descent direction. The working surface is the subset of the constraints that are currently active, i.e., the current active set.

Thus, at the current feasible point, one determines the active constraints and projects the negative gradient of the objective function onto the subspace tangent to the surface determined by these constraints. However, this may not be a feasible direction since the working surface may be curved. To deal with curvature, one searches for a feasible descent direction along an embedded curve within the constraint surface.

2.4.2.2 Penalty and Barrier Methods

Penalty and Barrier methods attempt to approximate constrained optimization problems with those that are unconstrained, and then apply standard unconstrained search techniques to obtain solutions. Penalty methods do this by adding a term to the objective function that penalizes violation of the constraints with a large factor. In the case of barrier methods, a term is added that favors points in the interior of the feasible region and penalizes those closer to the boundary.

The idea for penalty methods is to replace Eq. (2.4) with an unconstrained problem of the form

$$\min_{\mathbf{x}} f(\mathbf{x}) + \beta\sigma(\mathbf{x}) \tag{2.20}$$

where $\beta > 0$ and $\sigma : \mathbb{R}^n \to \mathbb{R}$ is a function satisfying

1. $\sigma(\mathbf{x})$ is continuous;
2. $\sigma(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$;
3. $\sigma(\mathbf{x}) = 0 \Leftrightarrow g_i(\mathbf{x}) \leq 0,\ h_j(\mathbf{x}) = 0\ \forall i = 1, \ldots, m_{\text{leq}},\ j = 1, \ldots, m_{\text{eq}}$, i.e., $\mathbf{x}$ is feasible.

That is, we set up an unconstrained optimization problem where we generate a new objective function that greatly increases in value as $\mathbf{x}$ moves out of the feasible region. A standard choice for $\sigma(\mathbf{x})$ is the *quadratic loss function* [26]:

$$\sigma(\mathbf{x}) = -r \sum_{i=1}^{m_{\text{leq}}} \max\bigl(0, g_i(\mathbf{x})\bigr) + \frac{1}{r} \sum_{i=1}^{m_{\text{eq}}} \bigl(h_i(\mathbf{x})\bigr)^2. \tag{2.21}$$

For $\mathbf{x}$ values inside the feasible region, $g_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$, giving a value of $\sigma = 0$. When $\mathbf{x}$ is outside of the feasible region, some of the $g_i > 0$ or $h_i \neq 0$, we begin to be penalized. To implement a penalty method, one needs to select a value for β. Standard techniques start with a relatively small value (and an infeasible point for $\mathbf{x}_0$) and monotonically increase β, solving subsequent unconstrained optimization problems (one for each β) and utilizing these intermediate solutions as the initial guess for the next problem. This *graduated optimization* method produces a sequence of solutions that converge to an optimal solution of the original constrained problem. Graduated optimization is a technique commonly used with hierarchical pyramid methods for matching objects within images [9].

Barrier methods are implemented when one does not wish to compute $f(\mathbf{x})$ outside of the feasible region. Thus, we would not be able to utilize a penalty function like Eq. (2.21). Instead, a selection would need to be made that was defined to converge for feasible points. A possible selection for problems with no equality constraints might be

$$\sigma(\mathbf{x}) = r \sum_{i=1}^{m} \frac{-1}{g_i(\mathbf{x})} \tag{2.22}$$

where $r > 0$ is the *barrier parameter*. As candidates get closer to the boundary of the feasible region, the value of the objective function becomes larger. The idea is to start with a feasible point and a relatively large value of the barrier parameter, preventing the candidates from nearing the boundary of the feasible set. Techniques then decrease the value of the barrier parameter monotonically until an optimum value for the original problem is achieved. Note that barrier methods require a feasible point from which to start. This can sometimes be difficult to find. Also, barrier methods do not work with equality constraints without cumbersome modifications to this basic approach, and by not allowing the method to ever leave the feasible region, much more computational effort is (usually) required.

Penalty methods are sometimes referred to as *external* methods since their augmented objective functions tend to utilize solutions in the exterior of the feasible region. Analogously, barrier methods are sometimes called *interior point* methods, for the opposite reason. There is a vast and vigorous field of research surrounding these methods, and we suggest utilizing the references to find current implementations. A great start would be [26].

These two types of methods are among the most powerful for attacking the general scalar problem in Eq. (2.4). Of the two, exterior methods are preferable (when applicable) as they can deal with equality constraints, they do not require a feasible starting point, and their computational effort is substantially lower than for the interior methods.

2.5 Gradient-Free Algorithms

Looking at the form of Eq. (2.4), we denote f as a function, and this is typically seen as an analytical expression. Most industrial applications of the general problem may involve formulations that do *not* encode f analytically, but have some type of *black box* that computes values of $f(\mathbf{x})$. That is, given a value $\mathbf{x}$, there is some process (numerical simulation, physical experiment, etc.) that computes the output $f(\mathbf{x})$. Furthermore, the constraint functions may also be black-box functions. Typically, these black-box functions will not have any derivative information associated with them (although in rare occasions, there may be derivative information available via another black-box function). In these cases, f is expensive to calculate in terms of time, and methods that require many evaluations of f rapidly become infeasible to use in many applications. In particular, to produce viable step lengths satisfying the Strong Wolfe conditions in Eq. (2.14), hundreds of function evaluations may be required per iterate.

Moreover, when evaluating the objective function via a numerical simulation or physical experiment, inaccuracies may arise in the value that f takes at a given point. This generates many difficulties approximating derivatives via finite differences. This line of thinking dismisses the use of many of the techniques from Sect. 2.4. Even in cases where derivative information is available, function inaccuracies adversely effect most of these methods [15].

2.5.1 Direct Methods

Direct methods are those that attempt to solve the general problem *directly* by utilizing objective function values. Here, we introduce a number of methods starting with a variant of gradient descent for the derivative-free case.

2.5.1.1 Coordinate Descent

Perhaps the simplest method to solve an unconstrained version of Eq. (2.4) without using gradients is to do successive line searches in each coordinate direction for each iteration. That is, one does a line search in a coordinate direction for each iteration, changing coordinates for each, and looping cyclically as the number of dimensions are reached. This process is called *Coordinate Descent* (CD). Iterations of a cycle of line search in all coordinate directions is equivalent to one gradient descent direction, but the number of function evaluations may prove to be prohibitive.

More efficient algorithms have been constructed in an attempt to limit the number of function evaluations made to reach convergence. In particular, choosing a random direction to do line search for each iteration, the so-called *Random Coordinate Descent,* was shown to converge, on average, in fewer iterations than CD [36, 42]. In general, one seeks an appropriate coordinate system where CD would operate optimally. The *Adaptive coordinate descent* algorithm [24] gradually builds a transformation of the coordinate system such that the new coordinates are as decorrelated as possible with respect to the objective function.

Instead of finding a pointwise trajectory to the minimum, other techniques attempt to locate a set wherein the optimal solution resides. The oldest and most famous of these is the simplex algorithm.

2.5.1.2 Nelder–Mead Simplex Algorithm

The Nelder–Mead (NM) algorithm [35] solves the general problem by containing the solution within a *simplex.* A simplex is the generalization of a polygon to n dimensions. The NM algorithm starts with a set of points in $\mathbb{R}^n$ forming a simplex and at each iteration, the objective function is evaluated at the vertices of the simplex.

The algorithm replaces the worst point on the simplex with a point reflected through the centroid of the remaining n points. If this point is better than the best current point, then the simplex is stretched exponentially out along this line. If not, then the simplex stretches across a valley, so the simplex is shrunk towards a (hopefully) better point. A few of the other means of replacing the chosen point include: reflection, expansion, inside and outside contractions.

The Nelder–Mead algorithm remains popular, mostly through its simplicity, but McKinnon [33] established analytically that convergence can occur to points with $\nabla f(\mathbf{x}) \neq 0$, even when the function is convex and twice continuously differentiable. Tseng [54] proposed a globally convergent simplex-based search method that considers an expanded set of candidate replacement points (besides those listed above). Other modifications are presented in [13].

2.5.1.3 Mesh Adaptive Direct Search (MADS)

The Mesh Adaptive Direct Search (MADS) [3] is a generalization of several existing direct search methods [25, 51–53]. MADS was introduced to extend direct search methods to deal with the constrained problem in Eq. (2.4), while improving both the practical and theoretical convergence results seen in previous methods.

MADS handles constraints $x \in \Omega$ by the so-called *extreme barrier method*, which simply consists in rejecting any trial point which does not belong to Ω. The term *extreme barrier method* comes from the fact that this approach can be implemented by solving the unconstrained minimization of

$$f_{\Omega}(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ \infty & \text{otherwise} \end{cases}$$

in place of Eq. (2.4). Note that this may impose severe discontinuities on the problem. A more subtle way of handling quantifiable constraints is presented in [4], and is summarized in Chap. 4.

Each MADS iteration proceeds as follows: Given a candidate solution $\mathbf{x}_k$, the SEARCH step produces a list of tentative trial points. Any mechanism can be used to create the list, as long as it contains a finite number of points located on a *conceptual mesh*. The conceptual mesh is defined by a *mesh parameter* $\Delta_k^M > 0$. This parameter, along with a finite set of positive spanning directions D, forms the mesh at iteration k:

$$M_k = \left\{\mathbf{x} + \Delta_k^M \mathbf{d} : \mathbf{x} \in V_k, \mathbf{d} \in D\right\} \tag{2.23}$$

where V_k is a set containing all previous points where the objective function has been evaluated. A *positive spanning set* of $\mathbb{R}^n$ is a set $D = \{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$ of vectors in $\mathbb{R}^n$ such that every vector in $\mathbb{R}^n$ is a linear combination of the $\mathbf{d}_i$ with nonnegative coefficients. Many methods exist for computing a set of points on the conceptual mesh: speculative search [21], Latin hypercube sampling [47], variable neighborhood searches [11], surrogates, and many others [48].

Having an initial set of points, the objective function is evaluated at each of the points until either a better candidate than $\mathbf{x}_k$ is found, or all of the points are evaluated. In the latter case, a POLL step is implemented that conducts a local exploration near the candidate point. Following an unsuccessful SEARCH step, the POLL step generates a list of mesh points near the incumbent x_k. The term *near* is tied to the so-called *poll size parameter* $\Delta_k^p > 0$. Similar to the SEARCH step, the POLL step may be interrupted as soon as an improvement point over the candidate is found.

Parameters are updated at the end of each iteration. There are two possibilities: If either the SEARCH or the POLL step generated a mesh point $\mathbf{p} \in M_k$ which is better than $\mathbf{x}_k$, then the candidate point $\mathbf{x}_{k+1}$ is set to $\mathbf{p}$ and both the mesh size and poll size parameters are increased or kept to the same value. For example, $\Delta_{k+1}^M \leftarrow \min\{1, 4\Delta_k^M\}$ and $\Delta_{k+1}^p \leftarrow 2\Delta_k^p$. Otherwise, $\mathbf{x}_{k+1}$ is set to $\mathbf{x}_k$ and the poll size is decreased and the mesh size parameter decreased or kept the same. For example, $\Delta_{k+1}^m \leftarrow \min\{1, \frac{1}{4}\Delta_k^m\}$ and $\Delta_{k+1}^p \leftarrow \frac{1}{2}\Delta_k^p$. At any iteration of the MADS algorithm, the poll size parameter Δ_k^p must be greater than or equal to the mesh size parameter Δ_k^M. Termination conditions arise when either the poll parameter matches the mesh size parameter or a predefined number of iterations have been reached.

2.5.2 Surrogate Methods

As mentioned above, surrogates may be used to determine a set of points for use in the SEARCH step for direct search. These methods build a model interpolating between the known points stored in V_k. This section looks at methods that do not restrict themselves to interpolation with a local search; rather, they utilize a *global* surrogate function to assist in the optimization.

There are many ways of employing surrogates. In particular, there is a standard engineering process [5] for using them:

1. Choose a surrogate s for the objective function f that is either
   (a) A simplified model of f (as is used in Chap. 6) or
   (b) A *response surface* of f generated from a set of points $\mathbf{x}_1, \dots, \mathbf{x}_q$ where f takes a finite value;
2. Minimize over the surrogate s, obtaining a candidate point $\mathbf{x}^s$;
3. Evaluate the objective function at $\mathbf{x}^s$ and repeat the process.

In cases where we do not have a simplified model for f and wish to generate a response surface (or metamodel) $\hat{f}$, the question arises as to which method to use. Barton [6] enumerates a list, including splines, radial basis functions, kernel smoothing, spatial correlation models, and frequency domain approaches. Regardless of the method employed, its quality depends crucially upon choosing an appropriate sampling technique [39]. The remainder of this subsection describes a state of the art response surface methodology known as *Gaussian Process Regression*. We will see its implementation in Chap. 3.

2.5.2.1 Gaussian Process Regression

Gaussian Process Regression (GPR) [41] is also known as Kriging prediction, Kolmogorov–Wiener prediction, or best linear unbiased prediction. It is a technique for estimating the objective function value at a new point $\mathbf{x}_*$ utilizing noisy observations $f(\mathbf{x})$ at points $\mathbf{x}_1, \dots, \mathbf{x}_m$. The surrogate is a process that generates data such that any finite subset follows a multivariate Gaussian distribution.

A typical assumption for the surrogate is that the mean of the data is zero everywhere (if not, we can subtract the mean and work with the transformed dataset). Then, pairs of points in GPR are related to each other by the *covariance function*. A popular choice is the *squared exponential*:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left[\frac{-\|\mathbf{x}_p - \mathbf{x}_q\|_2^2}{2L^2}\right] \tag{2.24}$$

where the maximum allowable covariance is σ_f^2.

Note, that the covariance between the outputs is written as a function of the inputs. For this particular covariance function, we see that the covariance is almost maximal between variables whose corresponding inputs are very close, and decreases as their distance in the input space increases. The covariance function has a characteristic length scale L, which informally can be thought of as roughly the distance you have to move in input space before the function value can change significantly. Alternatively, this relates how much influence distant points will have on each other.

We create the *covariance matrix* between all pairs of points

$$K(\mathbf{X},\mathbf{X}) = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_1,\mathbf{x}_2) & \dots & k(\mathbf{x}_1,\mathbf{x}_m) \\ k(\mathbf{x}_2,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_2) & \dots & k(\mathbf{x}_2,\mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m,\mathbf{x}_1) & k(\mathbf{x}_m,\mathbf{x}_2) & \dots & k(\mathbf{x}_m,\mathbf{x}_m) \end{bmatrix}. \tag{2.25}$$

Observations from the data are often noisy, for a various number of reasons. As is typical in most regression schemes, we model the observations as

$$y = f(\mathbf{x}) + \mathcal{N}\big(0,\sigma_v^2\big),$$

and the covariance between two points becomes

$$\operatorname{cov}(y_p,y_q) = k(\mathbf{x}_p,\mathbf{x}_q) + \sigma_v^2\delta_{pq} \quad \text{or} \quad \operatorname{cov}(\mathbf{y}) = K(\mathbf{X},\mathbf{X}) + \sigma_v^2 I, \tag{2.26}$$

where δ_{pq} is the Kronecker delta function which is 1 when $p = q$ and 0 otherwise. Here, I is the $m \times m$ identity matrix.

The purpose of generating the surrogate is to predict values of the observables at previously unseen points. The assumptions underpinning GPR state that the joint distribution of the observed data and unknown data point $\mathbf{x}_*$ is given by:

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma_v^2 I & K(\mathbf{X},\mathbf{x}_*) \\ K(\mathbf{x}_*,\mathbf{X}) & K(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}\right). \tag{2.27}$$

where y_* denotes the value of the surrogate at the unseen point $\mathbf{x}_*$. We seek the conditional probability $p(y_*|\mathbf{y})$, or "how likely is a certain prediction for y_* given the data?" As derived in [41], this probability follows the distribution

$$p(y_*|\mathbf{y}) \sim \mathcal{N}\big(K_* K^{-1}\mathbf{y},\, K_{**} - K_* K^{-1} K_*^T\big) \tag{2.28}$$

where T denotes transposition and we use the short hand notation of K being the covariance matrix, $K_* = [k(\mathbf{x}_*,\mathbf{x}_1)\ k(\mathbf{x}_*,\mathbf{x}_2)\ \dots\ k(\mathbf{x}_*,\mathbf{x}_m)]$ and $K_{**} = k(\mathbf{x}_*,\mathbf{x}_*)$.

Thus, the best estimate for y_* is the mean of this distribution

$$y_* = K_*\big(K + \sigma_v^2\big)^{-1}\mathbf{y}, \tag{2.29}$$

and the uncertainty is captured in the variance

$$\operatorname{var}(y_*) = K_{**} - K_*\big(K + \sigma_v^2\big)^{-1} K_*^T. \tag{2.30}$$

We note here for completeness that if our original data set did not have zero mean, but instead had mean $\mathbf{m}(\mathbf{X})$, then Eq. (2.29) would become

$$y_* = m(\mathbf{x}_*) + K_*\big(K + \sigma_\nu^2\big)^{-1}\big(\mathbf{y} - \mathbf{m}(\mathbf{X})\big) \tag{2.31}$$

where $m(\mathbf{x}_*)$ denotes the mean of the new data. The variance remains unchanged from Eq. (2.30).

For actual implementations of the above equations, we need to determine values for the parameters σ_f, L, σ_ν. This collection of parameters are referred to as *hyperparameters*. Most methods for determining the hyperparameters from data attempt to optimize the marginal likelihood of $p(y_*|\mathbf{y})$ with respect to the hyperparameters, given the data. This is itself a rich and interesting optimization problem having a long history in spatial statistics [31].

2.5.3 Stochastic Search Algorithms

In the above formulations, some assumption about the smoothness of the function, or continuity of the function is made. This is manifested either in the direct usage of gradients or in methods like the polling step of direct search, where shrinking the polling step parameter is assumed to lead to a better solution.

Sometimes, functions are not continuous for large swaths of the space over which we seek to optimize. This section presents approaches that rely on non-deterministic algorithmic steps. This is a more delicate way of saying that the algorithms "guess" which direction to search for a better candidate solution. Most algorithms of this type have a heuristic for choosing how to "guess." Some approaches occasionally allow new candidates that are "worse" (in terms of the objective function) than the current solution; the idea being that accepting a worse candidate at this iteration will lead to a better overall solution as the algorithm iterates. This idea allows the algorithm to theoretically find global solutions. The literature on stochastic algorithms is very extensive, especially on the applications side, since their implementation is rather straightforward compared to deterministic algorithms. See, for example, [22, 46, 58] for a general overview.

2.5.3.1 Random Search

The simplest algorithm of this type is random search. Random algorithms compare the current iterate $\mathbf{x}$ with a randomly generated candidate (no heuristic). The current iterate is updated only if the candidate is a better point (in terms of the objective function). The determination of new candidates is based on two random components: A direction $\mathbf{d}$ is generated using a uniform distribution over the unit sphere in $\mathbb{R}^n$, and a step α is generated from a uniform distribution over the set of steps S in a way that $\mathbf{x} + \alpha\mathbf{d}$ is feasible. Bélisle et al. [10] generalized these types of algorithms by allowing arbitrary distributions to generate both the direction **d** and step α, and proved convergence to a global optimum under mild conditions for continuous optimization problems. Unfortunately, the number of function evaluations for this type of method become prohibitive.

2.5.3.2 Genetic Algorithms

In an effort to chose points with less randomness than simply guessing, *Genetic Algorithms* (GA) were originally introduced by Holland [20] wherein a method was designed that mimics the process of natural evolution.

The GA operates on a population of individuals that are each represented by a chromosome $\mathbf{x}$. Initially, a random population is chosen and the objective function is evaluated on each member. The better performing members are chosen to mate and form a new generation, mimicking the process of natural selection. A mating pool is first formed by either sorting the population according to objective function value and then keeping the top performing members, or by using a threshold such as the mean or the median cost to eliminate any population members with a worse performance than the threshold value. Members in the mating pool are eligible for breeding. For each new solution to be produced, a pair of "parent" solutions is selected from the mating pool. These parents produce a "child" solution using *crossover*, creating a new solution which typically shares many of the characteristics of its "parents". New parents are selected for each new child, and the process continues until a new population of solutions of appropriate size is generated.

After selection and crossover have been performed to fill out the population for the next generation, a small percentage of elements in the new population are mutated in order to continue exploring new parts of the parameter space. If an individual is randomly selected for mutation, then its value is given a new random value within its allowed range. Typical mutation probabilities are on the order of a few percent, and different distributions are employed for new variates.

The final step in populating the new generation is to optionally enforce *elitism*. Elitism ensures that the best global fitness is maintained between generations by copying the chromosome with the best fitness from the previous generation into the new population. At this point, the new population is ready to be evaluated by the fitness function.

Different crossover methods and Nature-Inspired Optimization routines, including Genetic Algorithms, will be discussed in detail in Chap. 5.

2.5.3.3 Non-dominated Sorting Genetic Algorithm

We now introduce a method that attempts to produce the Pareto front for a general multi-objective optimization problem. We will see that Chap. 3 generates such a problem, and here we discuss the method used to solve it. This method, Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) [27] assumes our multi-objective function has k dimensions.

Again, we adopt the general idea of a genetic algorithm, but with some changes. The algorithm starts with a random parent population P of size N. Binary tournament selection, recombination, and mutation operators are used to create a child population of P of size N. We combine the parent and children populations then sort them via the principle of *non-domination.* An element $\mathbf{p} \in P$ dominates another element $\mathbf{q} \in P$ if there is an i with $p_i < q_i$ and $p_j \leq q_j$ for all other j. Here, the ith element of $\mathbf{p}$, denoted p_i represents the ith objective value for this population element.

Each solution is assigned a fitness equal to its non-domination level. Those elements with no dominating elements are given fitness 1. Those elements only dominated by elements with fitness $= 1$ are given fitness 2, etc. For each fitness level, we sort the elements in that level via *crowding comparison.* To do so, we first find the *local crowding distance* for each element. This distance is calculated by finding the average distance of the two nearest neighbors to this point along each of the objective axes.

We sort within each fitness level, giving preference to those solutions that are "more spread out," i.e., have a larger crowding distance. The new population is then generated by taking the first N elements of the sorted fitness levels. The process repeats itself (children are generated, combined with parents, sorted via non-domination, etc.) until either all elements of the population have fitness level 1 or a predetermined number of iterations are reached.

2.6 Summary

We have described a range of mathematical optimization problems and their respective solution techniques. Methods that utilize derivative information, both for constrained and unconstrained problems, were briefly introduced. These methods, combined with parametrized models of metamaterial structures to be simulated, are too often trapped in numerous local minima. As a result, their usefulness for metamaterial design is minimal, and they will not be covered further in the text.

Many methods for solving problems that do not take advantage of derivative information, either because it does not exist or is not available, were also discussed. These techniques, which will be covered over the next three chapters, are well-established methods of optimization. They are all robust against non-smooth optimization surfaces, and coincidentally are all direct search methods. Additionally, both Mesh Adaptive Direct Search in Chap. 4, and Nature Inspired Optimization in Chap. 5 work efficiently in high dimensions.

The last two chapters of the book do not focus solely on the optimization method itself. These chapters integrate both optimization routines with novel methods for calculating and representing the shapes of the individual resonant structures within a metamaterial. These approaches are both gradient-based, but they are able to circumvent the normal pitfalls of gradient-based optimization by transforming the space over which the optimization occurs. Both techniques are new to the field of metamaterial design; however, their applicability extends far beyond the focus of this book. This is clearly illustrated by the range of design examples that are covered throughout the last two chapters.

References

1. N. Andr'easson, A. Evgrafov, M. Patriksson, *An Introduction to Continuous Optimization: Foundations and Fundamental Algorithms*. Studentlitteratur (2005)
2. K.A. Atkinson, *An Introduction to Numerical Analysis*, 2nd edn. (Wiley, New York, 1988)
3. C. Audet, J.E. Dennis Jr., Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(2), 188–217 (2006)
4. C. Audet, J.E. Dennis Jr., A progressive barrier for derivative-free nonlinear programming. SIAM J. Optim. **20**(1), 445–472 (2009)
5. J.-F.M. Barthelemy, R.T. Haftka, Approximation concepts for optimum structural design—a review. Struct. Optim. **5**, 129–144 (1993)
6. R.R. Barton, Metamodeling: a state of the art review, in *Proceedings of the 1994 Winter Simulation Conference* (1994), pp. 237–244
7. D.P. Bertsekas, *Nonlinear Programming*, 2nd edn. (Athena Scientific, Belmont, 1999)
8. S.P. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2003)
9. P.J. Burt, Fast filter transformations for image processing. Comput. Graph. Image Process. **16**, 20–51 (1981)
10. H.E. Romeijn, C.J. Bélisle, R.L. Smith, Hit-and-run algorithms for generating multivariate distribution. Math. Oper. Res. **18**, 255–266 (1993)
11. V.B.C. Audet, S. Le Digabel, Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. J. Glob. Optim. **41**(2), 299–318 (2008)
12. E.K.P. Chong, S.H. Zak, *An Introduction to Optimization* (Wiley, New York, 1996)
13. A.R. Conn, K. Scheinberg, L.N. Vicentee, *Introduction to Derivative-Free Optimization* (SIAM, Philadelphia, 2009)
14. W.C. Davidon, Variable metric method for minimization. SIAM J. Optim. **1**, 1–17 (1991)
15. J.E. Dennis, H.F. Walker, Inaccuracy in quasi-Newton methods: local improvement theorems. Math. Program. Stud. **22**, 70–85 (1984)
16. R.T. Eckenrode, Weighting multiple criteria. Manag. Sci. **12**, 180–192 (1965)
17. G.F. Golub, C.F. Van Loan, *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences (1996)
18. W.W. Hager, H. Zhang, A survey of nonlinear conjugate gradient methods. Pac. J. Optim. **2**, 35–58 (2006)
19. B.F. Hobbs, A comparison of weighting methods in power plant siting. Decis. Sci. **11**, 725–737 (1980)
20. J.H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, 1975)
21. R. Hooke, T.A. Jeeves, Direct search solution of numerical and statistical problems. J. ACM **8**, 212–229 (1961)
22. H.H. Hoos, T. Stützlew, *Stochastic Local Search: Foundations and Applications* (Morgan Kaufmann/Elsevier, San Mateo, 2004)
23. C.-L. Hwang, K. Yoon, Multiple attribute decision making methods and applications: a state-of-the-art survey, in *Lecture Notes in Economics and Mathematical Systems*, ed. by M. Beckmann, H.P. Kunzi (Springer, Berlin, 1981)
24. M. Schoenauer, I. Loshchilov, M. Sebag, Adaptive coordinate descent, in *Genetic and Evolutionary Computation Conference (GECCO)* (ACM Press, New York, 2011), pp. 885–892

25. J.E. Dennis Jr., V. Torczon, Direct search methods on parallel machines. SIAM J. Optim. **1**(4), 448–474 (1991)
26. P.A. Jensen, J.F. Bard, *Operations Research Models and Methods* (Wiley, New York, 2003)
27. S. Agarwal, K. Deb, A. Pratap, T. Meyarivan, Physical programming: effective optimization for computational design. IEEE Trans. Evol. Comput. **6**(2), 182–191 (2002)
28. J. Koski, R. Silvennoinen, Norm methods and partial weighting in multicriterion optimization of structures. Int. J. Numer. Methods Eng. **24**, 1101–1121 (1987)
29. D.G. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, Reading, 1984)
30. H.D. Sherali, M.S. Bazaraa, C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd edn. (Wiley, New York, 1993)
31. K.V. Mardia, R.J. Marshall, Maximum likelihood estimation for models of residual covariance in spatial regression. Biometrika **71**(1), 135–146 (1984)
32. R.T. Marler, J.S. Arora, Survey of multi-objective optimization methods for engineering. Struct. Multidiscip. Optim. **26**, 369–395 (2004)
33. K.I.M. McKinnon, Convergence of the Nelder–Mead simplex method to a nonstationary point. SIAM J. Optim. **9**, 148–158 (1998)
34. A. Messac, Physical programming: effective optimization for computational design. AIAA J. **34**, 149–158 (1996)
35. J.A. Nelder, R. Mead, A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)
36. Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper (2010)
37. J. Nocedal, S.J. Wright, *Numerical Optimization*. Springer Series in Operations Research (Springer, New York, 1999)
38. N. Otto, K.N. Otto, A formal representational theory for engineering design. Technical report, Ph.D. Thesis, California Institute of Technology (1992)
39. M.R. Kirby, P.A. Barros Jr., D.N. Mavris, Impact of sampling techniques selection on the creation of response surface models. SAE Trans., J. Aerosp. **113**, 1682–1693 (2004)
40. M. Papadrakakis, G. Pantazopoulos, A survey of quasi-Newton methods with reduced storage. Int. J. Numer. Methods Eng. **36**, 1573–1596 (1993)
41. C. Rassmussen, C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Boston, 2006)
42. M. Richtárik, P. Takáč, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. arXiv:1107.2848 (2011)
43. M.J. Scott, Formalisms for negotiation in engineering design, in *The 1996 ASME Design Engineering Technical Conference and Computers in Engineering Conference*, 1996
44. M.J. Scott, E.K. Antonsson, Aggregation functions for engineering design tradeoffs, in *Fuzzy Sets and Systems* (1998), pp. 253–264
45. D. Shanno, Conditioning of quasi-Newton methods for function minimization. Math. Comput. **24**(111), 647–656 (1970)
46. J.C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control* (Wiley, New York, 2003)
47. M. Stein, Large sample properties of simulations using Latin hypercube sampling. Technometrics **29**, 143–151 (1987)
48. R.M. Lewis, T.G. Kolda, V. Torczon, Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. **45**(3), 385–482 (2003)
49. R. Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
50. A.N. Tikhonov, On the stability of inverse problems. Dokl. Akad. Nauk SSSR **39**(5), 195–198 (1943) (in Russian)
51. V. Torczon, Pattern search methods for nonlinear optimization. SIAG/OPT Views News **6**, 7–11 (1995)
52. V. Torczon, M.W. Trosset, From evolutionary optimization to parallel direct search: pattern search algorithms for numerical optimization. Comput. Sci. Stat. **29**, 396–401 (1998)

53. M.W. Trosset, I know it when I see it: toward a definition of direct search methods. SIAG/OPT Views News **9**, 7–10 (1997)
54. P. Tseng, Fortified-descent simplicial search method: a general approach. SIAM J. Optim. **10**, 269–288 (1999)
55. H. Voogd, *Multicriteria Evaluation for Urban and Regional Planning* (Pion, London, 1983)
56. L.A. Zadeh, Optimality and non-scalar-valued performance criteria. IEEE Trans. Autom. Control **AC-8**, 59–60 (1987)
57. W.I. Zangwill, *Nonlinear Programming: A Unified Approach* (Prentice-Hall, Englewood Cliffs, 1969)
58. A.A. Zhigljavsky, *Theory of Global Random Search* (Kluwer Academic, Dordrecht, 1991)

Chapter 3
Optimization with Surrogate Models

Tom Schaul

Abstract In this chapter, we show how artificial curiosity can be used to focus on the most pertinent search points in black-box optimization. We present a novel response surface method, which employs a memory-based model to estimate the interestingness of each candidate point using Gaussian process regression. For each candidate point this model estimates expected improvement and yields a closed-form expression of expected information gain. The algorithm continually pushes the boundary of a *Pareto-front* of candidates not dominated by any other known point according to both an information and a cost criterion. This makes the exploration–exploitation trade-off explicit, and permits maximally informed search point selection. We illustrate the robustness of our approach in a number of experimental scenarios.

3.1 Introduction

Within the field of design optimization, a wide variety of methods exist to move from an initial "guess" at a solution, to the final "best" solution. The most rigorous method for approaching such a multi-dimensional problem would be to search every possible permutation of all the design variables, and select the optimal solution after the fact. This method does have the advantage of determining the optimal design solution with no uncertainty; however, for any scenario where the time to evaluate/simulate a single design variant is non-trivial, the overall optimization time using this approach is prohibitively expensive.

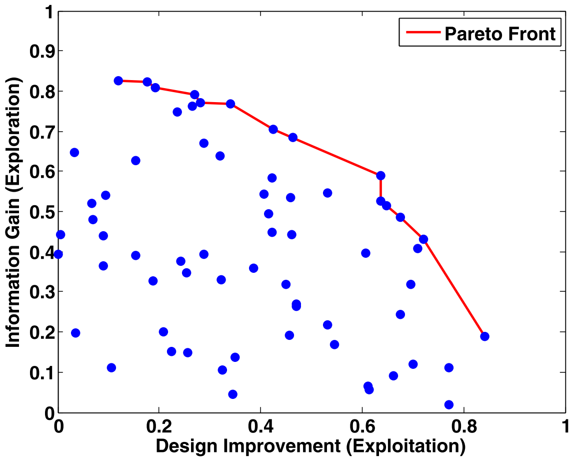
At the other end of the spectrum is a completely random search, with no methodology behind the selection of individual test points, and the total number of function calls is completely up to the researcher. This technique has obvious limitations in that, while the total optimization can be relatively short, there is little to no certainty that the optimal solution corresponds to the best possible solution.

The optimization method described in this chapter moves a step away from the most rigorous method and attempts to generate a model of the design performance

T. Schaul (✉)
Courant Institute of Mathematical Sciences, New York University, New York, NY 10003, USA
e-mail: schaul@gmail.com

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_3,

Fig. 3.1 Example of candidate test points in information gain/expected improvement space. The *Pareto front* is shown as the *solid red curve*



at every point in the optimization space using select test points throughout. The overall performance of this type of optimization is based on a delicate balance between the number of test points required to generate an accurate model at all points and the drive towards minimizing the total number of iterations required to find the optimal solution. In this sense, we attempt to generate a "maximally-predictive, yet minimally complex" model of the design space. This method of *artificial curiosity* represents one technique for directing exploration towards the most informative or most *interesting* data, and in this chapter, we show how it can be used to focus on the most pertinent search points required to generate an accurate model for black-box optimization problems.

We present a novel response surface method, which employs memory of all prior test points to estimate the "interestingness" of each candidate point using Gaussian process regression. For each candidate point, this model estimates expected improvement in the surrogate model from the new test point, and yields a closed-form expression of this expected information gain. This information gain represents the "interestingness" of every possible test point. The set of all candidate test points are plotted in the two-dimensional space described by the expected information gain and the expected improvement in design performance from the current optimal design. The candidate points that are Pareto optimal represent the subset of all points where it is not possible to improve either expected information gain or expected improvement without deteriorating the other. We refer to these candidate points as being non-dominated with respect to expected improvement and expected information gain. These points represent the *Pareto front* of candidates not dominated by any other known point. An example of such a plot, along with the *Pareto front,* is shown in Fig. 3.1. The algorithm continually pushes the boundary of the *Pareto front* of candidates not dominated by any other known point according to both an information and a cost criterion. This balance between information gain (exploration) and improvement in optimal design (exploitation) permits a maximally informed search point selection. We first describe quantitatively how the relevant parameters for this search are derived and defined, and then illustrate the ability of our approach in a number of experimental scenarios.

3.2 Background

For costly optimization, even a small reduction in the required number of function evaluations justifies a significant investment of computational resources. It is therefore common to store all previous evaluations in memory, and use them to prune the search space, estimate variability and if possible, predict which regions are likely to obtain the best values of the objective (cost) function.

A *surrogate model* is an auxiliary function that is built from those known points, and interpolates between them. It is called 'surrogate' because it can be evaluated cheaply, and many such evaluations can then be used to determine the next point at which to evaluate the costly objective function. That new evaluation is then incorporated into a refined surrogate model, and so on.

The main class of global optimization algorithms that use surrogate models are *response surface methods* (RSM; [3, 19]). They store all available evaluations (some possibly given in advance) and use them to model the cost function, which is useful for dimensionality reduction, visualization, assessing uncertainty in the surrogate model, and ultimately determining good points to explore [1, 15]. A multitude of regression techniques have been used for modeling the response surface, from the original polynomials [3] to more recent Gaussian processes that will be described here [14, 22].

3.3 Artificial Curiosity

The ability to focus on novel, yet learnable patterns in observations is an essential aspect of intelligence that has led mankind to explore its surroundings, all the way to our current understanding of the universe. When designing artificial agents, we have exactly this vision in mind; however, if an artificial agent is to exhibit some level of intelligence, or at least the ability to learn and adapt quickly in its environment, then it is essential to guide this agent to experience such patterns, a drive known as *artificial curiosity* [23, 24, 26]. This approach requires a principled way to judge and rank data, in order to drive itself towards observations exhibiting novel, yet learnable patterns. This property is compactly captured by the subjective notion of *interestingness*. Artificial learning agents are dependent on the interestingness of their observations. A number of formalizations of interestingness exist, although some of these have shortcomings. Our aim here is to find a formal measure of interestingness that can be used to guide exploration in the case of black-box optimization.

Curiosity is the drive to actively explore the interesting regions in search space that most improve the model's predictions or explanations of what is going on in the world. Originally introduced for reinforcement learning [23, 29], the curiosity framework has been used for active learning [9, 20], to explain certain patterns of human visual attention better than previous approaches [13], and to explain concepts such as beauty, attention and creativity [25, 26]. In this context, the amount that currently observable data can be compressed corresponds to its perceived beauty. The simpler the method for encoding a given set of data, the more beautiful it is, and obviously, the most simple method of encoding corresponds to the most beautiful or aesthetically pleasing. Additionally, "creative" behaviors are considered those that produce new types of interesting constructs. When actively selecting new observables to study, the most interesting data is that which most improves data compression, and correspondingly, things that are random or too difficult to understand are considered boring.

In this chapter, we use curiosity algorithms to explore the parameter space for a given design optimization problem, and determine the effectiveness of such an approach. The interestingness of a new observation is the difference between the performance of an adaptive model on the observation history before and after including the new point. In essence, this concept represents how much new information or knowledge is gained about the system being optimized from evaluating the objective function at a given point. The goal of the active data point selection mechanism is to maximize expected cumulative future interestingness. Various proposed distance measures to quantify the amount of information gain include:

- The difference in data compressibility before and after letting the learning algorithm take the new data into account [25, 26],
- The difference in mean squared prediction error on the observation history before and after re-training with the new point [23, 24],
- The Kullback–Leibler (KL) divergence which represents the change in information that is lost when our surrogate model is used to approximate the actual cost function space before and after the new test point has been evaluated [29].

Note that interestingness is observer-dependent and dynamic: a point that was interesting early on can become boring over time.

3.3.1 Artificial Curiosity as a Guide for Optimization

Motivated by *costly optimization* (as described in Sect. 3.2), we here aim to define an appropriate variant of artificial curiosity that can guide the exploratory drive of an optimization algorithm to pick the most interesting next search point, and thus minimize the number of evaluations required.

The optimization framework is a restricted case of the KL scenario for which artificial curiosity has been put forth, in that objective evaluations are atemporal (the order in which a set of points is evaluated does not affect the value). This allows for a simplified measure of interestingness that only captures the instantaneous informativeness of a search point.

Further, to permit the incorporation of a Bayesian prior, we will focus on probabilistic models and use a particular variant of the KL-based approach [29] to maximize *information gain* [4, 5, 10, 12, 18, 21]. The KL-divergence or relative entropy between prior and posterior (before and after seeing the new point) is invariant under any transformation of the parameter space.

3.3.2 Formal Framework

Formally, let Y_{env} be the environment of interest, and y_{pre} be our current knowledge.[1] The information gain (interestingness) $\psi(y|y_{\text{pre}})$ brought about by the observation y is defined as

$$\begin{aligned}\psi(y|y_{\text{pre}}) &= \mathbb{D}\big(\pi(Y_{\text{env}}|y_{\text{pre}};y)\|\pi(Y_{\text{env}}|y_{\text{pre}})\big)\\ &= \int \pi(y_{\text{env}}|y_{\text{pre}};y)\log\frac{\pi(y_{\text{env}}|y_{\text{pre}};y)}{\pi(y_{\text{env}}|y_{\text{pre}})}dy_{\text{env}},\end{aligned}$$

where $\pi(\cdot|\cdot)$ denotes a conditional probability and $\mathbb{D}(\cdot\|\cdot)$ denotes the KL-divergence. For a set of observations y_{pre}, it is also useful to define the leave-one-out (LOO) information gain for each observation y_o w.r.t. the remaining $y_{\text{pre}\backslash o}$ as

$$\psi_{\text{LOO}}(y_o) = \psi(y_o|y_{\text{pre}\backslash o}).$$

This method of cross-validation is used to quantify how accurately our predictive model is in terms of determining how much information can be gained from the untested candidate points. Here we remove one of the previously tested points from the set of measured data that generates our model. The surrogate is then computed based on the remaining points and used to estimate the value of the point that had been removed. Finally, this result is compared with the actual value of the removed, validation point.

The information gain $\psi(y|y_{\text{pre}})$ is defined a posteriori, meaning that it is only defined after we see the value y. However, in most cases, we want to assess the information gain of an observation a priori, i.e., before seeing the value. This leads to the *expected information gain* of random variable Y, defined by

$$\begin{aligned}\Psi(Y|y_{\text{pre}}) &= \mathbb{E}\big[\psi(Y|y_{\text{pre}})\big]\\ &= \int \pi(y|y_{\text{pre}})\int \pi(y_{\text{env}}|y_{\text{pre}};y)\log\frac{\pi(y_{\text{env}}|y_{\text{pre}};y)}{\pi(y_{\text{env}}|y_{\text{pre}})}\,dy_{\text{env}}\,dy\\ &= I(Y;Y_{\text{env}}|y_{\text{pre}}),\end{aligned}$$

which turns out to be the conditional mutual information between the observation and the environment.

[1] We use upper case letters for random variables, and the corresponding lower case letters for specific configurations.

3.4 Exploration–Exploitation Trade-Off

In this section, we propose a way of handling the fundamental exploration–exploitation trade-off. To make informed data selection decisions, they are postponed until the entire Pareto-optimal front—with respect to both an exploration and an exploitation objective—is known.

3.4.1 Exploration Objective: Information Gain

The previous section introduced maximal expected information gain as a possible objective for exploration, as an instantiation of the curiosity principle. It determines the points that will provide the most additional information about the surrogate model of the objective function. However, if this metric were the only basis for selecting optimization test points, the optimization routine would in no way be directed towards finding an optimal solution. As a result, we need a second objective.

3.4.2 Exploitation Objective: Expected Cost Improvement

The idea behind objective function "*exploitation*" is the more traditional concept of finding the ideal solution to an optimization problem. We note that in optimization, there is an asymmetry in utility: solutions that are better than the best currently found $f_{\min}$ largely outweigh those that are almost as good. Thus exploitation really aims at minimizing the *expected improvement* in cost with respect to $f_{\min}$. It can be shown [14] that the expected improvement takes the following form:

$$\Delta(x) = \sigma(Y|y_o)\big(s\Phi(s) + \phi(s)\big), \tag{3.1}$$

where

$$s = \frac{f_{\max} - \mathbb{E}[Y|y_o]}{\sigma(Y|y_o)},$$

while $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution and density functions of Gaussian distributions, respectively.

3.4.3 Combining the Objectives

Optimizing conflicting objectives necessarily involves some form of trade-off, many of which were discussed in Chap. 2. One of the most common ways to handle this problem is by using a weighted sum of both objectives, where the weights are set manually, or tuned to the problem. Combining two objectives of different scale into a single utility measure is common practice [4], but problematic [30], as one of the objectives can completely dominate in some regions of the cost landscape, while being dominated in others. As a result, which ever term is larger in magnitude within a given region of the objective function surrogate will tend to dominate the performance of the optimization, and yield sub-optimal results.

Therefore, we propose turning the problem around and only deciding on the trade-off after *first* having evaluated both objectives for a large set of candidate points. This means finding the Pareto front of candidates that are non-dominated w.r.t. expected improvement and expected information gain, which can be performed by any multi-objective optimization method, for example, the Non-dominated Sorting Genetic Algorithm version II (NSGA-II; [8]) which is used in the experiments in Sect. 3.6.4. This algorithm is a significant improvement over previously used sorting algorithms that suffer from poor scaling of the computational complexity, inability to select the best points (elitism) which directly affects the speed of the algorithm, and a dependence on user-defined inputs. The original NSGA algorithm [28] still suffered from its reliance on a user-defined sharing function approach, which maintained the spread of solutions in the search space and kept the population diverse. As a result, the performance of the algorithm was highly dependent on the chosen value of the sharing parameter. Additionally, scaling becomes an issue with this approach since each solution must be compared with every other solution. This results in an approach that scales as $\mathcal{O}(n^2)$ where n is the population size.

NSGA-II utilizes a crowded-comparison approach to solve the problems listed above. In this approach, the density of solutions around a given point is determined by looking at the average distance to two of the neighboring points around a given test point in the exploration-exploitation space. These distances form a rectangle around the given point and the average of these two distances is called the crowding distance. These values are then sorted by magnitude [8]. The Pareto front is generated by first ordering points based on the extent to which they are non-dominated, and then by the crowding distance. A new test point is then selected from the Pareto front and the process is repeated to determine the new Pareto front.

All non-dominated candidates are considered "good" solutions, and therefore each should be assigned a non-zero probability of being chosen. Ideally, this probability should favor candidates that stand out on the Pareto front, in terms of combining both objectives, but it should also be insensitive to quirks of the algorithm that builds the front (i.e., varying candidate densities), and to any smooth order-preserving transformation of the cost function. In addition, it can allow us to shift the focus from one objective to the other, e.g., focusing more on finding the optimal solution over time. In the absence of an optimal way of handling this decision, we opt for an unbiased solution, which consists in choosing the next point uniformly at random from the Pareto front.

3.5 Curiosity-driven Optimization

Algorithm 3.1 combines the components described in the previous section into a general framework for curiosity-driven optimization. In each iteration, it first fits a probabilistic model M_f to all known points X, and then uses M_f to determine the LOO-information gain at each point. This interestingness function is then approximated using a second model, M_ψ. The Pareto front of the candidate points is then computed using a multi-objective optimization algorithm, each model providing an estimate for one of the two objectives. Finally, a new point x^* is chosen, as described in Sect. 3.4.

Algorithm 3.1: Curiosity-Driven Optimization

```
input: cost function $f$, models $M_f$ and $M_\psi$, initial points $X$
repeat
    Fit $M_f$ to $X$
    for $s$ in $X$ do
        $\psi_{\mathrm{LOO}}(s) = \mathbb{D}(M_f(X) \| M_f(X_{-s}))$
    end
    Fit $M_\psi$ to $\psi_{\mathrm{LOO}}$
    Find a set $C$ of non-dominated candidate points
        maximizing information gain (estimated by $M_\psi$) and
        minimizing cost (estimated by $M_f$)
    Choose $x^*$ from $C$
    $X \leftarrow X \cup \{(x^*, f(x^*))\}$
until stopping criterion is met
```

3.5.1 Models of Expected Cost and Information Gain

The class of probabilistic models used for M_f and M_ψ should be general and flexible enough to fit multi-modal and highly nonlinear cost functions. Ideally, for every unknown point, such a model should be able to (efficiently) predict the expected value, the expected uncertainty associated with that prediction, and provide an analytical expression for computing information gain.

One option would be to use a mixture of Gaussians on the joint parameter–cost function space (as in [6]). However, this approach has the drawback of being sensitive to the number of Gaussians used, as well as giving poor interpolation in regions with few sampled points. Neural networks are another viable option, but come with a propensity of overfitting the scarce data.

3.5.2 A Good Model Choice: Gaussian Processes

In this section, we present an implementation of curiosity-driven optimization which satisfies all the above criteria by using *Gaussian processes* to model the cost function.

Gaussian processes (GP, [22]) can be seen as a probability distribution over functions, as evaluated on an arbitrary but finite number of points. Given a number of observations, a Gaussian process associates a Gaussian probability distribution to the function value for each point in the input space. Gaussian processes are capable of modeling highly complex cost landscapes through the use of appropriate covariance (kernel) functions, and are commonly used for regression and function modeling [22].

Formally, we consider the Gaussian process with zero mean and the kernel function

$$k\bigl(x, x'\bigr) + \sigma_n^2 \delta\bigl(x, x'\bigr),$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta function. Thus, for any values y, y' at x, x', $\mathbb{E}[yy'] = k(x, x') + \sigma_n^2$. We make the assumption that the function k is smooth and local in the sense that $k(x, x') \to 0$ when $|x - x'|$ goes to infinity.

3.5.3 *Derivation of Gaussian Process Information Gain*

The concept of information gain can easily be mapped onto Gaussian processes, but previous work has failed to provide a closed form expression for efficiently computing it for each candidate point [16]. Let us consider a collection of fixed reference points x_r, and view their value Y_r as the environment of interest. Our prior knowledge y_{pre} consists of all previously evaluated points x_o with value y_o. The expected information gain of the value Y at point x is thus defined by

$$\Psi_r(x|y_o) = I(Y_r; Y|y_o) = H(Y|y_o) - H(Y|Y_r, y_o),$$

where $H(a|b)$ is the conditional entropy, i.e., the amount of information needed to describe a given b. A major disadvantage of this definition is that the expected information gain depends on the reference points x_r. However, we may consider the situation where the number of reference points goes to infinity. By definition, $\pi(Y|Y_r, y_o)$ is a Gaussian distribution, and

$$H(Y|Y_r, y_o) = \frac{1}{2} \log 2\pi e \sigma^2(Y|Y_r, y_o).$$

Here $\sigma^2(Y|Y_r, y_o)$ is the predictive variance at x given by

$$\begin{aligned}\sigma^2(Y|Y_r, y_o) &= \sigma_n^2 + k(x, x) - k(x, x_{ro})\bigl(k(x_{ro}, x_{ro}) + \sigma_n^2 \mathbb{I}\bigr)^{-1} k(x_{ro}, x) \\ &= \sigma_n^2 + \sigma_e^2\end{aligned}$$

with $x_{ro} = [x_r, x_o]$. We take advantage of the fact that in GP, the predictive variance depends only on the location of observations. In particular, σ_e^2 is the variance of the predicted mean $\bar{y} = \mathbb{E}[Y]$, and

$$0 \le \sigma_e^2 = \sigma^2(\bar{y}|Y_r, y_o) \le \sigma^2(\bar{y}|Y_r) = \sigma_s^2$$

because conditioning always reduces the variance for Gaussian distributions. According to Rasmussen and Williams [22], σ_s^2 converges to 0 when the number of reference points x_r around x goes to infinity. This indicates that σ_e^2 converges to 0 uniformly w.r.t. y_o, thus $\sigma^2(Y|Y_r, y_o)$ converges to σ_n^2 uniformly w.r.t. y_o.

When the number of observation points is sufficiently large around any given point x, we have

$$\begin{aligned}\Psi_r(x|y_o) &= H(Y|y_o) - H(Y|Y_r, y_o)\\ &\to \frac{1}{2}\log 2\pi e\sigma^2(Y|y_o) - \frac{1}{2}\log 2\pi e\sigma_n^2\\ &= \frac{1}{2}\log \sigma^2(Y|y_o) - \frac{1}{2}\log \sigma_n^2.\end{aligned}$$

The limit no longer depends on the reference points, thus it can be used as an 'objective' measure for the expected information gain at point x:

$$\Psi(x|y_o) = \frac{1}{2}\log \sigma^2(Y|y_o) - \frac{1}{2}\log \sigma_n^2. \tag{3.2}$$

The second term is constant; therefore, there is a direct connection between the expected information gain and the predictive variance given the observation, which can be computed efficiently. Note that [27] found the predictive variance to be a useful criterion for exploration, without realizing that it is equivalent to information gain.

3.5.4 Curiosity-Driven Optimization with Gaussian Processes

Choosing a Gaussian process to model the cost function significantly simplifies the general algorithm introduced in Sect. 3.5. First, it allows us to compute the expected information gain Ψ instead of the less robust LOO-information gain. Second, the model M_ψ is no longer necessary, as Ψ can be computed for unknown points as well. The resulting algorithm (CO-GP) is shown in Algorithm 3.2. The remainder of this section discusses some practical considerations.

Algorithm 3.2: Curiosity-driven Optimization with Gaussian processes (CO-GP)

```
input: cost function $f$, kernel function, initial points $X$
repeat
    Fit a Gaussian process $\mathcal{G}$ to $X$
    Find a set $C$ of non-dominated candidate points $x$ simultaneously
    maximizing $\Psi_{\mathcal{G}}(x)$ (Eq. (3.2)) and $\Delta_{\mathcal{G}}(x)$ (Eq. (3.1))
    Choose $x^*$ from $C$
    $X \leftarrow X \cup \{(x^*, f(x^*))\}$
    Optionally optimize the kernel hyperparameters w.r.t. the marginal
    likelihood
until stopping criterion is met
```

Computational Complexity The computational complexity of each iteration of CO-GP is dominated by one matrix inversion of $\mathcal{O}(n^3)$, where n is the total number of evaluations. Building the Pareto front consumes most of the computation time early on, but scales with $\mathcal{O}(n^2)$. The computational complexity of Gaussian processes can be reduced, e.g., by implementing them online and using a reduced base vector set, containing only the most informative points [7]. We have not implemented these yet, as computation time was not a major concern in our experiments since it is relatively small compared to the time required for device simulations.

Choosing Kernel Parameters: Model Selection Gaussian process regression only gives reasonable results when the kernel hyperparameters are set properly. Depending on how much computation time is available, the hyperparameters could be optimized periodically with respect to the marginal likelihood. We use the exponential natural evolution strategies algorithm (xNES) for this purpose [11]. Potentially, we could also employ diagnostic methods [15] to determine whether the model is appropriate.

Informed Multi-objective Search At each iteration (line 2), an inner multi-objective optimization algorithm is used (in our case, NSGA-II from Sect. 3.4.3). We can make use of our available information to make this step more efficient. For example, we initialize the search with the Pareto front found in the previous iteration. Furthermore, as we want the search to roughly cover the range of known points, we adjust the scale (for step-sizes) accordingly.

3.6 Minimal Asymptotic Requirements

To demonstrate the practical viability of CO-GP, we first investigate how it handles a number of common but problematic scenarios and then illustrate it on a standard benchmark function. Following [22], all experiments use the common Gaussian kernel (also known as a radial basis function) with noise, which is a very robust choice in practice.

3.6.1 Reaching Optima at Arbitrary Distance

Many cost function landscapes contain large linear regions. Specifically, if the scale of the region covered by the initial points is too small, almost any landscape will appear linear. An ideal algorithm should be able to exploit the linearity of such regions. In particular, it is highly desirable to have the searched region grow exponentially with the number of points. Note that many well-known algorithms, such as estimation of distribution algorithms, do not have this property and instead rely either on correct initialization or heuristics [2]. In contrast, CO-GP does have this property, as our results on the linear function show (see Fig. 3.2).

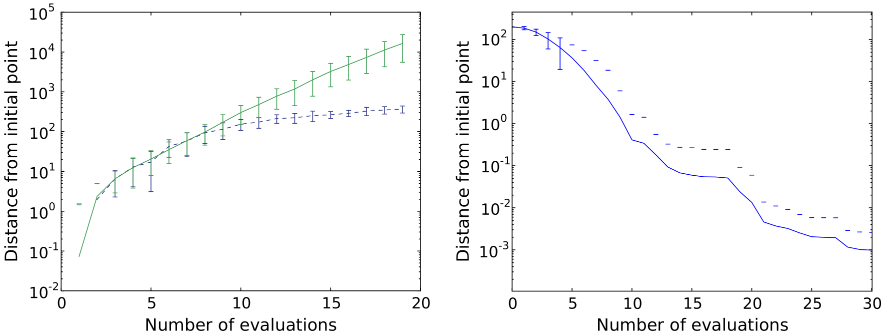


Fig. 3.2 *Left*: Escaping linear regions. The plot shows the performance of CO-GP on a linear surface, in terms of the distance from the best point found so far to the initial point (averaged over 20 independent runs). The *green* (*solid*) and *blue* (*dashed*) *curves* correspond to CO-GP with hyperparameter adaptation enabled and disabled, respectively. We observe that the distance from the initial point (and thus the decrease in cost) grows exponentially if the hyperparameters are adapted, but only linearly otherwise. *Right*: Precisely locating optima. The plot shows the performance of CO-GP on a unimodal, quadratic surface, in terms of the distance from the optimum to the best point found so far (averaged over 20 independent runs). This distance decreases exponentially with the number of points

3.6.2 Locating Optima with Arbitrary Precision

While designed primarily for multi-modal cost landscapes, we investigated how our approach handles simple cost landscapes with a single optimum. The success criterion for this case is to have the distance to the optimum decrease exponentially with the number of points. While we cannot prove that this is the case in general, Fig. 3.2 shows that it holds for the multi-dimensional sphere function. This indicates that CO-GP can locate optima up to a high precision, at least whenever, locally, the cost function is approximately quadratic.

3.6.3 Guaranteed to Find Global Optimum

Every global optimization algorithm should provide a guarantee that in the limit its chosen points will cover the search space densely, which is the only way to ensure that it will eventually find the global optimum. Optimization based on expected improvement has been shown to have this property [17]. It turns out that if we remove the information gain objective from CO-GP, the algorithms are equivalent. Therefore, as one extreme of the Pareto front will always correspond to the point maximizing expected improvement exclusively, and that point has a non-zero probability of being chosen, CO-GP inherits the property that it always finds the global optimum in the limit.

3.6.4 Proof-of-Concept

The Branin function [14, 15] is a commonly used benchmark for global optimization of the form:

$$f_{\text{Branin}}(x_1, x_2) = a\big(x_2 - bx_1^2 + cx_1 - d\big)^2 + e(1-f)\cos(x_1) + e,$$

where the standard parameter settings are $a = 1$, $b = \frac{5}{4\pi^2}$, $c = \frac{5}{\pi}$, $d = 6$, $e = 10$, $f = \frac{1}{8\pi}$. The function has three global optima, at $(-\pi, 12.275)$, $(\pi, 2.275)$ and $(9.42478, 2.475)$, with the value $f_{\text{Branin}}(x_\star) = 0.397887$, a bounded domain and a non-trivial structure. Figure 3.3 illustrates the behavior of CO-GP on the Branin function over the course of a single run, starting with four points on the boundary corners. The Gaussian process model produces a good fit of the true function after about 30 iterations. Locating one optimum (up to a precision of 0.1) requires only 28 ± 8 evaluations, locating all three requires 119 ± 31. The qualitative behavior of the algorithm is very intuitive, placing part of its search points spaced broadly within the domain, while the other part forms clusters of points ever closer around the optima. Although this experiment is intended as a proof of concept, not an empirical comparison to other global optimization algorithms, the quantitative results indicate that CO-GP is on par with the best results reported in the literature [14].

3.7 Discussion

The results in Sect. 3.6 demonstrate that CO-GP properly handles a number of typical optimization challenges, and the results bode well for applying the general template (Algorithm 3.2) to related domains such as constrained or discrete optimization, or even mixed-integer programming. Although based on the general and theoretically powerful principle of artificial curiosity, our current implementation exhibits certain weaknesses: Despite the derived closed-form expressions for the objectives, the method's computational cost is still high, limiting the application domain to (relatively) costly optimization problems.

Another drawback of the approach is that it is greedier than the original curiosity framework: it does not necessarily maximize cumulative future expected information gain, but greedily selects the next data point that is expected to be the immediately most interesting. More sophisticated reinforcement learning algorithms will be necessary to maximize the expected *sum* of future intrinsic rewards (each reward being proportional to the information gain of the corresponding observed data).

Further, as the dimensionality of the problem increases, the number of seed points required to build an accurate surrogate model increases exponentially. The current focus on exploration over exploitation is problematic not only because of the computational cost involved in the matrix inversions necessary to determine future test points, but also because of the total time required to evaluate all test points. Because we do not know the exact behavior of a response surface a priori, it is not possible to determine the exact number of test points that are required. The ability to confidently

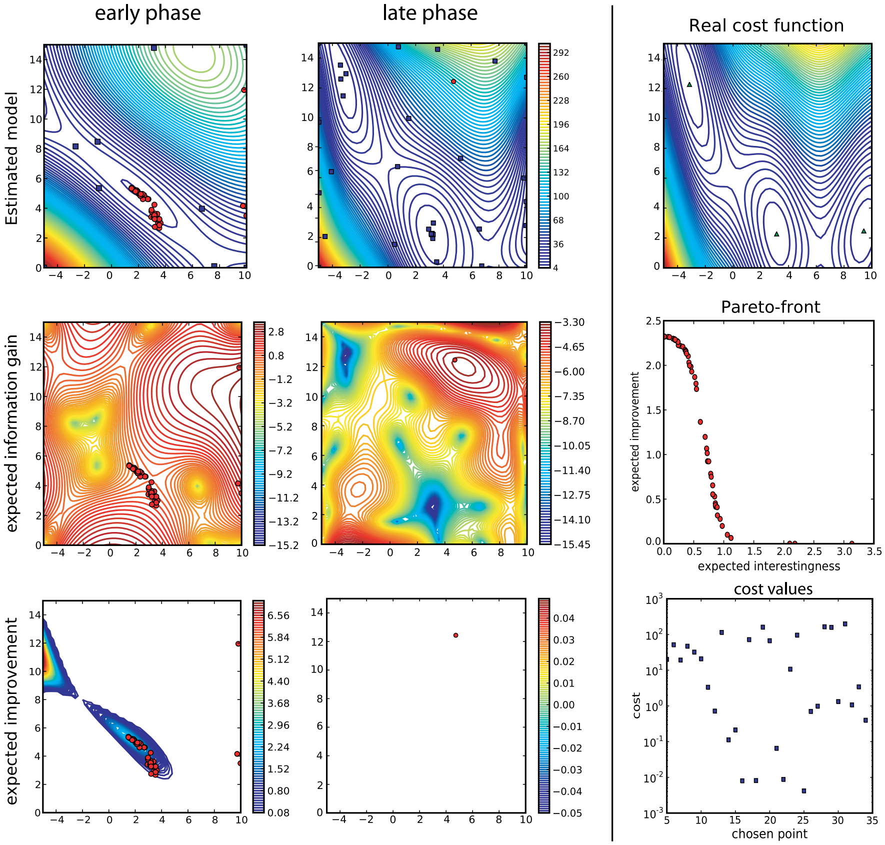


Fig. 3.3 Optimization on the Branin function. The *plot in the top right corner* shows a contour plot of the Branin function with the three global optima marked with triangles. The *left column* shows the estimated cost function model (*top*), and the two competing objectives, expected information gain (*middle*) and expected improvement (*bottom*), in an early phase after 10 points (*blue squares*) have been observed. The *red circles* are the points on the Pareto front being considered for the next choice. The *middle column* shows the same information after 30 iterations. Note that in this later stage the model is very close to the true function (*top*). The *plot in the middle of the right column* shows the shape of the Pareto front corresponding to the situation in the *left column* (10 points), and the *plot on the bottom right* shows the values of the cost function at all the chosen points (the initial 4 corner points are not shown). In the early phase, the Pareto front contains a continuum of points in the center that trade off improvement and information gain, plus a few isolated points with high information gain, but very low expected improvement. After 30 iterations, two of the global optima have been located precisely. The expected improvement is zero everywhere, so the Pareto front is collapsed to the single point with highest information gain. CO-GP now performs a purely exploratory step, and will continue to do so until it leads to non-zero expected improvement (e.g. around the third optimum). On average, CO-GP requires about 120 points to locate all three optima with high accuracy

determine that a predicted global optimum is the actual global optimum, is directly related to the maximum predictive variance throughout the surrogate, which again cannot be determined a priori. Using Gaussian Processes, the number of test points required increases in a strongly nonlinear manner based on each of the key regions of maximum, minimum, discontinuities, and stability within the response surface. As the number of optimization dimensions increases, surrogate models, including CDO, spend hundreds or thousands of function calls in the exploration phase of the optimization. This is simply the result of the fact that in the initial stages of the optimization, it is unlikely that the optimum points in the surrogate model accurately represent the true, global optimum.

As a result, ongoing efforts are focused on model simplification and dimensionality reduction techniques. A detailed review of such techniques, along with their scaling behavior in higher dimensions is given in [22]. As the dimensionality of the problem increases; there is an increasing shift towards finer-grained methods which trend towards selecting sample points one at a time. To that end, methods discussed in Chaps. 4 and 5 currently prove to be much more efficient at rapid convergence in higher dimensions. The trade-off essentially being that the speed of convergence to an optimum (within the convergence criteria) is improved by sacrificing both knowledge of the entire response surface and knowledge that the solution obtained is indeed the global optimum.

References

1. A.J. Booker, J.E. Dennis Jr., P.D. Frank, D.B. Serafini, Optimization using surrogate objectives on a helicopter test example, in *Computational Methods in Optimal Design and Control*, 1998
2. P. Bosman, J. Grahl, D. Thierens, Enhancing the performance of maximum-likelihood Gaussian EDAs using anticipated mean shift, in *Parallel Problem Solving from Nature (PPSN X)* (2008), pp. 133–143
3. G.E.P. Box, K.B. Wilson, On the experimental attainment of optimum conditions. J. R. Stat. Soc. **13**(1), 1–45 (1951)
4. K. Chaloner, I. Verdinelli, Bayesian experimental design: a review. Stat. Sci. **10**, 273–304 (1995)
5. D.A. Cohn, Neural network exploration using optimal experiment design, in *Advances in Neural Information Processing Systems* (1994), pp. 679–686
6. D.A. Cohn, Z. Ghahramani, M.I. Jordan, Active learning with statistical models. J. Artif. Intell. Res. **4**, 129–145 (1995)
7. L. Csató, M. Opper, Sparse on-line Gaussian processes. Neural Comput. **14** (2002)
8. K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**, 182–197 (2002)
9. M.P. Deisenroth, C.E. Rasmussen, J. Peters, Gaussian process dynamic programming, in *Neurocomputing* (2009), pp. 1508–1524
10. V.V. Fedorov, *Theory of Optimal Experiments* (Academic Press, New York, 1972)
11. T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, J. Schmidhuber, Exponential natural evolution strategies, in *Genetic and Evolutionary Computation Conference (GECCO)*, Portland, OR, 2010
12. J.-N. Hwang, J.J. Choi, S. Oh, R.J.I. Marks, Query-based learning applied to partially trained multilayer perceptrons. IEEE Trans. Neural Netw. **2**, 131–136 (1991)

13. L. Itti, P. Baldi, Bayesian surprise attracts human attention, in *Advances in Neural Information Processing Systems*, ed. by Y.W. Platt, B. Schölkopf (MIT Press, Cambridge, 2006), pp. 547–554
14. D.R. Jones, A taxonomy of global optimization methods based on response surfaces. J. Glob. Optim. **21**, 345–383 (2001)
15. D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**, 455–492 (1998)
16. A. Krause, C. Guestrin, Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach, in *Proceedings of the International Conference on Machine Learning*, 2007
17. M. Locatelli, Bayesian algorithms for one-dimensional global optimization. J. Glob. Optim., 57–76 (1997)
18. D.J.C. MacKay, Information-based objective functions for active data selection. Neural Comput. **4**, 550–604 (1992)
19. A.W. Moore, J. Schneider, Memory-based stochastic optimization, in *Advances in Neural Information Processing Systems*, 1996
20. T. Pfingsten, Bayesian active learning for sensitivity analysis, in *Machine Learning (ECML 2006)* (2006), pp. 353–364
21. M. Plutowski, G. Cottrell, H. White, Learning Mackey-glass from 25 examples, plus or minus 2, in *Advances in Neural Information Processing Systems* (1994), pp. 1135–1142
22. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006)
23. J. Schmidhuber, Curious model-building control systems, in *IEEE International Joint Conference on Neural Networks* (1991), pp. 1458–1463
24. J. Schmidhuber, Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. Connect. Sci. **18**, 173–187 (2006)
25. J. Schmidhuber, *Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity and Creativity*. Lecture Notes in Artificial Intelligence, vol. 4754 (2007)
26. J. Schmidhuber, Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes, in *Anticipatory Behavior in Adaptive Learning Systems, from Sensorimotor to Higher-Level Cognitive Capabilities*, 2009
27. S. Seo, M. Wallat, T. Graepel, K. Obermayer, Gaussian process regression: active data selection and test point rejection, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (IEEE, New York, 2000), pp. 241–246
28. N. Srinivas, K. Deb, Multiobjective function optimization using nondominated sorting genetic algorithms. Evol. Comput. **2**(3), 221–248 (1995)
29. J. Storck, J. Hochreiter, J. Schmidhuber, Reinforcement-driven information acquisition in non-deterministic environments, in *International Conference on Artificial Neural Networks (ICANN)*, Paris (1995), pp. 159–164
30. Y. Zhang, W. Xu, J. Callan, Exploration and exploitation in adaptive filtering based on Bayesian active learning, in *International Conference on Machine Learning (ICML)* (2003), pp. 896–903

Chapter 4
Metamaterial Design by Mesh Adaptive Direct Search

Charles Audet, Kenneth Diest, Sébastien Le Digabel, Luke A. Sweatlock, and Daniel E. Marthaler

Abstract In the design of optical metamaterials, some optimization problems require launching a numerical simulation. The Mesh Adaptive Direct Search algorithm is designed for such problems. The MADS algorithm does not require any derivative information about the problem being optimized, and no continuity or differentiability assumptions are made by MADS on the functions defining the simulation. A detailed discussion of the method is provided in the second section of the chapter, followed by a discussion of the NOMAD implementation of the method and its features. The final section of the chapter lists three instances of combining NOMAD with Finite-Difference Time-Domain electromagnetic simulations to tailor the broadband spectral response and near-field interactions of Split Ring Resonator metamaterials.

4.1 Introduction

This chapter describes methods for solving constrained optimization problems using the Mesh Adaptive Direct Search (MADS) algorithm, which belongs to the more broad class of *Derivative-Free Optimization* methods. Because small changes in the geometry of a metamaterial can results in large changes in the overall behavior of the structure, these techniques are well-suited for the design of optical metamaterials, and can handle the large discontinuities in the "cost function space" that often arise. The MADS algorithm does not require any derivative information about the problem being optimized, and no continuity or differentiability assumptions are made on the functions defining the simulation. Out of the many applicable techniques that can be used for metamaterial design, the NOMAD implementation of the MADS algorithm discussed in this chapter has many advantages, including built in capabilities to handle nonlinear constraints, bi-objective function optimization, sensitivity analysis, and up to 50 variables. For even larger problems, the PSD-MADS parallel version of the method is able to solve problems with 50–500 variables. Lastly,

C. Audet (✉)
GERAD and Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal, Québec H3C 3A7 Canada
e-mail: charles.audet@gerad.ca

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_4,

the NOMAD implementation has the added benefit that it is constantly being updated and improved within a number of programming languages including C++ and Matlab (http://www.gerad.ca/nomad).

4.1.1 Structuring the Optimization Problem

This chapter considers optimization problems that may be written in the following general form

$$\min_{x \in \Omega} f(x), \tag{4.1}$$

where f is a single-valued objective function, and Ω is the set of feasible solutions. The direct search methods described here can be applied without making any assumptions on the function f or on the set Ω. However, when analyzing the theoretical behavior of these methods, we will study them under various assumptions. Without any loss of generality, suppose that the set of feasible solutions is written as

$$\Omega = \left\{x \in X : c_j(x) \leq 0, \ j \in J\right\} \subset \mathbb{R}^n,$$

where X is a subset of $\mathbb{R}^n$ and $c_j : X \to \mathbb{R} \cup \{\infty\}$ for all $j \in J = \{1, 2, \ldots, m\}$ are quantifiable constraints. This means that for any x in X, the real value $c_j(x)$ provides a measure by which a constraint is violated or satisfied. This notation does not make the problem restrictive, as problems where $J = \emptyset$ are allowed.

The sets X and Ω define the feasible region and each one of them corresponds to a specific type of constraint for which different treatments are described in Sect. 4.2.2. The quantifiable constraints $c_j(x) \leq 0$, $j \in J$, defining Ω provide a distance to feasibility and/or to infeasibility. Violating these constraints is also permitted as long as these violations occur only at the intermediate candidates considered by an optimization method. The set X may contain any constraint for which a measure of the violation is not available, and/or constraints that cannot be relaxed. Typically, X contains bound constraints necessary to run the simulation, but can also include *hidden constraints* [20], which occur when the simulation fails to evaluate. In [5], the objective function failed to return a value on approximately 43 % of the simulations, and in [18], the failure rate climbed to 60 %. Such problems pose a challenge to optimization methods that use function evaluations to estimate derivatives.

Different approaches exist to tackle Problem (4.1), and this chapter discusses *Derivative-Free Optimization* (DFO) methods. This choice is justified by the fact that these methods are backed by rigorous convergence analyses based on different levels of assumptions on the nature of the functions defining the problem. This type of analysis marks the difference between DFO methods and heuristics. While this chapter focuses on the MADS algorithm to address these types of problems, a review of DFO methods may be consulted in the recent book [23], which does not focus on the constrained case. The present chapter aims at describing a practical and recent method and software for constrained blackbox optimization.

The chapter is divided as follows. Section 4.2 summarizes the general organization of the MADS algorithm, describes strategies to handle various types of constraints, discusses the use of surrogates and models to guide the optimization, and details the type of nonsmooth convergence analyses on which these methods rely. Section 4.3 describes our C++ implementation NOMAD of MADS and highlights some features that make the code versatile. Finally, Sect. 4.4 describes a metamaterial design optimization problem, shows how to formulate it as a blackbox optimization problem, and numerical experiments are conducted using the NOMAD software.

4.2 The Mesh Adaptive Direct Search Class of Algorithms

The Mesh Adaptive Direct Search (MADS) is presented in [9] as a generalization of several existing direct search methods.

The name of these methods comes from the fact that they are designed to work *directly* with the objective function values generated by the blackbox, and that they do not use or approximate derivatives or require their existence. MADS was introduced to extend the target class of problems to the constrained problem (4.1) while improving the practical and theoretical convergence results. That paper proposed a first instantiation of MADS called LTMADS, which was improved in ulterior work. The non-deterministic nature of LTMADS was corrected in the ORTHOMADS [3] instantiation. These algorithms were initially designed to handle the constraints of Ω by the so-called *extreme barrier*, which simply consists of rejecting any trial point which does not belong to Ω. The term *extreme barrier* comes from the fact that this approach can be implemented by solving the unconstrained minimization of

$$f_\Omega(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ \infty & \text{otherwise} \end{cases}$$

instead of the equivalent Problem (4.1). A more subtle way of handling quantifiable constraints is presented in [10], and is summarized in Sect. 4.2.2 below.

4.2.1 General Organization of the MADS *Algorithm*

In order to use a MADS algorithm, the user must provide an initial point denoted $x_0 \in \mathbb{R}^n$. It does not need to be feasible with respect to the quantifiable constraints $c_j(x) \leq 0$, $j \in J$, but must belong to the set X. MADS algorithms are iterative and the iteration counter is denoted by the index k. At each iteration, the algorithm deploys some effort to improve the *incumbent solution* x_k, i.e., the current *best* solution. For now, we will remain vague about the word *best* as it takes different meanings if it refers to a feasible or an infeasible solution. This clarification is made in Sect. 4.2.2

At each iteration, the algorithm tries to improve the incumbent solution by launching the simulation at a finite number of carefully selected trial points. This is done in two steps called the SEARCH and the POLL. The SEARCH is very flexible and allows the user to take advantage of his knowledge of the problem to propose some candidates where the simulation will be launched. Some SEARCH strategies are tailored to specific problems, while others are generic (e.g., speculative search [9], Latin hypercube sampling [36], variable neighborhood searches [6], surrogates). The POLL needs to satisfy more rigorous restrictions. It consists of a local exploration near the current incumbent solution, and its definition varies from one instantiation of MADS to another. In practice, the SEARCH can greatly improve the quality of the final solution, while the POLL structure allows a rigorous convergence analysis.

A fundamental requirement of both the SEARCH and POLL steps is that they must generate trial points belonging to a conceptual mesh M_k on the space of variables $\mathbb{R}^n$. The mesh is defined by a *mesh size parameter* $\Delta_k^m > 0$, by the set V_k of all trial points at which the simulation was launched before the start of iteration k, and by a finite set of positive spanning directions $D \subset \mathbb{R}^n$. Of these three elements, only D is fixed throughout the algorithm, while the two others vary from one iteration to another. In practice, the set D is often chosen to be the columns of the $n \times n$ identity matrix, together with their negatives: in matrix form, $D = [I_n \; -I_n] \in \mathbb{R}^{n \times 2n}$. Formally, the mesh at iteration k is the following enumerable subset of $\mathbb{R}^n$:

$$M_k = \left\{ x + \Delta_k^m Dz : x \in V_k, \; z \in \mathbb{N}^{n_D} \right\} \subset \mathbb{R}^n. \tag{4.2}$$

The set V_k is also called the *cache* as it contains the history of all evaluated trial points. For functions that are expensive to evaluate, the cache allows a reduction in computational time as the simulation at a previously evaluated trial point is not performed.

Figure 4.1 illustrates the mesh M_k on a problem with only two variables where the set of directions used to construct the mesh are the positive and negative coordinate directions: in matrix form,

$$D = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

The mesh points are represented by the intersections of the horizontal and vertical lines. The mesh M_k is conceptual as it is never generated, but the method must make sure that the trial points belong to M_k. The remaining elements of the figure are described below.

Each MADS iteration goes as follows. Given an incumbent solution $x_k \in X$, the SEARCH step produces a list of tentative trial mesh points. Any mechanism can be used to created the list, as long as it contains a finite number of points located on the mesh. The list may even be empty. Then, the simulation is launched at the trial points until all trial points are tested, or until one trial point is found to be better than

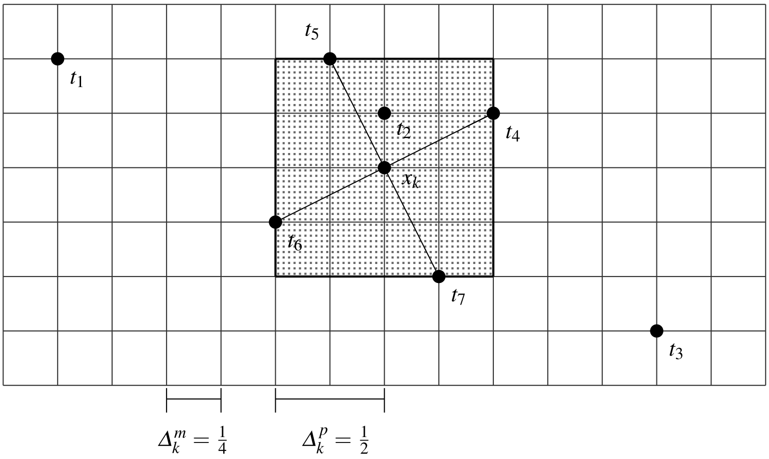


Fig. 4.1 Example MADS trial points in $\mathbb{R}^2$ consistent with the ones defined in [3]. The intersections of the *thin lines* represent the mesh of size Δ_k^m, and *thick lines* the points at distance Δ_k^p from x_k in the infinity norm. Examples of SEARCH $\{t_1, t_2, t_3\}$ and POLL trial points $\{t_4, t_5, t_6, t_7\}$ are illustrated

the incumbent x_k. In the latter case, the POLL step can be skipped, and the algorithm may continue directly to the updates.

Following an unsuccessful SEARCH step, the POLL step generates a list of mesh points near the incumbent x_k. The term *near* is tied to the so-called *poll size parameter* $\Delta_k^p > 0$. Again, the POLL step may be interrupted as soon as an improvement over the incumbent is found. During an iteration, the simulations can be launched sequentially or in parallel. Synchronous and asynchronous strategies are described in Sect. 4.3.3 when multiple processors are available.

Parameters are updated at the end of each iteration. There are two possibilities. If either the SEARCH or the POLL step generated a mesh point $t \in M_k$ which is better than x_k, then the next incumbent x_{k+1} is set to t and both the mesh size and poll size parameters are increased or kept to the same value. For example, $\Delta_{k+1}^p \leftarrow 2\Delta_k^p$ and $\Delta_{k+1}^m \leftarrow \min\{1, \sqrt{\Delta_k^p}\}$. Otherwise, x_{k+1} is set to x_k and the poll size parameter is decreased and the mesh size parameter decreased or kept the same. For example, $\Delta_{k+1}^p \leftarrow \frac{1}{2}\Delta_k^p$ and $\Delta_{k+1}^m \leftarrow \min\{1, \sqrt{\Delta_k^p}\}$.

At any iteration of a MADS algorithm, the poll size parameter Δ_k^p must be greater than or equal to the mesh size parameter Δ_k^m. In Fig. 4.1, $\Delta_k^p = \frac{1}{2}$ and $\Delta_k^m = \frac{1}{4}$, and the search points are $\{t_1, t_2, t_3\}$. In ORTHOMADS, the POLL points are obtained by generating a pseudo-random orthogonal basis H_k and by completing it to a maximal positive basis $D_k = [H_k \; -H_k]$. The poll points are then obtained from the incumbent x_k in the directions of the columns of D_k while remaining in the frame (the shaded region in the figure) defined by the poll size parameter Δ_k^p.

The iteration concludes by increasing the counter k by one. A new iteration is then initiated. Figure 4.2 summarizes the main steps of a MADS algorithm.

```
Initialization
    starting point: $x_0 \in X$
    initial mesh and poll size parameters: $\Delta_0^m, \Delta_0^p > 0$
    iteration counter: $k \leftarrow 0$
SEARCH and POLL steps
    Let $M_k$ be the mesh of Eq. (4.2)
    SEARCH(optional)
        launch the simulation on a finite number of mesh points
    POLL(optional if SEARCH was successful)
        launch the simulation on a finite number of mesh points
        within a radius of $\Delta_k^p$ of $x_k$
Updates
    determine success or failure of iteration $k$
    update the cache $V_{k+1}$
    update the incumbent $x_{k+1}$
    update the mesh and poll size parameters $\Delta_{k+1}^m$ and $\Delta_{k+1}^p$
    increase the iteration counter $k \leftarrow k+1$
    goto SEARCH and POLL steps if no stopping condition is met
```

Fig. 4.2 A general MADS algorithm. See Fig. 4.1 for some examples of search and poll points

4.2.2 Handling of Constraints

MADS possesses different techniques to handle constraints. The constraints $x \in X$ are handled by the extreme barrier discussed in the introduction of Sect. 4.2. The constraints $c_j(x) \leq 0$ are relaxable and quantifiable and this supplementary structure allows a potentially more efficient treatment. The *progressive barrier* [10] exploits this structure and allows the algorithm to explore the solution space around infeasible trial points. This treatment of constraints uses the constraint violation function originally devised for filter methods [25] for nonlinear programming:

$$h(x) = \begin{cases} \sum_{j \in J} (\max(c_j(x), 0))^2 & \text{if } x \in X, \\ \infty & \text{otherwise.} \end{cases}$$

The constraint violation function h is nonnegative, and $h(x) = 0$ if and only if $x \in \Omega$. It returns some kind of weighted measure of infeasibility.

The extreme barrier is essentially a mechanism that allows infeasible trial points whose constraint violation function value is below a threshold $h_k^{\max} > 0$. But as the algorithm is deployed and the iteration number increases, the threshold is progressively reduced. This is accomplished with the MADS algorithm by having two incumbent solutions around which polling is conducted. One poll center is the feasible incumbent solution x_k^F, i.e., the feasible solution found so far with the least objective function value. The second poll center is the infeasible incumbent solution x_k^I, i.e., the infeasible solution found so far with a constraint violation value under the threshold $h_k^{\max}$ having the least objective function value. Under this strategy, the in-

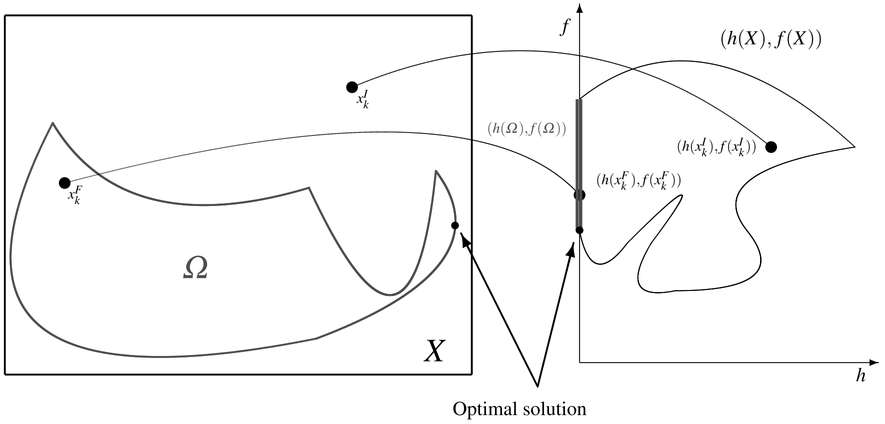


Fig. 4.3 The feasible region Ω and the domain X of an optimization problem, and their image under the mappings h and f

feasible trial points approach the feasible region by prioritizing the infeasible points with a low objective function value. This strategy differs from the ones in [2, 8, 24] where priority was given to feasibility at the expense of the objective function value.

Figure 4.3 represents an optimization problem as a tradeoff between the objective function f and the constraint violation function h. The left part of the figure depicts the domain X of a two-variable problem, as well as its feasible region Ω. The right part of the figure shows the image of both X and Ω under the mappings h and f. The mapping of X is delimited by the nonlinear curve, and the mapping of Ω is represented by the greyed region located on the f-axis. The optimal solution of the optimization problem corresponds to the feasible point ($h = 0$) with the least value of f, as indicated by the arrows. The figure also shows the feasible and infeasible incumbents, as well as their image.

With the progressive barrier, the iterations are categorized into more than two types. *Dominating* iterations are those that either generate a feasible trial point with a lower objective function value than that the feasible incumbent, or those that generate an infeasible trial point with better objective and constraint violation function values than the infeasible incumbent. *Improving* iterations are those that are not dominating, but generate an infeasible trial point with a better constraint violation value. Otherwise, the iteration is said to be *unsuccessful*.

At the end of an unsuccessful iteration the incumbents remain unchanged, and the poll size parameter is reduced as this suggests that we are near a locally optimal solution. At the end of a dominating solution, a new incumbent solution is identified, and the poll size parameter is increased to allow far-reaching explorations in the space of variables. Finally, after an improving iteration, the poll size parameter is kept unchanged, but the constraint violation threshold is reduce in such a way that the next iteration will not have the same infeasible incumbent as the one from the previous iteration.

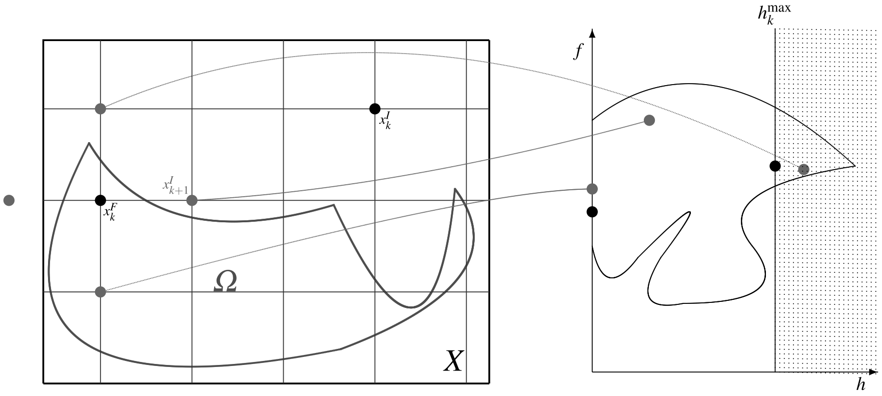


Fig. 4.4 Polling around the feasible incumbent x^F generates a new infeasible incumbent x_{k+1}^I

Figure 4.4 represents the poll step around the feasible incumbent x_k^F. The four poll points are represented by the light circles in the left part of the figure. The leftmost one is rejected by the extreme barrier, as it does not belong to the domain X. Only one of the poll points is feasible, but as illustrated in the right part of the figure, it is dominated by the feasible incumbent x_k^F. The two other poll points are infeasible. One of them is rejected by the progressive barrier, as its constraint violation function value exceeds the threshold $h_k^{\max}$. Any trial point that gets mapped in the shaded region gets rejected. The remaining poll point has a lower constraint violation value than the infeasible incumbent, but a worse objective function value. Therefore, the iteration is neither a dominating one nor an unsuccessful one, it is an improving iteration. At the end of the iteration, the mechanism of the progressive barrier updates the infeasible incumbent x_{k+1}^I to be the poll point located to the right of x_k^F, and the threshold $h_{k+1}^{\max}$ would be reduced to the constraint violation function value evaluated at the new infeasible incumbent solution.

Another strategy to handle relaxable quantifiable constraints is the *progressive to extreme barrier* [12]. As its name suggests, this strategy consists in first handling the constraint by the progressive barrier. But as soon as a trial point which satisfies the constraint is generated, then the treatment of the constraint is switched to the extreme barrier. This last strategy allows infeasible initial trial points, but forces the satisfaction of individual constraints as soon as they are satisfied for the first time.

4.2.3 Surrogates and Models

Surrogates are functions that can be considered as substitutes for the *true* functions defining the optimization problem, f and c_j, $j \in J$. A surrogate function shares some similarities with the original one, but has the advantage of being significantly less expensive to evaluate. Surrogate functions can be classified into *static surrogates* and *dynamic surrogates*.

4.2.3.1 Static Surrogates

Static surrogates are approximations that are provided by the user with some knowledge of the problem. A surrogate consists of a model of the true function, and is fixed during the optimization process. For example, static surrogates may be obtained by simplified physics models, or by allowing more relaxed stopping criteria within the blackbox simulation, or by replacing complicated subproblems by simpler ones. Straightforward uses of static surrogates within an optimization method are described in [15]. A first possibility is to order a list of tentative trial points by their surrogate values, and then to launch the expensive simulation defining the truth on the most promising trial points first. The process terminates as soon as a better point is found. This is called the opportunist strategy and can save a considerable amount of time.

A second strategy using static surrogates consists in defining a SEARCH step that optimizes the surrogates in order to determine one or two candidates for the true evaluations. In some situations, static surrogates may be parametrized with controllable precision. The use of such surrogates within a GPS framework is described in [39].

4.2.3.2 Dynamic Surrogates

In contrast to static surrogates, dynamic surrogates are not provided by the user. They correspond to models dynamically built within the optimization method, based on past evaluations from the cache. Any interpolation method can be used for this task, as, for example, quadratic models, neural networks, radial basis functions, or statistical methods. The Surrogate Management Framework [18] proposes ways to exploit such surrogates within direct search methods, and GPS in particular, for the unconstrained case. Successful applications of this framework include unsteady fluid mechanics problems [34, 35], helicopter rotor blade design [17] and multi-objective liquid-rocket injector design [40].

Recent developments propose the use of these surrogates for the constrained case within MADS. The first of these developments considers quadratic models and is summarized in the next section. The second approach is ongoing research and currently considers statistical methods [27], namely *tree Gaussian processes* [26]. Quadratic model and statistical surrogates share some similarities. They can be used to sort a list of trial points before launching the expensive true blackbox simulation, as proposed above for the static surrogates. Dynamic surrogates may also define a SEARCH step named the *model search*, enabled as soon as a sufficient number of true evaluations is available (typically $n + 1$). These points are denoted the *data points* and are used to build one model for f and m models for the quantifiable constraints c_j, $j \in J$. The model of the objective function is then optimized subject to the models of the constraints. This provides one or two mesh candidates, one feasible and possibly one infeasible, at which the true functions are evaluated. These are called *oracle points*. In addition to the nature of the surrogates, some subtleties remain: The quadratic models are kept local, while statistical surrogates consider the whole space and attempt to escape local solutions. They can also provide additional candidates based on statistics such as the *Expected Improvement* (EI) [29].

4.2.3.3 Quadratic Models in MADS

This section discusses the quadratic models described in [22] and currently used in NOMAD. The framework is inspired from the work of Conn et al. summarized in the DFO book [23] where more focus is put on model-based methods.

Quadratic models are employed at two different levels: First, the *model search* exploits the flexibility of the SEARCH by allowing the generation of trial points anywhere on the mesh. Candidates of the model search are the result of an optimization process considering the model of the objective constrained to the models of the constraints.

The other way of using models is to sort a list of candidates prior to their evaluations (*model ordering*), so that the most promising—from the model point of view—points will be evaluated first. The impact of this strategy is important because of the opportunistic strategy.

To construct a model, a set $Y = \{y^0, \ldots, y^p\}$ of $p+1$ data points is collected from the cache. The objective function $f(y)$ and the constraint functions $c_j(y)$, $j \in J$ are known and finite at each data point $y \in Y$. Since quadratic models are more suited for local interpolation, data points are collected in the neighborhood of the current iterate: $y^i \in B_\infty(x_k; \rho\Delta_k^p)$ with $B_\infty(x; r) = \{y \in \mathbb{R}^n : \|y - x\|_\infty \le r\}$, where the poll size parameter Δ_k^p bounds the distance between x_k and the POLL trial points, and ρ is a parameter called the *radius factor*, typically set to two.

Then, $m+1$ models are constructed: one for the objective f and one for each constraint $c_j \le 0$, $j \in J$. These models are denoted m_f and m_{c_j}, $j \in J$, and are such that

$$m_f(x) \simeq f(x) \quad \text{and} \quad m_{c_j}(x) \simeq c_j(x), \quad j \in J, \quad \text{for all } x \in B_\infty\big(x_k; \rho\Delta_k^p\big).$$

For one function (f or one of the constraints c_j), the model m_f is defined by $q+1$ parameters, $\alpha \in \mathbb{R}^{q+1}$, evaluated at x with $m_f(x) = \alpha^\top \phi(x)$ with ϕ the natural basis of the space of polynomials of degree less than or equal to two, which has $q+1 = (n+1)(n+2)/2$ elements:

$$\begin{aligned}\phi(x) &= \big(\phi_0(x), \ldots, \phi_q(x)\big)^\top \\ &= \left(1, x_1, \ldots, x_n, \frac{x_1^2}{2}, \ldots, \frac{x_n^2}{2}, x_1x_2, x_1x_3, \ldots, x_{n-1}x_n\right)^\top.\end{aligned}$$

The parameter α is selected in such a way that $\sum_{y\in Y}(f(y) - m_f(y))^2$ is as small as possible, by solving the system

$$M(\phi, Y)\alpha = f(Y) \qquad (4.3)$$

with $f(Y) = (f(y^0), f(y^1), \dots, f(y^p))^\top$ and

$$M(\phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \dots & \phi_q(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \dots & \phi_q(y^1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \dots & \phi_q(y^p) \end{bmatrix} \in \mathbb{R}^{(p+1)\times(q+1)}.$$

System (4.3) may possess one, several, or no solutions. If $p \geq q$, i.e., there are more interpolation points than necessary, it is overdetermined and regression is used in order to find a solution in the least squares sense. When $p < q$, i.e., there are not enough interpolation points, the system is underdetermined and there is an infinite number of solutions. Minimum Frobenius norm (MFN) interpolation is used in that case, which consists in choosing a solution that minimizes the Frobenius norm of the curvature subject to the interpolation conditions. This is captured in the quadratic terms of α. Thus writing $\alpha = \left[\begin{smallmatrix} \alpha_L \\ \alpha_Q \end{smallmatrix}\right]$ with $\alpha_L \in \mathbb{R}^{n+1}$, $\alpha_Q \in \mathbb{R}^{n_Q}$, and $n_Q = n(n+1)/2$, our model at x is given by $m_f(x) = \alpha_L^\top \phi_L(x) + \alpha_Q^\top \phi_Q(x)$ with $\phi_L = (1, x_1, \dots, x_n)^\top$ and $\phi_Q = (\frac{x_1^2}{2}, \dots, \frac{x_n^2}{2}, x_1x_2, x_1x_3, \dots, x_{n-1}x_n)^\top$. The corresponding MFN α vector is then found by solving

$$\min_{\alpha_Q \in \mathbb{R}^{n_Q}} \frac{1}{2}\|\alpha_Q\|^2 \quad \text{subject to } M(\phi_L, Y)\alpha_L + M(\phi_Q, Y)\alpha_Q = f(Y).$$

Once the $m+1$ models are available, the model search and the model ordering strategies differ slightly. The model ordering consists in evaluating the models at the candidates, and then sorting the candidates accordingly. The model search is more elaborated because the following optimization problem has to be solved:

$$\min_{x \in B_\infty(x_k; \rho\Delta_k^p).} m_f(x) \quad \text{subject to } m_{c_j}(x) \leq 0,\ j \in J. \tag{4.4}$$

After Problem (4.4) is solved (in practice, heuristically), its feasible and infeasible incumbent solutions define new candidates at which evaluate the true functions f and c_j, $j \in J$. In order to satisfy the MADS convergence analysis described in Sect. 4.2.4, these candidates are projected on the mesh before they are evaluated.

4.2.4 Convergence Analysis

Even if MADS is designed to be applied to the general optimization problem (4.1) without exploiting any of its structure, MADS is supported by a rigorous hierarchical convergence analysis. The analysis reveals that depending on the properties of the objective function f and the domain Ω, MADS will produce a limit point $\hat{x}$ at which some necessary optimality conditions are satisfied. Of course, we do not expect our target problems to satisfy any smoothness properties, but the convergence analysis can be seen as a validation of the behavior of the algorithm on smoother problems.

The entire convergence analysis relies on the following assumptions. Suppose that MADS was launched on a test problem, without any stopping criteria, and suppose that the union of all trial points generated by the algorithm belongs to a bounded subset of $\mathbb{R}^n$. The assumption that MADS is launched indefinitely is not realistic, as in practice it necessarily terminates after a finite amount of time. But for the analysis, we are interested in seeing where the iterates would lead, it the algorithm were not stopped. The second assumption on the bounded subset can be satisfied in multiple ways. For example, it is true when the variables are bounded, or when level sets of f are bounded. In practice, it is not frequent that real problems have unbounded solutions.

The convergence analysis then focuses on limits of incumbent solutions. Torczon [43] showed for pattern searches that the hypotheses on bounded trial points implies that there are infinitely many unsuccessful iterations and that the limit inferior of the mesh size parameters Δ_k^m converges to zero. These results were adapted in [9] in the context of the MADS algorithm. Let U denote the indices of the unsuccessful iterations and let $\hat{x}$ be an accumulation point $\{x_k\}_{k \in U}$. Such an accumulation point exists because of the assumption that the iterates belong to a bounded set.

An unsuccessful iteration occurs when the POLL step was conducted around the incumbent x_k, and no better solution was found. The mesh size parameter is reduced only after an unsuccessful iteration. We say that the incumbent solution x_k is a *mesh local optimizer*. At the low end of the convergence analysis [7], we have the zeroth order result: $\hat{x}$ is the limit of mesh local optimizers on meshes that get infinitely fine. At the other end of the analysis, we have that if f is strictly differentiable near $\hat{x}$ then $\nabla f(\hat{x}) = 0$ in the unconstrained case, and in the constrained case, the result ensures that the directional derivatives $f'(\hat{x}, ; d)$ are nonnegative for every direction $d \in \mathbb{R}^n$ that points in the contingent cone to the feasible region Ω, provided that Ω is regular [9]. The contingent cone generalizes the notion of tangent cone.

There are several intermediate results in the convergence analysis that involve different assumptions on f such as lower-semi continuity and regularity, and on the constraints such as properties of the hypertangent, Clarke tangent, and Bouligand cones. The fundamental theoretical result in the analysis was shown in [7, 9] and relies on Clarke's [21] generalization of the directional derivative f° for nonsmooth functions. The result states that $f^\circ(\hat{x}; d) \geq 0$ for every direction d in the hypertangent cone to Ω at $\hat{x}$. A generalization of this result for discontinuous functions was recently shown in [45]. In the case where the progressive barrier [10] fails to generate feasible solutions, the analysis ensures that the constraint violation function satisfies $h^\circ(\hat{x}; d) \geq 0$ for every direction d in the hypertangent cone to X at $\hat{x}$.

4.3 NOMAD: A C++ Implementation of the MADS Algorithm

This section describes the NOMAD software [32] which implements the MADS algorithm. We list several of its features, but do not expect to cover all of them in this chapter. NOMAD is a C++ code freely distributed under the LGPL license. The

```
# problem parameters:
DIMENSION 5
BB_EXE bb.exe
BB_OUTPUT_TYPE OBJ CSTR CSTR

# algorithmic parameters:
X0 * 0.0
MAX_BB_EVAL 100
```

Fig. 4.5 Example of a basic parameters file. The blackbox executable bb.exe takes five variables as input, and returns three outputs: one objective function value and two constraints. The initial point is the origin and NOMAD terminates after 100 evaluations

package is found at http://www.gerad.ca/nomad. It includes a complete documentation, a doxygen [44] interactive manual, and many examples and tools. As for other derivative-free codes, we expect as a rule of thumb that NOMAD will be efficient for problems with up to 50 variables.

4.3.1 Batch and Library Modes

NOMAD can be used in two different modes having various advantages. The user must choose with care the appropriate mode depending on its problem.

First, the *batch* mode, which launches the NOMAD executable in the command line with the name of a parameter file given as an argument. This text file contains the parameters that are divided into two categories: problem and algorithmic. Problem parameters are required while all the algorithmic parameters have default values. A simple parameter file is shown in Fig. 4.5, and the most important parameters are described in Sect. 4.3.2. The batch mode is simpler for beginners and non-programmers. The user must write a parameters file and design a wrapper for its application so that it is compatible with the NOMAD blackbox format. This format requires that the blackbox is callable from the command line with an input file containing the values of the variables, given as an argument.[1] The resulting outputs must be displayed to the standard output with a sufficient precision. The blackbox is disjoint from the NOMAD code, and consequently the application may be coded in any programming language, as long as a command-line version is available. A detailed description for one implementation of this command-line interface is covered in the Appendix. Finally, the batch mode is by definition resilient to the blackbox crashes that may occur when a hidden constraint is violated: NOMAD will simply reject the trial point that made the blackbox crash.

The second way to use the NOMAD algorithm is through the *library* mode. The user must write a C++ code which will be linked to the NOMAD static library included in the package. This way, interactions with NOMAD are directly performed

[1] Here, we note that for the purposes of metamaterial design and this book, the blackbox terminology refers to any electromagnetics solver used to simulate a given metamaterial design.

via C++ function calls and objects manipulations. The optimization problem is described as a class and is written in C++ or in a compatible language such as C, FORTRAN, R, etc. The problem and algorithmic parameters are given as objects, and no parameters file is necessary. The library mode must be considered only by users with basic C++ knowledge. The problem must also be C++-compatible in order to be expressed as a class, and hidden constraints need to be explicitly treated. If these points are addressed, the advantages of the library modes are numerous. First, when the blackbox is not costly, the execution will be much faster than the batch mode since no temporary files and no system calls are used. Second, numerical precision is not an issue because the communications between the algorithm and the problem occur at memory level. Third, more flexibility is possible with the use of callback functions that are user-defined and automatically called by NOMAD at key events, such as after a new success, or at the end of a MADS iteration. Last, the library mode is convenient when NOMAD is repeatedly called as a subroutine. Note finally that a hybrid use of the batch and library modes is possible. For example, one can define its problem as a C++ class and use a parameters file.

4.3.2 Important Algorithmic Parameters

The objective of this section is not to describe all the parameters, but to discuss the ones that may have a significative influence on the executions efficiency. The names of the parameters are not reported here but can be easily found in the user guide or by using the NOMAD command-line help (option -h).

Starting point(s): As for every optimization method, the starting point choice is crucial. The user has to provide his/her best guess so far for the method to be as efficient as possible. Within NOMAD, it is possible to define multiple starting points. This might be useful, for example, if in addition to a feasible initial solution, an infeasible one corresponding to a different and promising design is known.

Initial mesh size and scaling: In our opinion, the second most important algorithmic choice concerns the initial mesh size. Some defaults related to the scale of the starting point are used, but the user is encouraged to determine a good problem-related value. Some automatic scaling is performed, but there again, users should sometimes consider changing the scale of their variables and study the impact.

Poll directions: The default type of poll directions is ORTHOMADS [3] but sometimes other direction types, such as LTMADS [9], may perform well.

Surrogates: Static and dynamic surrogates can be used with NOMAD. Static surrogates are indicated by the user with the same format as the true blackbox. Concerning dynamic surrogates, the current NOMAD version 3.5 includes only quadratic models, but statistical surrogates will be available in a future release. Dynamic surrogates may be employed at two different levels: as a SEARCH, and as a way to sort a list of candidates before they are evaluated. The current NOMAD default is to use quadratic models at both places.

Projection to bounds: When generating candidates outside of the hyper-rectangle defined by the bounds on the variables, NOMAD projects these points to the boundary, by default. For some problems, this strategy might not be the appropriate one.

Seeds: If the context allows multiple executions, changing the random seed for LT-MADS, or the *Halton seed* for ORTHOMADS [3], will lead to different executions.

Termination criteria: We finish this expose by indicating that many termination criteria are available, in addition to the obvious choice of a budget of evaluations.

4.3.3 Extensions of MADS

This section describes some algorithmic extensions that are not covered in Sect. 4.2 on the basic description of MADS. These features may be useful in practice, to approach an optimization problem through different angles.

Parallel and variable decomposition methods: Three different parallel versions are available, using MPI. These methods are called P-MADS, COOP-MADS, and PSD-MADS. P-MADSsimply performs the evaluations in parallel, and two variants are available. First, the synchronous version waits for all ongoing parallel evaluations before iterating. This is opposed to the asynchronous variant inspired from [28] which iterates as soon as a new success is made, even if some evaluations are still not finished. In case one of these evaluations results in an improvement, the current iterate and the current mesh size are adjusted accordingly during a special update step. The two other parallel methods, COOP-MADS and PSD-MADS, are provided as tools in the package. The first executes several MADS instances in parallel with different seeds. Some cooperative actions are performed in order to guide the search. PSD-MADS performs the same collaborative process, but in addition, subgroups of variables are considered for each process. This technique is described in [11] and aims at solving larger problems (50 to 500 variables).

Groups of variables: The user with some knowledge of the problem can create groups of variables. Directions are then relative to these groups and variables from separate groups are not going to vary at the same time. This proved useful for localization problems and in particular the one presented in [4].

Different types of variables: It is possible to define integer and binary variables, which are treated by special meshes with a minimal size of one. Categorical variables may also be used. They are handled with the *extended poll* defined in [1, 31]. For such problems, the user must define a neighborhood structure, which may contain a different number of variables. NOMAD defines the concept of *signature* that allows such heterogeneous points. Other types of variables include *fixed* variables and *periodic* variables (angles, for example). The strategy used for the periodic variables is described in [14].

Bi-objective optimization: In some situations, the user is interested in considering the tradeoffs between two conflicting objective functions. The method from [16] executes a series of single-objective optimizations on reformulated versions of the original problem. These reformulations are not linear combinations of the objective, and ensure that non-convex Pareto fronts can be identified.

Variable Neighborhood Search (VNS): For problems with many local optima, it is possible to enable the generic VNS SEARCH. This strategy has been described in [6] and uses a *variable neighborhood search* metaheuristic in order to escape local optima.

Sensitivity analysis: The last feature described here is a tool that uses bi-objective optimization in order to conduct sensitivity analyses on the constraints, including bounds as well as non-relaxable constraints. This is described in [13] with plots illustrating the impact on the objective of changing the right-hand side of a given constraint.

4.4 Metamaterial Design Using NOMAD

This section describes the implementation of the NOMAD optimization routine [32] in combination with full-field electromagnetic simulations to tailor the broadband spectral response of gold and silver split ring resonator metamaterials. By allowing NOMAD to “drive” finite-difference time-domain simulations, the spectral position of resonant reflection peaks and near-field interactions within the metamaterial were tuned over a wide range of the near-infrared spectrum. While this section discusses the design problems studied and the optimized results, a detailed discussion of the implementation used to communicate between the different software packages is provided in the Appendix.

4.4.1 Split Ring Resonator Optimization

The first example of NOMAD driving the design of metamaterial device geometries involves the structure shown in Fig. 4.6. Here, the broad-band reflection spectrum from a single split-ring resonator/cross-bar structure (SRR) surrounded by air was studied as a function of the device dimensions. All of the electromagnetic simulations studied in this chapter were performed using the finite-difference time-domain method and the Lumerical software package.

In this section, the SRRs were illuminated with a broad-band plane wave source from 1–4 µm, and the structure was parameterized based on the height, width, ring thickness (t_1), bar thickness (t_2), and gap width. For all simulations, the thickness of the metal was 100 nm, the E-field was perpendicular to the arms of the SRR, and the width of the bar was kept the same as the width of the SRR. Also, linear constraints were imposed to ensure that the parameters made physical sense (e.g., $2t_1 \geq$ width), and a gap was always present between the two arms of the SRRs.

Because of the general, double peaked reflection spectrum that comes from the SRR/bar structure, a double Lorentzian was chosen as a plausible initial target for

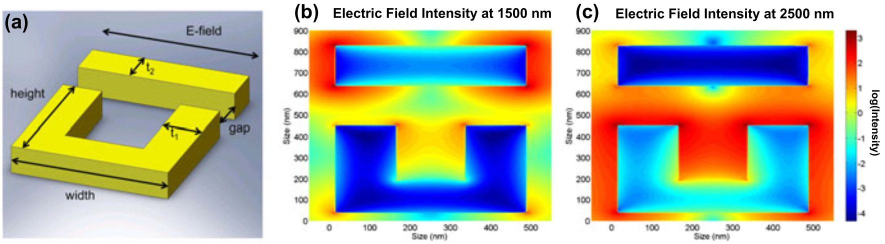


Fig. 4.6 Panel (**a**) shows a schematic of the SRR structure that was optimized in Sect. 4.4.1. The electric field intensity at 1500 nm, which corresponds to the resonance of the bar, is plotted in (**b**) and the electric field intensity at 2500 nm, which corresponds to the resonance of the SRR, is plotted in (**c**). The specific geometry in (**b**) and (**c**) corresponds to the optimized spectrum in Fig. 4.7(c) and Table 4.1

the optimization. Peaks were set at 1500 nm with a reflection intensity of 50 % and at 2500 nm with a reflection intensity of 35 %. Although the particular target wavelengths have been chosen arbitrarily, this double Lorentzian spectrum was chosen to correspond to the resonant modes of the bar and SRR shown in Fig. 4.6(b)–(c). This could be considered typical of an application in which the designer wishes to design a nanoantenna which simultaneously matches the center frequencies and line widths of both the absorption and the emission processes in a quasi-three level optical system, such as coupling to a photoluminescent quantum dot. The target spectrum is shown as the dashed green curve in Fig. 4.7(a)–(d). Simulations were done with multiple starting points, listed in Table 4.1, for both gold and silver SRRs. The upper and lower bounds for each of the five fit parameters are also listed in Table 4.1, and the bounds for t_1 are starred to indicated the imposed linear constraints. For these simulations, the objective function used to drive the optimization is listed in Eq. (4.5), with the first three terms scaled by $\frac{1}{150}$ to keep the magnitudes of all six terms comparable. Through experience, we have seen that objective functions that focus on a few key points in the broad band spectrum, almost always gives better results than metrics such as the mean squared error at every point in the reflection spectrum. If every point in the broad band spectrum is weighted equally, the key objectives in the cost function are essentially "swamped out" by the remaining hundreds or thousands of other, less important data points. As a result, the general type of objective function listed in Eq. (4.5) is used throughout the rest of this chapter:

$$\begin{aligned}\text{O.F.} = {} & \frac{|\lambda_{P1} - 1500|}{150} + \frac{|\lambda_{P2} - 2500|}{150} + \frac{|\lambda_V - 2025|}{150} + |I_{P1} - 0.5| \\ & + |I_{P2} - 0.35| + |I_V - 0.07|.\end{aligned} \tag{4.5}$$

Here "λ_{P1}" is the shorter wavelength peak position, "λ_{P2}" is the longer wavelength peak position, "λ_V" is the position of the reflection minima between λ_{P1} and λ_{P2}, "I_{P1}" is the intensity of the shorter wavelength peak, "I_{P2}" is the intensity of the longer wavelength peak, and "I_V" is the intensity of the reflection minima between λ_{P1} and λ_{P2}.

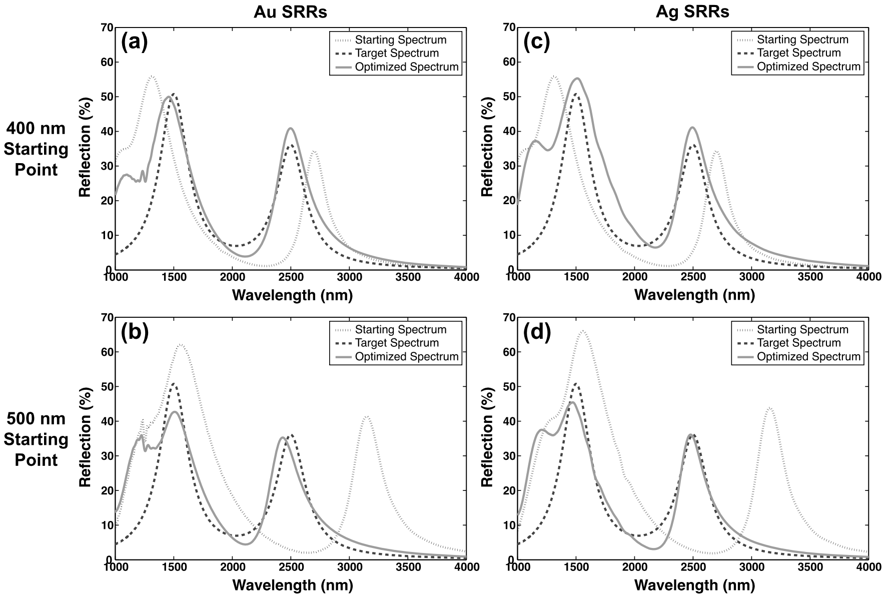


Fig. 4.7 SRR spectrum optimization using NOMAD with FDTD. Simulations using gold are shown on the *left* and simulations using silver are shown on the *right*. The *top row* corresponds to simulations using the first starting point in Table 4.1, and the *bottom row* corresponds to simulations using the second starting point

Table 4.1 Starting and optimized dimensions (in nm) for the SRR structures tested in Sect. 4.4.1. The variables correspond to those listed in Fig. 4.6. Boundary conditions for t_1 were linearly constrained so that [width − $2t_1$] > 0 for all optimizations

	SRR initial and optimized values				
	Initial value	Gold optimized value	Silver optimized value	Minimum	Maximum
Width	400	450	474	200	600
Height	400	394	412	200	600
t_1	100	125	152	50*	200*
t_2	100	114	193	50	200
Gap	100	192	184	50	200
Width	500	455	438	200	600
Height	500	476	480	200	600
t_1	125	190	182	50*	200*
t_2	125	94	105	50	200
Gap	125	190	173	50	200

Figure 4.7 shows the optimization results for a matrix of initial conditions. The left panels (a & b) are simulations of gold SRRs, while the right panels (c & d) are silver SRRs. Table 4.1 details two sets of initial values for each of the resonators' geometrical parameters. The top row of Fig. 4.7, panels (a & c), correspond to the first set of initial conditions. The bottom row, panels (b & d), correspond to the second starting point. For all four panels, the dashed curve represents the double Lorentzian target spectrum, the dotted curve represents the reflection spectrum from the starting point, and the solid curve represents the reflection spectrum from the optimized result. A close look at the results in Table 4.1 shows that for a each metal, there is some variability in the optimized results based on the starting point. This is not necessarily surprising considering the degeneracy in potential solutions for this five-parameter optimization, and a relatively relaxed convergence tolerance specified during the optimizations. For all four cases, the results are especially encouraging based on the fact that the ideal double Lorentzian to which the curves were fit was arbitrarily chosen, and a perfect fit cannot necessarily be obtained with the given geometrical constraints and materials.

4.4.2 Split Ring Resonator Filters

The second example involves an array of individual SRRs with the same unit cell design shown in Fig. 4.6(a). In this case, the array was on top of a sapphire substrate and was used as a "notch filter". A seven-parameter NOMAD optimization was performed on this array which included the five parameters from Fig. 4.6(a), as well as the spacing between scattering elements along the x-axis, parallel to the E-field, and y-axis, perpendicular to the E-field. Here the objective was to minimize the reflectivity and pass band at a pre-specified wavelength, while maximizing the reflectivity on either side of the pass band. The objective function used to drive the optimization is listed in Eq. (4.6):

$$\text{O.F.} = 100 * \left[(1 - I_{P1}) + (1 - I_{P2}) + (I_V)\right] + |\lambda_{P1} - \lambda_{P2}| + |\lambda_V - \lambda_T| \quad (4.6)$$

where "λ_T" is the pass band target wavelength, and the remaining terms are identical to those used in Eq. (4.5). Target wavelengths of $\lambda = 1310, 1550,$ and 1800 nm were chosen and the optimization was run with the same starting conditions every time. The resulting spectra from the three optimizations are shown in Fig. 4.8. All three optimized spectra show an ~45 % change in reflectivity at the pass band and corresponding linewidths of ~90 meV. The starting and optimized dimensions for each solution are given in Table 4.2. Figure 4.8 clearly shows successfully optimized designs for all three target wavelengths and the wide range of tunability this technique can offer in terms of metamaterials design.

As a final check of the solutions' robustness in the previous sections, a systematic variation of each parameter of the 400 nm Au SRR in Sect. 4.4.1 and the 1500 nm SRR filter in Sect. 4.4.2 was performed near the optimized value. Variations from 1–3 % produced a corresponding change in the objective function of <0.3 % for the

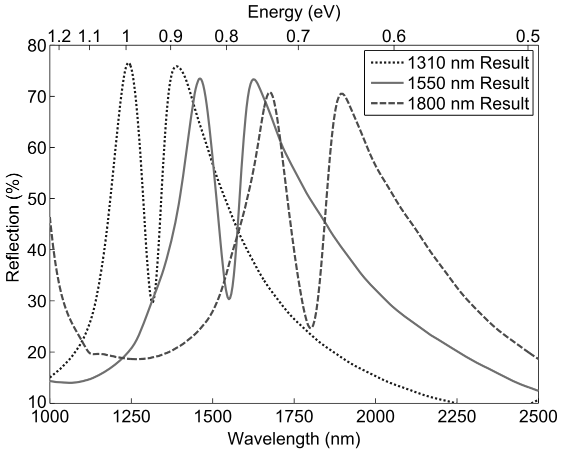


Fig. 4.8 Optimized reflection spectra for arrays of SRRs on sapphire substrates. The array was designed to act as a notch filter at three target wavelengths of $\lambda = 1310, 1550,$ and 1800 nm, respectively. The dimensions that produced each spectrum are given in Table 4.2

Table 4.2 Starting and optimized dimensions (in nm) for the SRR structures in Sect. 4.4.2 designed to act as a notch filter at $\lambda = 1310, 1550,$ and 1800 nm

	SRR filter initial and optimized values					
	Initial value	1310 nm	1550 nm	1800 nm	Minimum	Maximum
Width	500	348	424	508	200	600
Height	500	356	420	492	200	600
t_1	200	100	164	208	50*	200*
t_2	100	124	140	104	50	200
Gap	100	51	60	96	50	200
x_{spacing}	200	128	120	192	50	1000
y_{spacing}	200	330	392	436	50	1000

400 nm Au SRR and <4 % for the 1500 nm notch filter. We conclude from this that the method is robust to local perturbations (which is an attribute of the relationship between the SRR geometry and the objective function, not the optimization method used).

4.4.3 Coupling Quantum Dots to Split-Ring Resonators

In the third and final example of this chapter, we examine the design requirements involved in coupling the previously analyzed resonances of SRRs to the electronic transition states of quantum dots. The nonlinear nature of metals at optical frequencies makes them strong candidates for nonlinear mixing experiments. As an example, at these frequencies gold exhibits a strong, third-order nonlinear susceptibility where ($\chi^{(3)} \sim 1\ \text{nm}^2\ V^{-2}$) [41]. Further, one of the biggest strengths of

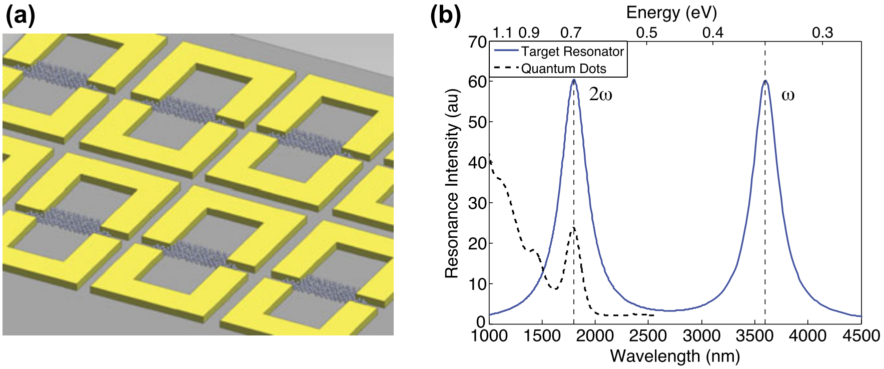


Fig. 4.9 The SRR design that was optimized to couple incident light at $\lambda = 3600$ nm and $\lambda = 1800$ nm to quantum dots lithographically patterned within the high-field regions of the resonant structure, (**a**). The plot in (**b**) shows an idealized resonant spectrum with the 2ω of the gold matched to a typical absorption spectrum of quantum dots. The optimized dimensions for this structure are listed in Table 4.4

these nanoscale resonant structures is there ability to manipulate and couple light in the near-field to nanostructures such as quantum dots. Combining lithography techniques with surface chemistry modification of the quantum dots presents an interesting opportunity to pattern these dots within the high-field regions of an array of resonators, Fig. 4.9(a).[2] This portion of the process has previously been reported in the literature [33, 38, 42, 46], and nonlinear mixing within gold nanostructures has already been demonstrated by Kim et al. [30]. By using the techniques described in this chapter, we can tailor the exact device geometry to have a resonance at the electronic transition energy of a specific batch of dots [19].

While there exist a wide range of energy transitions for this type of system, for this example we consider second harmonic generation studies with resonances at both $\lambda = 3600$ nm and $\lambda = 1800$ nm, which would enhance both the absorption of light $\lambda = 3600$ nm (ω_{Au}) and the coupling of light between the resonator and quantum dots at $\lambda = 1800$ nm ($\sim 2\omega_{\mathrm{Au}}$), Fig. 4.9(b). For the case of gold resonators patterned on a sapphire substrate, the resonance wavelengths were chosen to closely match those of the lead selenide quantum dots.[3]

The diagram in Fig. 4.9(b) shows typical broadband absorption spectra of lead selenide quantum dots (dot-dashed line) overlaid on top of a target resonance spectrum of a metamaterial tuned to the wavelengths of interest (solid line). While a wide range of resonator geometries would satisfy the performance requirements specified above, throughout this chapter, variations on the basic SRR geometry have been studied, and will again be used in this section. Using $\lambda = 3600$ nm and $\lambda = 1800$ nm as the target spectrum, initial simulations showed that the basic SRR

[2] Figures 4.9 and 4.10 were produced using the resources of MIT Lincoln Laboratory.

[3] Lead selenide quantum dot spectra courtesy of Dr. Seth Taylor.

Table 4.3 Objective functions used for the design optimization of an array of Split Ring Resonators with resonances at $\lambda = 3600$ nm and $\lambda = 1800$ nm, coupled to quantum dots. For the above functions, $I^*_{P1,2} = 500(1 - \text{Intensity}_{P1,2})$. The resulting broadband resonance spectra are shown in Fig. 4.10

O.F. #	SRR initial and optimized values
	Objective function
1	$\lvert\lambda_{P1} - 1800\rvert + \lvert\lambda_{P2} - 3600\rvert/2$
2	$\lvert\lambda_{P1} - 1800\rvert + \lvert\lambda_{P2} - 3600\rvert/2 + I^*_{P1}$
3	$\lvert\lambda_{P1} - 1800\rvert + \lvert\lambda_{P2} - 3600\rvert/2 + I^*_{P2}$
4	$\lvert\lambda_{P1} - 1800\rvert + \lvert\lambda_{P2} - 3600\rvert/2 + I^*_{P1} + I^*_{P2}$
5	$\lvert\lambda_{P1} - 1800\rvert \cdot \lvert\lambda_{P2} - 3600\rvert$
6	$\lvert\lambda_{P1} - 1800\rvert \cdot \lvert\lambda_{P2} - 3600\rvert \cdot I^*_{P2}$
7	$\lvert\lambda_{P1} - 1800\rvert \cdot \lvert\lambda_{P2} - 3600\rvert \cdot I^*_{P1}$
8	$\lvert\lambda_{P1} - 1800\rvert \cdot \lvert\lambda_{P2} - 3600\rvert \cdot I^*_{P1} \cdot I^*_{P2}$

structure shown in Sects. 4.4.1 and 4.4.2 was unable to span the wavelength region of interest; however, the SRR variant shown in Fig. 4.9(a) would. In this case, the array was on a sapphire substrate and a seven-parameter NOMAD optimization was performed. While the array optimization in this situation was similar to that in Sect. 4.4.2; here the gap between the two SRRs was a design parameter with the lower bound set by fabrication constraints (50 nm), and the height of each SRR was varied independently. This last parameter is not to be confused with resonator thickness, which was set at 50 nm for all structures studied in this example. Finally, the total width and arm width for both SRRs was kept equal for all designs. As in Sects. 4.4.1 and 4.4.2, a nonlinear constraint was imposed within NOMAD to maintain a minimum gap between the arms of each SRR to maintain the general shape of the broadband reflection spectra.

In an effort to study the convergence behavior of NOMAD using different objective functions, eight separate multi-objective cost functions were used. These variants are shown in Table 4.3. For all eight functions, the absolute difference between the peak wavelengths and the two target wavelengths were included. In addition, the resonant intensity of one or both of the two broadband peaks were added. Lastly, the first four objective functions take the sum of all the individual terms, while the last four take the product of all the individual terms. Here again, the variables in Table 4.3 match those in Eq. (4.5).

For each optimization, the initial conditions as well as upper and lower bounds were kept constant. For all the optimizations studied here, Table 4.4 shows the initial conditions (Column 2), the corresponding dimensions of the optimized designs for all eight objective functions (Columns 3–5), and the upper and lower boundary conditions (Columns 6 & 7).

As in Sects. 4.4.1 and 4.4.2, the idea was to maximize the reflection intensity at the two wavelengths of interest while at the same time, set the peak resonance wavelengths as close to the target wavelengths as possible. From Fig. 4.10 we can see strong resonances from the metamaterial array at both $\lambda = 3600$ nm and $\lambda = 1800$ nm; however, there are clearly small differences between the results. These results illustrate a point that was made in Chap. 2. Even when the key features

Table 4.4 Starting and optimized dimensions (in nm) for an array of SRRs shown in Fig. 4.9(a) with dimensions optimized to resonate at both λ = 3600 nm and λ = 1800 nm. The columns "O.F. 1–6", "O.F. 7", and "O.F. 8", correspond to the optimized dimensions obtained using Objective Functions #1–6, #7, and #8 from Table 4.3. The resulting broadband resonance spectrum is shown in Fig. 4.10

	SRR initial and optimized values					
	Initial value	O.F. 1–6	O.F. 7	O.F. 8	Minimum	Maximum
Width	500	467 ± 2	579	592	300	600
Length 1	500	494 ± 1	530	573	300	600
t_1	100	134 ± 1	198	196	50*	200*
Length 2	500	494 ± 1	494	549	300	600
Gap	100	114 ± 1	183	113	50	200
x_{spacing}	100	306 ± 8	374	296	100	500
y_{spacing}	100	299 ± 1	299	298	100	500

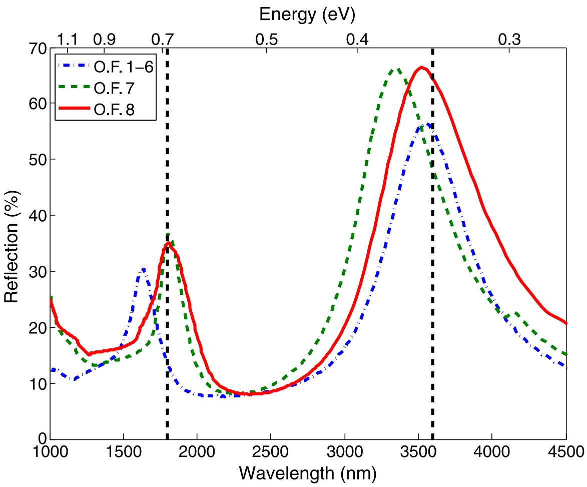


Fig. 4.10 The SRR design that was optimized to couple incident light at λ = 3600 nm and λ = 1800 nm to quantum dots lithographically patterned within the high-field regions of the resonant structure, (**a**). The plot in (**b**) shows an idealized resonant spectrum with the 2ω of the gold matched to the measured absorption spectrum of the quantum dots. The optimized dimensions for these spectra are listed in Table 4.4

of the broadband spectrum are well known, the way in which these terms are combined can prove to be one of the most challenging parts of the design optimization. Figure 4.10 clearly illustrates that while the three optimized spectrum that resulted from the eight objective functions all closely match the intended optimized spectrum, objective function #8 is clearly better than the other two.

While these resonances are close, but not a perfect match to the desired resonances, the full-width half-maxima of the resonances for O.F. #8 at λ = 1800 nm and λ = 3600 nm are ∼240 nm and ∼675 nm respectively, and that of quantum dots such is ∼150 nm. Hence, while the inherently broad nature of the resonances from metamaterial arrays is normally considered a drawback; in this situation, it can compensate for some amount of discrepancy between the optimized peaks and the ω and 2ω targets.

Finally, it should be noted that this example is only a first-order result. Here, the resonances of the quantum dots and the metamaterial are considered independently when setting the objective function targets. The addition of quantum dots into the near-field of the SRRs, will introduce perturbations in the local dielectric environment and as a result, shift the actual resonances from those predicted in the simulations. While non-negligible, this change is a second-order effect and combined with the substantial bandwidth of the individual resonances, should not significantly affect the results of the example.

Acknowledgements Work of the first author was supported by NSERC Discovery Grant 239436-05 and AFOSR FA9550-09-1-0160.

References

1. M.A. Abramson, C. Audet, J.W. Chrissis, J.G. Walston, Mesh adaptive direct search algorithms for mixed variable optimization. Optim. Lett. **3**(1), 35–47 (2009)
2. M.A. Abramson, C. Audet, J.E. Dennis Jr., Filter pattern search algorithms for mixed variable constrained optimization problems. Pac. J. Optim. **3**(3), 477–500 (2007)
3. M.A. Abramson, C. Audet, J.E. Dennis Jr., S. Le Digabel, OrthoMADS: a deterministic MADS instance with orthogonal directions. SIAM J. Optim. **20**(2), 948–966 (2009)
4. S. Alarie, C. Audet, V. Garnier, S. Le Digabel, L.A. Leclaire, Snow water equivalent estimation using blackbox optimization. Pac. J. Optim. **9**(1), 1–21 (2013)
5. C. Audet, V. Béchard, J. Chaouki, Spent potliner treatment process optimization using a MADS algorithm. Optim. Eng. **9**(2), 143–160 (2008)
6. C. Audet, V. Béchard, S. Le Digabel, Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. J. Glob. Optim. **41**(2), 299–318 (2008)
7. C. Audet, J.E. Dennis Jr., Analysis of generalized pattern searches. SIAM J. Optim. **13**(3), 889–903 (2003)
8. C. Audet, J.E. Dennis Jr., A pattern search filter method for nonlinear programming without derivatives. SIAM J. Optim. **14**(4), 980–1010 (2004)
9. C. Audet, J.E. Dennis Jr., Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(1), 188–217 (2006)
10. C. Audet, J.E. Dennis Jr., A progressive barrier for derivative-free nonlinear programming. SIAM J. Optim. **20**(1), 445–472 (2009)
11. C. Audet, J.E. Dennis Jr., S. Le Digabel, Parallel space decomposition of the mesh adaptive direct search algorithm. SIAM J. Optim. **19**(3), 1150–1170 (2008)
12. C. Audet, J.E. Dennis Jr., S. Le Digabel, Globalization strategies for mesh adaptive direct search. Comput. Optim. Appl. **46**(2), 193–215 (2010)
13. C. Audet, J.E. Dennis Jr., S. Le Digabel, Trade-off studies in blackbox optimization. Optim. Methods Softw. **27**(4–5), 613–624 (2012)
14. C. Audet, S. Le Digabel, The mesh adaptive direct search algorithm for periodic variables. Pac. J. Optim. **8**(1), 103–119 (2012)
15. C. Audet, D. Orban, Finding optimal algorithmic parameters using derivative-free optimization. SIAM J. Optim. **17**(3), 642–664 (2006)
16. C. Audet, G. Savard, W. Zghal, Multiobjective optimization through a series of single-objective formulations. SIAM J. Optim. **19**(1), 188–210 (2008)
17. A.J. Booker, J.E. Dennis Jr., P.D. Frank, D.B. Serafini, V. Torczon, Optimization using surrogate objectives on a helicopter test example, in *Optimal Design and Control*, ed. by J. Borggaard, J. Burns, E. Cliff, S. Schreck. Progress in Systems and Control Theory (Birkhäuser, Basel, 1998), pp. 49–58

18. A.J. Booker, J.E. Dennis Jr., P.D. Frank, D.B. Serafini, V. Torczon, M.W. Trosset, A rigorous framework for optimization of expensive functions by surrogates. Struct. Multidiscip. Optim. **17**(1), 1–13 (1999)
19. A. Chipouline, V. Fedotov, Coupling plasmonics with quantum systems, in *SPIE Newsroom* (2011)
20. T.D. Choi, C.T. Kelley, Superlinear convergence and implicit filtering. SIAM J. Optim. **10**(4), 1149–1162 (2000)
21. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983). Reissued in 1990 by SIAM Publications, Philadelphia, as Vol. 5 in the series Classics in Applied Mathematics
22. A.R. Conn, S. Le Digabel, Use of quadratic models with mesh adaptive direct search for constrained black box optimization. Optim. Methods Softw. **28**(1), 139–158 (2013)
23. A.R. Conn, K. Scheinberg, L.N. Vicente, *Introduction to Derivative-Free Optimization*. MOS/SIAM Series on Optimization (SIAM, Philadelphia, 2009)
24. J.E. Dennis Jr., C.J. Price, I.D. Coope, Direct search methods for nonlinearly constrained optimization using filters and frames. Optim. Eng. **5**(2), 123–144 (2004)
25. R. Fletcher, S. Leyffer, Nonlinear programming without a penalty function. Math. Program., Ser. A **91**, 239–269 (2002)
26. R.B. Gramacy, tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. J. Stat. Softw. **19**(9), 1–46 (2007). http://www.jstatsoft.org/v19/i09
27. R.B. Gramacy, S. Le Digabel, The mesh adaptive direct search algorithm with treed Gaussian process surrogates. Technical Report G-2011-37, Les cahiers du GERAD, 2011
28. G.A. Gray, T.G. Kolda, Algorithm 856: APPSPACK 4.0: asynchronous parallel pattern search for derivative-free optimization. ACM Trans. Math. Softw. **32**(3), 485–507 (2006)
29. D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black box functions. J. Glob. Optim. **13**(4), 455–492 (1998)
30. S. Kim, J. Jin, Y.J. Kim, I.Y. Park, Y. Kim, S.W. Kim, High-harmonic generation by resonant plasmonic field enhancement. Nature **453**, 757–760 (2008)
31. M. Kokkolaras, C. Audet, J.E. Dennis Jr., Mixed variable optimization of the number and composition of heat intercepts in a thermal insulation system. Optim. Eng. **2**(1), 5–29 (2001)
32. S. Le Digabel, Algorithm 909: NOMAD: nonlinear optimization with the MADS algorithm. ACM Trans. Math. Softw. **37**(4), 44 (2011)
33. L.Y. Lin, C.J. Wang, M.C. Hegg, L. Huang, Quantum dot nanophotonics—from waveguiding to integration. J. Nanophotonics **3**, 031603 (2009)
34. A.L. Marsden, J.A. Feinstein, C.A. Taylor, A computational framework for derivative-free optimization of cardiovascular geometries. Comput. Methods Appl. Mech. Eng. **197**(21–24), 1890–1905 (2008)
35. A.L. Marsden, M. Wang, J.E. Dennis Jr., P. Moin, Trailing-edge noise reduction using derivative-free optimization and large-eddy simulation. J. Fluid Mech. **572**, 13–36 (2007)
36. M.D. McKay, W.J. Conover, R.J. Beckman, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)
37. E.D. Palik, *Handbook of Optical Constants of Solids* (Academic Press, New York, 1997)
38. E. Plum, V.A. Fedotov, P. Kuo, D.P. Tsai, N.I. Zheludev, Towards the lasing spaser: controlling metamaterial optical response with towards the lasing spaser: controlling metamaterial optical response with semiconductor quantum dots. Opt. Express **17**(10), 8548–8551 (2009)
39. E. Polak, M. Wetter, Precision control for generalized pattern search algorithms with adaptive precision function evaluations. SIAM J. Optim. **16**(3), 650–669 (2006)
40. N. Queipo, R. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P. Kevintucker, Surrogate-based analysis and optimization. Prog. Aerosp. Sci. **41**(1), 1–28 (2005)
41. J. Renger, R. Quidant, N.V. Hulst, L. Novotny, Surface enhanced nonlinear four-wave mixing. Phys. Rev. Lett. **104**(4), 046803 (2010)

42. K. Tanaka, E. Plum, J.Y. Ou, T. Uchino, N.I. Zheludev, Multifold enhancement of quantum dot luminescence in plasmonic metamaterials. Phys. Rev. Lett. **105**(22), 227403 (2010)
43. V. Torczon, On the convergence of pattern search algorithms. SIAM J. Control Optim. **7**(1), 1–25 (1997)
44. D. van Heesch, Doxygen manual (1997). http://www.doxygen.org/manual.html
45. L.N. Vicente, A.L. Custódio, Analysis of direct searches for discontinuous functions, in *Mathematical Programming* (2010)
46. C.J. Wang, L.Y. Lin, B.A. Parviz, 100-nm quantum dot waveguides by two-layer self-assembly, in *Lasers and Electro-Optics Society, LEOS 2005. the 18th Annual Meeting of the IEEE* (2005), pp. 194–195

Chapter 5
Nature Inspired Optimization Techniques for Metamaterial Design

Douglas H. Werner, Jeremy A. Bossard, Zikri Bayraktar, Zhi Hao Jiang, Micah D. Gregory, and Pingjuan L. Werner

Abstract This chapter considers a class of optimization techniques that were developed to imitate processes found in nature. Nature is a wonderful source of inspiration for global optimization because so many aspects of natural phenomenon can be mimicked and employed for solving challenging design problems, from the very process of evolution to the coordinated search behavior of various swarming organisms. Nature inspired search algorithms have played an important role in electromagnetic design, as they have proven to be very robust at solving complex problems with many design parameters. Also, as the field of metamaterials has developed, optimization has become an important tool in the quest to overcome performance limitations such as high loss and narrow bandwidth, which have limited the widespread use of metamaterials in practical device applications. In the first part of this chapter, three prominent nature inspired optimization algorithms are described in detail, including the genetic algorithm (GA), particle swarm optimization (PSO), and the covariance matrix adaptation evolutionary strategy (CMA-ES). Following this, several examples of metamaterial surfaces are presented that have each been optimized by one of the three nature inspired techniques. Finally, two homogenization techniques that can be employed to invert scattering parameters for a slab of metamaterial to obtain isotropic or anisotropic effective medium parameters are examined and used in conjunction with a GA to overcome previous limitations in terms of loss and angular stability in metamaterials.

5.1 Introduction

Nature inspired optimizers generally fall under the category of global optimization techniques as illustrated in Fig. 5.1. Unlike local optimizers, global optimization techniques are more robust for solving complex engineering design problems in which there may be multiple minima in the parameter space where a local optimization could become stuck. Here, the *cost* function is treated as a black box, where

D.H. Werner (✉)
Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: dhw@psu.edu

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_5,

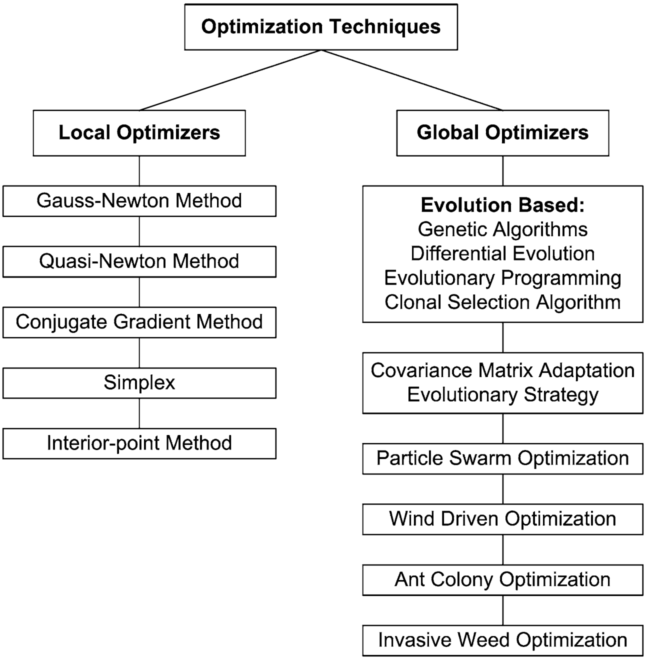


Fig. 5.1 Flowchart showing the family tree of optimization methods

the algorithm probes discrete locations in the parameter space, and higher order knowledge of the *cost* surface such as the surface gradient, which may be difficult or impossible to obtain, does not need to be known.

A primary drawback of nature inspired optimizers is the number of function evaluations that are required to converge to a result. However, because many independent designs are evaluated at each iteration of the algorithm, these techniques are good candidates for parallel computing. Hence, efforts have been made to parallelize these algorithms to reduce the total optimization time. Furthermore, newer algorithms such as CMA-ES are more efficient in terms of speed of convergence, requiring fewer *cost* evaluations than their older counterparts.

A variety of nature inspired optimization techniques have been introduced and applied to solving engineering design problems. Probably the most well known is the genetic algorithm (GA), which is inspired by the Darwinian notion of natural selection in evolution [31]. GAs have been widely studied and applied to electromagnetic optimization problems [43]. Differential evolution (DE) is another nature-inspired algorithm that is simple and straightforward to implement and can provide fast convergence as compared with other evolutionary strategies [88]. The covariance matrix adaptation evolutionary strategy (CMA-ES) is a newer method that operates by moving and reshaping a Gaussian search distribution within the parameter space [33, 40]. CMA-ES has been gaining in popularity because of its fast convergence and ease of use. The clonal selection algorithm (CLONALG) mimics the natural response of the immune system in vertebrates to stimulus from antigens [18]. A parallel version of CLONALG has recently been introduced into the electromagnetics community as an effective alternative to the GA for multimodal problems [4]. Particle swarm optimization (PSO) is an artificial implementation of the social intelligence of insect swarms, where particles share information as they work together to search for the problem solution [11]. Wind driven optimization (WDO) is a related algorithm that is based on the movement of air particles through the atmosphere [8]. Ant colony optimization (ACO) is inspired by the foraging behavior of ants [23] and has been effectively used for electromagnetics design [44]. Invasive weed optimization (IWO) is another new stochastic optimization algorithm inspired from colonizing weeds that has also been applied to electromagnetic problems including array synthesis and antenna design [49]. These algorithms indicate the wide range of optimizers that have been inspired by natural systems. Throughout this chapter, we will focus on the implementation and application of three of the nature inspired search algorithms, namely the GA, PSO, and CMA-ES.

In the first part of this chapter, three prominent nature inspired optimization algorithms will be described in detail, including the genetic algorithm (GA), particle swarm optimization (PSO), and the covariance matrix adaptation evolutionary strategy (CMA-ES). Following this, several examples of metamaterial surfaces will be presented that have each been optimized by one of the three nature inspired techniques. Finally, the last part of the chapter will examine two homogenization techniques for metamaterials that can be employed to invert scattering parameters for a slab of metamaterial to obtain isotropic or anisotropic effective medium parameters. Both of these inversion techniques will be used in conjunction with a GA to overcome previous limitations in terms of loss and angular stability in metamaterials.

5.2 Nature Inspired Optimization Methods

While there are many nature inspired optimization methods that have been developed and introduced to the electromagnetics community, the following sections will focus on explaining the operation of three techniques. The first technique described in Sect. 5.2.1 is the genetic algorithm (GA) which seeks to evolve designs according to the principles of natural selection. The GA is an early nature inspired optimization method that has a proven track record for electromagnetic design. Particle swarm optimization (PSO), which will be detailed in Sect. 5.2.2, is an artificial implementation of the social intelligence of insect swarms that has also been extensively used by the electromagnetics community. The last technique described in Sect. 5.2.3 is a relative newcomer. The covariance matrix adaptation evolutionary strategy (CMA-ES) is a self-adaptive search algorithm that can act as a “black box” for the end user. In contrast to GA and PSO, which both have many knobs that can be tweaked to tune the optimization, CMA-ES tunes itself during the optimization process and also has the advantage of requiring relatively fewer function evaluations before convergence.

5.2.1 Genetic Algorithms

The basis for one of the most popular nature inspired optimization techniques in computational electromagnetics is the evolutionary process. Evolution combined with genetics can be seen as an analogy to numerical optimization, where genes are the parameters defining living organisms, and evolution is the process by which organisms are optimized for surviving in their environment. A *genetic algorithm* (GA) mimics the evolutionary process by casting the parameters of a given design problem as genes and then evolving the design features over multiple generations while applying the principle of survival of the fittest until the parameters are optimized to a most fit design.

The GA was originally introduced in 1975 by Holland [45] and then later applied to many practical problems by Goldberg [31]. In the field of electromagnetics, the GA has been employed to solve a wide variety of problems ranging from antenna element design and phased array synthesis to scattering control of frequency selective surfaces and absorbers [43]. GAs have also been successfully employed to synthesize a variety of metamaterials as will be discussed in more detail when some specific examples are considered later in this chapter.

The GA operates on a population of individuals that are each represented by a chromosome containing all of the genes, or parameters, that describe the design:

$$population = \begin{bmatrix} chromosome_1 \\ chromosome_2 \\ \vdots \\ chromosome_N \end{bmatrix} = \begin{bmatrix} gene_{11} & gene_{12} & \cdots & gene_{1M} \\ gene_{21} & gene_{22} & \cdots & gene_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ gene_{N1} & gene_{N2} & \cdots & gene_{NM} \end{bmatrix} \tag{5.1}$$

where the population contains N chromosomes, and M parameters are defined in the chromosome. In the GA implementation, the genes can be represented by either binary bits or by real numbers and are each mapped to design parameters. These parameters can be discrete values, such as an index to a table of materials, or continuous values, such as the dimension of a feature in the design. The user must define the parameters in the design as well as how the genes will be mapped to the design parameters. The user must also define a fitness function f that accepts the chromosome as input and provides the *cost* as output:

$$f\left\{\begin{bmatrix} chromosome_1 \\ chromosome_2 \\ \vdots \\ chromosome_N \end{bmatrix}\right\} = \begin{bmatrix} cost_1 \\ cost_2 \\ \vdots \\ cost_N \end{bmatrix}. \tag{5.2}$$

The *cost* is a measure of the design performance, where low *cost* indicates a high fitness. The fitness evaluation of the population can be done sequentially using a single processing thread, or the chromosomes can be fed to multiple processors and be evaluated in parallel in order to greatly speed up the total execution time of the GA.

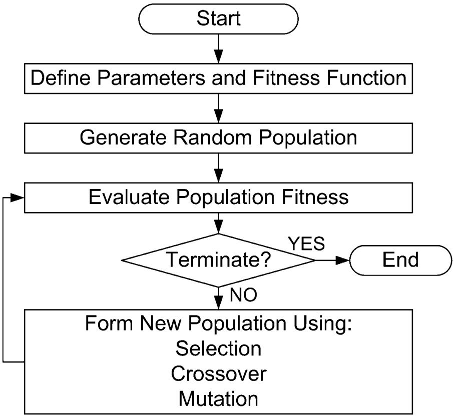

Start
Define Parameters and Fitness Function
Generate Random Population
Evaluate Population Fitness
Terminate?
YES
End
NO
Form New Population Using:
Selection
Crossover
Mutation

Fig. 5.2 Flowchart showing the operation of the genetic algorithm

A flowchart showing the operation of the GA is given in Fig. 5.2. After the user defines the design parameters and the fitness function, the GA starts by initializing the population with random chromosome values. This initial population spreads random "guesses" of the optimal solution across the M dimensional parameter space. The initial population members are then fed into the user provided fitness function as shown in (5.2) to determine the *cost*, or performance, of each design.

Once the population fitness has been evaluated, the better performing members are chosen to mate and fill out a new generation, mimicking the process of natural selection. A mating pool is first formed by either sorting the population according to *cost* and then keeping the top N_{sel} members, or by using a threshold such as the mean or the median *cost* to eliminate any population members with a worse performance than the threshold value. Members in the mating pool have survived natural selection and are eligible for breeding. Two common methods for implementing mate selection are using a roulette wheel and tournament selection. In the first method, the mating pool must first be sorted in order from lowest to highest *cost*. A roulette wheel is generated at the beginning of the optimization, giving each chromosome in the mating pool the following probability of being selected:

$$p_n = \frac{N_{\text{sel}} - n + 1}{\sum_{i=1}^{N_{\text{sel}}} i} \tag{5.3}$$

where n is the index to the chromosome array sorted from lowest to highest *cost*. For instance, a mating pool size of four would generate the following probabilities $P = [0.4, 0.3, 0.2, 0.1]$. For each parent, a uniform random number in the range [0, 1] is generated, which would map as follows to a four parent mating pool:

$$\begin{aligned} 0.0 \le r \le 0.4 &\rightarrow \mathit{chromosome}_1, \\ 0.4 \le r \le 0.7 &\rightarrow \mathit{chromosome}_2, \\ 0.7 \le r \le 0.9 &\rightarrow \mathit{chromosome}_3, \\ 0.9 \le r \le 1.0 &\rightarrow \mathit{chromosome}_4. \end{aligned} \tag{5.4}$$

If the random number $r = 0.4378$ is generated, then $chromosome_2$ will be used as a parent for mating.

The tournament selection method does not require the mating pool to be sorted. In this method, a small group (typically two or three) of chromosomes are randomly selected from the mating pool for each tournament. In each group, the two chromosomes with the lowest *cost* are mated to produce offspring. Tournament selection can be used with thresholding so that neither the population nor the mating pool require sorting. According to [42], using either tournament selection or sorting and a roulette wheel results in similar probabilities of selection for a given chromosome.

Once two parents ($parent_1$ and $parent_2$) are selected for mating, offspring are generated using a crossover operation. In the case of a binary-valued chromosome, a binary mask is generated with the same length as the chromosome. Each bit in the mask indicates from which parent the corresponding bit in the offspring will come. The binary mask can be generated using several methods, including single point crossover, double point crossover, or uniform crossover. Uniform crossover is the most general case, in which each bit in the mask is randomly populated. A single point crossover mask with 10 bits and the crossover point chosen after the sixth bit would be:

$$mask = [\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ ]. \tag{5.5}$$

If the parents have the values:

$$\begin{aligned} parent_1 &= [\ 1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ ], \\ parent_2 &= [\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ ], \end{aligned} \tag{5.6}$$

then the offspring generated using this mask would be:

$$\begin{aligned} offspring_1 &= [\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ ], \\ offspring_2 &= [\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ ] \end{aligned} \tag{5.7}$$

where the bits coming from $parent_1$ are highlighted in gray. In the case of a real-valued chromosome, several techniques for performing crossover on continuous values are described in [43].

After selection and crossover have been performed to fill out the population for the next generation, a small percentage of bits in the new population are mutated in order for the algorithm to continue exploring new parts of the parameter space. If a bit is randomly selected for mutation, then its binary value is flipped from 0 to 1 or from 1 to 0. In the case of a real-valued chromosome, it is common to mutate a gene by giving it a new random value within its allowed range. Typical mutation probabilities are on the order of a few percent, but one variation on a GA called a micro GA uses a mutation probability of zero. A micro GA will converge to a solution much quicker than a GA, but the lack of mutation does not allow the micro GA to search the parameter space beyond the range afforded by the initial population.

The final step in populating the new generation is to optionally enforce elitism. Elitism ensures that the global best fitness is maintained between generations by copying the chromosome with the best fitness from the previous generation into the new population. At this point, the new population is ready to be evaluated by the fitness function.

This generational cycle of evaluating fitness and filling out a new population using natural selection, mating, and mutation continues until the algorithm is stopped according to a predefined termination criterion or manually by the user. Some common conditions for terminating the run include reaching a *cost* value lower than an acceptable minimum, the *cost* has not improved after a set number of iterations, or completing a set number of generations.

5.2.2 Particle Swarm Optimization

The artificial implementation of the social intelligence of insect swarms [11] has been of interest to many researchers over the last few decades, resulting in various successful applications from robotics to optimization [26, 92]. To the untrained eye, swarm behavior may appear very chaotic, yet it can execute highly coordinated and sophisticated working structures that perform different tasks ranging from searching for food to defending against predators, each of which are crucial to the survival of all members in the population.

The particle swarm optimization (PSO) technique was introduced in 1995 by Kennedy and Eberhart as an artificial implementation of swarm intelligence [25, 50] to mimic the decentralized but coordinated movements of the members of a swarm over an N-dimensional search space. In essence, PSO is a heuristic, population-based, iterative global optimization algorithm. In PSO, the population, i.e., the swarm, consists of a predetermined number of infinitesimally small members, which are also called particles. The coordinates of a particle over the search space are mapped to the optimization parameters, which correspond to a unique solution candidate for the optimization problem at hand. As particles traverse the search space, they try to improve their location by remembering their personal best locations and sharing this information among the rest of the population. This information sharing approach sets PSO apart from the GA, where the GA relies on competition among its population members.

As shown in the flowchart in Fig. 5.3, PSO starts with the initialization of its parameters, including specifying the constraints on the search boundaries for each dimension $[x^d_{\min}, x^d_{\max}]$ and on the maximum allowed velocity ($\upsilon_{\max}$). The swarm of infinitesimally small particles are each assigned a position vector ($\boldsymbol{x}$) and randomly distributed within the boundaries of the N-dimensional search space. The movement of the swarm over the search space can be analogous to a flock of birds flying over a hilly terrain in search of the best position to land. The hills and valleys of the terrain are analogues to the maxima and minima *cost* locations within the search domain, where each particle evaluates its current location based on the user-defined

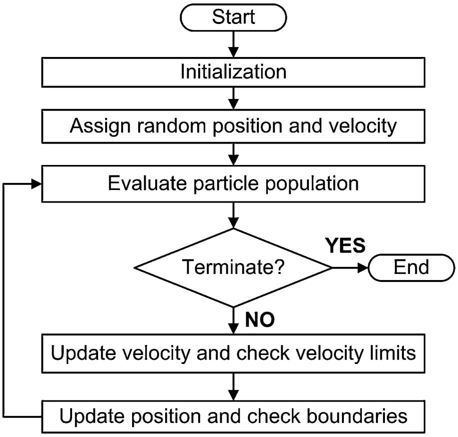

Start
Initialization
Assign random position and velocity
Evaluate particle population
Terminate?
YES
End
NO
Update velocity and check velocity limits
Update position and check boundaries

Fig. 5.3 Flowchart showing the operation of the particle swarm optimization

cost function. The location vector that provides the lowest *cost* among all particles is assigned to be the global best location ($\boldsymbol{x}^{\mathrm{gbest}}$). If the global best location provides a desired minimum *cost* defined by the user, the optimization would terminate. If the targeted *cost* has not been achieved, the iterative procedure continues by updating the velocity ($\boldsymbol{v}$) and the position of each particle. The velocity update formula is given by:

$$\boldsymbol{v}(t+\delta t) = \big(\omega \cdot \boldsymbol{v}(t)\big) + c_1 \cdot \big(\boldsymbol{x}^{\mathrm{pbest}} - \boldsymbol{x}(t)\big) + c_2 \cdot \big(\boldsymbol{x}^{\mathrm{gbest}} - \boldsymbol{x}(t)\big) \tag{5.8}$$

where $\boldsymbol{v}(t+\delta t)$ is the updated velocity vector for each particle, ω is the nostalgia term, c_1 and c_2 are randomly-generated constants, $\boldsymbol{x}(t)$ is the position vector of the particle at the current iteration, and $\boldsymbol{x}^{\mathrm{pbest}}$ and $\boldsymbol{x}^{\mathrm{gbest}}$ are personal best location of the particle and the global best location of the whole swarm, respectively. The position of each particle is then updated according to:

$$\boldsymbol{x}(t+\delta t) = \boldsymbol{x}(t) + \big[\delta t \cdot \boldsymbol{v}(t+\delta t)\big] \tag{5.9}$$

where for each particle, i, the position vector of the current iteration, $\boldsymbol{x}(t)$, is updated to the position vector of the next iteration, $\boldsymbol{x}(t+\delta t)$. The time step, δt, is usually chosen to be $\delta t = 1$ for simplicity.

Since the dimensions of the search space are limited by upper and lower boundaries, constraints must be placed on the particle velocities and positions, so that the PSO can efficiently search for good solutions. Fast particles can take large steps at each iteration and easily overshoot good regions of the search space, which, in turn, hinders the performance of the optimization algorithm. Taking large steps would also result in the particles traveling quickly and accumulating at the boundaries without completely exploring the search space. Hence, exceedingly fast particles must be velocity limited according to:

$$\text{if } \big(\big|\boldsymbol{v}(t+\delta t)\big| > \nu_{\max}\big) \text{ then } \boldsymbol{v}(t+\delta t) = \nu_{\max} \cdot \frac{\boldsymbol{v}(t+\delta t)}{|\boldsymbol{v}(t+\delta t)|} \tag{5.10}$$

Table 5.1 Description of the particle swarm optimization terminology

Term	Description
Particle	Representation of a candidate solution to the N-dimensional optimization problem
Population (i.e., Swarm)	A predetermined number of particles, which iteratively alter their dimensions in search of the best parameter values for the optimization problem
Generation	Successive iterations, where velocity and position of particles are updated
Position; $\boldsymbol{x}(t)$	The coordinates of a particle, which corresponds to the parameter values for the N-dimensional optimization problem
Velocity; $\boldsymbol{v}(t)$	The main operator that determines the position change every iteration
Cost Function	A scalar value assigned to each particle, i.e., candidate solution, based on their proximity to the desired design goals
$\boldsymbol{x}^{\text{gbest}}$	Best location found among all particles up until the current iteration
$\boldsymbol{x}^{\text{pbest}}$	Personal best location found by each particle up until the current iteration
$v_{\max}$	The maximum allowable velocity in one dimension

to prevent particles from skipping over large areas of the search space. Furthermore, if a particle passes the boundary of a given dimension, the position is reset along that dimension according to:

$$\text{if } \left(x^d > x^d_{\max}\right) \text{ then } x^d = x^d_{\max} \text{ or if } \left(x^d < x^d_{\min}\right) \text{ then } x^d = x^d_{\min} \tag{5.11}$$

to prevent the particle from flying out of the boundaries.

This iterative procedure of evaluating the swarm *cost* and then updating particle velocities and positions continues as the PSO is searching for the optimum coordinates. Finally, the PSO can be terminated according to some preset criteria, such as reaching a global best *cost* value lower than an acceptable minimum or completing a certain number of time steps. A summary of the PSO terminology along with a description is given in Table 5.1. The PSO algorithm is an effective tool for optimizing metamaterials as will be demonstrated through the synthesis of a double-sided artificial magnetic conducting (DSAMC) ground plane presented in Sect. 5.3.2.

5.2.3 *Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) Optimization*

The most commonly used techniques for global function minimization in the electromagnetics community have been the previously described genetic algorithm (GA) [43, 45] and particle swarm optimization (PSO) methods [10, 25, 26, 30, 50, 77]. These techniques have worked well for most problems facing engineers in the field thus far, yielding designs with suitable performance in reasonable amounts of time. However, electromagnetics simulation software and computing platforms have substantially advanced in the past few decades, nearly to the point where if a design can be conceptualized, it can likely be simulated, giving the designer a great deal of flexibility and potentially leading to optimization problems with dozens of parameters, or even more. Even with modern high-performance computing platforms, full-wave simulation of such designs can require substantial amounts of time, and optimizations using these analysis methods often require many function calls (i.e., simulations) to reach a desired performance goal. For this reason, it is always of interest to those using evolutionary design to select an algorithm which offers the fastest optimization times (in the form of the fewest *cost* function calls) with a reasonable certainty that an acceptable solution will be found.

The covariance matrix adaptation evolutionary strategy (CMA-ES) is an increasingly popular method for global optimization of real-valued electromagnetics design problems [33]. The current algorithm has been developed incrementally in the evolutionary computation (EC) community [39–41], where it has proven itself to be a very competent and competitive strategy [37, 38]. Self-adaptive strategies such as CMA-ES are a popular and powerful technique for global function minimization. In addition, to the designer it is relatively easy to use as a "black-box optimizer" requiring only the selection of the population size before beginning an optimization. In this section, the inner workings of the CMA evolutionary strategy are explained. Illustrations with a simple test function are used to demonstrate how the algorithm performs.

CMA-ES operates by moving and reshaping a Gaussian distribution about the search space in an attempt to find the global function minimum. The distribution is completely defined by the mean, $\boldsymbol{m}$, and the shape, $\sigma^2\mathbf{C}$. Several internal strategy parameters, such as evolution paths, are utilized to give the algorithm its self-adaptive properties. In addition, the algorithm makes use of *cumulation* to dampen the adaptation of the covariance matrix to effectively work with small population sizes. Evolutionary strategy parameters such as mutation and crossover rates with the GA, or nostalgia and social constants with PSO are commonly chosen beforehand and remain fixed during the course of optimization. However, not only does this leave the decision up to a "best guess" for the user (although there are usually suggested values for each algorithm), but the ideal set of strategy parameters is likely problem dependent and may also change throughout the course of the optimization. Modern self-adaptive strategies such as CMA-ES internally adjust search characteristics according to progress, accounting for changing function landscapes and attempting to make the most progress in the fewest number of algorithm iterations.

CMA-ES uses several strategy parameters and internal operating scalars, vectors, and matrices for its operation. The strategy parameters along with their symbols and descriptions are provided in Table 5.2. Although there seems to be many to choose, all but the population size λ are determined by the properties of the problem. Once the population size has been selected, many of the other parameters are then determined by the choice of λ. The internal operating parameters are given in Table 5.3.

Table 5.2 Symbols, values, and descriptions of the strategy parameters of CMA-ES

Symbol	Value	Description
n	Determined by problem	Number of optimizable parameters
λ	$\lambda \geq 4 + \lfloor 3 \ln n \rfloor$	Population size or number of children
μ	$\mu = \lfloor \lambda/2 \rfloor$	Number of parents, typically $\lambda/2$
w_i	$w_i = \frac{\ln(\mu'+0.5)-\ln i}{\sum_{j=1}^{\mu} \ln(\mu'+0.5)-\ln j}$ where $\mu' = \lambda/2$	Selection/recombination weights
μ_{eff}	$\mu_{\text{eff}} = (\sum_{i=1}^{\mu} w_i^2)^{-1}$	Variance effective selection mass
c_σ	$c_\sigma = \frac{\mu_{\text{eff}}+2}{n+\mu_{\text{eff}}+5}$	Learning rate for cumulation of the step size control
d_σ	$d_\sigma = 1 + c_\sigma + 2\max(0, \sqrt{\frac{\mu_{\text{eff}}-1}{n+1}} - 1)$	Damping parameter for the step-size update
c_c	$c_c = \frac{4+\mu_{\text{eff}}/n}{n+4+2\mu_{\text{eff}}/n}$	Learning rate for cumulation of the rank-one update of the covariance matrix
c_1	$c_1 = \frac{2}{(n+1.3)^2+\mu_{\text{eff}}}$	Learning rate for the rank-one update of the covariance matrix
c_μ	$c_\mu = \min(1 - c_1, \frac{2\mu_{\text{eff}}-4+2/\mu_{\text{eff}}}{(n+2)^2+\mu_{\text{eff}}})$	Learning rate for the rank-μ update of the covariance matrix

Table 5.3 Symbols and descriptions of the internal operating parameters of CMA-ES

Symbol	Description
$\sigma \in \mathbb{R}^+$	Step-size
$\boldsymbol{m} \in \mathbb{R}^n$	Distribution mean
$\mathbf{C} \in \mathbb{R}^{n\times n}$	Covariance matrix
$\mathbf{B} \in \mathbb{R}^{n\times n}$	Columns are eigenvectors of $\mathbf{C}$
$\mathbf{D} \in \mathbb{R}^{n\times n}$	Diagonal elements are eigenvalues of $\mathbf{C}$
$\boldsymbol{p}_\sigma \in \mathbb{R}^n$	Conjugate evolution path
$\boldsymbol{p}_c \in \mathbb{R}^n$	Evolution path
$\boldsymbol{z}_k \in \mathbb{R}^n$	A sample from the standard Normal multivariate distribution
$\boldsymbol{y}_k \in \mathbb{R}^n$	A sample from the distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$
$\boldsymbol{x}_k \in \mathbb{R}^n$	An offspring or potential candidate solution

These contain the information about the search distribution, evolution paths to determine past properties of the distribution, and information about population members.

For illustration purposes, a simple canyon test function shown in Fig. 5.4 will be used to demonstrate the operation of the algorithm. The test function was specifically designed to have diagonal trenches to show how CMA-ES effectively traverses irregular and inseparable search spaces. The algorithm is initialized by setting the initial distribution position ($\boldsymbol{m}$) and shape ($\sigma^2 \boldsymbol{C}$) as well as choosing a population size. After the problem is formulated, the initial position is typically set randomly inside the search boundary, although, like most evolutionary strategies, it can be

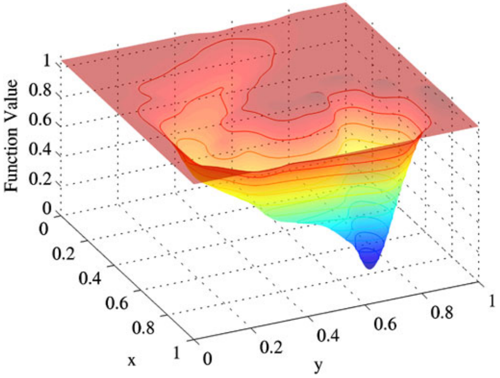


Fig. 5.4 The canyon test function used for illustrating the operation of CMA-ES. The minimum function value and goal is located at $(x, y) = (0.764, 0.724)$

preset to a specific position if information is known about the problem. The initial shape is usually set such that for each parameter, the distribution has a standard deviation of one-third of its range. This leaves the user to choose only the population size, for which

$$\lambda \geq 4 + \lfloor 3 \ln n \rfloor \tag{5.12}$$

is recommended [41], with larger populations resulting in increased robustness and global search capacity at the cost of slower optimization (in the form of more function calls). Once λ is selected, all of the strategy parameters in Table 5.2 can then be computed. The evolution paths in Table 5.3 are both set to **0** upon initialization as well. The initial distribution configured to operate on the test function in Fig. 5.4 is shown in Fig. 5.5 (at iteration 0). A small initial step-size is used in this example since the function is simple and it illustrates the ability of CMA-ES to easily traverse the search spaces. Additionally, a large population size is again used for illustrative purposes, as it tends to generate a more regular movement of the distribution. Smaller population sizes tend to have more sporadic movement of the mean about the search space, yet result in fewer total function evaluations.

With the initial distribution and evolution paths configured, the algorithm is ready to begin the first round of sampling. In order to sample from the distribution $\mathcal{N}(\boldsymbol{m}, \sigma^2\boldsymbol{C})$, the covariance matrix must first be broken up into its eigenvectors **B** and eigenvalues **D**. This is commonly done through principle component analysis (PCA), also called eigendecomposition. For optimizations where *cost* function calls are very fast, PCA can consume a significant percentage of the total CPU time. However, for electromagnetics design problems, *cost* function calls are usually the primary time consumer due to the need for computation-intensive simulations. Once **B** and **D** are obtained, $\mathcal{N}(\boldsymbol{m}, \sigma^2\boldsymbol{C})$ can be sampled by first drawing from a standard Normal distribution (a simple procedure for computers)

$$\boldsymbol{z}_k^g \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{5.13}$$

and then transforming to the desired distribution through

$$\boldsymbol{y}_k^g = \mathbf{BDz}_k \tag{5.14}$$

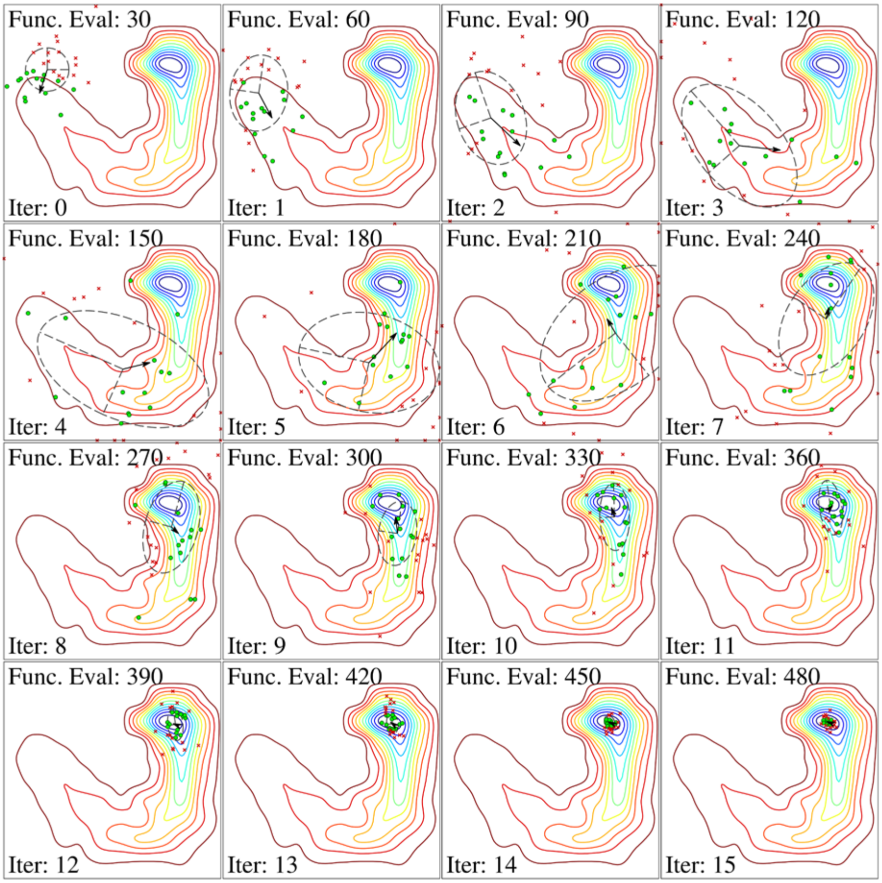


Fig. 5.5 Operation of CMA-ES on the two-dimensional cavern test function shown in Fig. 5.4. A population size of $\lambda = 30$ is used with the algorithm initialized to $\boldsymbol{m} = (0.2, 0.7)$, $\sigma = 0.1$, and $\mathbf{C} = \mathbf{I}$. The *dashed ellipse* represents a contour of equal likelihood of selection. The μ selected children from each iteration are represented by *green circles*; the $(\lambda - \mu)$ discarded children with a *red* $\times$. The *arrow* represents the movement of the mean at each iteration. A smaller than nominal initial step-size is used to highlight the ability of the algorithm to traverse valleys of low cost in inseparable search spaces

and

$$\boldsymbol{x}_k^g = \boldsymbol{m} + \sigma\, \boldsymbol{y}_k^g, \tag{5.15}$$

giving our first set of candidate solutions (or population members) at iteration $g = 0$. Now the population is evaluated according to the user-defined *cost* function to return a single *cost* value for each member. It is not uncommon here for a large cluster of computers to each share in the burden of computing the *cost* of one or more of the population members, thus saving the user a significant amount of total optimization time. After the entire population is evaluated, the members $\boldsymbol{y}_k^g$ and $\boldsymbol{x}_k^g$ are sorted (where the sorted members are identified by $\boldsymbol{y}_{i:\lambda}^{g}$ and $\boldsymbol{x}_{i:\lambda}^{g}$) according to cost value and are used to form the new mean given by

$$\langle \boldsymbol{y}\rangle_w = \sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}^{g} \tag{5.16}$$

and

$$\boldsymbol{m}^{g+1} = \boldsymbol{m}^{g} + \sigma^{g}\langle \boldsymbol{y}\rangle_w, \tag{5.17}$$

which is also equivalent to simply adding a weighted average of the μ best members in

$$\boldsymbol{m}^{g+1} = \sum_{i=1}^{\mu} w_i \boldsymbol{x}_{i:\lambda}^{g}. \tag{5.18}$$

After the new mean is computed, the conjugate evolution path and evolution path are updated using

$$\boldsymbol{p}_{\sigma}^{g+1} = (1-c_{\sigma})\boldsymbol{p}_{\sigma}^{g} + \sqrt{c_{\sigma}(2-c_{\sigma})\mu_{\text{eff}}}\left(\mathbf{C}^{g}\right)^{-\frac{1}{2}}\langle \boldsymbol{y}\rangle_w \tag{5.19}$$

and

$$\boldsymbol{p}_{c}^{g+1} = (1-c_{c})\boldsymbol{p}_{c}^{g} + \sqrt{c_{c}(2-c_{c})\mu_{\text{eff}}}\langle \boldsymbol{y}\rangle_w, \tag{5.20}$$

respectively. These contain normalized (5.19) and non-normalized (5.20) distribution movement history that is used for updating the step size and covariance matrix, respectively. Note that $\mathbf{C}^{-1/2}$ can be found through the identity

$$\mathbf{C}^{-1/2} = \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{\mathrm{T}}, \tag{5.21}$$

for which $\mathbf{D}^{-1}$ can be computed easily as it contains only diagonal terms. Next, the step size is updated using

$$\sigma^{g+1} = \sigma^{g} e^{\frac{c_{\sigma}}{d_{\sigma}}\left(\frac{\|p_{\sigma}\|}{\mathrm{E}\|\mathcal{N}(\mathbf{0},\boldsymbol{I})\|}-1\right)} \tag{5.22}$$

where $\mathrm{E}\|\mathcal{N}(\mathbf{0},\boldsymbol{I})\|$ is the expected value of the n-dimensional standard normal distribution. Lastly, the covariance matrix is updated using

$$\mathbf{C}^{g+1} = (1-c_1-c_{\mu})\mathbf{C}^{g} + c_1\boldsymbol{p}_c\boldsymbol{p}_c^{\mathrm{T}} + c_{\mu}\sum_{i=1}^{\mu} w_i \boldsymbol{y}_{i:\lambda}\boldsymbol{y}_{1:\lambda}^{\mathrm{T}} \tag{5.23}$$

where the first term is the historical contribution (cumulation), the second term is the rank-one update (elongation of distribution along the direction of search), and the third term is the rank-μ update (formation of distribution from successful search steps). Note that the update signals h_{σ} and $\delta(h_{\sigma})$ are omitted for simplicity from (5.20) and (5.23) in the aforementioned implementation of CMA-ES [35]. With the updated mean, step-size, and covariance matrix, the next round of sampling can begin and the process repeats until the desired function value is reached, the algorithm converges, or time is expired.

The remaining blocks in Fig. 5.5 show the algorithm progressing along the search space, eventually finding the desired function goal of 0.001 over 15 iterations. It is easy to observe how the algorithm effectively operates by extending the search distribution along the path of movement, thus ensuring that future steps will yield solutions with lower *cost*. Conversely, when the movement beings to slow, the step-size shrinks, and the algorithm starts to search locally.

While implementing CMA-ES is much more complex as compared with a simple genetic algorithm or particle swarm technique, the implementation challenges are balanced by the considerable advantage that CMA-ES offers in terms of performance improvement over the simpler strategies [33, 37, 38]. For common electromagnetics problems, where *cost* function calls can take minutes or longer per evaluation, the optimization time can be reduced to a fraction of what was previously required, sometimes subtracting hours or days from the optimization process. For example, in [33] the time to design a simple stacked-patch antenna with CMA-ES was reduced to 38 % of what was required with PSO (18 hours versus 47 hours).

In addition to being a fast algorithm, CMA-ES also tends to be robust even with the smallest recommended population size given by Eq. (5.12). The stacked-patch antenna optimization reported in [33] was carried out reliably with a population size of 10 compared to the population of 40 particles needed by PSO to be reliable. Also in [33], several different ultra-wideband (UWB) non-uniformly spaced antenna array layouts were optimized, demonstrating the ability of the algorithm to quickly generate a large number of high performance designs with as many as 100 elements.

The effectiveness of CMA-ES for optimization of metamaterials will be demonstrated later in Sect. 5.3.3 of this chapter, where it is utilized for synthesizing a matched magneto-dielectric absorber. Readers interested in implementing CMA-ES are highly encouraged to visit Hansen's webpage to obtain source code and to read the provided tutorial [35, 36]. More details on the application of CMA-ES to specifically solving electromagnetics problems can be found in [33].

5.3 Metamaterial Surface Optimization Examples

There are many types of artificial surfaces with novel electromagnetic properties that have been proposed and demonstrated. Three examples of metamaterial surfaces will be considered in the following sections that have been synthesized by each of the nature inspired optimizers described in this chapter. In Sect. 5.3.1, a dual-band metamaterial absorber (MMA) with wide angular stability is synthesized using a GA. Then, in Sect. 5.3.2, PSO is utilized to design a double-sided artificial magnetic conducting (DSAMC) ground plane. Finally, in Sect. 5.3.3, another type of absorber based on a matched-impedance magneto-dielectric metamaterial (MIMDM) is synthesized using CMA-ES.

5.3.1 Metallo-Dielectric Metamaterial Absorbers for the Infrared

Most metamaterial development work has focused on engineering the real part of the effective permittivity and/or permeability [56]. However, a recently proposed perfect metamaterial absorber (MMA) has drawn attention to the oft-overlooked imaginary part of the optical constant, which can be manipulated to create high absorption [63]. The availability of such MMAs could provide significant performance improvements for diverse applications including microwave-to-infrared signature control [16, 51, 63, 95], bio-chemical spectroscopy [9, 17, 66, 89–91, 97, 99], and thermal imaging [19, 62, 67]. Compared with classical absorber designs such as Salisbury screens, metamaterial absorbers can possess much thinner structures, rendering them more suitable for radar and tracking applications. Nevertheless, most up-to-date metamaterial absorber designs, covering the RF [16, 51, 63, 95] and THz [9, 62, 89–91, 97] up to IR [62, 66, 67, 99] wavelengths, function only at either near-normal incidence or for a single polarization. In this section, the electromagnetic design and nanofabrication of a conformal MMA is effectively optimized by a genetic algorithm (GA) to have two nearly perfect, narrow absorption bands centered at mid-infrared (mid-IR) wavelengths of 3.3 and 3.9 μm with polarization-independent absorptivity greater than 90 % over a ±50° angular range [48].

This dual-band mid-IR MMA employs a three-layer metallodielectric stack composed of two gold (Au) layers—a doubly periodic array of electrically isolated nanoresonators at the top and a solid ground plane at the bottom—separated by a thin dielectric layer. Kapton was chosen for the dielectric layer because it is a highly durable and flexible polymer that can easily conform to the topography of most practical curved surfaces. The array of Au nanostructures on the top screen create a resonant electric response, while the Au ground plane functions together with the top screen to produce strong coupling to the magnetic component of the incident light radiation. The continuous Au ground plane, which is thicker than the penetration depth of light in the mid-IR wavelength regime, prevents transmission of incident radiation through the structure. Therefore, strong absorption is achieved by minimizing the in-band reflection. Importantly, the Au ground plane also decouples the electromagnetic properties of the MMA coating from the surface it protects, allowing integration onto curved surfaces of arbitrary materials.

A robust GA coupled with a full-wave electromagnetic solver was employed to optimize the geometry and dimensions of the structure to best satisfy the user-defined requirements. During the GA evolution, the wavelength-dependent scattering parameters of each candidate design were calculated using the Ansoft High Frequency Structure Simulator (HFSS) full-wave finite-element solver with appropriate boundary conditions assigned to approximate a TEM wave incident on the structure at different angles. The absorptivity was calculated by $A = 1 - T_{\mathrm{TE,TM}} - R_{\mathrm{TE,TM}}$, where $T_{\mathrm{TE,TM}} = |S_{21}|^2$ and $R_{\mathrm{TE,TM}} = |S_{11}|^2$ represent the TE and TM reflectivity and transmittance, respectively. The absorptivity was evaluated against an ideal dual-band absorber response to determine its *cost*, given by

$$Cost = \sum_{\lambda} \sum_{\theta_i} \left[(1 - A_{\theta_i,\mathrm{TE}}) + (1 - A_{\theta_i,\mathrm{TM}}) \right] \qquad (5.24)$$

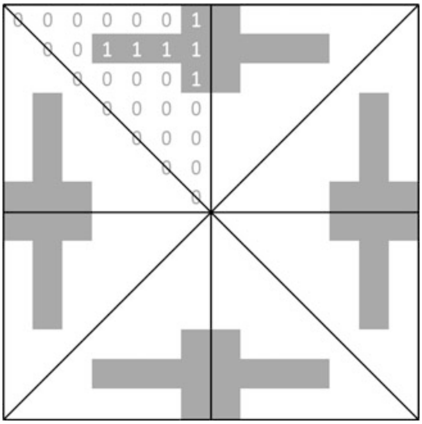


Fig. 5.6 Pixellized unit cell geometry with 8-fold symmetry and a 14 × 14 array of pixels. One triangular fold is encoded into the chromosome with "0" representing no metal and "1" representing metal

where λ is the wavelength of the arbitrarily selected target bands (3.3 μm, 3.9 μm) and θ_i is the desired angle of incidence range ($\theta_i = 0°, 10°, 20°, 30°, 40°, 50°$). The GA evolved the top Au screen nanoresonator geometry, unit cell size, and Kapton thickness until it converged to a sufficiently low *cost* solution (i.e., the optimization goal was achieved).

The unit cell geometry was pixellized into a 14 × 14 grid with 8-fold symmetry as shown in Fig. 5.6, so that the normal incidence response would be polarization insensitive. One triangular fold of the unit cell was encoded into the chromosome, where each row was represented by a single parameter. Two additional 8-bit parameters were used to encode the unit cell size and the Kapton thickness, while the Au screen thicknesses were fixed to be 50 nm. The GA operated on a population of 64 members over 50 generations, using tournament selection, uniform crossover, and a mutation probability of 0.1.

The final GA-optimized dual-band MMA design is displayed in Fig. 5.7(a) (right top). A detailed single unit cell illustration is also shown in Fig. 5.7(a), including its geometry and dimensions. The calculated angular dispersion of the absorption for both polarizations is shown in Fig. 5.8(a). The two vertical red strips demonstrate that the two absorption peaks remain centered at 3.3 and 3.9 μm over a broad range of incidence angles for both polarizations. The absorptivity in both bands remains $>91\ \%$ over a wide field-of-view of ±50° due to the efficient excitation of both electric and magnetic resonances. Further investigation shows that this MMA still achieves absorptivity $>60\ \%$ for TE polarization and $>85\ \%$ for TM polarization in both bands over an incidence angle range of 160°.

The GA-optimized MMA coating was fabricated by evaporating the Au ground plane layer and spin coating the thin Kapton dielectric layer on a handle substrate. The periodic array of H-shaped nanoresonators was patterned on top of the Kapton layer using electron beam lithography followed by a Au lift off procedure as shown in the field emission scanning electron microscope (FESEM) image in Fig. 5.7(a). The three-layer metallodielectric structure was then removed from the handle substrate to demonstrate its mechanical flexibility and durability (see Fig. 5.7(b)). The absorptivity of the fabricated MMA, calculated from the reflectivity measured using

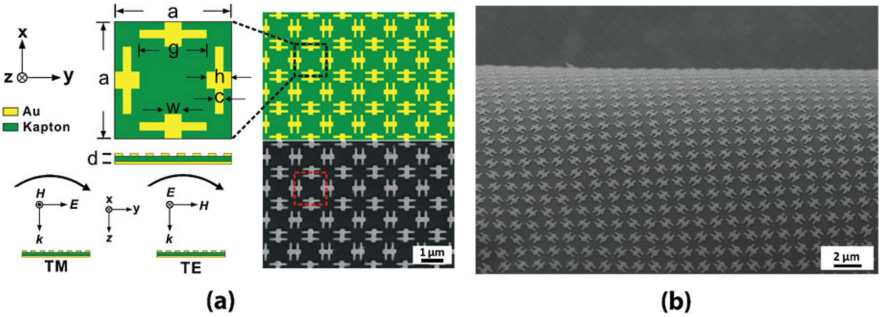


Fig. 5.7 (**a**) *Top*: Doubly periodic array of H-shaped nanoresonators with magnified view of one unit cell with dimensions $a = 1475$ nm, $h = 315$ nm, $w = 210$ nm, $g = 840$ nm, $c = 105$ nm, and $d = 200$ nm (*top* and *bottom* Au: 50 nm, Kapton: 100 nm). *Bottom*, *right*: FESEM image of a portion of the fabricated MAA. Scale = 1 µm. *Bottom*, *left*: The orientation of the incident fields with respect to the MMA. (**b**) Low magnification FESEM image of the freestanding fabricated conformal MMA coating showing its mechanical flexibility. Scale = 2 µm

a Fourier Transform IR (FTIR) spectrometer, is shown in Fig. 5.8(b) as a function of both the wavelength and the angle of incidence. The two absorption peaks of the MMA remain above 90 % for incidence angles up to 50° for both polarizations, with the −10 dB bandwidth of both bands exhibiting a maximum broadening of 0.06 µm over the entire angular range compared to the simulated results.

5.3.2 Double-Sided AMC Ground Planes for RF

Another type of practical metamaterial that has been extensively explored for RF applications are artificial magnetic conducting (AMC) surfaces, which are engineered to mimic Perfect Magnetic Conducting (PMC) ground planes. Here, the effectiveness of particle swarm optimization (PSO) for synthesizing AMC metamaterials will be explored. Unlike Perfect Electric Conducting (PEC) surfaces (e.g., simple metallic sheets at RF), which have an out of phase reflection ($R = -1$), PMCs do not occur naturally. Rather, composite metallo-dielectric structures must be synthesized to behave as PMCs, at least within a certain frequency range. These are usually constructed as high-impedance resonance surfaces, and since their introduction by Sievenpiper et al. over a decade ago [84], there have been various proposed configurations discussed in the literature [53, 54, 68]. PMCs are frequently utilized in low profile antenna applications, where a horizontal antenna element is placed in close proximity to the AMC ground plane ($h \ll \lambda_0/4$) and can still operate efficiently. Such interaction can be explained through image theory, which states that a horizontal electric source near a PMC will see an in-phase image (virtual source) at an equivalent distance below the PMC. Hence, the horizontal antenna element will couple constructively with its image below the PMC to allow the system to radiate efficiently [2].

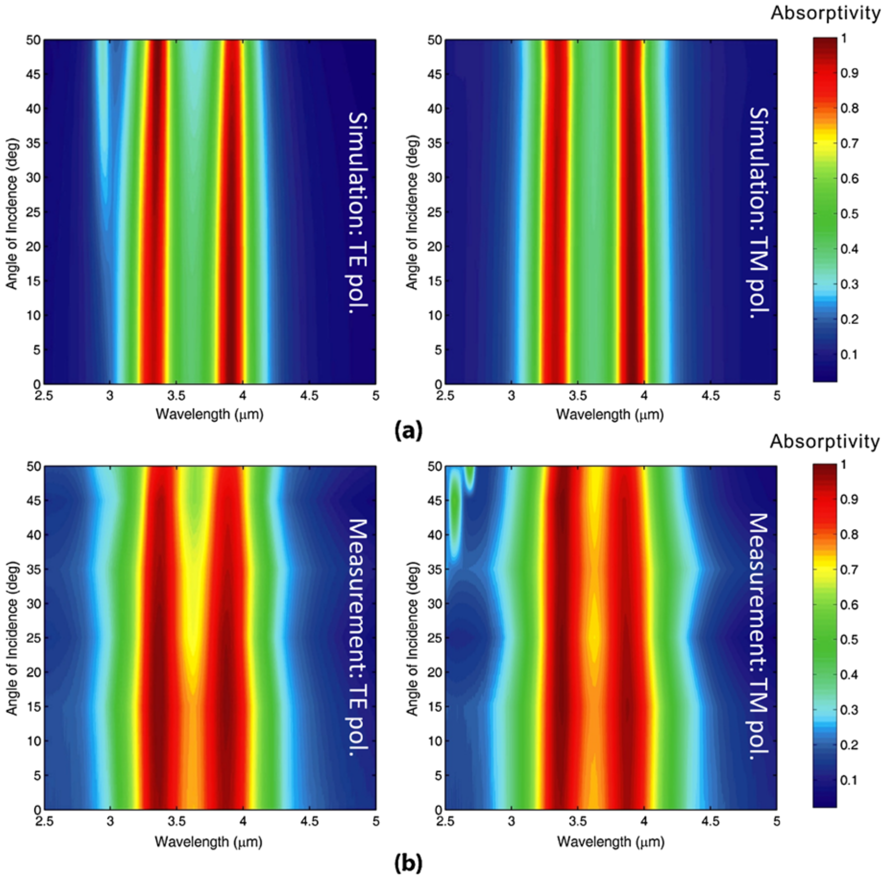


Fig. 5.8 (**a**) Contour plot of simulated absorptivity as a function of wavelength and angle of incidence under TE (*left*) and TM (*right*) incident radiation. The *two vertical red strips* clearly show two strong absorption bands nearly independent of the incident angle. (**b**) Contour plot of measured absorptivity as a function of wavelength and angle of incidence under TE (*left*) and TM (*right*) incident radiation. The *two vertical red strips* exhibit the two angularly-independent absorption bands in strong agreement with the theoretical prediction. Near-unity absorptivities were achieved in both bands experimentally, confirming a successful realization of the designed structure

AMC metasurfaces are composite metallo-dielectric structures consisting of a periodic metallic top layer printed on a thin dielectric substrate, which is backed by a PEC ground plane. The low profile antenna is placed in close proximity to the top of the AMC metasurface, which is synthesized to exhibit a high-impedance, where the tangential magnetic field at the surface is small resulting in an artificial magnetic behavior. Depending on the configuration, some designs may also have vertical metallic vias that connect the top metallic patches on the surface with the PEC ground plane backing. Erentok et al. introduced the first volumetric AMC configuration for antenna applications in [29] with a design that eliminated the need

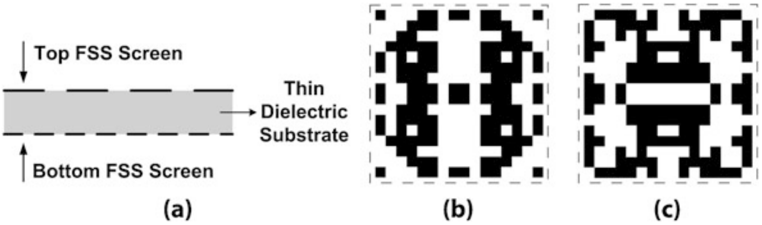


Fig. 5.9 (**a**) Cross-sectional diagram of the proposed AMC structure. (**b**) The top and (**c**) bottom FSS screen geometries were optimized by the PSO design procedure. Material properties and thickness of the thin dielectric substrate are predefined, so that only the FSS screen geometries are optimized. The black pixels represent metallic patches and dielectric exposed areas are represented by the white pixels

for a complete PEC ground plane while still supporting AMC behavior at the upper surface. This volumetric AMC design consisted of vertically oriented metallic capacitively loaded loops (CLL) printed and stacked on dielectric layers. While one face of the surface provided AMC behavior at the design frequency, the opposite surface provided a PEC response. However, the stacked layers could potentially make the fabrication difficult and the electrical thickness of the metamaterial was approximately $\lambda_r/3$, where λ_r is the wavelength within the dielectric. To simplify the fabrication process and reduce the overall thickness, a novel, thin double-sided AMC (DSAMC) design was introduced in [3]. The proposed DSAMC structure consists of two different metallic frequency selective surface (FSS) screens printed on either side of a thin dielectric as depicted in Fig. 5.9. It was shown in [5, 7] that by independently optimizing the geometries of the top and bottom FSS screens, it is possible to design a DSAMC separator that achieves an AMC response from top surface at design frequency f_1 and an AMC response from bottom surface at another frequency f_2. Such design flexibility allows two antennas operating at different frequencies to be placed very close together when utilizing a DSAMC separator in between.

To illustrate the versatility of the design technique, a DSAMC separator is optimized via the PSO algorithm. The targeted operation frequencies are chosen to be at Wi-Fi bands, where $f_1 = 2.4$ GHz and $f_2 = 5.2$ GHz. The FSS geometries and the dielectric are discretized and evaluated numerically by a full-wave finite-element boundary integral (FEBI) numerical solver [27] at the two target frequencies. The square unit cell dimensions are set to be 1.345 cm, with a dielectric thickness of 0.254 cm. For the substrate, commercially available Rogers RT/Duroid® 6010 [1] with a relative permittivity of $\varepsilon_r = 10.2 - j0.0253$ is used. To simplify the unit cell geometry parametrization, four-fold symmetry is applied, so that each of the two 16×16 unit cells are defined by 8 parameters, where each parameter represents a row with eight binary digits. The two geometries are mirrored across the horizontal and vertical axes as illustrated in Fig. 5.9. The design is optimized for a TE polarized, normally incident wave, and the *cost* function is given by

$$Cost = \left|1 - R_{\text{Top}}^{2.4\ \text{GHz}}\right| + \left|1 - R_{\text{Bottom}}^{5.2\ \text{GHz}}\right| \quad (5.25)$$

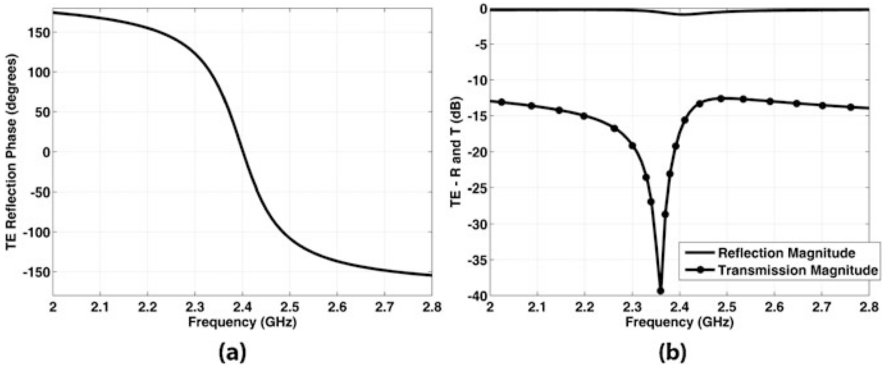


Fig. 5.10 (**a**) TE reflection phase from the top surface, (**b**) reflection and transmission magnitudes

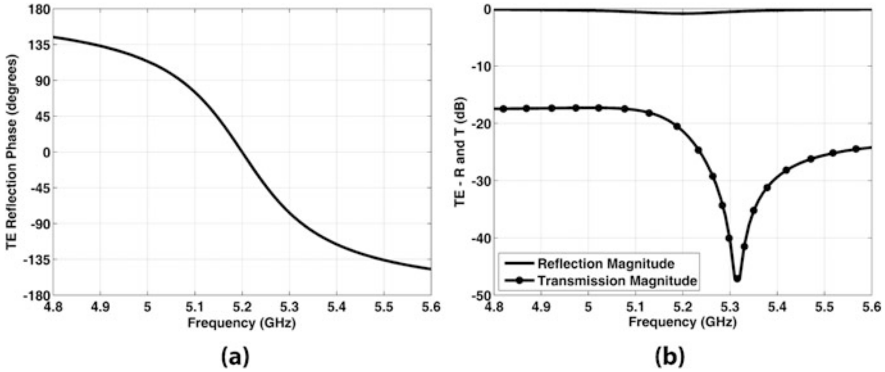


Fig. 5.11 (**a**) TE reflection phase from the bottom surface, (**b**) reflection and transmission magnitudes

where $R_{\text{Top}}^{2.4\ \text{GHz}}$ and $R_{\text{Bottom}}^{5.2\ \text{GHz}}$ are the reflection coefficients from the top and bottom surfaces, respectively, at the desired AMC frequencies. A population of 30 particles was evolved over a maximum of 400 iterations to arrive at the optimized FSS geometries illustrated in Fig. 5.9. A fine frequency sweep of the doubly-periodic DSAMC structure yields the reflection phase and magnitudes for the top and bottom surfaces shown in Figs. 5.10 and 5.11, respectively. These curves demonstrate that the reflection coefficient approaches unity at 2.4 GHz for the top surface and at 5.2 GHz for the bottom surface, illustrating that the PSO was able to effectively design the metasurface to achieve the desired performance goals.

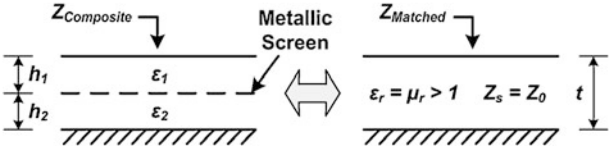


Fig. 5.12 (*Left*) Cross-sectional view of an engineered PEC backed composite metamaterial structure and (*Right*) its equivalent PEC backed ideal homogeneous matched ($\mu_r = \varepsilon_r > 1$) magneto-dielectric material

5.3.3 Matched Magneto-Dielectric RF Absorbers

In this section, the theory will be developed for synthesizing matched-impedance magneto-dielectric metamaterials (MIMDM), which can be utilized effectively as thin electromagnetic absorbing surfaces [6]. Using this theory, the effectiveness of CMA-ES is demonstrated for designing thin RF absorbers with a polarization independent response and a wide field of view.

High permittivity dielectric loading of electromagnetic devices has been shown to have many advantages, including device miniaturization [2]. Similar effects can also be achieved by utilizing magnetic loading [52] or using low-loss magneto-dielectric materials with both magnetic ($\mu_r > 1$) and dielectric ($\varepsilon_r > 1$) properties. Such low-loss materials would pave the way to realizing miniaturized and more advanced RF and microwave devices. Unfortunately, at RF and higher frequencies, naturally occurring homogeneous magnetic materials exhibit high losses. Hence, at these frequencies magnetic and magneto-dielectric materials need to be engineered using composite structures to mimic homogeneous materials with low-loss effective constitutive parameters over a desired bandwidth.

At RF and microwave frequencies, we can synthesize composite metasurface structures that emulate a PEC backed slab of homogeneous magneto-dielectric material with matched impedance as illustrated in Fig. 5.12. A propagating plane wave, normally incident upon this composite metamaterial structure, encounters a surface impedance, $Z_{\text{Composite}}$, represented by

$$Z_{\text{Composite}} = R_{\text{Composite}} + jX_{\text{Composite}} \tag{5.26}$$

where $R_{\text{Composite}}$ and $X_{\text{Composite}}$ are the respective resistive and reactive parts of the impedance. In the case of a thin slab of PEC backed homogeneous magneto-dielectric material with thickness t, permeability $\mu_r = \mu_r' - j\mu_r''$ and permittivity $\varepsilon_r = \varepsilon_r' - j\varepsilon_r''$, the surface impedance, Z_{in}, can be written [94] as

$$Z_{\text{in}} = Z_s \tanh(j\omega\sqrt{\mu_0\varepsilon_0} n_s t) \tag{5.27}$$

where

$$Z_s = Z_0\sqrt{\mu_r/\varepsilon_r} \tag{5.28}$$

is the impedance of the PEC backed slab. If the homogeneous magneto-dielectric material is matched to free space (i.e., $\mu_r = \varepsilon_r$) such that $Z_s = Z_0$, then (5.27) simplifies to

$$Z_{\text{Matched}} = Z_0 \tanh(j\beta_0 n_s t). \tag{5.29}$$

The index of refraction, n_s, for a matched magneto-dielectric material can also be written as

$$n_s = \sqrt{\mu_r \varepsilon_r} = \sqrt{\left(\mu_r' - j\mu_r''\right)\left(\varepsilon_r' - j\varepsilon_r''\right)} = \varepsilon_r = \mu_r. \tag{5.30}$$

Equating $Z_{\text{Composite}}$ and Z_{Matched} using (5.27) and (5.29) leads to

$$\tanh(j\beta_0 n_s t) = Z_{\text{Composite}}/Z_0, \tag{5.31}$$

and solving for the index of refraction, n_s, yields

$$n_s = \varepsilon_r = \mu_r = \frac{1}{j\beta_0 t}\tanh^{-1}\left(\frac{R_{\text{Composite}} + jX_{\text{Composite}}}{Z_0}\right). \tag{5.32}$$

This suggests that the real and imaginary parts of the complex permeability (or permittivity) can be expressed in terms of $Z_{\text{Composite}}$ as

$$\varepsilon_r' = \mu_r' = \text{Re}\left\{\frac{1}{j\beta_0 t}\tanh^{-1}\left(\frac{R_{\text{Composite}} + jX_{\text{Composite}}}{Z_0}\right)\right\} \tag{5.33a}$$

and

$$\varepsilon_r'' = \mu_r'' = \text{Im}\left\{\frac{1}{j\beta_0 t}\tanh^{-1}\left(\frac{R_{\text{Composite}} + jX_{\text{Composite}}}{Z_0}\right)\right\}. \tag{5.33b}$$

These equations relate the surface impedance of the composite metamaterial to the permeability (and permittivity) of the homogeneous magneto-dielectric material. By utilizing (5.26), it is also possible to derive independent equations for $R_{\text{Composite}}$ and $X_{\text{Composite}}$ in terms of the desired homogeneous magneto-dielectric material permittivity and permeability, which are given by

$$R_{\text{Composite}} = \frac{Z_0}{2}\frac{\sinh(2\beta_0 t\mu_r'')}{\cosh^2(\beta_0 t\mu_r'') - \sin^2(\beta_0 t\mu_r')} \tag{5.34a}$$

and

$$X_{\text{Composite}} = \frac{Z_0}{2}\frac{\sin(2\beta_0 t\mu_r')}{\cosh^2(\beta_0 t\mu_r'') - \sin^2(\beta_0 t\mu_r')} \tag{5.34b}$$

where $\mu_r' = \varepsilon_r'$ and $\mu_r'' = \varepsilon_r''$.

Equations (5.26)–(5.34a), (5.34b) can be utilized to synthesize low-loss matched impedance magneto-dielectric metamaterials (MIMDM). However, these thin composite structures can also be employed in electromagnetic absorber applications, if they are designed to have high loss. In the case of high loss MIMDM, the

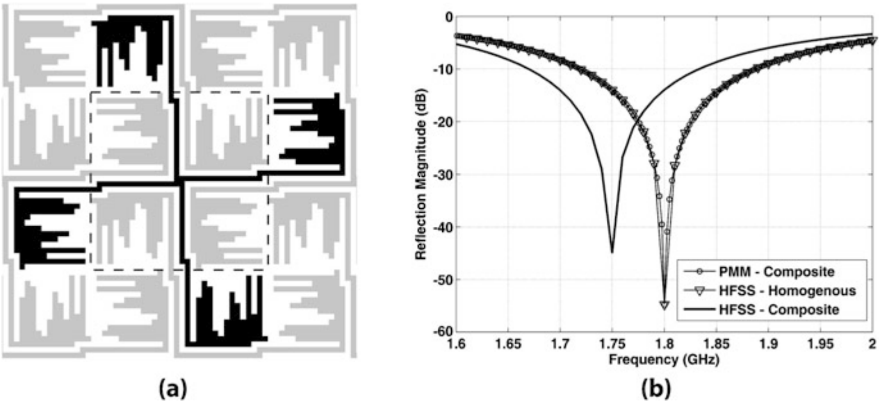


Fig. 5.13 (**a**) Metallic unit cell geometry for the optimized magneto-dielectric metamaterial absorber. The *black* and *gray shaded regions* represent PEC pixels with two tones used to distinguish the interwoven unit cell geometry, which spans multiple unit cells. *Dashed lines* represent the boundaries of a single unit cell. (**b**) Simulated reflection magnitude comparing the composite metamaterial structure with a slab of homogeneous effective material

above equations need to be modified as follows. To achieve zero reflection at the free space–absorber interface, the surface impedance of the composite structure is matched to free space ($Z_{\text{Composite}} = R_{\text{Composite}} + jX_{\text{Composite}} = Z_0$). Using (5.33a), (5.33b), the surface impedance of the composite structure can be matched to the impedance of free space by

$$\begin{aligned} Z_{\text{Composite}} &= R_{\text{Composite}} + jX_{\text{Composite}} \\ &= Z_0 \to R_{\text{Composite}} = Z_0 \quad \text{and} \quad X_{\text{Composite}} = 0. \end{aligned} \tag{5.35}$$

Based on this requirement, the desired values for μ_r' and μ_r'' and ε_r' and ε_r'' can be derived from (5.33a), (5.33b) such that

$$X_{\text{Composite}} = 0 \to \mu_r' = n\lambda_0/4d \quad \text{where } n = 1, 2, 3, \ldots, \tag{5.36}$$

and

$$R_{\text{Composite}} = Z_0 \to \mu_r'' = m\lambda_0/2d \quad \text{where } m = 1, 2, 3, \ldots. \tag{5.37}$$

The CMA-ES is employed here to optimize composite metamaterial structures utilizing the above equations to synthesize thin absorbers with an excellent field of view at a target frequency of 1.8 GHz. These equations ensure a matched impedance metamaterial with zero reflection at the interface of the absorber with free space as well as high electric and magnetic losses. The geometry of the metallic screen in the composite structure can be seen in Fig. 5.13(a), where CMA-ES optimizes only one quarter of the entire unit cell, and then rotational symmetry is applied to complete the structure. CMA-ES adjusts the lengths of eleven rows of pixel strips, which constitute the first 11 dimensions of the optimiza-

Table 5.4 Optimized design parameters for the magneto-dielectric metamaterial absorber targeting 1.8 GHz

Unit cell dimensions (cm × cm)	0.50 × 0.50
Superstrate permittivity, ε_1	$1.22 - j1.0304$
Substrate permittivity, ε_2	$6.99 - j0.0029$
Superstrate thickness, h_1	0.425 cm
Substrate thickness, h_2	0.494 cm

tion problem. CMA-ES also optimizes the unit cell dimension, the thicknesses of the bottom and top dielectric layers as well as the real and imaginary part of the dielectric materials, resulting in a total of $N = 18$ dimensions for optimization.

CMA-ES is linked with a full-wave Periodic Method of Moments (PMM) code for fast numerical analysis of each candidate composite structure. Since PMM is a periodic solver, the simulation must only be carried out on a single unit cell, which makes it extremely efficient. The PMM code employs layered media Green's functions, periodic boundary conditions based on Floquet's theorem, and rooftop basis functions [98]. Once each candidate unit cell geometry is generated by CMA-ES, it is evaluated by the PMM, and the corresponding scattering coefficients are returned for the *cost* function computations. The following *cost* function is utilized to achieve the desired metamaterial properties:

$$Cost = (R_{\text{Composite}} - R_{\text{Matched}})^2 + (X_{\text{Composite}} - X_{\text{Matched}})^2 \tag{5.38}$$

where $R_{\text{Composite}}$ and $X_{\text{Composite}}$ represent the surface resistance and surface reactance values of the composite structure and $R_{\text{Matched}} = Z_0$ and $X_{\text{Matched}} = 0$ represent the real and imaginary parts of the surface impedance associated with the desired homogeneous MIMDM. Once an optimized design was reached using PMM, full-field simulations of the structure were done using HFSS to confirm the results.

CMA-ES was allowed to run for a maximum of 200 iterations utilizing a population size of 20 members. The optimized geometry of the final design is shown in Fig. 5.13(a). The optimized design parameters for the MIMDM absorber are also provided in Table 5.4. A plot of the reflection magnitude at normal incidence versus frequency is shown in Fig. 5.13(b), where attenuation of more than 50 dB at the design frequency of 1.8 GHz is attained. The angular dependency of the reflection for this design is also computed and illustrated in Fig. 5.14. The ultra-small subwavelength unit cell ($0.03\lambda_0$) of the metamaterial absorber provides an extraordinarily stable angular response, with strong absorption for incident angles up to 50° from normal. Figure 5.15 illustrates that the surface impedance $Z_{\text{Composite}}$ approaches a matched value at the designed absorption frequency.

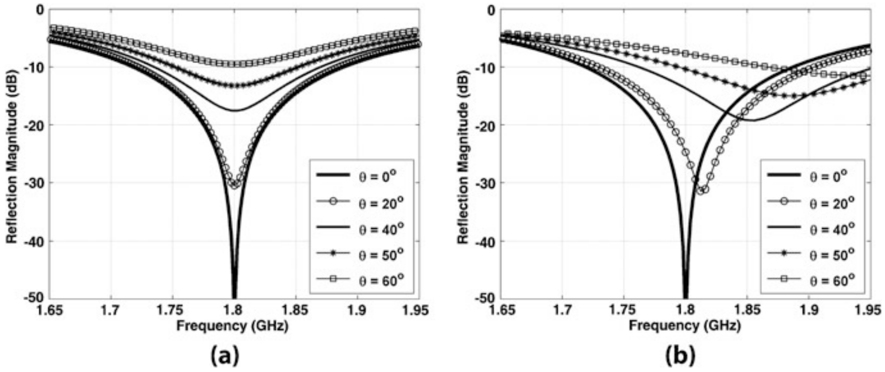


Fig. 5.14 Angular dependency of the (**a**) TE and (**b**) TM reflection magnitude, where a good absorber response is observed up to $\theta = 50°$ from normal

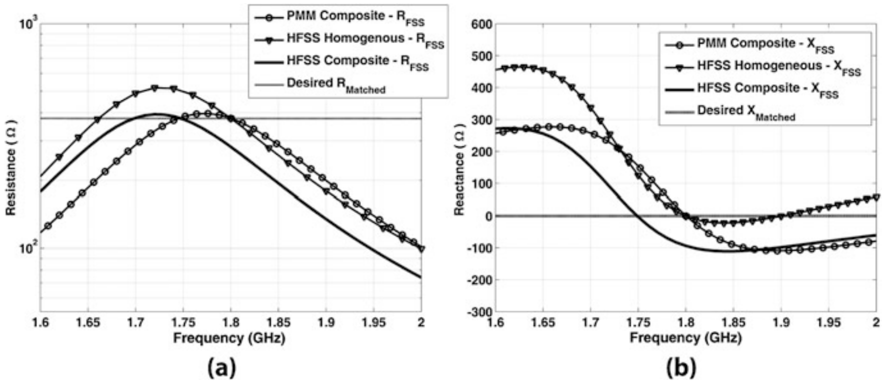


Fig. 5.15 Real (**a**) and imaginary (**b**) parts of the surface impedance ($Z_{\text{Composite}}$) of the composite metamaterial absorber design, computed using PMM and HFSS, along with the surface impedance of a homogeneous slab with the dispersive constitutive parameter values assigned to it. The *dashed lines* represent the surface resistance of an ideal non-dispersive magneto-dielectric material with the same thickness

5.4 Homogenized Metamaterial Optimization Examples

In addition to their use as artificial surfaces, metamaterials can be homogenized using an inversion algorithm to obtain the effective parameters (refractive index n and impedance Z or permittivity ε and permeability μ) of an equivalent homogeneous slab of material. In the following sections, two homogenization techniques will be described. In Sect. 5.4.1, the standard inversion technique for an isotropic planar slab is described, and in Sect. 5.4.2, a technique for inverting anisotropic effective parameters using more than one incidence angle is described. These two techniques are then coupled with a GA in order to synthesize a low-loss multilayer negative index metamaterial (NIM) in Sect. 5.4.3 and a zero index metamaterial (ZIM) with a wide field of view (FOV) in Sect. 5.4.4. Finally, a dispersion engineering approach is combined with a GA in order to synthesize broadband metamaterials for practical applications. Using dispersion engineering a broadband filter with suppressed group delay in the pass band is introduced in Sect. 5.4.5.

5.4.1 Homogenization Technique for an Isotropic Planar Slab

Homogenization techniques for metamaterials are used to calculate the effective parameters (i.e., permittivity and permeability) based on the simulated or measured scattering parameters (S-parameters). The standard inversion procedure that is widely used throughout the metamaterials community [87] is based on an algorithm originally proposed by Nicolson, Ross, and Weir (NRW) [73, 96]. This algorithm assumes that a wave is normally incident upon a slab of an isotropic medium. The NRW procedure begins by calculating the S-parameters (S_{11} and S_{21}) from the metamaterial, which are then manipulated as follows:

$$V_1 = S_{21} + S_{11}, \tag{5.39}$$

$$V_2 = S_{21} - S_{11}. \tag{5.40}$$

V_1 and V_2 are used to calculate

$$X = \frac{1 - V_1 V_2}{V_1 - V_2}, \tag{5.41}$$

$$Y = \frac{1 + V_1 V_2}{V_1 + V_2} \tag{5.42}$$

which are then used to determine the parameters Γ and P from

$$\Gamma = X \pm \sqrt{X^2 - 1}, \tag{5.43}$$

$$P = Y \pm \sqrt{Y^2 - 1}. \tag{5.44}$$

The signs in (5.43) and (5.44) are chosen such that $|\Gamma| \leq 1$ and $|P| \leq 1$. Finally, the normalized impedance and refractive index for the equivalent slab can be obtained from

$$Z = \frac{1 + \Gamma}{1 - \Gamma}, \tag{5.45}$$

$$n = \left(\frac{jc}{\omega d}\right)\left[\ln(P) + j2\pi m\right], \quad m = \pm 0, 1, 2, \ldots, \tag{5.46}$$

where ω is the angular frequency, c is the speed of light, and m indicates the root for n. The solution for n is multi-valued and must be determined using additional information. The constitutive parameters can be calculated from n and Z according to

$$\varepsilon = \frac{n}{ZZ_0}, \tag{5.47}$$

$$\mu = nZZ_0 \tag{5.48}$$

where Z_0 is the intrinsic impedance of free space.

The root for n is determined by enforcing n' to be continuous across frequency and by tracing n' from a known point. For the negative index metamaterial (NIM) example described in Sect. 5.4.3, the root for n is determined in the low frequency limit, where the metamaterial is not magnetic and the permeability is expected to approach unity:

$$\lim_{f \to 0} \mu = 1 - 0j. \tag{5.49}$$

Thus, the S-parameters are collected from low frequencies up to the frequency range of interest, choosing m such that n' remains continuous.

5.4.2 Anisotropic Inversion Technique

As discussed in the previous section, most of the scattering parameter retrieval methods that have been applied in the literature assume the metamaterial has isotropic effective parameters (e.g., permittivity and permeability values the same in every direction) [14, 69, 81, 87]. However, it has been shown that many metamaterials have anisotropic properties that are dependent on the orientation and arrangement of their unit cell structures, such as split-ring resonators (SRRs) and wire dipoles. In view of this, inversion algorithms that are able to characterize the anisotropy of metamaterials are highly beneficial in designing more practical devices. When applied to retrieve the anisotropic effective material parameter tensors, conventional homogenization techniques require that the scattering parameters in three orthogonal directions be collected, which is difficult to achieve in experiment, especially for measurements performed in the infrared or optical wavelength regimes. Moreover, the angular dependent response of anisotropic metamaterials to obliquely incident waves, an important characteristic of metamaterials, has only been considered in a few references [24, 72]. In [72], the effective wave parameters for metamaterials were retrieved as functions of incidence angle. However, because a conventional isotropic material model (i.e., isotropic permittivity and permeability) was used, the inversion procedure could not capture the full anisotropic tensor quantities. Retrieval techniques have also been introduced to obtain bi-isotropic [61] and bi-anisotropic [15, 64] effective medium parameters. In [15], the authors invert the scattering parameters from all orthogonal directions to retrieve anisotropic permittivity and permeability tensors as well as bi-anisotropy in one plane, whereas in [64]

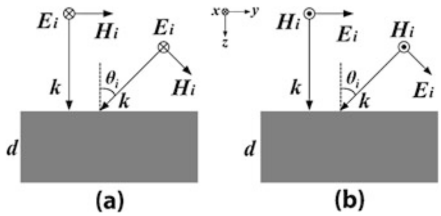


Fig. 5.16 Schematics of a homogeneous anisotropic slab placed in free space and illuminated by (**a**) TE polarized and (**b**) TM polarized normally and obliquely incident plane waves

the scattering parameters from only a single direction are used to retrieve a subset of the bi-anisotropic terms.

In this section, a methodology is presented for retrieving the anisotropic effective permittivity and permeability of a metamaterial slab using a combination of transmission and reflection coefficients calculated or measured at several angles of incidence with respect to only one face of the metamaterial slab. The analytical retrieval expressions used to determine the constitutive parameters of a homogeneous anisotropic slab are first described. This method is then applied to analyze a composite SRR-wire array, and the physical relevance of the retrieved parameters will be discussed.

A periodic metamaterial can be approximated as a homogeneous medium under the long wavelength condition. Let us first consider the forward problem of calculating the scattering parameters based on a simplified model of a homogeneous anisotropic material slab, which has diagonal constitutive permittivity and permeability tensors given by

$$\bar{\bar{\varepsilon}} = \varepsilon_0 \bar{\bar{\varepsilon}}_r = \varepsilon_0 \,\mathrm{diag}[\varepsilon_{xx}, \varepsilon_{yy}, \varepsilon_{zz}], \tag{5.50a}$$

$$\bar{\bar{\mu}} = \mu_0 \bar{\bar{\mu}}_r = \mu_0 \,\mathrm{diag}[\mu_{xx}, \mu_{yy}, \mu_{zz}] \tag{5.50b}$$

where ε_0 and μ_0 are the permittivity and permeability of free space, respectively [55]. In this model, the harmonic time dependence is assumed to be $e^{-j\omega t}$. Figure 5.16 shows schematics of a homogeneous anisotropic slab with thickness d illuminated by a plane wave at an angle θ_i with respect to the free-space slab interface normal $\hat{z}$. Without loss of generality in the case of a homogeneous slab, it is assumed that the incident plane wave vectors are in the y–z plane for both transverse electric (TE) and transverse magnetic (TM) polarized waves. The TE waves satisfy the conditions $\mathbf{k} \cdot \mathbf{E} = 0$ and $E_z = 0$, whereas the TM waves satisfy $\mathbf{k} \cdot \mathbf{H} = 0$ and $H_z = 0$. Notice that the six tensor parameters can be divided into two groups: ε_{xx}, μ_{yy}, and μ_{zz}, which are active when the slab is illuminated by TE waves, and ε_{yy}, μ_{xx}, and ε_{zz}, which are active when the slab is illuminated by TM waves. Only the expressions for TE polarization will be shown here, since the TM polarization is a straightforward dual case of the TE solution. The dispersion relation inside the material for TE polarization is

$$\frac{\beta_y^2}{\mu_{zz}} + \frac{\beta_{z\mathrm{TE}}^2}{\mu_{yy}} = k_0^2 \varepsilon_{xx} \tag{5.51}$$

where the y-component of the wave number satisfies $\beta_y = k_y = k_0 \sin\theta_i$, and k_0 is the free space wave number [46]. By assigning boundary conditions on both interfaces of the slab, the scattering parameters of the slab can be calculated for an illuminating plane wave with an arbitrary incident angle. The expressions for the scattering parameters corresponding to TE waves can be written as

$$S_{11} = \frac{\Gamma_{\mathrm{TE}}(1 - e^{j2\beta_{z\mathrm{TE}}d})}{1 - \Gamma_{\mathrm{TE}}^2 e^{j2\beta_{z\mathrm{TE}}d}}, \tag{5.52a}$$

$$S_{21} = \frac{(1 - \Gamma_{\mathrm{TE}}^2) e^{j2\beta_{z\mathrm{TE}}d}}{1 - \Gamma_{\mathrm{TE}}^2 e^{j2\beta_{z\mathrm{TE}}d}} \tag{5.52b}$$

where

$$\Gamma_{\mathrm{TE}} = \frac{Z_{\mathrm{TE}} - 1}{Z_{\mathrm{TE}} + 1} \tag{5.53}$$

is the reflection coefficient from the top interface. The normalized wave impedance for TE waves is given by

$$Z_{\mathrm{TE}} = \frac{k_z \mu_{yy}}{\beta_{z\mathrm{TE}}} \tag{5.54}$$

where $k_z = k_0 \cos\theta_i$ is the z-component of the free space wave number. Because the structures considered in this chapter possess vertical symmetry, the scattering parameters will be reciprocal, so that $S_{11} = S_{22}$ and $S_{21} = S_{12}$.

Now, the inverse problem will be solved. First, let us consider the scattering parameters for two TE incident waves with different angles of incidence, θ_{i1} and θ_{i2}, which provide four equations ($S_{\mathrm{TE}11-1}$, $S_{\mathrm{TE}21-1}$, $S_{\mathrm{TE}11-2}$, $S_{\mathrm{TE}21-2}$) given by (5.52a) and (5.52b). These four equations can be used to determine the three unknowns (ε_{xx}, μ_{yy}, and μ_{zz}). During the solution process, the z-components of the refractive indices of the slab ($n_{z\mathrm{TE}-1}$, $n_{z\mathrm{TE}-2}$) and the wave impedances ($Z_{\mathrm{TE}-1}$, $Z_{\mathrm{TE}-2}$) for both incidence angles are inverted, leading to the following two equations:

$$\cos(n_{z\mathrm{TE}-l} k_0 d) = \frac{1 - S_{\mathrm{TE}11-l}^2 + S_{\mathrm{TE}21-l}^2}{2S_{\mathrm{TE}21-l}}, \tag{5.55}$$

$$Z_{\mathrm{TE}-l} = \pm\sqrt{\frac{(1 + S_{\mathrm{TE}11-l}^2) + S_{\mathrm{TE}21-l}^2}{(1 - S_{\mathrm{TE}11-l}^2) + S_{\mathrm{TE}21-l}^2}}, \quad l = 1, 2. \tag{5.56}$$

Proper care should be exercised to select the correct branch for the real part of the z-components of the refractive indices. Similar to approaches that have been employed for isotropic inversion procedures, the imaginary parts of both $n_{z\mathrm{TE}-1}$ and $n_{z\mathrm{TE}-2}$ must obey the conditions $n''_{z\mathrm{TE}-1} \le 0$, $n''_{z\mathrm{TE}-2} \le 0$. Likewise, for passive materials, the real parts of $Z_{\mathrm{TE}-1}$ and $Z_{\mathrm{TE}-2}$ must satisfy $Z'_{\mathrm{TE}-1} \ge 0$ and $Z'_{\mathrm{TE}-2} \ge 0$. Then, making use of the four inverted parameters and the dispersion relation

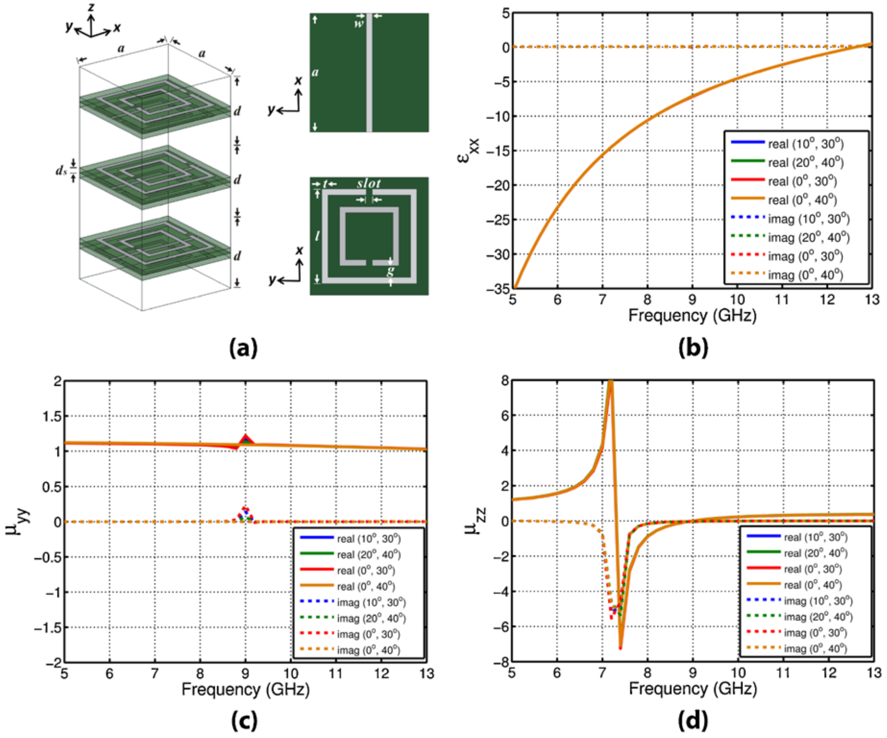


Fig. 5.17 (**a**) Unit cells ($a = 2.6$ mm, $d = 1.82$ mm) of a composite SRR-wire array. *Left*: 3D isometric view of the three layer unit cell. *Right*: top views of the wire and SRR structures with dimensions given by $g = 0.39$ mm, $l = 2.08$ mm, slot $= 0.13$ mm, $t = 0.13$ mm, and $w = 0.13$ mm. The dielectric slabs (FR4: $\varepsilon_r = 4.4$, $\delta_{\tan} = 0.02$) have thickness $d_s = 0.13$ mm. (**b**) Retrieved effective ε_{xx}. (**c**) Retrieved effective μ_{yy}. (**d**) Retrieved effective μ_{zz}

(5.51) as well as the wave impedance (5.54), the three tensor parameters active for the TE case can be retrieved by:

$$\mu_{yy} = n_{z\mathrm{TE}-l} Z_{\mathrm{TE}-l} / \cos\theta_{il}, \quad l = 1, 2, \tag{5.57}$$

$$\varepsilon_{xx} = \frac{n_{z\mathrm{TE}-1} \frac{\cos\theta_{i1}}{Z_{\mathrm{TE}-1}} \sin^2\theta_{i2} - n_{z\mathrm{TE}-2} \frac{\cos\theta_{i2}}{Z_{\mathrm{TE}-2}} \sin^2\theta_{i1}}{\sin^2\theta_{i2} - \sin^2\theta_{i1}}, \tag{5.58}$$

$$\mu_{zz} = \frac{\sin^2\theta_{i2} - \sin^2\theta_{i1}}{n_{z\mathrm{TE}-1} \frac{\cos\theta_{i1}}{Z_{\mathrm{TE}-1}} - n_{z\mathrm{TE}-2} \frac{\cos\theta_{i2}}{Z_{\mathrm{TE}-2}}}. \tag{5.59}$$

The validity of the anisotropic retrieval method will be demonstrated using an interesting type of metamaterial, a composite SRR-wire array. The unit cell geometry for the metamaterial under consideration is illustrated in Fig. 5.17(a). Periodic boundary conditions are assigned to the lateral walls in both the x- and y-directions. A plane wave (contained in the y–z plane) is assumed to be incident from the upper half-space at an angle θ_i ($0° \leq \theta_i \leq 90°$) with respect to the free-space metamaterial interface normal $\hat{z}$. Three layers of unit cells are used in the z-direction in order to take into account the coupling between adjacent unit cells, thus enabling the acquisition of more accurate effective medium parameters. For each layer, the infinite wires are sandwiched by two dielectric slabs of thickness d_s. A pair of broadside coupled SRRs are printed on each side of the dielectric substrate as shown in Fig. 5.17(a) to eliminate the bi-anisotropy associated with conventional arrangements of SRRs [70, 71]. The geometrical parameters of the SRRs and wires are defined in Fig. 5.17(a). Since the unit cells are much smaller than the wavelength of interest, these two metamaterials can be approximated as homogeneous anisotropic materials under the effective medium theory with diagonal permittivity and permeability tensors as described by (5.50b), provided the geometrical axes of the metamaterials coincide with the principal axes of the effective parameter tensors [34].

Figure 5.17(b)–(d) shows the retrieved effective electromagnetic parameters of the composite SRR-wire array under TE polarization. It can be observed that the retrieved results using scattering parameters calculated at different incidence angles (with and without using normal incidence) agree well with each other, and as a result, the curves essentially lie on top of each other. Figure 5.17(b) shows a Drude-type response in ε_{xx} due to the infinite-wire array. The retrieved μ_{yy} has no strong resonances except for a small anti-resonance in μ_{yy} within the same frequency region as the magnetic resonance observed in μ_{zz}. This anti-resonance phenomenon has been widely discussed in the literature and is attributed to the intrinsic periodicity of the metamaterial [57, 58]. The longitudinal magnetic resonance excited by the component of the incident H-field, which is perpendicular to the plane of the SRRs, occurs at around 7.4 GHz. This resonance can be easily characterized by the inversion method described here with the utilization of obliquely incident waves. The region where both ε_{xx} and μ_{zz} are negative forms a negative refractive index band, which can be experienced with in-plane propagating waves. The good agreement of the retrieved effective electromagnetic tensor parameters for the composite SRR-wire array confirms the validity of the retrieval method as well as the particular homogeneous anisotropic model assumed for the metamaterials (i.e., diagonal effective permittivity and permeability tensors). This method thus provides a useful tool for retrieving the effective anisotropic tensor parameters of metamaterials with the angular response taken into account and can be linked with global optimizers to synthesize metamaterials with improved angular stability [47]. In Sect. 5.4.4, this method is used in conjunction with a GA to synthesize a zero index metamaterial with a wide field of view.

5.4.3 Low-Loss, Multilayer Negative Index Metamaterials for the Infrared and RF

There has been a substantial research effort into demonstrating metamaterials with a negative refractive index since 2000, when Pendry first proposed that a negative index metamaterial (NIM) slab with $n = -1$ could form a "perfect" flat focusing lens that overcomes the diffraction limit of conventional optics [75]. Since Pendry's observation, NIMs have been demonstrated from the RF [32, 83, 86] through optical wavelengths [20, 22, 82], and experiments have shown the "super-resolution" that Pendry predicted. However, many of these demonstrations also exhibited high losses due to absorption within the metamaterial and to reflection arising from an impedance mismatch between the metamaterial and the surrounding medium. Furthermore, most optical NIMs are very thin with respect to the wavelength, which is impractical in terms of realizing a flat, near-field focusing lens. Recent experimental efforts have been made to increase the thickness of optical metamaterials [93]. In this section, a genetic algorithm is employed to first synthesize a multilayer NIM for the mid-infrared (mid-IR) regime with minimum intrinsic and impedance mismatch losses [13] and then to synthesize multilayer low-loss NIMs for RF operation at 10 GHz using a look up table of potential constituent materials.

The primary metamaterial losses for a NIM can be quantified using the following two figures of merit (FOM):

$$FOM_n = \left| \frac{n'}{n''} \right|, \tag{5.60}$$

$$FOM_Z = \left| \frac{1}{Z - 1} \right| \tag{5.61}$$

where n is the effective index and Z is the effective impedance normalized to the free space impedance Z_0. FOM_n is the magnitude of the ratio of the real and imaginary parts of the index. Because the intrinsic loss of the metamaterial is captured in n'', FOM_n is a measure of the loss due to absorption. FOM_Z, on the other hand, measures the difference between the effective impedance and the impedance of free space, which gives rise to reflection losses. As FOM_n and FOM_Z become large, the losses are reduced, leading to a highly transmissive NIM.

The first metamaterial structure considered here is composed of a stack of metallic screens separated by dielectric insulators. This freestanding stack is perforated by periodic air holes that are defined by a single unit cell geometry. The constituent materials selected for the metamaterial are Ag and polyimide because of their low-loss properties in the mid-IR wavelength range from 2 to 5 µm. This metamaterial configuration can give rise to negative refraction by way of electric and magnetic resonances in the structure. The negative permittivity comes from the natural Drude dispersion in the Ag metal [76], which is diluted such that the plasma frequency occurs in the mid-IR. On the other hand, the impinging magnetic field can excite loop currents between neighboring metallic layers, giving rise to a Lorentz-shaped resonance in the permeability.

The metamaterial geometry is defined in the GA by pixellizing the unit cell into a grid of 13 × 13 pixels. Eight-fold symmetry is enforced on the geometry, such that one triangular fold is encoded into the chromosome. Each of the seven rows in the triangle is represented by one parameter in the chromosome. Fabrication rules

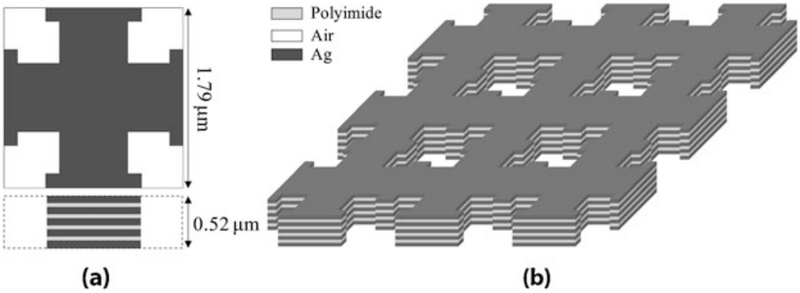


Fig. 5.18 Geometry for an optimized NIM stack with five metal layers. (**a**) Top and cross-section views of the structure. (**b**) 3D isometric view of the metamaterial

were imposed on the pixellized unit cell during optimization such that diagonal pixels were eliminated. Also, because the metamaterial is intended to be free-standing, any islands of patches that were not fully connected were removed from the screen. The unit cell dimension and the inter-layer spacing were also encoded into the chromosome as 8-bit numbers and allowed to vary from 0.8 to 2.0 μm and from 130 to 500 nm, respectively. Finally, the Ag screen thickness was fixed to be 75 nm, and the number of screens was chosen to be five.

During the *cost* evaluation of each population member, the scattering parameters for the geometry are predicted using an efficient full-wave finite-element boundary integral (FEBI) numerical solver [27]. The S-parameters are then fed into the NRW inversion procedure described in Sect. 5.4.1 to retrieve n and Z. Each structure is evaluated at 63 frequency points starting from the low frequency of 10 THz and extending to 95 THz, so that the correct root of n can be selected at 10 THz and then traced up to the mid-IR range. Then, a fine sweep of 62 frequencies from 95 to 105 THz is performed to select the optimal frequency point according to the following *cost* function:

$$Cost = \min_{\text{freqs}}\left[|n - n_{\text{target}}|^2 + |Z - Z_{\text{target}}|^2\right] \tag{5.62}$$

where $n_{\text{target}} = -1 + 0j$ is the target refractive index, $Z_{\text{target}} = 1 + 0j$ is the target normalized impedance, and *freqs* are frequency sweep sample points around 3 μm wavelength in the mid-IR.

The GA operated on a population of 32 members and employed tournament selection with a crossover probability of 0.5 and a mutation rate of 0.02. Elitism was enforced so that the population fitness would always be maintained or increased. Evolving the population over 220 generations resulted in the NIM geometry shown in Fig. 5.18. The inverted index and impedance curves in Fig. 5.19 show optimum values of $n = -0.99 - 0.13j$ and $Z = 1.01 + 0.08j$ at 105 THz, which indicate low absorption loss ($FOM_n = 7.6$) and a good match to free space ($FOM_Z = 12.4$). The effective parameters shown in Fig. 5.20 reveal the expected Drude dispersion for the permittivity and a Lorentz resonance in the permeability, both of which give rise to the negative index band. At the optimum frequency, the S-parameters

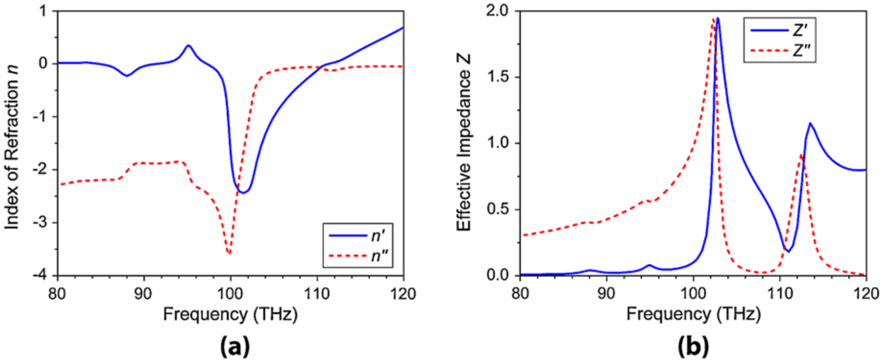


Fig. 5.19 Inverted effective medium parameters for the NIM in Fig. 5.18. (**a**) Refractive index n and (**b**) normalized impedance Z

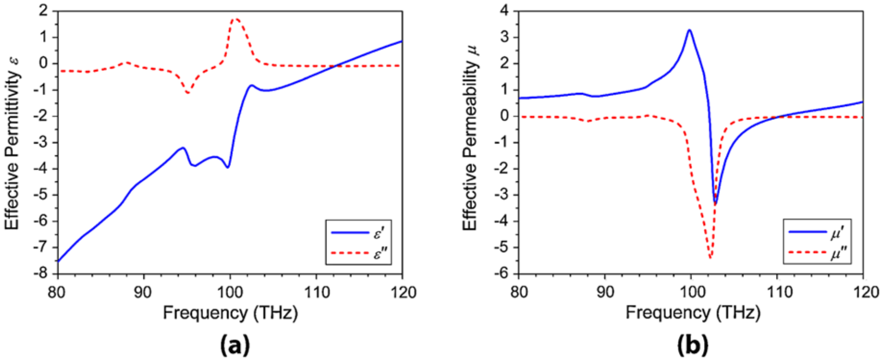


Fig. 5.20 Inverted effective medium parameters for the NIM in Fig. 5.18. (**a**) Effective permittivity ε and (**b**) effective permeability μ

plotted in Fig. 5.21 indicate a high transmission $|T| = -1.3$ dB, with low reflection $|R| = -23.6$ dB and an absorption $|A| = -5.9$ dB. This low-loss performance demonstrates that GA optimization can be used to synthesize practical, multilayer NIMs at optical wavelengths.

At microwave frequencies, multilayer fishnet structures can also be employed to construct NIMs. The planar multilayer NIM structure considered here for RF operation consists of seven cascaded metal fishnet screens sandwiching six dielectric slabs with perforated air holes (see Fig. 5.22). Under TE polarized, normally incident waves, the metal strips along the direction of the electric field (x-axis) produce a Drude-type behavior with a negative permittivity below the effective plasma frequency. Likewise, a negative effective permeability is produced by the metal strips along the magnetic field (y-axis), which form a series of parallel plate magnetic resonators. Two design configurations were optimized utilizing this structure. The first design is restricted to uniformly thick dielectric layers with Rogers Ultralam 2000 chosen as the dielectric material. The second design was optimized with a

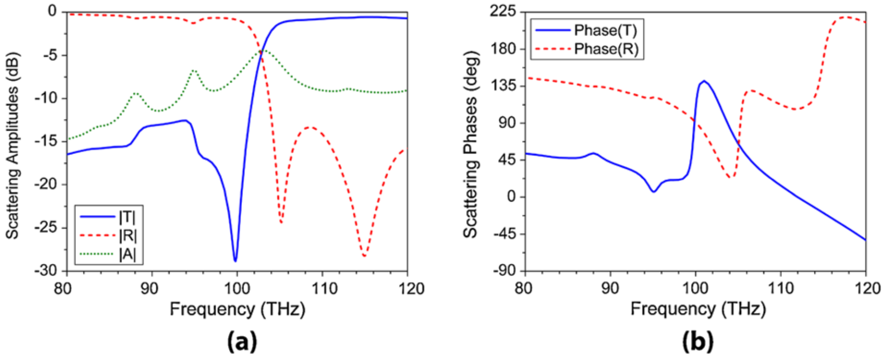


Fig. 5.21 Scattering parameters for the optimized NIM in Fig. 5.18. (**a**) Transmission T, reflection R, and absorption A magnitudes for a normally incident wave and (**b**) transmission and reflection phases

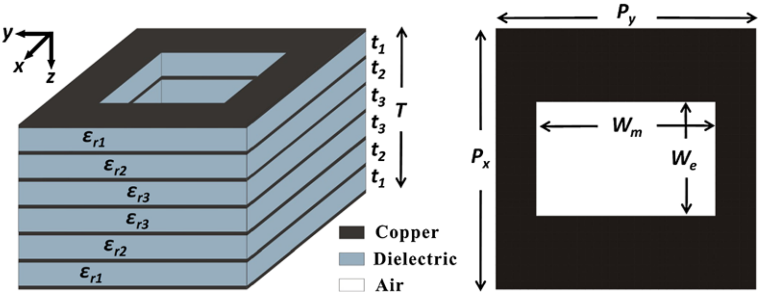


Fig. 5.22 (*left*) Unit cell geometry of NIM structure. (*right*) Top view of the metal screen

constituent material parameter lookup table containing the permittivity, loss tangent and available thicknesses (or the sum of available thickness values) of several Rogers dielectric materials [1], as listed in Table 5.5. Each dielectric layer was limited to have a permittivity and thickness corresponding to those available in the material table as a fabrication constraint. The metal layers in both designs consist of 0.035 mm thick copper. The metamaterial design parameters, including unit cell size, widths of the metal strips in both x- and y-directions, thicknesses of the dielectric slabs, and dielectric material types for the second design were encoded into the binary chromosome and simultaneously optimized using a GA. The *cost* function given by (5.62) was used to evaluate the design performance during the GA optimization with $n_{\text{target}} = -1 + 0j$, $Z_{\text{target}} = 1 + 0j$, and a design frequency chosen to be in the X-band at 10 GHz.

The GA evolved a population of 32 over 500 generations to obtain the optimized design result for both cases. The design dimensions for the two optimized structures are listed in Table 5.6, and the dielectric properties for both designs are listed in Table 5.7. The impedances and the inverted effective refractive indexes for both designs are shown in Figs. 5.23 and 5.24. As can be seen in Fig. 5.23, the pre-

Table 5.5 Material look-up table with various rogers materials and thickness values

Name	ε_r	$\delta_{\tan}$	t1 (mm)	t2 (mm)	t3 (mm)	t4 (mm)	t5 (mm)	t6 (mm)
RO3003	3	0.0013	0.127	0.254	0.508	0.762	1.524	–
RO3035	3.5	0.0017	0.127	0.254	0.508	0.762	1.524	–
RO3203	3.02	0.0016	0.254	0.508	0.762	1.524	–	–
RO4003C	3.55	0.0021	0.203	0.305	0.406	0.508	0.813	1.524
RO4350B	3.66	0.0031	0.168	0.254	0.338	0.422	0.508	0.762
RT5870	2.33	0.0012	0.127	0.254	0.381	0.508	0.787	1.575
RT5880	2.2	0.0009	0.127	0.254	0.381	0.508	0.787	1.575
RT6002	2.94	0.0012	0.127	0.254	0.508	0.787	1.575	–
Ultralam 2000	2.5	0.0019	0.101	0.256	0.373	0.482	0.762	1.524

Table 5.6 Structure dimensions optimized by GA for both RF designs (dimensions in mm)

	p_x	p_y	w_e	w_m	T	t_1	t_2	t_3
Design 1	18.55	17.6	8.12	13.2	8.58	1.43	1.43	1.43
Design 2	18.1	15.7	6.79	12.76	10.26	1.702	1.726	1.702

Table 5.7 Dielectric properties for both RF designs

	ε_{r1}	ε_{r2}	ε_{r3}
Design 1	$2.5 - 0.00475j$	$2.5 - 0.00475j$	$2.5 - 0.00475j$
Design 2	$2.33 - 0.002796j$	$2.5 - 0.00475j$	$2.33 - 0.002796j$

dicted effective metamaterial parameters for the first design at the target frequency of 10 GHz were $n = -0.999 - 0.01j$ and $Z = 1.009 + 0.017j$, showing a negative index with a low transmission loss ($|T| = -0.156$ dB) and a good impedance match to free space. The inverted parameters at 10 GHz for the second design, seen in Fig. 5.24, were $n = -1.007 - 0.009j$ and $Z = 1.014 + 0.015j$ with a transmission loss of only -0.107 dB. The total thickness of the second design is greater than $\lambda/3$ at the target frequency. Both of the NIM designs possess very high figures of merit as defined previously in (5.60) and (5.61). The FOMs for the first design are $FOM_n = 99.9$ and $FOM_Z = 52.0$, while the FOMs for the second design are $FOM_n = 111.9$ and $FOM_Z = 48.7$, indicating that both designs have extremely low intrinsic loss and a very good impedance match to free space at the target frequency. Both of these NIM designs are able to achieve much better FOMs than the infrared design in Fig. 5.18 due to the lower material losses in the microwave. Compared with other conventional volumetric NIMs for the RF described in the literature, the FOMs for n achieved here using GA optimization are significantly higher, and the FOMs for Z are comparable or better [78, 79].

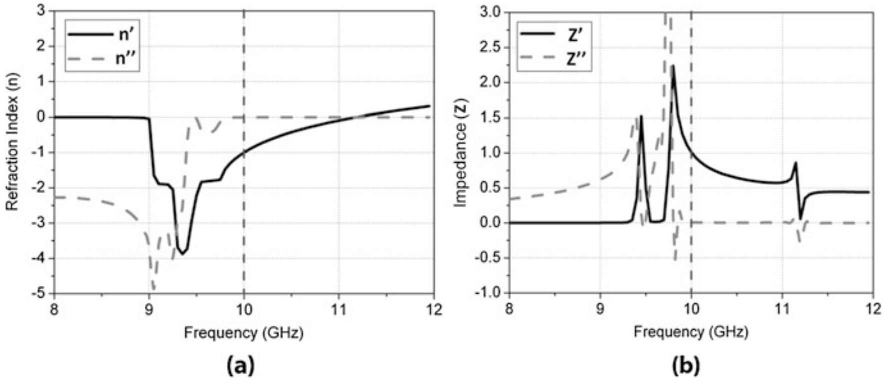


Fig. 5.23 Inverted effective medium parameters for the first RF NIM design. (**a**) Refractive index n and (**b**) normalized impedance Z

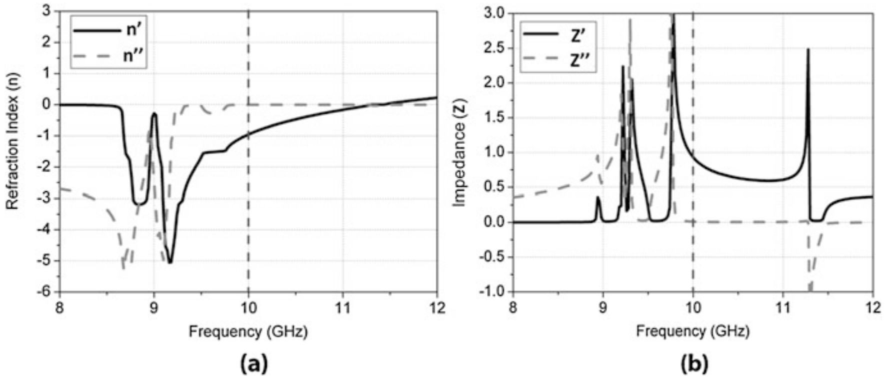


Fig. 5.24 Inverted effective medium parameters for the second RF NIM design. (**a**) Refractive index n and (**b**) normalized impedance Z

5.4.4 Wide-Angle Zero Index Metamaterials for the IR

Compared with negative index metamaterials (NIM), zero index metamaterials (ZIMs) have received less attention in recent years, but hold promise for a wide variety of possible practical applications [12, 28, 100]. A zero index of refraction condition can be achieved under three different cases of permittivity and permeability: permittivity approaching zero, permeability approaching zero, and permittivity and permeability simultaneously approaching zero. In the first case, the permittivity approaching zero results in a large value for the normalized impedance and a corresponding reflection coefficient approaching positive unity, meaning that the reflected wave is in-phase with the incident wave. Such an epsilon-near-zero (ENZ) material can be used as either an artificial magnetic conducting (AMC) surface [53] or for subwavelength electromagnetic energy tunnels [85]. In the second case, when the permeability approaches zero, the material acts like a perfect electric conductor (PEC), with the reflection coefficient approaching negative unity. Hence, in both of these first two cases the ZIM acts as either a perfect magnetic mirror (in-phase reflection) or a perfect electric mirror (out-of-phase reflection). The final and perhaps most interesting case is when the permittivity and permeability simultaneously approach zero at the same rate, resulting in a ZIM that is impedance matched to free space. This type of "matched" ZIM is an essential component in many transformation optics devices including the well-known electromagnetic cloak [80]. Another important property of ZIMs is their ability to act as effective collimators that convert cylindrical or spherical waves emanating from a source embedded in the metamaterial to plane waves at the interface between the metamaterial and free space. ZIMs can thus be utilized as flat lenses to achieve highly directive far-field radiation from embedded antennas, as extremely convergent nanolenses and in other imaging applications [100]. ZIMs have been experimentally demonstrated at both microwave frequencies and optical wavelengths [59, 60]. While the structures chosen for these ZIM experiments are inherently anisotropic, most reported work only considers the metamaterial response to normally incident plane waves. Here, the angular response of ZIMs is considered in order to gain an improved picture of the metamaterial electromagnetic properties.

In this section, the generalized anisotropic retrieval method previously described in Sect. 5.4.2 is coupled with a robust genetic algorithm (GA) optimizer in order to synthesize an infrared (IR) ZIM with a wide field-of-view (FOV). In many applications (e.g., flat lenses), metamaterials must be capable of properly responding to illumination from obliquely incident waves in addition to those that are normally incident. Hence, in such cases, the ability to customize the metamaterial properties via optimization is of significant practical importance. Furthermore, in order to demonstrate the superiority of the anisotropic retrieval method, the ZIM optimized by a GA coupled with the anisotropic retrieval technique is compared with a second GA optimized design based on using the conventional isotropic retrieval method.

The planar metamaterial structure targeting infrared wavelengths to be optimized by the GA consists of two stacked metallic screens sandwiching a dielectric layer. To reduce absorption loss, the stack is perforated with air holes in both the metal and dielectric layers in a periodic pattern. Silver (Ag) and polyimide were chosen as the screen and dielectric materials, respectively, because they have low losses in the mid-IR range from 2 to 5 µm. The structure is defined by a pixilated unit cell geometry and is constrained to possess an eight-fold symmetry, so that the metamaterial response is polarization insensitive to normally incident waves. The eight-fold symmetry also minimizes ϕ dependence at oblique incidence, where ϕ represents the azimuth angle between the x-axis and the projection of the wave vector of obliquely incident waves onto the x–y plane. The geometrical parameters of the metamaterial structure described by the chromosome include the unit cell dimension, the thickness of the polyimide layer and the binary pixilated pattern where "1" represents metal/dielectric and "0" represents an air hole. When generating each structure, fabrication constraints are also applied to the pixilated geometry, eliminating diagonal connections that are difficult to fabricate. Measured permittivities of both Ag and

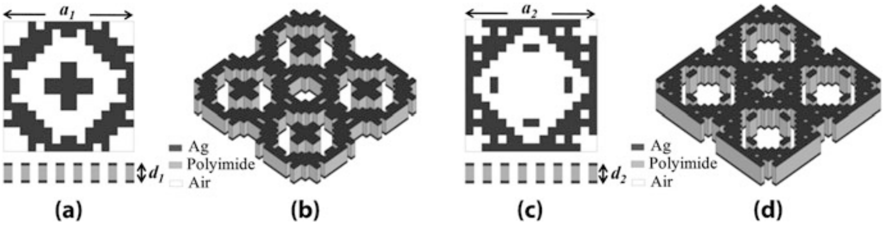


Fig. 5.25 16 × 16 pixel geometry for a metamaterial stack with two Ag screens. (**a**) Optimized geometry for the first design, with a unit cell size $a_1 = 1.42$ μm, total thickness $d_1 = 476$ nm, and a Ag screen thickness of 75 nm. (**b**) 3D isometric view of the first ZIM design. (**c**) Optimized geometry for the second design, with a unit cell size $a_2 = 1.58$ μm, total thickness $d_2 = 735$ nm, and a Ag screen thickness of 75 nm. (**d**) 3D isometric view of the second ZIM design

polyimide are incorporated in the simulation. The calculated scattering parameters are then inverted to produce effective medium parameters using either the conventional isotropic retrieval method described in Sect. 5.4.1 or the anisotropic retrieval method described in Sect. 5.4.2. The *cost* function employed in the GA is given by [64], where $n_{z\text{TE\_}\theta i} = 0$ (but restricted to have a positive real part) is the desired $\hat{z}$-component of the normalized refractive index and $Z_{\text{tar}} = 1$ is the desired normalized wave impedance for the metamaterial:

$$Cost = \sum_{\theta_i} \left[|n_{z\text{TE\_}\theta i} - n_{z\text{\_tar}}| + |Z_{\text{TE\_}\theta i} - Z_{\text{tar}}| \right]. \tag{5.63}$$

The target frequency for both designs was chosen to be 100 THz (3 μm). For the first design, four sample angles of incidence ($\theta_i = 0°, 10°, 20°, 30°$) were considered, whereas for the second design, only the response of the metamaterial to normally incident waves, i.e., $\theta_i = 0°$, was taken into account. This was done to demonstrate the improved wide-angle metamaterial performance achieved when including these simulations in the optimization.

The geometries and dimensions of the two optimized structures are shown in Fig. 5.25. For each design, the GA optimized a population of 64 members over 50 generations. While the anisotropic inversion was used during only the first optimization, both optimized designs were analyzed using the anisotropic inversion in order to study their angular responses. The retrieved $\hat{z}$-component of the effective refractive index and wave impedance at 100 THz as a function of the incident angle are shown in Figs. 5.26(a) and (b), respectively. It can be observed from Figs. 5.26(a)–(b) that, for the first design, a positive near-zero $\hat{z}$-component of the effective refractive index with low loss and a wave impedance matched to free space are achieved throughout the range of incidence angles from 0° to 30°, thus ensuring good transmission. The imaginary part of the wave impedance has a very small value from 0° to 23° and then increases slowly. Compared to the first design, the second design has a smaller effective refractive index and a better matched effective impedance for normal incidence, but its angular response is inferior. The imaginary part of the $\hat{z}$-component of the effective refractive index increases significantly at large oblique

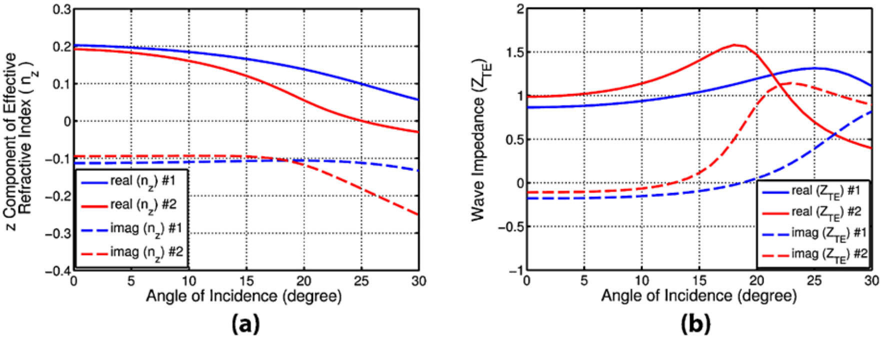


Fig. 5.26 (**a**) Retrieved $\hat{z}$-component of the effective refractive index and (**b**) wave impedance at 100 THz versus angle of incidence

angles, resulting in large absorption loss. Also, due to the drastic changes in both the real and imaginary parts of its wave impedance, the match to free space deteriorates at larger angles.

In order to verify the consistency of the anisotropic inversion for these examples, the three retrieved effective anisotropic constitutive parameters for the first design that are active under TE polarized illumination are shown in Fig. 5.27. All three parameters retrieved using the scattering parameters calculated at different angles of incidence show very good agreement, indicating that the homogenization procedure is accurate.

5.4.5 Dispersion Engineering Broadband Negative–Zero–Positive Index Metamaterials

As described in the previous sections, metamaterials can be optimized for low-loss, custom negative and zero/low indices of refraction. Fundamentally, these effective medium properties arise from sub-wavelength inclusions in a periodic lattice and are frequency dependent. The dispersive behavior in the effective medium properties and the group delay results in signal distortion and can lead to narrow operational bandwidths, which represent major roadblocks to metamaterials being incorporated into practical devices. One approach to overcome the limitations of dispersion is to employ dispersion engineering in order to exploit the frequency-dependent properties of the metamaterial by tailoring them to specific device needs. Dispersion engineering has been applied in the RF for enhancing horn antennas [65]. Here, dispersion engineering is utilized to tailor the dispersive properties of an optical metamaterial to realize a broadband filter for the infrared.

The target device is a high performance optical band pass filter for the mid-infrared with properties as described in Fig. 5.28. Dispersion engineering can first be applied to a theoretical material to realize this optical band pass filter by controlling the effective medium properties, μ and ε, which give rise to the refractive

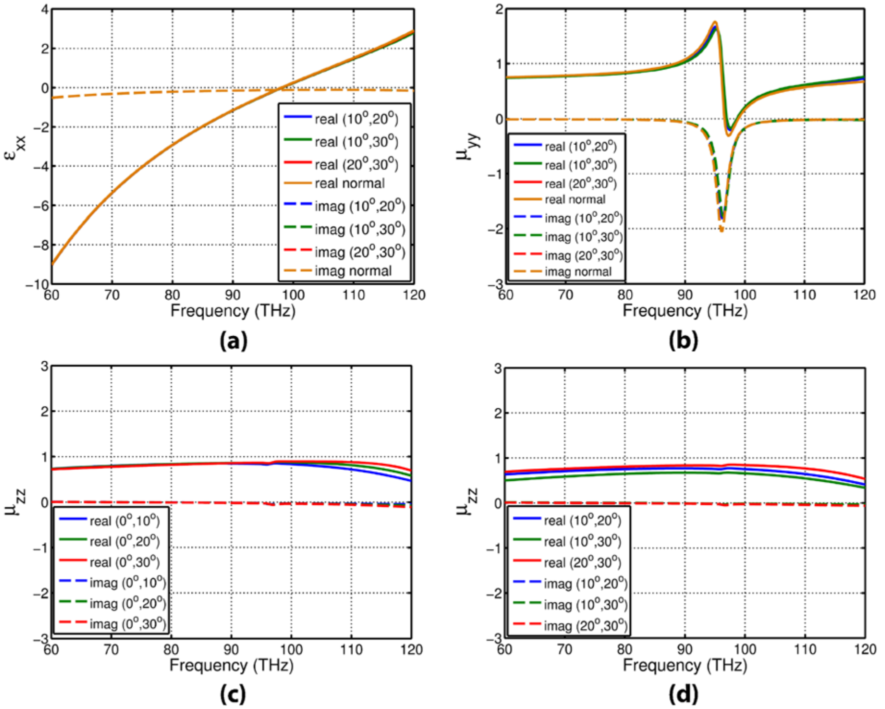
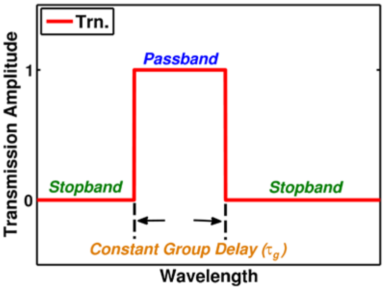


Fig. 5.27 Retrieved effective parameters of the first ZIM design shown in Fig. 5.25(a): (**a**) ε_{xx}, (**b**) μ_{yy}, and (**c**) μ_{zz} using normal and oblique incidence angles; (**d**) μ_{zz} using two oblique incidence angles

Fig. 5.28 Ideal response of a pass band filter with a flat transmission window and a flat group delay within the pass band



index, impedance, and group delay. Figure 5.29 shows the real parts of the desired material parameters that will produce the band pass filter response. In this plot the permittivity possesses a Drude type dispersion, and the permeability exhibits two Lorentzian resonances, one on either side of the plasma wavelength λ_e in the permeability. Given an $e^{j\omega t}$ time dependence, these two models can be expressed

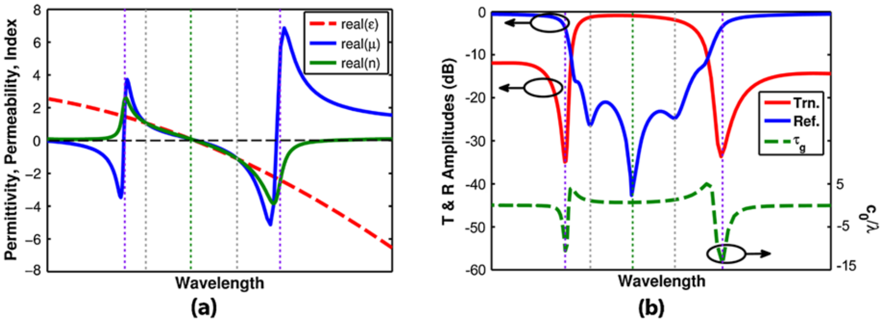


Fig. 5.29 Theoretical metamaterial pass band filter. (**a**) Real parts of the dispersive permittivity, permeability, and refractive index profiles for the theoretical material. (**b**) Scattering magnitudes and group delay for the slab of theoretical material showing a transmission passband with constant group delay

as

$$\varepsilon(\lambda) = \varepsilon_0 \left(\varepsilon_\infty - \frac{c_0^2/\lambda_e^2}{c_0^2/\lambda^2 - jc_0\gamma_e/\lambda} \right), \tag{5.64}$$

$$\mu(\lambda) = \mu_0 \left(1 - \sum_{i=1}^{2} \frac{F_i c_0^2/\lambda_e^2}{c_0^2/\lambda^2 - c_0^2/\lambda_{mi}^2 - jc_0\gamma_{mi}/\lambda_{mi}} \right) \tag{5.65}$$

where c_0 is the speed of light, F_1 and F_2 are the filling factors, γ_e, γ_{m1}, and γ_{m2} are the damping factors, and λ_{m1} and λ_{m2} are the wavelengths associated with the two magnetic resonances [74]. With properly tuned damping factors and resonance wavelengths, the theoretical material exhibits a gradually changing refractive index from negative unity at λ_n to positive unity at λ_p. Between these wavelengths, ε and μ are balanced such that the impedance matches that of free space and the transmission is high. Outside of this pass band, the impedance is no longer matched, and the metamaterial effectively blocks the transmission of waves. In addition, the slope of the refractive index, which is determined by the slopes of ε and μ can be controlled in the pass band to minimize group delay fluctuations. The group delay is calculated according to

$$\tau_g = \frac{L}{v_g} = \frac{L(\mathrm{Re}(n) + \omega\frac{d(\mathrm{Re}(n))}{d\omega})}{c_0} \tag{5.66}$$

where v_g is the group velocity and L is the total thickness of the material slab. The transmission and reflection magnitudes and group delay for a theoretical metamaterial slab of $\sim 0.15\lambda$ thickness are shown in Fig. 5.29, illustrating that the metamaterial possesses the targeted scattering properties as well as a near constant group delay in the pass band.

In order to realize a photonic metamaterial with the desired dispersion in the permittivity and permeability, we employ a modified fishnet structure consisting of a

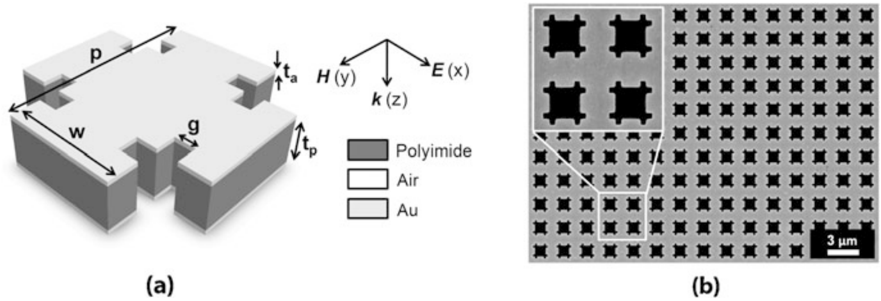


Fig. 5.30 (**a**) Modified fishnet structure employed to realize metamaterial filter. (**b**) Top-view SEM image of a portion of the metamaterial structure with magnified view inset. Scale = 3 µm

metal-dielectric-metal stack with square air holes that are loaded with nanonotches as shown in Fig. 5.30. As discussed earlier in Sect. 5.4.3, the fishnet structure gives rise to a Drude permittivity profile and a Lorentzian magnetic resonance. The added notch loads provide additional control for tuning the permittivity as well as the primary and secondary magnetic resonances. A genetic algorithm was employed to optimize the metamaterial structure, including the unit cell size, air hole notch size, and layer thicknesses for the desired permittivity and permeability profiles. In the GA fitness evaluation, the complex scattering parameters are calculated using Ansoft High Frequency Structure Simulator (HFSS) finite-element solver with periodic boundary conditions assigned to the walls of a single unit cell geometry. The effective permeability and permittivity were then retrieved using the inversion technique described in Sect. 5.4.1 and compared with the target effective medium profile requirements in the following *Cost* function:

$$\begin{cases} Cost_1 = \sum_{f_{\text{pass}}} [|\varepsilon - \varepsilon_{\text{tar},i}| + |\mu - \mu_{\text{tar},i}|] + \sum_{f_{\text{stop}}} \left[\left| \frac{1}{\log_{10}(\mu/\varepsilon)} \right| \right], \\ Cost_2 = \sum_{f_{\text{pass}}} [|\tau_g - \tau_{g,\text{mean}}|], \end{cases} \tag{5.67}$$

$$Cost = Cost_1 + Cost_2 \tag{5.68}$$

where $\varepsilon_{\text{tar},i} = \{-1, 0, 1\}$ and $\mu_{\text{tar},i} = \{-1, 0, 1\}$ are the targeted permittivity and permeability values, respectively, and $\tau_{g,\text{mean}}$ is the average group delay of the sample frequency points within the pass band. In this *Cost* function, $Cost_1$ minimizes the difference between the target and predicted effective parameters in the pass band and maximizes the impedance mismatch in the stop band, while $Cost_2$ flattens the group delay across the pass band.

The GA optimized a population of 32 over 40 generations resulting in the optimum dimensions given by $p = 2113$ nm, $w = 1123$ nm, $g = 1.98$ nm, $t = 30$ nm, and $d = 450$ nm. The predicted scattering parameters shown in Fig. 5.31 show that the optimized structure possesses a strong pass band over the 3.0 to 3.5 µm band with a maximum insertion loss of 1.1 dB over the band as well as stop bands with an average transmission attenuation of 10.2 dB. A steep roll off is also seen in the transition from the pass band to the stop bands, with ∼93 dB/µm on the shorter

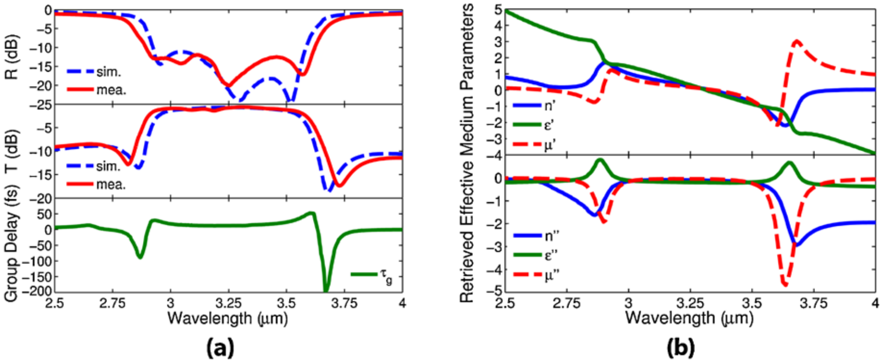


Fig. 5.31 (**a**) Simulated and measured scattering magnitudes and simulated group delay for the optimized metamaterial filter shown in Fig. 5.30. (**b**) Real and imaginary parts of the inverted refractive index, permittivity, and permeability showing a balanced permittivity over the transmission band as well as low intrinsic absorption loss

wavelength side and ∼101 dB/μm on the longer wavelength side. The predicted effective medium parameters for the optimized structure shown in Fig. 5.31 match the theoretical example, with a Drude permittivity profile and two Lorentzian resonances in the permeability. The imaginary part of the refractive index also has a magnitude less than 0.15 across the entire pass band, indicating a low intrinsic absorption loss in the metamaterial. The permittivity and permeability transition simultaneously through zero, preventing a spike in intrinsic absorption at the zero-index wavelength. The predicted group delay shows only a small variation from 15 to 27 fs within the pass band from 3.0 to 3.5 μm. This fluctuation amounts to about 1 period within a 20 % bandwidth and is much lower than the fluctuations of 3 periods within 5 % bandwidth reported for other photonic metamaterials [21, 22].

The GA-optimized metamaterial filter was fabricated by patterning the notch-loaded square air hole array in a deposited tri-layer Au–polyimide–Au stack. The pattern was defined using electron-beam lithography and then transferred into the stack with high-aspect-ratio dry-etching. The metamaterial structure was released from the handle substrate prior to characterization in order to avoid substrate-induced reflection loss. A field emission scanning electron microscope (FESEM) image of the final, freestanding filter in Fig. 5.30 shows that the fabricated structure accurately replicates the design geometry. The freestanding filter was characterized using a Fourier transform infrared spectrometer that could obtain transmission and reflection at normal incidence. The measured scattering parameter amplitudes shown in Fig. 5.31 agree well with the simulation results, showing high, flat transmission in the pass band and strong out-of-band rejection, with only slight discrepancies in the filter bandwidth and roll-off rates. In summary, the dispersion engineering approach enabled the optimization of a broadband optical filter by tailoring the metamaterial effective parameters to those required by the target device.

Acknowledgements The authors would like to thank Theresa S. Mayer and Seokho Yun for their contributions on the fabrication and characterization of the infrared metamaterial devices that are presented in this chapter.

References

1. Rogers Corporation, Advanced circuit materials division high frequency laminates and flexible circuit materials datasheet. http://www.rogerscorp.com/products/index.aspx
2. C.A. Balanis, *Advanced Engineering Electromagnetics* (Wiley, New York, 1989)
3. Z. Bayraktar, J. Bossard, D.H. Werner, AMC metamaterials for low-profile antennas mounted on or embedded in composite platforms, in *Proceedings of the 2007 IEEE Antennas and Propagation Society International Symposium* (2007), pp. 1305–1308
4. Z. Bayraktar, J.A. Bossard, X. Wang, D.H. Werner, A real-valued parallel clonal selection algorithm and its application to the design optimization of multi-layered frequency selective surfaces. IEEE Trans. Antennas Propag. **60**, 1831–1843 (2012)
5. Z. Bayraktar, M. Gregory, D.H. Werner, Composite planar double-sided AMC surfaces for MIMO applications, in *Proceedings of the 2009 IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting*, 2009
6. Z. Bayraktar, M.D. Gregory, X. Wang, D.H. Werner, Matched impedance thin planar composite magneto-dielectric metasurfaces. IEEE Trans. Antennas Propag. **60**, 1910–1920 (2012)
7. Z. Bayraktar, M.D. Gregory, X. Wang, D.H. Werner, A versatile design strategy for thin composite planar double-sided high-impedance surfaces. IEEE Trans. Antennas Propag. **60**, 2770–2780 (2012)
8. Z. Bayraktar, M. Komurcu, Z. Jiang, D.H. Werner, P.L. Werner, Stub-loaded inverted-F antenna synthesis via wind driven optimization, in *Proceedings of the 2011 IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting*, 2011
9. C.M. Bingham, H. Tao, X. Liu, R.D. Avertti, X. Zhang, W.J. Padilla, Planar wallpaper group metamaterials for novel terahertz applications. Opt. Express **16**, 18565–18575 (2008)
10. D.W. Boeringer, D.H. Werner, Particle swarm optimization versus genetic algorithms for phased array synthesis. IEEE Trans. Antennas Propag. **52**, 771–779 (2004)
11. E. Bonabeau, M. Dorigo, G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*. Santa Fe Institute Studies in the Science of Complexity (Oxford University Press, Oxford, 1999)
12. D. Bonefacic, S. Hrabar, D. Kvakan, Experimental investigation of radiation properties of an antenna embedded in low permittivity thin-wire-based metamaterial. Microw. Opt. Technol. Lett. **48**, 2581–2586 (2006)
13. J.A. Bossard, S. Yun, D.H. Werner, T.S. Mayer, Synthesizing low loss negative index metamaterial stacks for the mid-infrared using genetic algorithms. Opt. Express **17**, 14771–14779 (2009)
14. X. Chen, T.M. Grzegorczyk, B.-I. Wu, J. Pacheco, J.A. Kong, Effective parameters of resonant negative refractive index metamaterials: interpretation and validity. J. Appl. Phys. **98**, 063505 (2005)
15. X. Chen, B.-I. Wu, J.A. Kong, T.M. Grzegorczyk, Retrieval of the effective constitutive parameters of bianisotropic metamaterials. Phys. Rev. E **71**, 046610 (2005)
16. Q. Cheng, T.J. Cui, W.X. Jiang, B.G. Cai, An onmidirectional electromagnetic absorber made of metamaterials. New J. Phys. **12**, 063006 (2010)
17. E. Cubukcu, S. Zhang, Y.-S. Park, G. Bartal, X. Zhang, Split ring resonator sensors for infrared detection of single molecular monolayers. Appl. Phys. Lett. **95**, 043113 (2009)
18. L.N. de Castro, F.J.V. Zuben, Learning and optimization using the clonal selection principle. IEEE Trans. Evol. Comput. **6**, 239–251 (2002)

19. M. Diem, T. Koschny, C.M. Soukoulis, Wide-angle perfect absorber/thermal emitter in the terahertz regime. Phys. Rev. B **79**, 033101 (2009)
20. G. Dolling, C. Enkrich, M. Wegener, C.M. Soukoulis, S. Linden, Low-loss negative-index metamaterial at telecommunication wavelengths. Opt. Lett. **31**, 1800–1802 (2006)
21. G. Dolling, C. Enkrich, M. Wegener, C.M. Soukoulis, S. Linden, Simultaneous negative phase and group velocity of light in a metamaterial. Science **312**, 892–894 (2006)
22. G. Dolling, M. Wegener, C.M. Soukoulis, S. Linden, Negative-index metamaterial at 780 nm wavelength. Opt. Lett. **32**, 53–55 (2007)
23. M. Dorigo, V. Maniezzo, A. Colorni, Ant system: optimization by a colony of cooperative agents. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **26**, 29–41 (1996)
24. T. Driscoll, D.N. Basov, W.J. Padilla, J.J. Mock, D.R. Smith, Electromagnetic characterization of planar metamaterials by oblique angle spectroscopic measurements. Phys. Rev. B **75**, 115114 (2007)
25. R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in *IEEE Proceedings of the Sixth International Symposium on Micro Machine and Human Science* (1995), pp. 39–43
26. R.C. Eberhart, Y. Shi, J. Kennedy, *Swarm Intelligence*. Morgan Kaufmann (Academic Press, New York, 2001)
27. T.F. Eibert, J.L. Volakis, D.R. Wilton, D.R. Jackson, Hybrid FE/BI modeling of 3-D doubly periodic structures utilizing triangular prismatic elements and an MPIE formulation accelerated by the Ewald transformation. IEEE Trans. Antennas Propag. **47**, 843–850 (1999)
28. S. Enoch, G. Tayeb, P. Sabouroux, N. Guérin, P. Vincent, A metamaterial for directive emission. Phys. Rev. Lett. **89**, 213902 (2002)
29. A. Erentok, P.L. Luljak, R.W. Ziolkowski, Characterization of a volumetric metamaterial realization of an artificial magnetic conductor for antenna applications. IEEE Trans. Antennas Propag. **53**, 160–172 (2005)
30. D. Gies, Particle swarm optimization: applications in electromagnetic design. Master's thesis, University of California Los Angeles, Los Angeles, CA, USA, 2004
31. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, 1989)
32. A. Grbic, G.V. Eleftheriades, Overcoming the diffraction limit with a planar left-handed transmission-line lens. Phys. Rev. Lett. **92**, 117403 (2004)
33. M.D. Gregory, Z. Bayraktar, D.H. Werner, Fast optimization of electromagnetic design problems using the covariance matrix adaptation evolutionary strategy. IEEE Trans. Antennas Propag. **59**, 1275–1285 (2011)
34. T.M. Grzegorczyk, Z.M. Thomas, J.A. Kong, Inversion of critical angle and Brewster angle in anisotropic left-handed metamaterials. Appl. Phys. Lett. **86**, 251909 (2005)
35. N. Hansen, The CMA evolution strategy. http://www.lri.fr/~hansen/cmaesintro. Accessed July 24, 2011
36. N. Hansen, CMA evolution strategy source code. http://www.lri.fr/~hansen/cmaes_inmatlab. Accessed July 24, 2011
37. N. Hansen, A. Auger, R. Ros, S. Finck, P. Posik, Comparing results of 31 algorithms from the black-box optimization benchmarking, in *Workshop Proceedings of the GECCO Genetic and Evolutionary Computation Conference* (2010), pp. 1689–1696
38. N. Hansen, S. Kern, Evaluating the CMA evolution strategy on multimodal test functions, in *Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature* (2004), pp. 282–291
39. N. Hansen, S.D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. **11**, 1–18 (2003)
40. N. Hansen, A. Ostermeier, *Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: The Covariance Matrix Adaptation* (1996), pp. 312–317
41. N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolutionary strategies. Evol. Comput. **9**, 159–195 (2001)

42. R.L. Haupt, S.E. Haupt, *Practical Genetic Algorithms*, 2nd edn. (Wiley, Hoboken, 2007)
43. R.L. Haupt, D.H. Werner, *Genetic Algorithms in Electromagnetics* (Wiley, New York, 2004)
44. S.L. Ho, S. Yang, H.C. Wong, K.W.E. Cheng, G. Ni, An improved ant colony optimization algorithm and its application to electromagnetic device designs. IEEE Trans. Magn. **41**, 1764–1767 (2005)
45. J.H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, 1975)
46. L.H. Hu, S.T. Chui, Characteristics of electromagnetic wave propagation in uniaxially anisotropic left-handed materials. Phys. Rev. B **66**, 085108 (2002)
47. Z.H. Jiang, J.A. Bossard, X. Wang, D.H. Werner, Synthesizing metamaterials with angularly independent effective medium properties based on an anisotropic parameter retrieval technique coupled with a genetic algorithm. J. Appl. Phys. **109**, 013515 (2011)
48. Z.H. Jiang, S. Yun, F. Toor, D.H. Werner, T.S. Mayer, Conformal dual-band near-perfectly absorbing mid-infrared metamaterial coating. ACS Nano **5**, 4641–4647 (2011)
49. S. Karimkashi, A.A. Kishk, Invasive weed optimization and its features in electromagnetics. IEEE Trans. Antennas Propag. **58**, 1269–1278 (2010)
50. J. Kennedy, R. Eberhart, Particle swarm optimization, in *Proceedings of the Ninth International Conference on Neural Networks*, vol. 4 (1995), pp. 1942–1948
51. D. Kern, D.H. Werner, A genetic algorithm approach to the design of ultra-thin electromagnetic bandgap absorbers. Microw. Opt. Technol. Lett. **38**, 61–64 (2003)
52. D.J. Kern, D.H. Werner, Magnetic loading of EBG AMC ground planes and ultra-thin absorbers for improved bandwidth performance and reduced size. Microw. Opt. Technol. Lett. **48**, 2468–2471 (2006)
53. D.J. Kern, D.H. Werner, A. Monorchio, L. Lanuzza, M. Wilhelm, The design synthesis of multiband artificial magnetic conductors using high impedance frequency selective surfaces. IEEE Trans. Antennas Propag. **53**, 8–17 (2005)
54. D.J. Kern, D.H. Werner, P.L. Werner, Optimization of multi-band AMC surfaces with magnetic loading, in *Proceedings of the IEEE Antennas Propagation International Symposium* (2004), pp. 823–826
55. J.A. Kong, *Electromagnetic Wave Theory* (EMW, New York, 2000)
56. T. Koschny, M. Kafesaki, E.N. Economou, C.M. Soukoulis, Effective medium theory of left-handed materials. Phys. Rev. Lett. **93**, 107402 (2004)
57. T. Koschny, P. Markoš, E.N. Economou, D.R. Smith, D.C. Vier, C.M. Soukoulis, Impact of inherent periodic structure on effective medium description of left-handed and related metamaterials. Phys. Rev. E **71**, 245105 (2005)
58. T. Koschny, P. Markoš, D.R. Smith, C.M. Soukoulis, Resonant and antiresonant frequency dependence of the effective parameters of metamaterials. Phys. Rev. E **68**, 065602 (2003)
59. D.-H. Kwon, L. Li, J.A. Bossard, M.G. Bray, D.H. Werner, Zero index metamaterial with checkerboard structure. IEEE Electron Device Lett. **43**, 9–10 (2007)
60. D.-H. Kwon, D.H. Werner, Low-index metamaterial designs in the visible spectrum. Opt. Express **15**, 9267–9272 (2007)
61. D.-H. Kwon, D.H. Werner, A.V. Kildishev, V.M. Shalaev, Material parameter retrieval procedure for general bi-isotropic metamaterials and its application to optical chiral negative-index metamaterial design. Opt. Express **16**, 11822–11829 (2008)
62. N.I. Landy, C.M. Bingham, T. Tyler, N. Jokerst, D.R. Smith, W.J. Padilla, Design, theory, and measurement of polarization-insensitive absorber for terahertz imaging. Phys. Rev. B **79**, 125104 (2009)
63. N.I. Landy, S. Sajuyigbe, J.J. Mock, D.R. Smith, W.J. Padilla, Perfect metamaterial absorber. Phys. Rev. Lett. **100**, 207402 (2008)
64. Z. Li, K. Aydin, E. Ozbay, Determination of the effective constitutive parameters of bianisotropic metamaterials from reflection and transmission coefficients. Phys. Rev. E **79**, 026610 (2009)
65. E. Lier, D.H. Werner, C.P. Scarborough, Q. Wu, J.A. Bossard, An octave-bandwidth negligible-loss radiofrequency metamaterial. Nat. Mater. **20**, 216–222 (2011)

66. N. Liu, M. Mesch, T. Weiss, M. Hentschel, H. Giessen, Infrared perfect absorber and its application as plasmonic sensor. Nano Lett. **10**, 2342–2348 (2010)
67. X. Liu, T. Starr, A.F. Starr, W.J. Padilla, Infrared spatial and frequency selective metamaterial with near-unity absorbance. Phys. Rev. Lett. **104**, 207403 (2010)
68. K.-P. Ma, K. Hirose, F.-R. Yang, Y. Qian, T. Itoh, Realisation of magnetic conducting surface using novel photonic bandgab structure. IEEE Electron Device Lett. **34**, 2041–2042 (1998)
69. P. Markos, C.M. Soukoulis, Transmission properties and effective electromagnetic parameters of double negative metamaterials. Opt. Lett. **11**, 649–661 (2003)
70. R. Marques, F. Medina, R. Rafii-El-Idrissi, Role of bianisotropy in negative permeability and left-handed metamaterials. Phys. Rev. B **65**, 0144440 (2002)
71. R. Marques, F. Mesa, J. Martel, F. Medina, Comparative analysis of edge- and broadside-coupled split ring resonators for metamaterial design—theory and experiments. IEEE Trans. Antennas Propag. **51**, 2572–2581 (2003)
72. C. Menzel, C. Rockstuhl, T. Paul, F. Lederer, T. Pertsch, Retrieving effective parameters for metamaterials at oblique incidence. Phys. Rev. B **77**, 195328 (2008)
73. A.M. Nicolson, G.F. Ross, Measurement of the intrinsic properties of materials by time-domain techniques. IEEE Trans. Instrum. Meas. **19**, 377–382 (1970)
74. K.E. Oughstun, S. Shen, Velocity of energy transport for a time-harmonic field in a multiple-resonance Lorentz medium. J. Opt. Soc. Am. B **5**, 2395–2398 (1988)
75. J.B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966–3969 (2000)
76. A.D. Rakić, A.B. Djurišić, J.M. Elazar, M.L. Majewski, Optical properties of metallic films for vertical-cavity optoelectronic devices. Appl. Opt. **37**, 5271–5283 (1998)
77. J. Robinson, Y. Rahmat-Samii, Particle swarm optimization in electromagnetics. IEEE Trans. Antennas Propag. **52**, 397–407 (2004)
78. S.M. Rudolph, A. Grbic, Super-resolution focusing using volumetric, broadband NRI media. IEEE Trans. Antennas Propag. **56**, 2963–2969 (2008)
79. S.M. Rudolph, A. Grbic, The design and performance of an isotropic negative-refractive-index metamaterial lens, in *Proc. of XXX General Assembly of the International Union of Radio Science*, 2011
80. 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**, 977–980 (2006)
81. D. Seetharamdoo, R. Sauleau, K. Mahdjoubi, A.-C. Tarot, Robust method to retrieve the constitutive effective parameters of metamaterials. Phys. Rev. E **70**, 016608 (2004)
82. V.M. Shalaev, W. Cai, U.K. Chettiar, H.-K. Yuan, A.K. Sarychev, V.P. Drachev, A.V. Kildishev, Negative index of refraction in optical metamaterials. Opt. Lett. **30**, 3356–3358 (2005)
83. R.A. Shelby, D.R. Smith, S. Schultz, Experimental verification of a negative index of refraction. Science **292**, 77–79 (2001)
84. D. Sievenpiper, L. Zhang, R.F.J. Broas, N.G. Alexopolous, E. Yablonovitch, High-impedance electromagnetic surfaces with a forbidden frequency band. IEEE Trans. Microw. Theory Tech. **47**, 2059–2074 (1999)
85. M. Silveirinha, N. Engheta, Tunneling of electromagnetic energy through sub-wavelength channels and bends using epsilon-near-zero materials. Phys. Rev. Lett. **97**, 157403 (2006)
86. D.R. Smith, W.J. Padilla, D.C. Vier, S.C. Nemat-Nasser, S. Schultz, Composite medium with simultaneously negative permeability and permittivity. Phys. Rev. Lett. **84**, 4184–4187 (2000)
87. D.R. Smith, S. Schultz, P. Markoš, C.M. Soukoulis, Determination of effective permittivity and permeability of metamaterials from reflection and transmission coefficients. Phys. Rev. B **65**, 195104 (2002)
88. R. Storn, K. Price, Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. **11**, 341–359 (1997)
89. H. Tao, C.M. Bingham, D.V. Pilon, K. Fan, A.C. Strikwerda, D. Shrekenhamer, W.J. Padilla, X. Zhang, R.V. Avertti, A dual band terahertz metamaterial absorber. J. Phys. D, Appl. Phys. **43**, 225102 (2010)

90. H. Tao, C.M. Bingham, A.C. Strikwerda, D. Pilon, D. Shrekenhamer, N.I. Landy, K.X. Fan, X. Zhang, W.J. Padilla, R.V. Avertti, Highly flexible wide angle of incidence terahertz metamaterial absorber: design, fabrication, and characterization. Phys. Rev. B **78**, 241103 (2008)
91. 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**, 7181–7188 (2008)
92. V. Trianni, *Evolutionary Swarm Robotics: Evolving Self-Organising Behaviours in Groups of Autonomous Robots* (Springer, Berlin, 2008)
93. J. Valentine, S. Zhang, T. Zentgraf, E. Ulin-Avila, D.A. Genov, G. Bartal, X. Zhang, Three-dimensional optical metamaterial with a negative refractive index. Nature **455**, 376–379 (2008)
94. K.J. Vinoy, R.M. Jha, *Radar Absorbing Materials: From Theory to Design and Characterization* (Kluwer, Dordrecht, 1996)
95. B. Wang, T. Koschny, C.M. Soukoulis, Wide-angle and polarization-independent chiral metamaterial absorber. Phys. Rev. B **80**, 033108 (2009)
96. W.B. Weir, Automatic measurement of complex dielectric constant and permeability at microwave frequencies. Proc. IEEE **62**, 33–36 (1974)
97. Q.-Y. Wen, H.-W. Zhang, Y.-S. Xie, Q.-H. Yang, Y.-L. Liu, Dual band terahertz metamaterial absorber: design, fabrication, and characterization. Appl. Phys. Lett. **95**, 3276072 (2009)
98. T.K. Wu (ed.), *Frequency Selective Surface and Grid Array* (Wiley, New York, 1995)
99. W. Zhu, X. Zhao, Metamaterial absorber with dendritic cells at infrared frequencies. J. Opt. Soc. Am. B **26**, 2382–2385 (2009)
100. R.W. Ziolkowski, Propagation in and scattering from a matched metamaterial having a zero index of refraction. Phys. Rev. E **70**, 046608 (2004)

Chapter 6
Objective-First Nanophotonic Design

Jesse Lu and Jelena Vuckovic

Abstract We introduce an "objective-first" strategy for designing nanophotonic devices, and we demonstrate the design of nanophotonic coupler, cloak, and mimic devices. Simply put, our objective-first method works by prioritizing the performance of the device even above satisfying Maxwell's equations. We show how this is accomplished starting from Maxwell's equations, applying numerical discretization, and then solving not only for the field variables but the structure variables as well. We then demonstrate the ability to quickly produce designs for both traditional devices such as waveguide couplers, as well as more exotic devices such as optical cloaks and mimics. Finally, we point the reader to future improvements and extensions of our method.

6.1 Introduction

Our initial foray into design methods for nanophotonic devices began with a very simple and naive question: Could we make an inverse solver which, when given the electromagnetic fields we desire, returns the nanophotonic structure that will produce them [1]? In other words, since we already know how to solve for E and H in Maxwell's equations, why can't we solve for ε or even μ instead? Not surprisingly, it did not take us long to find that such a simple strategy would inevitably run into many problems.

Over the subsequent years, we were able to come up with a better solution, which we call an "objective-first" strategy for nanophotonic design, and which we present in this chapter. Although it is more advanced than our original idea, objective-first design still carries the same fundamental concept, which is to specify the electromagnetic fields, and then to solve for a structure to produce them.

One of the strengths of this approach is that it recasts the optimization problem into separably convex problems whose global optima can be reliably determined. Additionally, we show that our formulation is general enough to encompass the design of all linear nanophotonic structures and is not limited to the design of optical

J. Lu (✉)
Department of Electrical Engineering, Stanford University, Palo Alto, CA 94305, USA
e-mail: jesselu@stanford.edu

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_6,

metamaterials. While this approach typically produces designs with very high efficiencies (typically 99 %) our solutions do feature continuously varying values of the dielectric constant. Future approaches for dealing with this constraint are discussed in Sect. 6.7.3.

In this chapter, we present the simple theoretical underpinnings of objective-first design in Sects. 6.2 and 6.3—namely the numerical discretization of the electromagnetic wave equation, and the formulation of the objective-first design problem. Then, in Sects. 6.4–6.6, we show examples of the method in action in designing nanophotonic devices in three broad categories: waveguide couplers, optical cloaks, and optical mimics; the source code for which is available online [2]. In Sect. 6.7, we conclude by commenting on possible extensions of our method.

Lastly, we note that in contrast to many of the other chapters in this book, the designs proposed in this chapter are all permittivity and permeability profiles in one and two-dimensions. The additional task of choosing specific material sets to fabricate actual devices is left as an additional step.

6.2 The Electromagnetic Wave Equation

In this section, we outline the wave equation that is central to the application of our method, with the end-result being to show that it is separably linear (bi-linear) in the field and structure variables. We do this by first formulating this wave equation in the language of physics, and then discretizing it in order to achieve numerical solutions. We then show how one can not only obtain the solution for the fields, but also obtain the solution for the structure using simple, standard numerical tools.

6.2.1 Physics Formulation

First, let's derive our wave equation, starting with the differential form of Maxwell's equations,

$$\nabla \times E = -\mu_0 \frac{\partial H}{\partial t}, \tag{6.1}$$

$$\nabla \times H = J + \varepsilon \frac{\partial E}{\partial t}, \tag{6.2}$$

where E, H, and J are the electric, magnetic and electric current vector fields, respectively, ε is the permittivity and μ_0 is the permeability, which we assume to be that of vacuum everywhere.

Assuming the time dependence $\exp(-i\omega t)$, where ω is the angular frequency, these become

$$\nabla \times E = -i\mu_0\omega H, \tag{6.3}$$

$$\nabla \times H = J + i\varepsilon\omega E, \tag{6.4}$$

which we can combine to form our (time-harmonic) wave equation,

$$\nabla \times \varepsilon^{-1} \nabla \times H - \mu_0 \omega^2 H = \nabla \times \varepsilon^{-1} J. \tag{6.5}$$

In this chapter, we are only going to consider the two-dimensional form of this equation, and specifically the two-dimensional transverse electric (TE) mode [3]. In this case (6.5) is simplified because only the z-component of H is nonzero. Nevertheless, a single equation (6.5), represents all the physics which we take into account in this chapter.

6.2.2 Numerical Formulation

On top of the analytical formulation of the wave equation (6.5), we will now add a numerical, or discretized, formulation. This will be needed in order to solve for arbitrary structures for which there are not analytical solutions.

The salient step in order to do so is to use the Yee grid [4], which allows us to easily define the curl ($\nabla\times$) operators in (6.5). Since both the individual curl operators and the equation as a whole is linear in H, we can reformulate (6.5) with a change of variables, as

$$A(p)x = b(p), \tag{6.6}$$

where $H \to x$, $\varepsilon^{-1} \to p$; and where

$$A(p) = \nabla \times \varepsilon^{-1} \nabla \times -\mu_0 \omega^2 \tag{6.7}$$

and

$$b(p) = \nabla \times \varepsilon^{-1} J. \tag{6.8}$$

Note that our use of $A(p)$ and $b(p)$ instead of A and b simply serves to clarify the dependence of both A and b on p.

Apart from using the Yee grid, which at some length scale requires that our designs conform to a rectangular grid, the only other salient implementation detail is the use of stretched-coordinate perfectly matched layers [5] where necessary, in order to prevent unwanted reflections at the boundaries of the simulation domain. The effect of such layers is to modify the curl operators, although their linear property is still maintained.

6.2.3 Solving for H

With our numerical formulation, we can now solve for the H-field (the E-field can be computed from the H-field using (6.4)) by applying general linear algebra solvers to (6.6). Recall that since we have chosen a time-harmonic formulation, solving for

x in (6.6) is actually performing what is simply known as a time-harmonic or a finite-difference frequency-domain (FDFD) simulation [6]. Furthermore, since we have limited ourselves to the two-dimensional case, (6.6) is easily solved using the standard sparse solver included in Matlab on a single desktop computer.

We call the routine that solves for x in (6.6) given p a field-solver, or a simulator.

6.2.4 Solving for ε^{-1}

After having built a field-solver or simulator (which finds x given p) for our wave equation, the next step is to build a structure-solver for it. In other words, we need to be able to solve for p given x.

To do so, we return to (6.5) and remark that $\varepsilon^{-1}(\nabla \times H) = (\nabla \times H)\varepsilon^{-1}$ and $\varepsilon^{-1}J = J\varepsilon^{-1}$ since scalar multiplication is commutative. This allows us to rearrange (6.5) as

$$\nabla \times (\nabla \times H)\varepsilon^{-1} - \nabla \times J\varepsilon^{-1} = \mu_0\omega^2 H \tag{6.9}$$

which we now write as

$$B(x)p = d(x), \tag{6.10}$$

where

$$B(x) = \nabla \times (\nabla \times H) - \nabla \times J \tag{6.11}$$

and

$$d(x) = \mu_0\omega^2 H. \tag{6.12}$$

With this extremely simple trick, we have shown that we can seemingly solve for p given x with approximately the same ease as solving for x given p! We see this because the dimensions and complexity of $B(x)$ are basically equivalent to that of $A(p)$, and this implies that the same simple tools used in our field-solver should be applicable to solving (6.10). This is indeed what we find, although the later addition of constraints on p will require the use of more powerful (but just as dependable) numerical tools.

6.2.5 Bi-linearity of the Wave Equation

Although additional mathematical machinery must still be added to obtain a useful design tool, we have shown so far that the wave equation is separately linear in x and p (i.e., bilinear). Namely,

$$A(p)x - b(p) = B(x)p - d(x). \tag{6.13}$$

In other words, fixing p makes solving the wave equation for x a linear problem, and vice versa. Note that the joint problem, where both x and p are allowed to vary, is not linear.

The bi-linearity of the wave equation is *absolutely fundamental* in our objective-first strategy because it relies on the fact that, although simultaneously solving for x and p is very difficult, we already know how to solve linear systems (x and p separately) well. In fact, it is this very property that forms the natural division of labor which our objective-first method exploits.

6.3 The Objective-First Design Problem

We now describe the remaining machinery used in the objective-first method, in addition to the field-solver and the structure-solver, as previously outlined. Specifically, we introduce the idea of a design objective and a physics residual, and we reference the mathematical notion of convexity in order to motivate the need to divide the objective-first problem into two separately convex sub-problems.

6.3.1 Design Objectives

Our design objective or objective function, $f(x)$, is simply defined as a function we wish to be minimal for the design to be produced. For instance, in the design of a device which must transmit efficiently into a particular mode, we could choose $f(x)$ to be the negative power flow into that mode. Or, if the device was to be a low-loss resonator, we could choose $f(x)$ to be the amount of power leaking out of the device. In general, there are multiple choices of $f(x)$ which can be used to describe the same objective. For example, $f(x)$ for a transmissive device may not only be the negative power transmitted into the desired output mode, but it could also be the amount of power lost to other modes, or even the error in the field values at the output port relative to the field values needed for perfect transmission. These design objectives are equivalent in the sense that, if minimized, all would produce structures with good performance. At the same time, we must consider that the computational cost and complexity of using one $f(x)$ over another may indeed vary greatly.

6.3.2 Convexity

Before formulating the design problem, we would like to add a note regarding the complexity of various optimization problems. Specifically, we want to introduce the notion of *convexity* [7] and to note the difference between problems that are convex and those which are not. The difference is simply this: Convex problems have a single optimum point (only one local optimum, which is therefore the global optimum) which we can reliably find using existing numerical software, whereas non-convex problems typically have multiple optima and are thus much more difficult to reliably solve.

That a convex problem can be reliably solved, in this case, means that regardless of the starting guess, convex optimization software will always arrive at the globally optimal solution and will be able to numerically prove global optimality as well. Thus, the advantage in formulating a design problem in terms of convex optimization problems is to eliminate both the need to circumvent local optima and any notion of randomness. On a practical level, there exist mature convex optimization software packages among which is CVX, a convex optimization package written for Matlab [8], which we use for the examples in this chapter.

6.3.3 Typical Design Formulation

We now examine the typical, and most straightforward formulation of the design problem, in order to relate and contrast it to the objective-first formulation. The design problem for a physical structure is typically formulated as

$$\begin{aligned} &\underset{x,p}{\text{minimize}}\ f(x) \\ &\text{subject to } A(p)x - b(p) = 0, \end{aligned} \tag{6.14}$$

which states that we would like to vary x and p simultaneously in order to decrease $f(x)$ while always satisfying physics (e.g., the electromagnetic wave equation).

Since solving (6.14) is quite difficult in the general sense (simultaneously varying x and p makes the problem non-convex), traditional optimization approaches, such as those described in previous chapters, have relied on either a brute-force parameter search, or a gradient-descent method utilizing first-order derivatives. In the gradient-descent case, solving (6.14) results in the well-known adjoint optimization method [9].

6.3.4 Objective-First Design Formulation

In contrast with the typical formulation, the objective-first formulation simply switches the roles of the wave equation and the design objective with one another:

$$\underset{x,p}{\text{minimize}}\ \left\| A(p)x - b(p) \right\|^2 \tag{6.15}$$

$$\text{subject to } f(x) = f_{\text{ideal}}. \tag{6.16}$$

Although such a switch may seem trivial, and even silly at first, we show that it fundamentally changes the nature of the design problem and actually gives us advantages in our efforts at finding a solution.

This first fundamental change, as seen from (6.15), is that we allow for a nonzero residual in the electromagnetic wave equation. This literally means that we allow for *non-physical* x and p, since $A(p)x - b(p) \neq 0$ is permissible. And since $A(p)x - b(p)$ can now be a nonzero entity, we choose to call it the *physics residual*. The second fundamental change is that we always force the device to exhibit ideal performance, as seen from (6.16). This, of course, ties in very closely with (6.15) since ideal performance is usually not obtainable unless one allows for some measure of error in the underlying physics (nonzero physics residual). As such, our strategy will be to iteratively vary x and then p in order to decrease the physics residual (6.15) to zero, while always maintaining ideal performance.

The primary advantage of the objective-first formulation is that, although the full problem is still non-convex, it allows us to form two convex sub-problems, as outlined below. In contrast to an adjoint method, here we can still access information regarding second-order derivatives, which decreases the time it takes to obtain a solution. An additional advantage of this approach is that our insistence that ideal performance be always attained provides a mechanism which can potentially "override" local optima in the optimization process.

To this end, we have found that such a strategy results in surprisingly nonintuitive devices which exhibit highly efficient performance, even when the starting point of the design problem is completely non-functional. Furthermore, we have found this to be true even when the physics residual fails to be completely removed.

In practice, we add an additional constraint to the original formulation, [10] which is to set hard-limits on the allowable values of p, namely $p_0 \leq p \leq p_1$. This is actually a relaxation of the ideal constraint, which would be to allow p to only have discrete values, $p \in [p_0, p_1]$, but such a constraint would essentially force us to only perform brute force trial-and-error.

Our objective-first formulation is thus:

$$\begin{aligned} &\underset{x,p}{\text{minimize}} \; \left\| A(p)x - b(p) \right\|^2 \\ &\text{subject to } f(x) = f_{\text{ideal}} \\ &\qquad\qquad p_0 \leq p \leq p_1, \end{aligned} \tag{6.17}$$

which is still non-convex, but can be broken down into two convex sub-problems, the motivation being that each of these will be able to be easily and reliably solved.

6.3.5 Field Sub-problem

The first of these is the field sub-problem, which simply involves fixing p and independently optimizing x,

$$\begin{aligned} &\underset{x}{\text{minimize}} \ \left\| A(p)x - b(p) \right\|^2 \\ &\text{subject to } f(x) = f_{\text{ideal}}. \end{aligned} \tag{6.18}$$

This problem is convex, and actually quadratic, which means that it can even be solved using standard numerical tools, in the same way as a simple least-squares problem.

The field sub-problem can be thought of as an update to x (H-field) where we try to "fit" the electromagnetic fields to the structure (p). Of course, if it were not for the hard-constraint on the design objective, the field sub-problem would be able to perfectly fit x to p.

6.3.6 Structure Sub-problem

The second sub-problem is formulated by fixing x and independently optimizing p. At the same time, we use the bi-linearity property of the physics residual from (6.13) to rewrite the problem in a way that makes its convexity explicit:

$$\begin{aligned} &\text{minimize} \ \left\| B(x)p - d(x) \right\|^2 \\ &\text{subject to } p_0 \leq p \leq p_1. \end{aligned} \tag{6.19}$$

The structure sub-problem is also convex, but not quadratic because of the inequality constraints on p. However, use of the CVX package still allows us to obtain results quickly and reliably.

Note that in an analogous fashion to the field sub-problem, the structure sub-problem attempts to fit p to x, and is prevented from perfectly doing so by its own constraint. Because neither sub-problem is capable of completely reducing the physics residual to zero, they must be used in an iterative manner in order to gradually decrease the physics residual. To this end, we employ the alternating directions optimization method.

6.3.7 Alternating Directions

We use a simple alternating directions scheme to piece together (6.18) and (6.19), which is to say that we simply continually alternate between solving each equation until we reach some stopping point, normally measured by how much the physics residual has decreased.

$$
\begin{aligned}
&\text{Loop:}\\
&\underset{x}{\text{minimize}}\ \left\|A(p)x - b(p)\right\|^2\\
&\text{subject to } f(x) = f_{\text{ideal}};\\
&\underset{p}{\text{minimize}}\ \left\|B(x)p - d(x)\right\|^2\\
&\text{subject to } p_0 \le p \le p_1.
\end{aligned}
\tag{6.20}
$$

The alternating directions scheme is extremely simple and does not require additional processing of x or p outside of the two sub-problems, nor does it require the use of auxiliary variables.

The advantage of the alternating directions method is that the physics residual is guaranteed to monotonically decrease with every iteration, which is useful in that no safeguards are needed to protect against "rogue" steps in the optimization procedure. Note that this robustness stems from the fact that, among other things, each sub-problem does not rely on previous values of the variable which is being optimized, but only on the variable which is held constant.

The disadvantage of such a simple scheme is that the convergence is quite slow, although we have found it to be sufficient in our cases. Related methods, such as the Alternating Directions Method of Multipliers [11], exhibit far better convergence.

6.4 Waveguide Coupler Design

We first apply the objective-first formulation with the alternating directions algorithm to the design of nanophotonic waveguide couplers in two dimensions, where our goal is to couple light from a single input waveguide mode to a single output waveguide mode with as close to unity efficiency as possible. We would also like to allow the user to choose arbitrary input and output waveguides, as well as to select arbitrary modes within those waveguides (as opposed to allowing only the fundamental mode, for example).

This problem is very general and, in essence, encompasses the design of all linear nanophotonic components because the function or performance of all such components is simply to convert a defined set of input modes into a defined set of output modes. Such a broad, general problem is ideally suited for an objective-first strategy since no approximations or simplifications of the electromagnetic fields are required; we only make the simplification of working in two dimensions (transverse magnetic mode) and dealing only with a single input and output mode.

Since the electromagnetic wave equation is scale-invariant (e.g., double the length scale and half the frequency and you obtain the same equation), we state all dimensions terms of the vacuum wavelength itself. Therefore, our solutions are applicable to regions of the electromagnetic spectrum where the dielectric constants

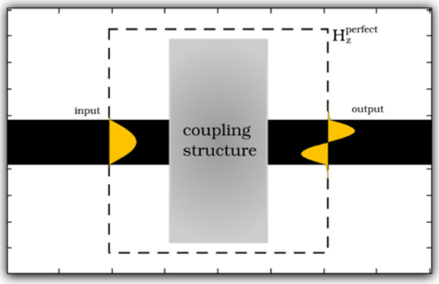


Fig. 6.1 Formulation of the design objective

used are achievable. In most cases, values between 1 and 12.25 are chosen because these are realizable for semiconductor devices operating at telecommunication frequencies. Finally, note that dispersive effects are ignored for all results in this chapter since we always consider device performance at a single, fixed frequency.

6.4.1 Choice of Design Objective

As mentioned in Sect. 6.3.1, multiple equivalent choices of design objective exist which should allow one to achieve the same device performance; however, we will choose, for generality, the following design objective:

$$f(x) = \begin{cases} x - x_{\text{perfect}} & \text{at boundary,} \\ 0 & \text{elsewhere.} \end{cases} \tag{6.21}$$

That is, $f(x)$ simply selects the outermost values of the field in the design space and compares them to values of a perfect device.

Furthermore, we choose $f_{\text{ideal}} = 0$ so that when placed into the objective-first problem (6.17), this will result in fixing the boundary values of the field at the edge of the design space to those of an ideal device, as shown in Fig. 6.1. In this case, we choose such an ideal device to have perfect (unity) coupling efficiency, and these ideal fields are simply obtained by using the input and output mode profiles at the corresponding ports and using values of zero at the remaining ports.

Such a design objective is general in the sense that the boundary values of the device contain all the information necessary to determine how the device will interact with its environment, when excited with the input mode in question. In other words, we only need to know the boundary field values, and not the interior field values to determine the performance of the device; and thus, it would be conceivable that such a scheme might be generally applied to linear nanophotonic devices beyond just waveguide mode couplers.

In our case, we only need to know the value of H_z and its derivative along the normal direction, $\partial H_z/\partial n$, along the design boundary in order to completely characterize its performance. Alternatively, one can, of course, use the outermost two layers of the H_z instead of calculating a spatial derivative.

6.4.2 Application of the Objective-First Strategy

Having chosen our design objective we apply the alternating directions algorithm to (6.17) which results in solving the following two sub-problems iteratively:

$$\begin{aligned} &\underset{x}{\text{minimize}} \ \left\| A(p)x - b(p) \right\|^2 \\ &\text{subject to } x = x_{\text{perfect}}, \quad \text{at boundary}; \end{aligned} \tag{6.22}$$

$$\begin{aligned} &\underset{p}{\text{minimize}} \ \left\| B(x)p - d(x) \right\|^2 \\ &\text{subject to } p_0 \leq p \leq p_1. \end{aligned} \tag{6.23}$$

For the results throughout this chapter, we uniformly choose $p_0 = 1/12.25$ and $p_0 = 1$, corresponding to ε^{-1} of silicon and air, respectively. Additionally, since a starting value for p is initially required, we always choose to use a uniform value of $p = 1/9$ across the entire design space. There is nothing really unique about such a choice, although we have noticed that an initial value of p near 1 often results in poor designs. Note that unlike p we do not require an initial guess for x. The only other significant value that needs to be set initially is the frequency, or wavelength of light. We use free space wavelengths in the range of 25 to 63 grid points for the results in this chapter.

Lastly, for all the examples presented in the chapter, we run the alternating directions algorithm for 400 iterations. In terms of convergence, the physics residual never fully vanishes, and seems to asymptotically approach a nonzero value. Even so, we seem to obtain good performance from the produced designs. Note that although we do not present the convergence results here, such information can be obtained by inspecting the source code [2].

6.4.3 Coupling to a Wide, Low-Index Waveguide

As a first example, we design a coupler from the fundamental mode of a narrow, high-index waveguide to the fundamental mode of a wide, low-index waveguide. Such a coupler would be useful for coupling from an on-chip nanophotonic waveguide to an off-chip fiber, for example.

The input and output mode profiles used as the ideal fields are shown in the upper-left corner of Fig. 6.2. The final structure is shown in the upper right plot, and the simulated H_z fields, under excitation of the input mode in this final structure, are shown in the bottom plots.

Figure 6.2 then shows that the design structure has nearly unity efficiency (99.8 %) and converts between the input and output modes within a very small footprint (roughly 1.5 square vacuum wavelengths).

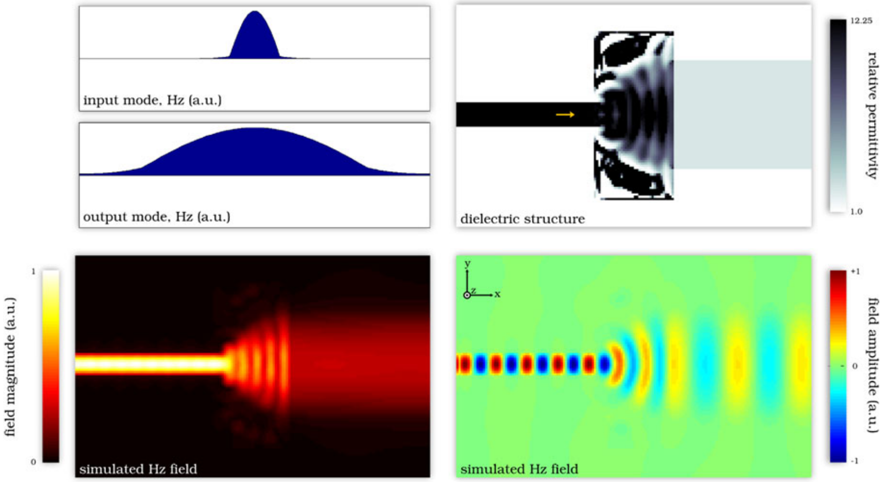


Fig. 6.2 Coupler to a wide low-index waveguide. Efficiency: 99.8 %, device footprint: 36 × 76 grid points, wavelength: 42 grid points

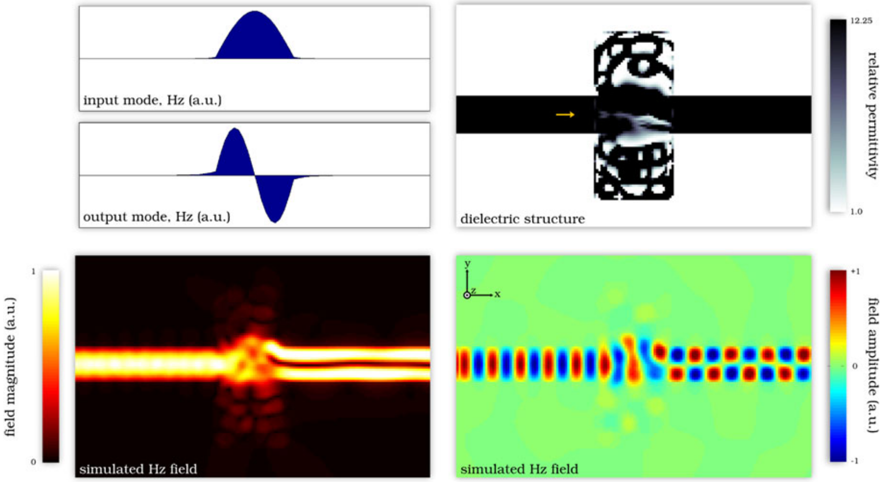


Fig. 6.3 Mode converter. Efficiency: 98.0 %, device footprint: 36 × 76 grid points, wavelength: 42 grid points

6.4.4 Mode Converter

In addition to coupling to a low-index waveguide, we show that we can successfully apply the objective-first method to convert between modes of a waveguide. We do this by simply selecting the output mode in the design objective to be the second-order waveguide mode, as seen in Fig. 6.3. Note that the design of this coupler

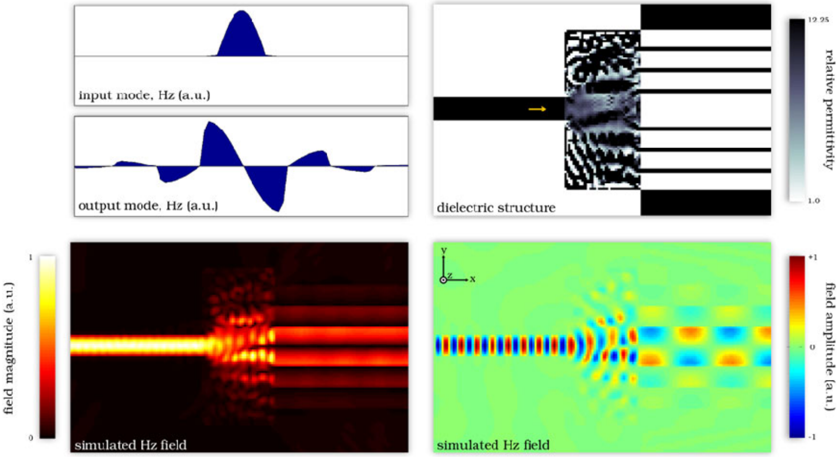


Fig. 6.4 Coupler to a wide low-index waveguide. Efficiency: 98.9 %, device footprint: 36 × 76 grid points, wavelength: 25 grid points

is made challenging because of the opposite symmetries of the input and output modes. Moreover, because our initial structure is symmetric, we initially have 0 % efficiency to begin with. Fortunately, the objective-first method can still design an efficient (98.0 %) coupler in this case as well in a footprint of less than two square vacuum wavelengths.

6.4.5 Coupling to an Air-Core Waveguide Mode

We can then continue to elucidate the generality of our method by coupling between waveguides which confine light in completely different ways. Figure 6.4 shows a high-efficiency coupling device between an index-guided input waveguide and a "air-core" output waveguide, in which the waveguiding effect is achieved using distributed Bragg reflection (instead of total internal reflection as in the input waveguide).

In this case, the device footprint is increased to 4.3 square vacuum wavelengths and the final efficiency is 98.9 %.

6.4.6 Coupling to a Metal–Insulator–Metal and Metal Wire Plasmonic Waveguides

Additionally, our design method can also generate couplers between different material systems such as between dielectric and metallic (plasmonic) waveguides, as

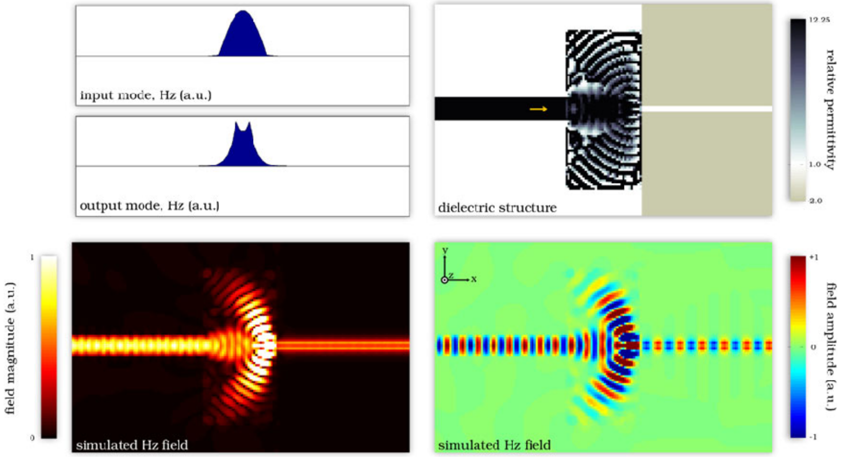


Fig. 6.5 Coupler to a plasmonic metal–insulator–metal waveguide. Efficiency: 97.5 %, device footprint: 36 × 76 grid points, wavelength: 25 grid points

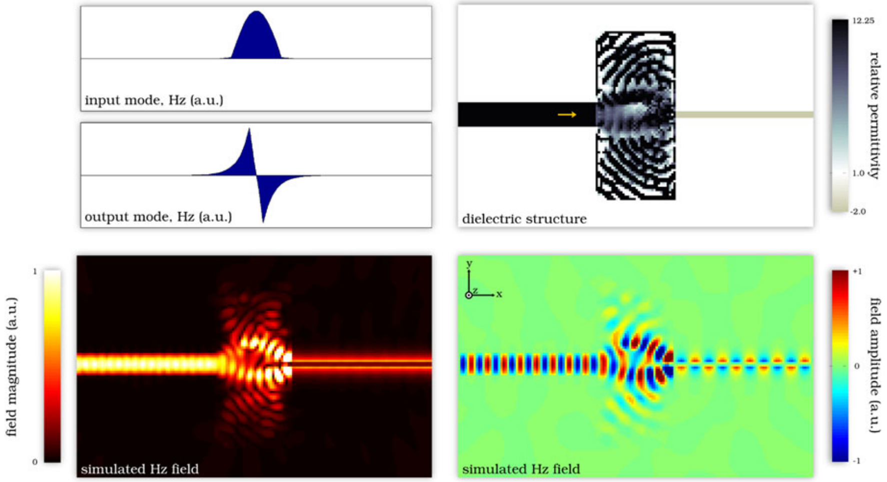


Fig. 6.6 Coupler to a plasmonic wire waveguide. Efficiency: 99.1 %, device footprint: 36 × 76 grid points, wavelength: 25 grid points

shown in Fig. 6.5 (97.5 % efficiency). In this case, the permittivity of the metal ($\varepsilon = -2$) is chosen to be near the plasmonic resonance ($\varepsilon = -1$).

Extending this method to include plasmonic wire waveguides, Fig. 6.6 shows that efficiently coupling to this type of structure is achievable as well (99.1 % efficiency).

6.5 Optical Cloak Design

In the previous section, we showed that couplers between virtually any two waveguide modes could be constructed using the objective-first design method, and based on the generality of the method one can guess that it may also be able to generate designs for any linear nanophotonic device. Now, we extend the applicability of our method to the design of metamaterial devices which operate in free-space. In particular, we adapt the waveguide coupler algorithm to the to the design of optical cloaks.

6.5.1 Application of the Objective-First Strategy

Adapting the method used in Sect. 6.4 to the design of optical cloaks really only requires one to change the simulation environment to allow for free-space modes. This is accomplished by modifying the upper and lower boundaries of the simulation domain from absorbing boundary conditions to periodic boundary conditions, which allows for plane-wave modes to propagate without loss until reaching the left or right boundaries, where absorbing boundary conditions are still maintained.

In terms of the design objective, we allow the device to span the entire height of the simulation domain, and thus consider only the leftmost and rightmost planes as boundary values. Specifically, for this section the input and output modes are plane waves with normal incidence, as can be expected for good cloaking devices. The achieved results all yield high efficiency, although we note that the cloaking effect is only measured for a specific input mode. That is to say, just as the waveguide couplers previously designed were single-mode devices, so the cloaks designed in this section are also “single-mode” cloaks. An additional modification, as compared to Sect. 6.4, is that we now prevent the structure from being modified in certain areas which contain the object to be cloaked. With these simple changes we continue to solve (6.17) with the alternating directions algorithm in order to design optical cloaks instead of waveguide couplers. Once again, as in Sect. 6.4, each design is run for 400 iterations with a uniform initial value of $p = 1/9$ for the structure (where the structure is allowed to vary), and the range of p is limited to $1/12.25 \leq p \leq 1$, implying a dielectric cloak.

6.5.2 Anti-reflection Coating

As a first example, we attempt to design the simplest and most elementary “cloaking” device available, which, we argue, is a simple anti-reflection coating; in which case the object to be cloaked is nothing more than the interface between two dielectric materials. In this case, we use the interface between air and silicon, as shown in Fig. 6.7.

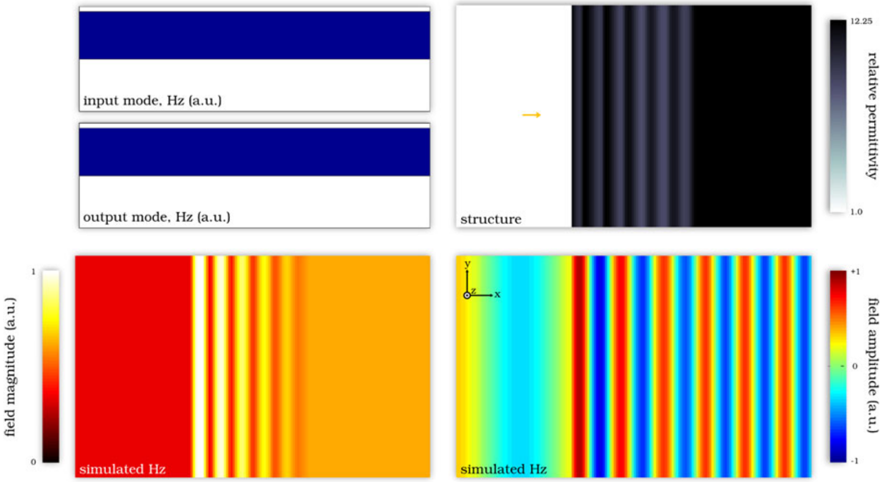


Fig. 6.7 Anti-reflection coating. Efficiency: 99.99 %, device footprint: 60 × 100 grid points, wavelength: 63 grid points

Unsurprisingly for such a simple case, we achieve a very high efficiency device. Note also that the efficiency of the device can be deduced by eye, based on the absence of reflections or standing waves in bottom two plots of Fig. 6.7.

6.5.3 Wrap-Around Cloak

Next, we design a cloak for a plasmonic cylinder, which is quite effective at scattering light as can be seen from Fig. 6.8, where we show that the uncloaked cylinder, although sub-wavelength in size, scatters the majority of light away from the desire output (plane-wave) mode.

In designing the wrap-around cloak, we allow the structure to vary at all points within the design area except in the immediate vicinity of the plasmonic cylinder. Application of the objective-first strategy results in an efficient (greater than 99 %) device as seen in Fig. 6.9. Note that our cloak employs only isotropic, non-magnetic materials, and at the same time it is specific to a particular input and to a particular object. That is to say, it is a single-frequency, single-mode, and single-object cloaking device.

6.5.4 Open-Channel Cloak

With a simple modification, from the previous example, we can design a cloak which features an open channel to the exterior electromagnetic environment (Fig. 6.10).

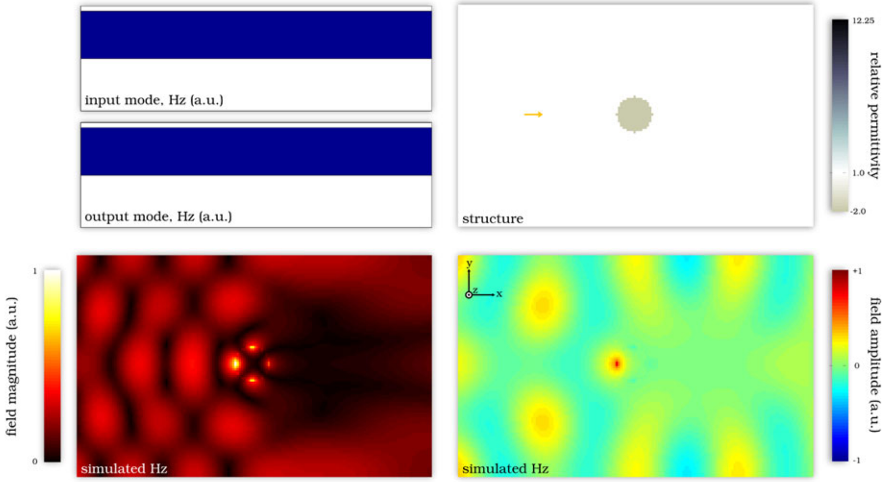


Fig. 6.8 Plasmonic cylinder to be cloaked. 68.5 % of light is diverted away from the desired output mode

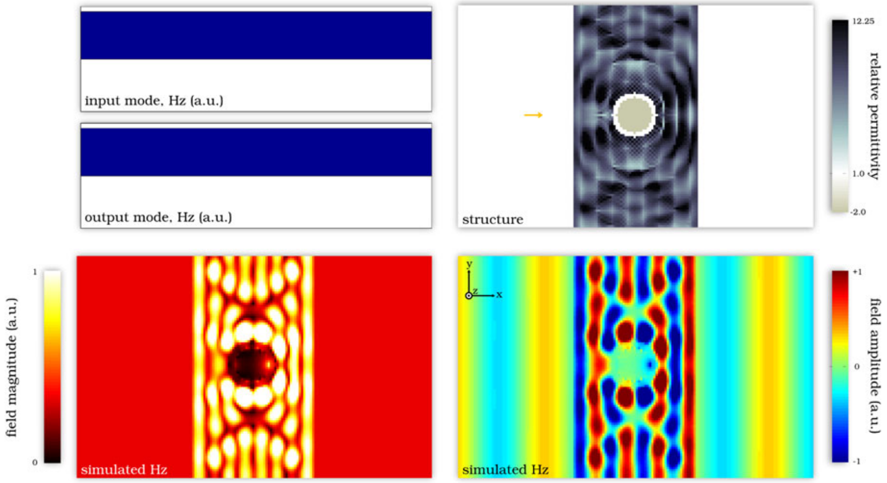


Fig. 6.9 Wrap-around cloak. Efficiency: 99.99 %, device footprint: 60 × 100 grid points, wavelength: 42 grid points

This simple modification creates an air gap that connects the cylinder to the outside world both from the front and back and may be useful in the case where one would like to remove or replace the cloaked object.

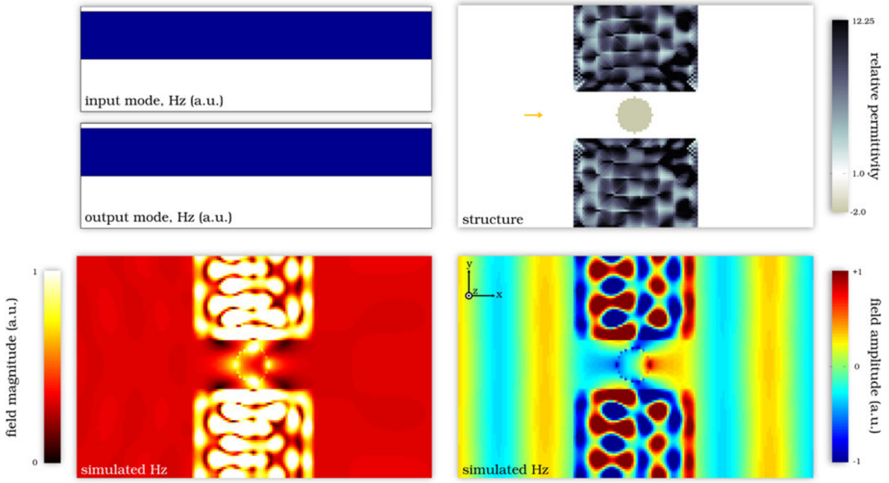


Fig. 6.10 Open-channel cloak. Efficiency: 99. 8 %, device footprint: 60 × 100 grid points, wavelength: 42 grid points

Such a design is still very efficient (greater than 99 % efficiency) and demonstrates the usefulness of the objective-first strategy in cases where other methods, such as transformation optics, may require use of the entire space around the object to be cloaked.

6.5.5 Channeling Cloak

Our last cloaking example replaces the plasmonic cylinder with a thin metallic wall in which a sub-wavelength channel is etched. Such a metallic wall is very effective at blocking incoming light (as can be seen from Fig. 6.11 where more than 99 % of the incoming light is blocked) because of its large negative permittivity ($\varepsilon = -20$), meaning that any cloaking device would be forced to channel all the input light into a very small aperture and then to flatten that light out into a plane wave again.

Once again, our method is able to produce a design with efficiency greater than 99 %, as shown in Fig. 6.12.

6.6 Optical Mimic Design

We now apply our objective-first strategy to the design of optical mimics. We define an optical mimic to be a linear nanophotonic device which mimics the output field of another device. In this sense optical mimics are anti-cloaks; where cloaks strive to make an object's electromagnetic presence vanish, mimics strive to implement an object's presence without that object actually being there.

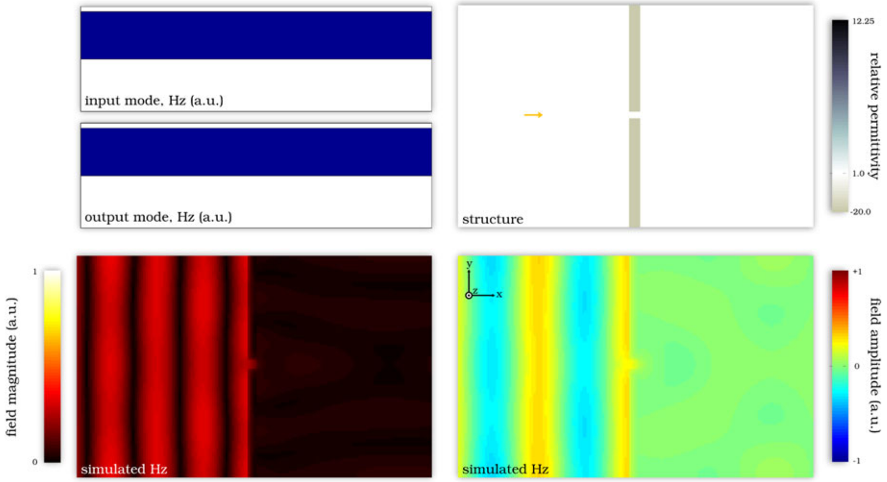


Fig. 6.11 Metallic wall with sub-wavelength channel to be cloaked. 99.9 % of the light is blocked from the desired output plane-wave

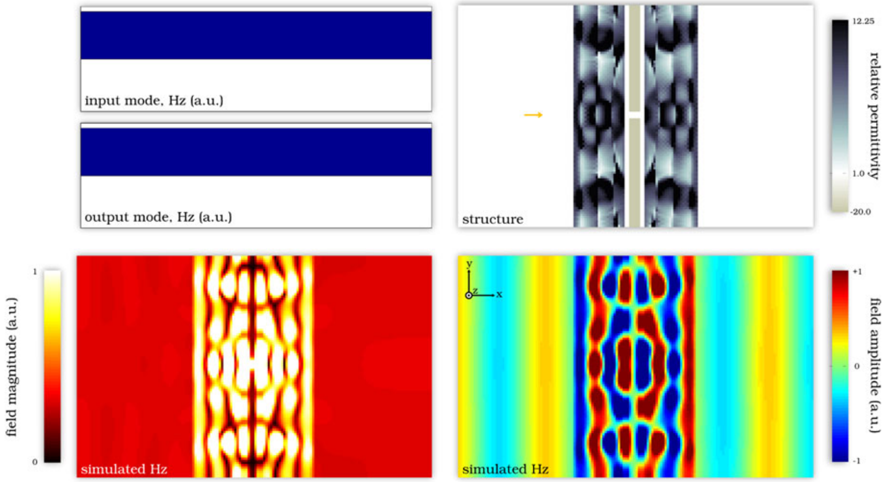


Fig. 6.12 Channeling cloak. Efficiency: 99.9 %, device footprint: 60 × 100 grid points, wavelength: 42 grid points

As such, the design of optical mimics provides a tantalyzing approach to the realization of practical metamaterial devices. That is to say, if one can reliably produce practical optical mimics, then producing metamaterials can be accomplished by simply producing an optical mimic of that material. In a more general sense, designing optical mimics is really just a recasting of the thrust of the objective-first design strategy in its purest form, namely the design of a nanophotonic device based

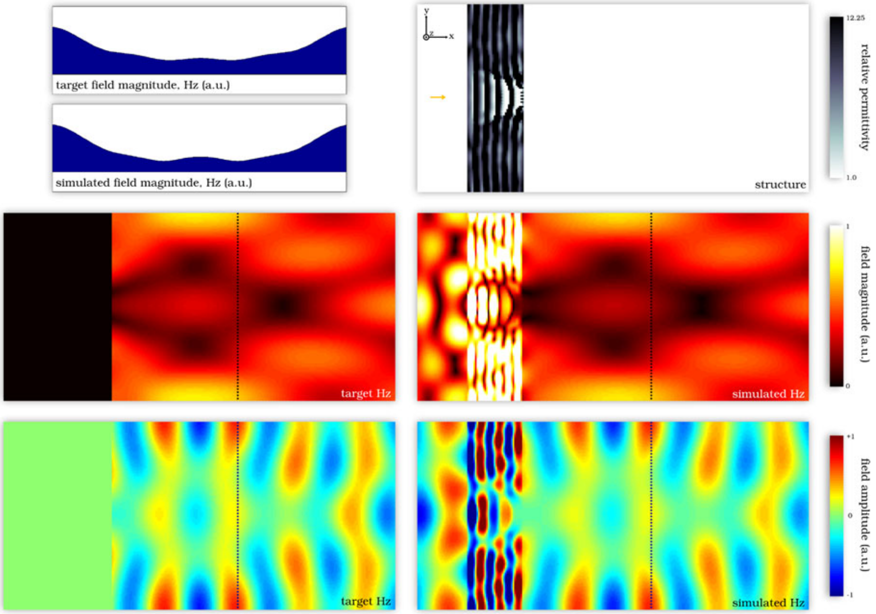


Fig. 6.13 Plasmonic cylinder mimic (see Fig. 6.8 for the original object). Error: 8.1 %, device footprint: 40 × 120 grid points, wavelength: 42 grid points

purely on the electromagnetic fields one wishes to produce. As such, devices which perform well-known optical functions (e.g., focusing, lithography) can also be designed.

6.6.1 Application of the Objective-First Strategy

The objective-first design of optical mimics proceeds in virtually an identical way to the design of optical cloaks, the only difference being that the output modes are specifically chosen to be those that produce the desired function. For most of the examples provided, the input illumination is still an incident plane wave. Lastly, instead of measuring efficiency, we measure the relative error of the simulated field against that of a perfect target field at a relevant plane some distance away from the device. The location of this plane is identified as a dotted line in the subsequent figures.

6.6.2 Plasmonic Cylinder Mimic

Our first design is simply to mimic the plasmonic cylinder which we cloaked in the previous section. Figure 6.13 shows the result of the design.

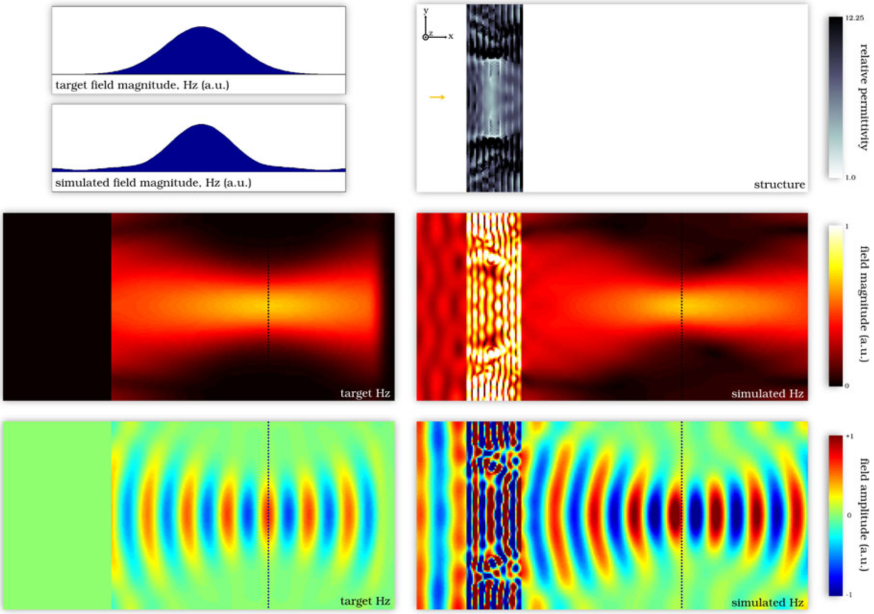


Fig. 6.14 Full-width-half-max at focus: 1.5 λ, focus depth: 100 grid points. Error: 12.0 %, device footprint: 40 × 120 grid points (1.6 λ thick), wavelength: 25 grid points

The final structure is shown in the upper right plot, while the ideal field and the simulated field are shown in the middle and bottom plots. Note that the ideal field is cut off to emphasize the fields to the right of the device (the output fields). Also, the magnitude of the fields are compared at the dotted black line at which point the relative error is also calculated. For this simple, initial mimic, the simulated field quite closely imitates that produced by a single plasmonic cylinder (8.1 % error)

6.6.3 *Diffraction-Limited Lens Mimic*

We now design a mimic for a typical diffraction-limited lens. In this case, the object which we wish to mimic does not require simulation since the fields of a lens can be readily computed. For the three figures below, the computed ideal fields are shown as the target fields.

Figure 6.14 shows the mimic of a lens with relatively moderate focusing. In such a lens, the focusing action is gradual and easily discernible by eye. The computed error in this case is 12.0 %.

In contrast, Figs. 6.15 and 6.16 are both mimics of a lens with a smaller half-wavelength spot size. Such a lens is much harder to design because of the high-frequency spatial components involved; and yet, we show that an objective-first

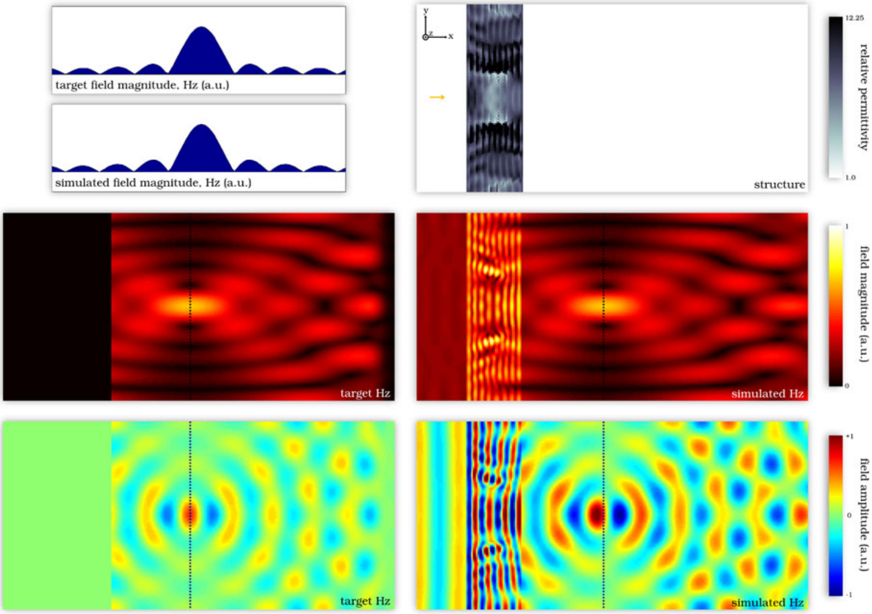


Fig. 6.15 Full-width-half-max at focus: 0.5 λ, focus depth: 50 grid points. Error: 5.6 %, device footprint: 40 × 120 grid points (1.6 λ thick), wavelength: 25 grid points

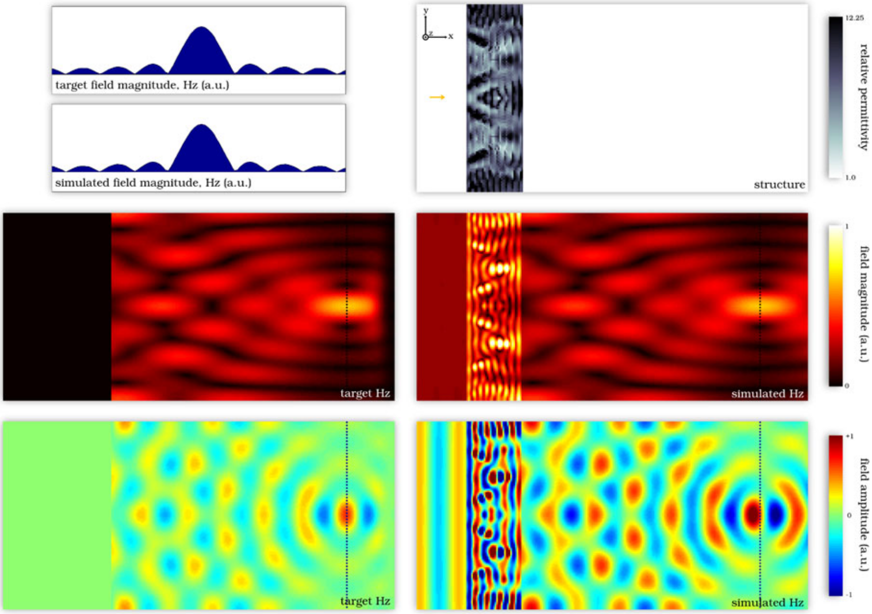


Fig. 6.16 Full-width-half-max at focus: 0.5 λ, focus depth: 150 grid points. Error: 1.4 %, device footprint: 40 × 120 grid points (1.6 λ thick), wavelength: 25 grid points

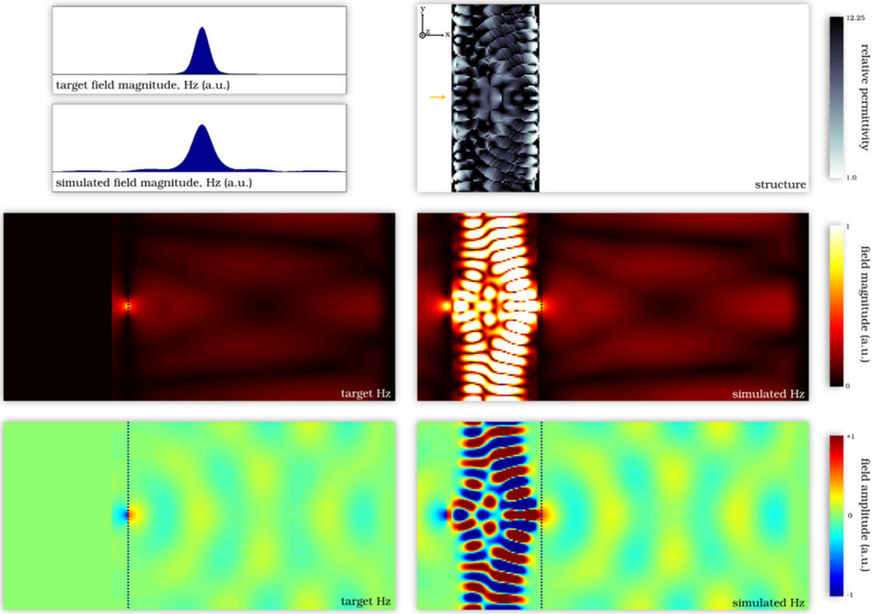


Fig. 6.17 Sub-diffraction lens mimic. The target field has a full-width half-maximum of 0.14 λ. Error: 28.6 %, device footprint: 60 × 120 grid points (1.43 λ thick), wavelength: 42 grid points

strategy can produce successful designs (5.6 % and 1.4 % error) and that this is achievable at both shorter and longer focal depths.

6.6.4 Sub-diffraction Lens Mimic

Our method is now employed to mimic the effect of a sub-diffraction lens. Since such a lens can be created using a negative-index material [12], this mimic can be viewed as an imitation of a negative-index material, in that the following device recreates the sub-diffraction target-field at the output plane (dotted line) when illuminated by the same target field at the input of the device. In other words, this device is an image-specific sub-diffraction imager, which is another way of saying that it is a single-mode imager.

As Fig. 6.17 shows, we are able to recreate the target field at the output, albeit with higher error (28.6 %). Although the error in this example is larger, the field produced by the device has a full-width half-maximum nearly equal to that of the target field.

Note that the target field is created simply by placing the imaging field at the output plane. Also note that, as expected, the output field decays very quickly since, for such a deeply sub-wavelength field, it is composed primarily of evanescently decaying modes.

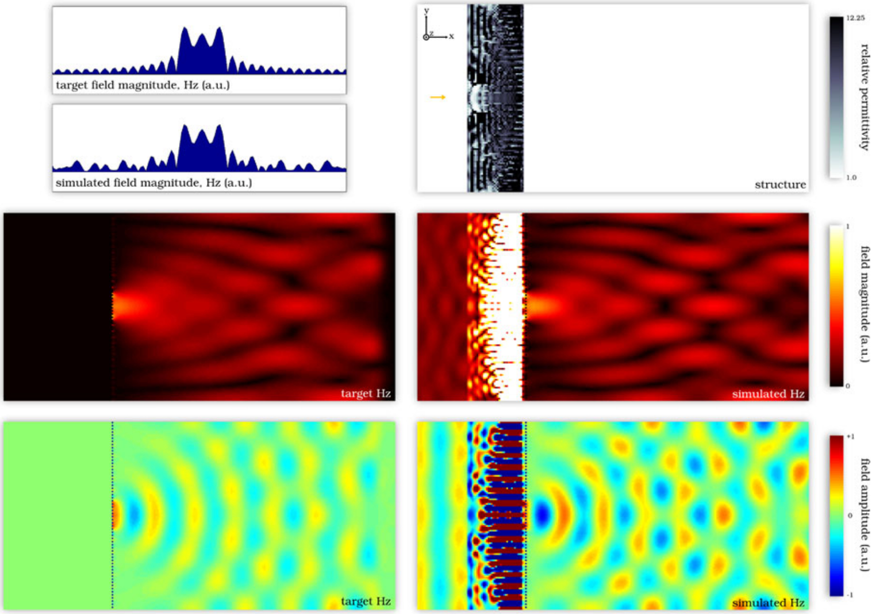


Fig. 6.18 Sub-diffraction optical mask. The three central peaks in the target field are each separated by 0.28 λ. Error: 19.8 %, device footprint: 40 × 120 grid points, wavelength: 25 grid points

6.6.5 Sub-diffraction Optical Mask

Lastly, we extend the idea of a sub-diffraction lens mimic one step further and design a sub-diffraction optical mask. Such a device takes a plane wave as its input and produces a sub-diffraction image at its output plane. Of course, akin to its lens counterpart, this output plane must lie within the near-field of the device (specifically, two computational cells away) because of its sub-wavelength nature. Figure 6.18 shows the design of a simple mask which successfully produces three peaks at its output with an error of 19.8 %.

6.7 Extending the Method

The objective-first method, as applied in the examples in this chapter, represents only a small foray into the area of nanophotonic design. Several key extensions to what is presented here are needed to fully address real-world nanophotonic design challenges.

6.7.1 Three-Dimensional Design

The first of these is the need to design fully three-dimensional structures. Doing so provides no inherent difficulties aside from the matrices in (6.6) becoming very large. This is not insurmountable as electromagnetic simulation software for three-dimensional nanophotonic structures already exists.

In fact, for certain choices of the design objective (i.e., those of low-rank) (6.18) can be efficiently solved by a small number of calls to unmodified simulation software. Of course, for general design objectives, such software will need to be modified in order to solve (6.18).

On the other hand, specialized software to solve (6.10) in any number of dimensions does not exist, although this was not a problem in two dimensions since generic linear algebra solvers are more than accurate. In three dimensions, the large size of matrix $B(x)$ can be greatly compressed by considering only fabrication processes which modify a structure in-plane. In this way, the degrees of freedom in p can be greatly reduced and the original methods used in this chapter can still be applied. This work-around is especially appealing since in-plane structures are significantly less challenging to fabricate.

6.7.2 Multi-mode

A second necessary extension is to be able to consider the multiple fields that a structure produces in response to input fields of differing frequency and spatial distribution. Such an extension is straightforward in the objective-first formulation and results in the following modified problem statement,

$$\begin{aligned} \underset{x_i, p}{\text{minimize}} \quad & \sum_i \left\| A(p)x_i - b(p) \right\|^2 \\ \text{subject to} \quad & f(x_i) = f_{i,\text{ideal}}, \quad i = 1, \ldots, n, \\ & p_0 \le p \le p_1, \end{aligned} \tag{6.24}$$

which can be separated into field and structure sub-problems as in the single-mode formulation. In the multi-mode case, this results in one structure sub-problem and n field sub-problems. Interestingly, the n field sub-problems lend themselves naturally to parallelization since they can be solved independently, leading to the possibility that a multi-mode design completing in roughly the same time as a single-mode design.

6.7.3 Binary Structure

Another necessary extension of our method is to force the values of p to be discrete. This is not trivial since a naive restatement of (6.17) which includes such a constraint,

$$\begin{aligned} &\underset{x,p}{\text{minimize}} \ \left\| A(p)x - b(p) \right\|^2 \\ &\text{subject to } f(x) = f_{\text{ideal}}, \\ &\qquad\qquad p \in \{p_0, p_1\}, \end{aligned} \tag{6.25}$$

results in a very difficult combinatorial problem.

Tractable approaches include penalizing intermediate values of p [13] or even transferring to a level-set method [9] where the distinction between materials is explicit.

6.7.4 Robustness

Lastly, the design of structures which are robust to both fabrication imperfections and fluctuations in environmental parameters is also a necessity for practical real-world devices.

It seems likely in this case that a heuristic approach may be most successful in this case, rather than tackling the problem head-on. For instance, to account for fluctuating material parameters induced by temperature changes one may design a device that operates over a larger bandwidth than is actually required.

6.8 Conclusions

We have introduced an objective-first approach to the design of nanophotonic components, and applied it to the design of waveguide couplers, optical cloaks, and optical mimics. In doing so, we hope to have exhibited both the simplicity and the breadth of our method to the design of a broad class of linear, single-mode devices. In addition to posting the source code for all the examples online [2], we have outlined the necessary extensions of our method in order to design practical, three-dimensional devices.

References

1. J. Lu, J. Vuckovic, Objective-first design of high-efficiency, small-footprint couplers between arbitrary nanophotonic waveguide modes. Opt. Express **20**, 7221–7236 (2012)
2. https://github.com/JesseLu/objective-first
3. A. Taflove, S.C. Hagness, *Computational Electrodynamics*, 3rd edn. (Artech House, Norwook, 2005), Sect. 3.3.2
4. K.S. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 802–807 (1966)

5. S.G. Johnson, Notes on perfectly matched layers (PMLs) (2010). http://math.mit.edu/~stevenj/18.369/pml.pdf
6. W. Shin, S. Fan, Choice of the perfectly matched layer boundary condition for frequency-domain Maxwells equations solvers. J. Comput. Phys. **231**, 3406–3431 (2012)
7. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
8. M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 1.21 (2011). http://cvxr.com/cvx
9. O.D. Miller, *Photonic Design: From Fundamental Solar Cell Physics to Computational Inverse Design* (U. C. Berkeley, Berkeley, 2012)
10. J. Lu, J. Vuckovic, Inverse design of nanophotonic structures using complementary convex optimization. Opt. Express **18**, 3793–3804 (2011)
11. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–1122 (2011)
12. J.B. Pendry, Negative refraction makes a perfect lens. Phys. Rev. Lett. **85**, 3966–3969 (2000)
13. M.P. Bendsoe, O. Sigmund, Material interpolation schemes in topology optimization. Arch. Appl. Mech. **69**, 635–654 (1999)

Chapter 7
Gradient Based Optimization Methods for Metamaterial Design

Weitao Chen, Kenneth Diest, Chiu-Yen Kao, Daniel E. Marthaler, Luke A. Sweatlock, and Stanley Osher

Abstract The gradient descent/ascent method is a classical approach to find the minimum/maximum of an objective function or functional based on a first-order approximation. The method works in spaces of any number of dimensions, even in infinite-dimensional spaces. This method can converge more efficiently than methods which do not require derivative information; however, in certain circumstances the "cost function space" may become discontinuous and as a result, the derivatives may be difficult or impossible to determine. Here, we discuss both level set methods and eigenfunction optimization for representing the topography of a dielectric environment and efficient techniques for using gradient methods to solve different material design problems. Numerous results are shown to demonstrate the robustness of the gradient-based approach.

7.1 Introduction

In this chapter, we introduce methods for representing the topography of a dielectric environment and efficient techniques for using gradient methods to solve different material design problems. We first discuss the level set method introduced by Osher and Sethian [19]. This method is well known for its flexibility in capturing the interface between two phases and its ability to handle topology changes, such as breaking one component into several, merging several components into one, and forming sharp corners. An example of this is shown in Fig. 7.1. In panel (a), we see the level set function represented by the multi-colored surface, along with the yellow, horizontal plane that could be set at any height. In panel (b), we see the corresponding boundary regions which represent the zero level set of the function.

The level set method has been used on numerous occasions to study shape optimization problems. This approach is fundamentally different than those discussed in Chaps. 4 and 5, which utilize discrete parameterization of the geometries under consideration. Such methods inherently exhibit extreme discontinuities in "cost

K. Diest (✉)
Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA 02420, USA
e-mail: diest@mit.edu

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8_7,

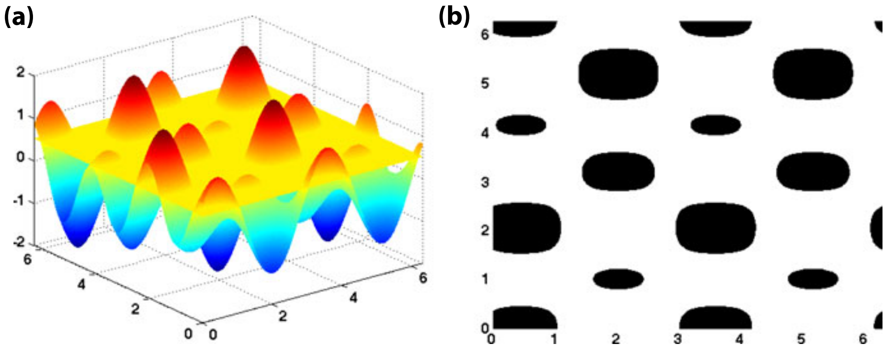


Fig. 7.1 Example of a level set function (**a**) and a corresponding zero level set cross-section (**b**)

function space" when such topology changes occur. Examples include antenna elements within an array merging together or the gap in a split ring resonator completely closing. Instead of using a physically driven velocity, the level set method typically moves surfaces by the gradient flow of an objective energy functional. For this approach to be successful, this method inherently requires the determination of such a gradient flow. For certain types of physical systems, such as fluid flow, these types of gradients can be calculated analytically; however, within the field of electromagnetics, determining such gradients is still an ongoing field of research.

Previously this gradient flow was computed based on shape derivatives [4, 11, 14, 18, 21]; however, as pointed out in [1, 2], the level set approach based on shape sensitivity may converge to a domain with fewer holes than the optimal geometry in some structure design applications. To address this issue, modified level set approaches based on topological derivatives are proposed in [3, 5]. By incorporating topological derivatives into the level set method, one provides an alternative way to create holes efficiently and thereby avoids converging to a domain with fewer holes than optimal.

The approach based on shape and/or topological derivatives has been applied to the study of extremum problems of eigenvalues of inhomogeneous structures, including identification of composite membranes with extremum eigenvalues [10, 18, 27], design of composite materials with a desired spectral gap or maximal spectrum gap [14], finding optical devices that have a high quality factor (low loss of energy) [15], and principle eigenvalue optimization in population biology [13]. For simplicity, we will only discuss the approach based on shape derivative here. As an example of how the method can be employed, we will discuss the optimization of band gaps in photonic crystals.

The second design method discussed in this chapter is based on the localization of energy with Dirichlet boundary conditions, as proposed by Dobson and Santosa [9]. The optimized two-dimensional configurations shown in their paper possess "bang-bang" structure which suggests that the optimal structure has two phases. In order to demonstrate the approach and illustrate the optimal configuration more precisely, we first study the same problem in one dimension and show that the optimal structure can not only have a point (oddly symmetric) defect, but also an evenly symmetric defect structure. The defect type depends on whether the eigenvalue is an odd or even mode. In two dimensions, these two different defect structures are also observed when the weighting function is chosen as the square of the distance function from the localization location.

This approach is extended to electromagnetic designs where the eigenfunction is localized at several points, or on a specified curve, by adjusting the weight function to be the square of the distance function from the points or the specified curves. We also consider more complicated cases; structures that can simultaneously support multiple resonances at multiple wavelengths. The multiple eigenmodes are confined at different locations and/or the corresponding eigenvalues are resonant near predetermined wavelengths by incorporating penalty methods. We further extend the problem to include Dirichlet–Bi-Laplacian problems. Many numerical results are shown to demonstrate the robustness of the gradient approach. These results directly translate into the design of two-dimensional frequency selective devices. Also, the methods are applicable for engineering nanoscale resonant antennas where the ability to confine and focus light at specific locations and wavelengths is critical to the overall device performance. The optimal configurations possess “bang-bang” structure which suggests that the optimal structures have two phases. As a result, we limit our analysis of design solutions to those that satisfy this criteria.

7.2 Level Sets and Dynamic Implicit Surfaces

The level set method is well known for its flexibility in capturing shape and topology changes. Since the dielectric distribution is of bang-bang control, i.e., $\varepsilon(x) = \varepsilon_1 \chi_{D\setminus\Omega} + \varepsilon_2 \chi_\Omega$ (χ denotes the indicator function of a set), the level set method can be easily adapted to represent the two-phase dielectric distribution. The zero level set is used to represent the material boundary $\partial\Omega$. With an initial configuration, we compute the corresponding eigenvalues and eigenfunctions. Based on this input, we can use the shape derivative to move the level set function in the direction of the gradient of the objective function $\mathcal{F}(\Omega)$, where Ω is a simply connected, bounded domain in $\mathbb{R}^N$. After the level set is updated, the same process is repeated iteratively until the level set reaches the optimal shape. In the following section, we provide a detailed discussion of the approach with an example of the optimization of bandgaps in photonics crystals.

7.2.1 Finding the Maximal Bandgap in Photonic Crystals

Photonic crystals are periodic dielectric structures which manipulate photons and control optical properties, such as completely preventing the propagation of light, allowing it only in certain directions at certain frequencies, or localizing light in specified areas. Photonic crystals with bandgaps have many applications including: reflection coatings on lenses and mirrors, color changing paints and inks, waveguides, and light confinement devices. They were first studied by Rayleigh in 1887 for one-dimensional layered structures. Later, Yablonovitch [25] and John [12] in 1987 introduced the concepts of photonic bandgaps in two and three dimensions. Yablonovitch [25] investigated the influence of spontaneous emission in the solid state devices, while John [12] studied the localization of light.

In this section, we summarize the methodology used in solving the maximal spectrum gap problem for photonic crystal design. We assume that there are no electric currents or charge and the electromagnetic waves are monochromatic, i.e., $E(x,t) = E(x)e^{-i\omega t}$ and $H(x,t) = H(x)e^{-i\omega t}$. In this context, Maxwell's equations are reduced to the following decoupled system:

$$\frac{1}{\varepsilon(x)} \nabla \times \big(\nabla \times E(x)\big) = \frac{\omega^2}{c^2} E(x), \tag{7.1}$$

$$\nabla \times \frac{1}{\varepsilon(x)} \big(\nabla \times H(x)\big) = \frac{\omega^2}{c^2} H(x), \tag{7.2}$$

where ε is the dielectric function, $\varepsilon(x) = \varepsilon_1 \chi_{D\setminus\Omega} + \varepsilon_2 \chi_\Omega$, and c is the speed of light. We denote the wavelength of operation as $\lambda = \frac{w^2}{c^2}$. The maximum spectral gap problem for photonic crystal design is to find the structures having the maximal spectrum gap between bands k and $k+1$:

$$\max_{\Omega} (\lambda_{k+1} - \lambda_k)$$

while the desired spectral gap problem is

$$\min\ \text{area}(\Omega) \quad \text{s.t. } |\lambda_{k+1} - \lambda_k| = const.$$

These two problems are very challenging and require three-dimensional simulations. The physical meaning of a bandgap is that for such a range of frequencies, light from any incident angle is totally reflected. Under the assumption that the medium is isotropic, the magnetic permeability is constant, and dielectric function $\varepsilon(x)$ is periodic (i.e., $\varepsilon(x + R_i) = \varepsilon(x)$ for some primitive lattice vectors R_i), Bloch's theorem can be applied, and then the governing equations for both TM (transverse magnetic field) and TE (transverse electric field) light can be written as

$$-\frac{1}{\varepsilon(x)} (\nabla + i\alpha) \cdot \big((\nabla + i\alpha) E_\alpha\big) = \frac{\omega_{\text{TM}}^2}{c^2} E_\alpha, \tag{7.3}$$

$$-(\nabla + i\alpha) \cdot \left(\frac{1}{\varepsilon(x)} (\nabla + i\alpha) H_\alpha\right) = \frac{\omega_{\text{TE}}^2}{c^2} H_\alpha, \tag{7.4}$$

where α is a wave vector in the irreducible Brillouin zone. The optimization problems for the bandgap G are:

(i) Maximize the bandgap in TM

$$\sup_{\Omega} G_{\mathrm{TM}} = \sup_{\Omega}\Big(\inf_{\alpha} \omega_{\mathrm{TM}}^{k+1} - \sup_{\alpha} \omega_{\mathrm{TM}}^{k}\Big);$$

(ii) Maximize the bandgap in TE

$$\sup_{\Omega} G_{\mathrm{TE}} = \sup_{\Omega}\Big(\inf_{\alpha} \omega_{\mathrm{TE}}^{k+1} - \sup_{\alpha} \omega_{\mathrm{TE}}^{k}\Big).$$

7.2.2 *Definition of the Level Set Function*

Following the work in [19], we use a level set function $\phi(x)$ to represent the unknown set Ω:

$$\Omega := \big\{x : \phi(x) < 0\big\}$$

and then

$$\varepsilon(x) = \begin{cases} \varepsilon_2 & \text{if } \phi(x) < 0, \\ \varepsilon_1 & \text{if } \phi(x) > 0. \end{cases}$$

In this way, we can optimize the objective with respect to the level set function. If the normal velocity of the boundary Ω is V_n, the level set function satisfies

$$\phi_t + V_n|\nabla\phi| = 0. \tag{7.5}$$

In order to start from an initial guess for the shape and morph it to the optimal shape via the level set method, we need to determine what kind of shape and topological changes can lead to the optimal shape and how to derive the normal velocity. This leads to the study of shape derivatives and topological derivatives. For simplicity, we only discuss the shape derivative which measures the sensitivity of boundary perturbation.

7.2.3 *Shape Derivatives*

The shape derivative [17, 23] is defined as the following: Let $\Omega \subset D \subset R^N$ be a reference domain. Consider the perturbation under the map $\theta \in W^{1,\infty}(R^N, R^N)$ s.t. $\|\theta\|_{W^{1,\infty}} < 1$

$$\Omega_\theta = (I + \theta)\Omega,$$

where I is the identity map. The set Ω_θ is defined as

$$\Omega_\theta = \big\{x + \theta(x) \mid x \in \Omega\big\}.$$

The shape derivative of an objective shape functional, $\mathcal{F}: R^N \to R$, at Ω is defined as the Frechet differential of $\theta \to \mathcal{F}(\Omega_\theta)$ at 0. Here θ can be viewed as a vector field affecting the reference domain Ω. The shape derivative $d_S\mathcal{F}(\Omega)(\theta)$ depends only on $\theta \cdot n$ on the boundary $\partial\Omega$ because the shape of Ω does not change at all if θ is lying on the tangential direction of the domain Ω. For example, when an objective functional that is the integral on the volume of Ω, i.e.,

$$\mathcal{F}(\Omega) = \int_\Omega f(x)\,dx, \tag{7.6}$$

the shape derivative is

$$d_S\mathcal{F}(\Omega)(\theta) = \int_{\partial\Omega} \theta \cdot n f(x)\,dx. \tag{7.7}$$

In order to maximize the volume integral (7.6), one can choose the gradient flow as

$$V_n = \theta = n f(x)$$

where V_n is the normal velocity to evolve the level set function in (7.5) which results in $d_S\mathcal{F}(\Omega)(\theta) > 0$ if $f(x) \neq 0$ before reaching the extremum.

The second example that we give here is the maximization of the first eigenvalue of the elliptic eigenvalue problem

$$\begin{cases} -\Delta u(x) = \lambda\varepsilon(x)u(x) & \text{for } x \in D, \\ u(x) = 0 & \text{for } x \in \partial D, \end{cases}$$

where $\varepsilon(x) = \varepsilon_1\chi_{D\setminus\Omega} + \varepsilon_2\chi_\Omega$. The first eigenvalue can be represented by the Rayleigh Quotient formula

$$\lambda_1 = \min_u \frac{\int_D |\nabla u_1|^2}{\int_D \varepsilon u_1^2} = \min_u \frac{\int_D |\nabla u_1|^2}{(\int_{D\setminus\Omega} \varepsilon_1 u_1^2 + \int_\Omega \varepsilon_2 u_1^2)},$$

where u_1 is the first eigenfunction. The shape derivative becomes

$$d_S\lambda_1(\Omega)(\theta) = -\frac{\lambda_1(\varepsilon_2 - \varepsilon_1)}{\int_D \varepsilon u_1^2} \int_{\partial\Omega} \theta \cdot n u_1^2\,dx. \tag{7.8}$$

Thus the gradient direction to increase the first eigenvalue is

$$V_n = \theta = -\lambda_1(\varepsilon_2 - \varepsilon_1)u_1^2 n.$$

In general, the shape derivative depends only on the boundary $\partial\Omega$. After the shape derivative is computed, the gradient flow can be chosen to optimize the objective functional. Suppose the shape derivative of an objective function is

$$d_S\mathcal{F}(\Omega)(\theta) = \int_{\partial\Omega} \theta \cdot n\, W(\Omega)(x)\,dx, \tag{7.9}$$

then to maximize the objective functional $\mathcal{F}(\Omega)$, we choose the gradient flow as $\theta = W(\Omega)(x)n(x)$. This means that the normal velocity of the shape is $W(\Omega)(x)$. When we use the zero level set of function ϕ to represent the boundary of Ω, the motion under the normal velocity $W(\Omega)(x)$ is simply

$$\phi_t + W(\Omega)(x)|\nabla\phi| = 0. \tag{7.10}$$

Notice that the shape derivative is only defined on the boundary $\partial\Omega$; however, under the level set framework, it has to be defined on the whole domain D. We naturally extend it to D by using the $W(\Omega)(x)$.

7.2.4 Normal Velocity Formulas in Photonic Crystals

To solve the optimization problems (i) and (ii) in Sect. 7.2.1, there are several obstacles in deriving the classical gradient due to nondifferentiability of the bandgap functions. In these circumstances, the generalized gradients are used. Denote the convex hull by co. The generalized gradients [6–8] with respect to Ω can be written as follows

$$\partial_\Omega \omega_{\mathrm{TM}}^k \subset co\left\{-\frac{1}{2}(\varepsilon_2-\varepsilon_1)\omega_{\mathrm{TM}}^k|u|^2 : u \in \Upsilon_{\mathrm{TM}}^k(\varepsilon,\alpha)\right\}, \tag{7.11}$$

$$\partial_\Omega \omega_{\mathrm{TE}}^k \subset co\left\{\frac{1}{2\omega_{\mathrm{TE}}^k}\left(\frac{1}{\varepsilon_2}-\frac{1}{\varepsilon_1}\right)\left|(\nabla+i\alpha)v\right|^2 : v \in \Upsilon_{\mathrm{TE}}^k(\varepsilon,\alpha)\right\}, \tag{7.12}$$

where Υ_{TM}^k (and Υ_{TE}^k) are the span of all eigenfunctions u (and v) associated with the eigenvalues λ_{TM}^k (and λ_{TE}^k, respectively) and satisfying the normalization $\int_D \varepsilon|u|^2 = 1$ and $\int_D |v|^2 = 1$. The corresponding velocities which give the ascent direction for the optimization are

$$\begin{aligned} V_{\mathrm{TM}} = {} & co\left\{-\frac{1}{2}(\varepsilon_2-\varepsilon_1)\omega_{\mathrm{TM}}^{n+1}|u|^2 : u \in \Upsilon_{\mathrm{TM}}^{n+1}(\varepsilon,\alpha)\right\} \\ & - co\left\{-\frac{1}{2}(\varepsilon_2-\varepsilon_1)\omega_{\mathrm{TM}}^{n}|u|^2 : u \in \Upsilon_{\mathrm{TM}}^{n}(\varepsilon,\alpha)\right\}, \end{aligned} \tag{7.13}$$

$$\begin{aligned} V_{\mathrm{TE}} = {} & co\left\{\frac{1}{2\omega_{\mathrm{TE}}^{m+1}}\left(\frac{1}{\varepsilon_2}-\frac{1}{\varepsilon_1}\right)\left|(\nabla+i\alpha)v\right|^2 : v \in \Upsilon_{\mathrm{TE}}^{m+1}(\varepsilon,\alpha)\right\} \\ & - co\left\{\frac{1}{2\omega_{\mathrm{TE}}^{m}}\left(\frac{1}{\varepsilon_2}-\frac{1}{\varepsilon_1}\right)\left|(\nabla+i\alpha)v\right|^2 : v \in \Upsilon_{\mathrm{TE}}^{m}(\varepsilon,\alpha)\right\}. \end{aligned} \tag{7.14}$$

7.2.5 The Level Set Optimization Algorithm

We summarize the level set method for optimization in Algorithm 7.1. To implement the algorithm above, we choose the relative permittivity $\varepsilon = \varepsilon_2/\varepsilon_1 = 11.4$

Algorithm 7.1: Level set optimization algorithm

```
input: Initial guess for the level set function $\phi(x)$ s.t. $\Omega := \{x : \phi(x) < 0\}$
repeat
    Solve the elliptic eigenvalue problem (7.3) or (7.4)
    Compute the gradient direction by using shape derivative (7.13) or (7.14)
    Update $\phi(x)$ by using (7.5)
until stopping criterion is met
```

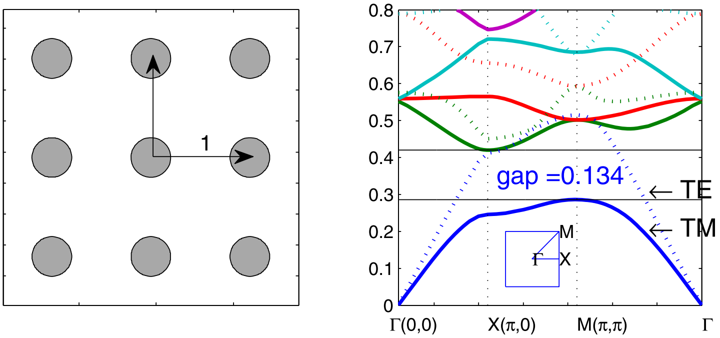


Fig. 7.2 *Left*: A two-dimensional square lattice of dielectric columns with relative permittivity $\varepsilon_2/\varepsilon_1 = 11.4$. *Right*: The corresponding bandgap structure. The *solid lines* represent TM modes and the dashed lines represent TE modes. The *horizontal axis* is wave vector α and the *vertical axis* is $\omega/2\pi c$. The *bottom inset* shows the Brillouin zone, with the irreducible zone as the triangular wedge. The three special points Γ, X, and M correspond to $\alpha = (0, 0)$, $\alpha = (\pi, 0)$ and $\alpha = (\pi, \pi)$, respectively

which corresponds to gallium arsenide in the air. We consider a photonic crystal which is made using a square lattice and has rotation, mirror-reflection, and inversion symmetry. In all numerical simulations, the computational domain is a unit square $\Omega = [-0.5, 0.5] \times [-0.5, 0.5]$ and the mesh sizes are a $\frac{1}{64}(64 \times 64$ grid). A 3×3 array of the unit lattice is shown for clarity. The white color indicates the low dielectric constant $\varepsilon = 1$ while the gray color indicates the high dielectric constant $\varepsilon = 11.4$.

Figure 7.2 shows a two-dimensional square lattice of dielectric columns with a relative permittivity $\varepsilon_2/\varepsilon_1 = 11.4$. Since the lattice is square, the irreducible Brillouin zone is the triangular wedge in the upper-right corner of the first Brillouin zone $K = [-\pi, \pi]^2$, which is the inset in the bottom. The frequency for the TM and TE modes with respect to different α is plotted as solid and dashed lines, respectively. We can see that there is a bandgap in the TM mode between ω^1_{TM} and ω^2_{TM}. By changing the dielectric distribution, a bandgap in the TE mode can also be generated. The extensive numerical results listed in [14] show that a lattice of isolated high ε regions is preferred for the TM mode while a lattice of connected high ε regions is preferred for TE mode. It is also possible to design a photonic crystal that

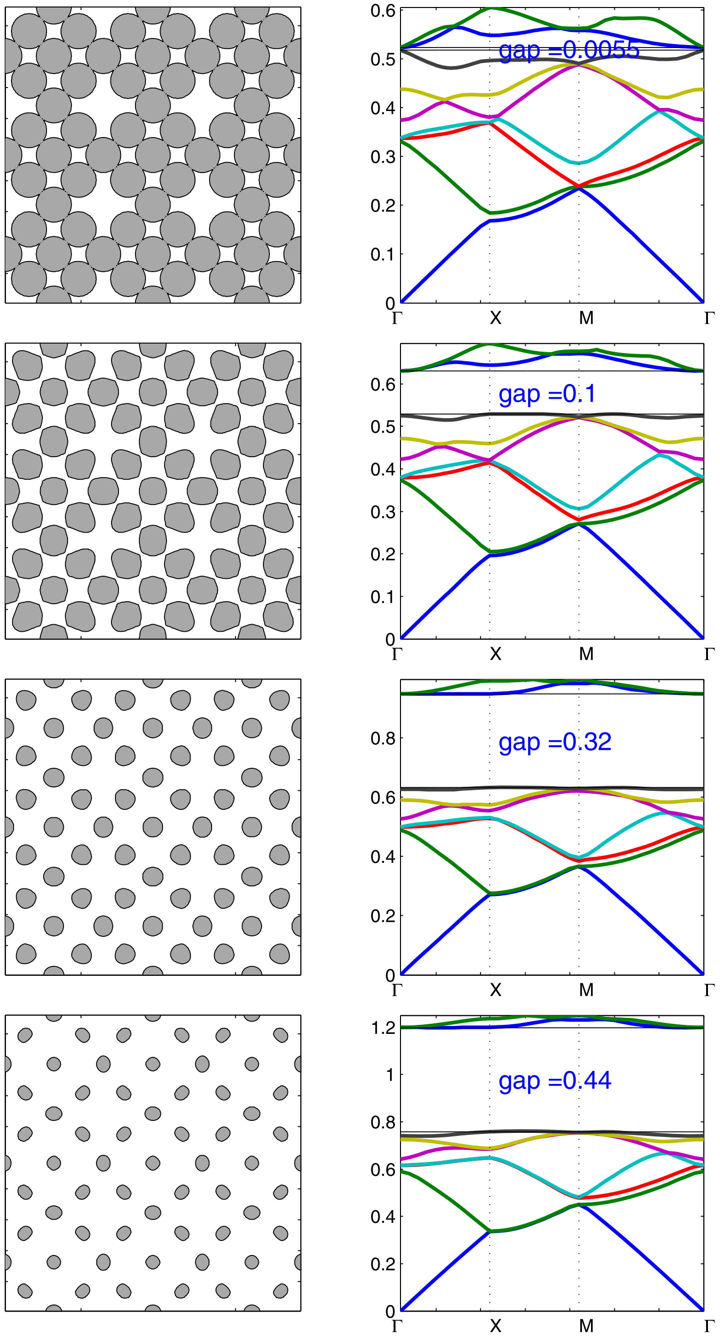


Fig. 7.3 The evolution of the dielectric distribution and its corresponding band structure for maximizing the bandgap between ω^7_{TM} and ω^8_{TM}

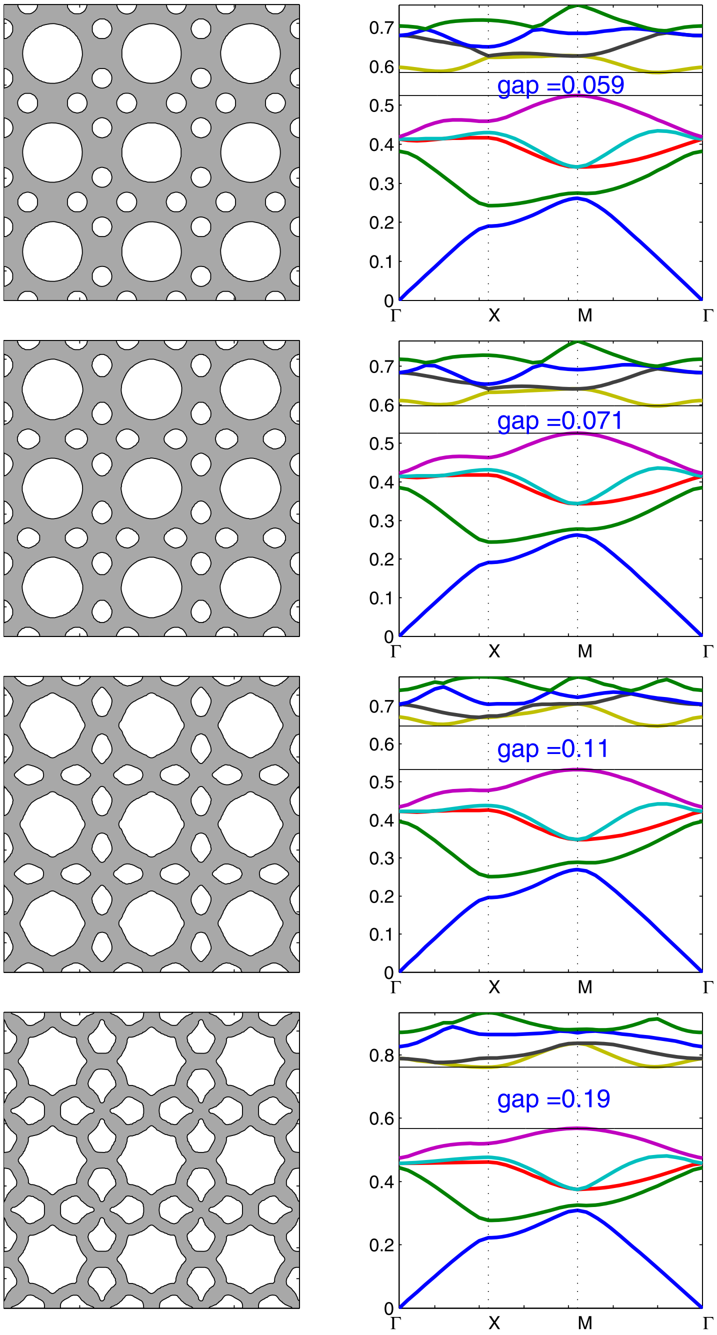


Fig. 7.4 The evolution of the dielectric distribution and its corresponding band structure for maximizing the bandgap between ω_{TE}^{5} and ω_{TE}^{6}

has bandgaps for both TM and TE modes. By adjusting the geometry of the lattice, one can even arrange for the bandgaps to overlap, resulting in a complete bandgap for all polarizations. This optimization problem can be formulated as:

Maximize the complete bandgap

$$\sup_{\Omega} G_{\text{complete}} = \sup_{\Omega}\left\{\inf\left(\inf_{\alpha}\omega_{\text{TE}}^{k+1}, \inf_{\alpha}\omega_{\text{TM}}^{k+1}\right) - \sup\left(\sup_{\alpha}\omega_{\text{TE}}^{k}, \sup_{\alpha}\omega_{\text{TM}}^{k}\right)\right\}. \tag{7.15}$$

In Fig. 7.3, we demonstrate the process of optimizing the bandgap for the seventh and eighth eigenvalues of the TM mode. As the number of iterations increases, the bandgap gradually increases from 0.0055 until it reaches a stable value of 0.44. The high ε region separates and evolves into smaller regions. The topological change in the dielectric distribution is well captured via the level set method.

In Fig. 7.4, we demonstrate the process of optimizing the bandgap for the fifth and sixth eigenvalues of the TE mode. As the number of iterations increases, the bandgap gradually increases from 0.059 until it reaches a stable value of 0.19. Comparing with the previous example, we observe that a lattice of isolated high ε regions is preferred for the TM mode while a lattice of connected high ε regions is preferred for the TE mode.

7.3 Eigenfunction Optimization

7.3.1 Finding the Optimal Localization of Eigenfunctions

In prior work by Dobson and Santosa [9], the localization of the eigenfunction in the TM mode for an electromagnetic wave in two dimensions is considered for the wave vector $\alpha = (0, 0)$ (the Γ point) in Eq. (7.3). Dobson and Santosa derived a numerical approach to localize a single eigenmode of the Laplacian problem using the gradient descent method, and their numerical results showed that the optimal configuration has a point defect structure. Here we generalize the approach to study eigenfunction localization at multiple specified points and on curved boundaries.

The governing equation is

$$\begin{cases} -\Delta u(x) = \lambda\varepsilon(x)u(x) & \text{for } x \in \Omega, \\ u(x) = 0 & \text{for } x \in \partial\Omega, \end{cases}$$

Here we limit our discussion to $N = 1$ or $N = 2$ dimensions, even though the method works in higher dimensions. The goal is to localize the eigenfunctions $u_{k_1}, u_{k_2}, \ldots, u_{k_n}$ that correspond to the electromagnetic field distributions, with specific eigenvalues $\mu_1, \mu_2, \ldots, \mu_n$ that correspond to the resonant frequencies of these eigenfunctions at given locations $x_1, x_2, \ldots, x_n$.

We formulate our problem as

$$\min J(\varepsilon, u) = \frac{1}{2}\sum_{i=1}^{n}\left[\int_{\Omega} w_i \varepsilon u_{k_i}^2\right] + \frac{1}{2}\sum_{i=1}^{n}\nu_i(\lambda_{k_i} - \mu_i)^2, \tag{7.16}$$

where w_i are the weight functions which penalize nonlocalization of the fields, e.g., $w_i = |x - x_i|^2$ subject to the constraints that

$$\begin{cases} -\Delta u_{k_i} = \lambda_{k_i} \varepsilon u_{k_i} & \text{for } x \in \Omega, \\ u = 0 & \text{on } x \in \partial\Omega, \end{cases}$$

and $\frac{1}{2}\sum_{i=1}^{n} \nu_i (\lambda_{k_i} - \mu_i)^2$ is the penalty for the difference between the corresponding eigenvalues and the given constants μ_i, $i = 1, \ldots, n$. In our simulations, we simply take $\nu_i = 1$ for all i. In some cases, it may be necessary to determine ν_i numerically, e.g., by the Lagrange-multiplier method, to achieve better results. At the discrete level, we have $-\Delta u_{k_i} = \lambda_{k_i} \varepsilon u_{k_i}$ approximated by $A_{\mathrm{L}} U = E U \Lambda$ where A_{L}, E, U, Λ are discrete approximations of the Laplacian operator, $\varepsilon(x)$, eigenfunctions u_{k_i}, and eigenvalues λ_{k_i}, respectively. The quality $\varepsilon u_{k_i}^2$ is proportional to energy density in the medium. By normalizing the eigenfunctions to have unit energy in Ω, each eigenfunction will satisfy

$$\int_{\Omega} \varepsilon u_{k_i}^2 \, dx = 1,$$

for the discrete case, this is equivalent to

$$U^T E U = 1.$$

For the localization of eigenfunctions on curved boundaries, we replace the weight function w by the square of the distance function to curves, which will be discussed in detail in Sect. 7.3.4.

For Bi-Laplacian problems, we consider the equation

$$\begin{cases} \Delta^2 u(x) = \lambda \varepsilon(x) u(x) & \text{for } x \in \Omega, \\ u(x) = 0 & \text{for } x \in \partial\Omega, \end{cases}$$

with a hinged boundary condition $\frac{\partial u}{\partial n} = 0$ or a clamped boundary condition $\Delta u = 0$ for $x \in \partial\Omega$, where n is the vector normal to the boundary. Notice the differences between the Laplacian problem and the Bi-Laplacian problem are in the differential operators and boundary conditions. In the discrete sense, we can extend all formulas for Laplacian problems by simply using the approximated Laplacian matrix A_{L} instead of the matrices A_{BL}^h and A_{BL}^c which approximate the Bi-Laplacian operator with hinged and clamped boundary conditions in the discrete case. Due to this extension, we drop the subscript and superscript of A when we discuss the optimization approach in the following sections.

7.3.2 Localization of Single Eigenmodes with Eigenvalue Constraints

In this section, we follow the same method as Dobson and Satosa [9] to deduce the gradient descent approach of energy minimization for Eq. (7.16). We first use the finite difference method to discretize the equation and enforce boundary conditions. The eigenfunction u in two dimensions is reorganized into a column vector, as is the dielectric constant ε. Let $b = \varepsilon^{-\frac{1}{2}}$ and B be the diagonal matrix with vector b on the main diagonal. Set $v = B^{-1}u$, and the discretized problem becomes

$$\begin{aligned} BABv &= \lambda v, \\ \langle v, v\rangle &= 1, \\ \langle v, BABv\rangle &= \lambda. \end{aligned}$$

Given any vector b, an eigenmode, say $v(b)$, is chosen to be localized at some specified point. Without loss of generality, the point is chosen as $(0, 0)$, with the associated eigenvalue staying close to prescribed constant μ_1. Our objective function (7.16) becomes

$$J(b) = \frac{1}{2}\big\langle v(b), Wv(b)\big\rangle + \frac{1}{2}\nu\big(\lambda(b) - \mu_1\big)^2,$$

where W is a diagonal matrix with the vector $w = x^2 + y^2$ on its main diagonal. To avoid complication, we assume that the eigenvalue is simple. Let δb be a small perturbation in b. From the calculus of variations, we have:

$$\delta J = \langle \delta v, wv\rangle + \nu(\lambda - \mu_1)\delta\lambda. \tag{7.17}$$

Now define an adjoint vector q as the solution to the equation

$$BABq - \langle v, BABv\rangle q - 2\langle q, v\rangle BABv = Wv,$$

then after simplification

$$\langle \delta v, Wv\rangle = \big\langle \delta b, -\operatorname{diag}(q)ABv - \operatorname{diag}(v)ABq + 2\langle q, v\rangle \operatorname{diag}(v)ABv\big\rangle. \tag{7.18}$$

Furthermore, the linearized response $\delta\lambda$ can be derived as

$$\delta\lambda = 2\big\langle \delta b, \operatorname{diag}(v)ABv\big\rangle.$$

Combining (7.17) and (7.18) gives

$$\delta J = \langle \delta b, g\rangle, \tag{7.19}$$

where

$$g = -\operatorname{diag}(q)ABv - \operatorname{diag}(v)ABq + 2\langle q, v\rangle \operatorname{diag}(v)ABv + 2\nu(\lambda - \lambda_1)\operatorname{diag}(v)ABv.$$

Thus the normalized descent direction is

$$\delta b = -g/\|g\|, \tag{7.20}$$

for a single mode to minimize the objective function.

7.3.3 Localization of Multiple Eigenmodes

The result above works for the localization of a single eigenmode; however, since the objective function for the localization of multiple eigenmodes is accumulated from many single-mode cases, it is natural to extended this to eigenmode localization at different points for multiple eigenfunctions by summing up the effects from each mode. Let $v_{k_i} = B^{-1}u_{k_i}$, then we have the eigenproblems

$$\begin{aligned} BABv_{k_i} &= \lambda_{k_i} v_{k_i}, \\ \langle v_{k_i}, v_{k_i}\rangle &= 1, \\ \langle v_{k_i}, BABv_{k_i}\rangle &= \lambda_{k_i}, \end{aligned}$$

for $i = 1, \ldots, n$. The objective function is

$$J(b) = \sum_{i=1}^{n} \left[\frac{1}{2}\big\langle v_{k_i}(b), W_i v_{k_i}(b)\big\rangle + \frac{1}{2}\nu\big(\lambda_{k_i}(b) - \lambda_i\big)^2 \right],$$

where W_i are diagonal matrices associated with the square of the distance function to different points on the main diagonal. We then calculate the variation mode by mode and denote the gradient for each mode by g_i, i.e.,

$$\begin{aligned} g_i = &-\operatorname{diag}(q_i)ABv_{k_i} - \operatorname{diag}(v_{k_i})ABq_i \\ &+ 2\langle q_i, v_{k_i}\rangle \operatorname{diag}(v_{k_i})ABv_{k_i} + 2\nu(\lambda_{k_i} - \lambda_i)\operatorname{diag}(v_{k_i})ABv_{k_i}, \end{aligned} \tag{7.21}$$

then $g = \sum_{i=1}^{n} g_i$ will be the gradient of J with respect to δb. Let g be a normalized unit vector. Then the descent gradient direction to minimize the objective function is $-g$.

7.3.4 Weight Functions

It is of interest to develop the localization of an eigenmode not only at one point, but also at several points, or even on an arbitrary curved boundary. To realize more complicated confinement of eigenfunctions, we only need to change the weight function accordingly. As mentioned before, for the single point localization, the square of the distance function to that point is used as the weight function. If we want to localize a single mode at several points, we can use the square of the minimal distance function at all points as the weight function, i.e., let $w = \min_i (x - x_i)^2$.

For complex curves, the distance function to the curve satisfies the eikonal equation with the prescribed Dirichlet boundary condition on the curve. Fast marching [20, 22] or fast sweeping methods [16, 24, 26] can be applied to compute the distance function very efficiently. It is not necessary to choose the square of distance function as the weight function to confine the eigenfunction. The weight function can be adjusted slightly, e.g., applying different powers, to obtain much sharper localization.

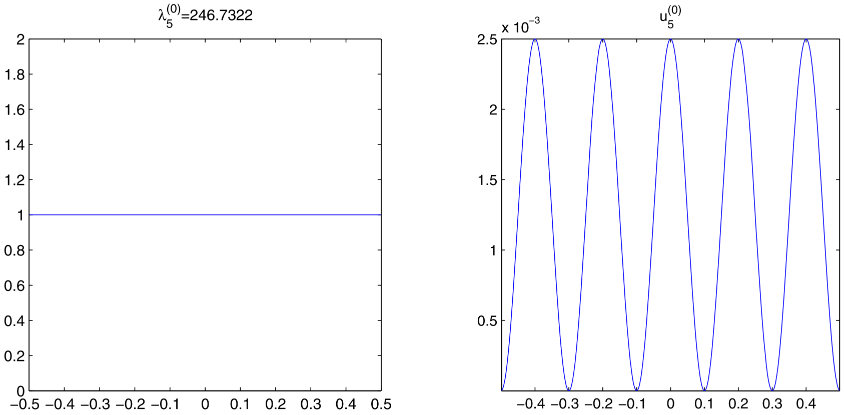
7.3.5 Numerical Tests

In this section, we apply the gradient descent method formulated in Eqs. (7.20) and (7.21) to obtain shape designs that yield localized eigenfunctions. During the implementation, it is possible that adjacent eigenvalues cross each other and, at certain frequencies, can possibly result in a faulty choice of eigenfunction. To avoid this situation, we choose the eigenmode which is closest to the result in the previous iteration as shown in Algorithm 7.2.

Algorithm 7.2: Eigenmode algorithm

```
input: Initial design $b_0$ and a distinct eigenfunction $v_0$ with corresponding
       eigenvalue $\lambda$. Give a reasonable constant $\lambda_0$ that the eigenvalue will be
       fixed at. Set the iteration number $n = 0$. Choose a step size parameter
       $\tau > 0$.
Compute the normalized descent gradient $g$ of $J(b_n)$.
repeat
    Let $v = \max\{|v \cdot v_n| : v$ is an eigenvector of $B_n A B_n$, with $\langle v, v \rangle = 1\}$,
    where $B_n = \mathrm{diag}(b_n - \tau g)$. Let $\lambda_v$ be the associated eigenvalue.
    If $J(v) < J(v_n)$, then
        $v_{n+1} = v$
        $b_{n+1} = T(b_n - \tau g)$
        $\lambda_{n+1} = \lambda_v$,
    else let $\tau = \tau/2$.
until stopping criterion is met and $\tau$ is too small
```

All the numerical simulations in the following sections start with a homogeneous initial density, $\varepsilon(x) = 1$, and stop when the step size in the gradient descent direction is less than 10^{-15}. In all one-dimensional problems, the domain $[-0.5, 0.5]$ is discretized into 800 cells; the meshes for all two-dimensional problems are 112×112.



(a) initial profile of ε

(b) energy density of the initial 5th eigenvector

Fig. 7.5 Initial profile for Example 7.1: localization of the fifth Laplacian eigenmode of the one-dimensional problem

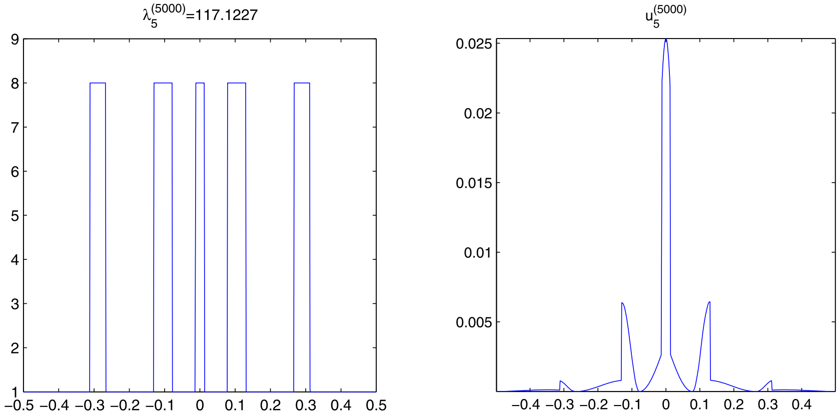
We show the final density profiles of ε with information about eigenvalue in the title, where the subscript of λ denotes the eigenmode and the superscript is the iteration number. Scaled images of energy density for confined eigenvectors are also provided. Here energy density is computed as v^2, which is equal to εu^2. Note that our gradient descent method assumes the eigenmode that we localize is distinct, so the 11th and 20th eigenmodes are picked to test our algorithm in 2-dim problems. Finally, as was the case in Chap. 6, since the electromagnetic wave equation is scale-invariant, we state all dimensions terms of the vacuum wavelength itself.

7.3.6 One-Dimensional Problems Without Eigenvalue Constraints

We start with one-dimensional Laplacian examples without eigenvalue constraints, i.e., $\nu = 0$. In one dimension, we see two different types of symmetry results which depend on odd or even modes.

Example 7.1 The initial density and corresponding fifth eigenfunction are shown in Fig. 7.5. We aim to confine the fifth eigenmode at the origin. After the optimization, the results are shown in Fig. 7.6. We see that fifth eigenmode stays oddly symmetric and there is a point defect of material distribution at the origin in the optimal result.

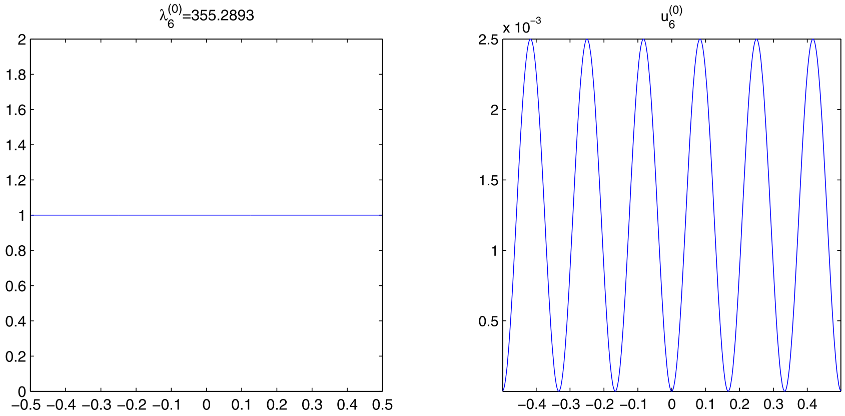
Example 7.2 The setup is the same as the previous example. The sixth eigenmode is chosen to be localized. The initial and final profiles are shown in Figs. 7.7



(a) optimal profile of ε

(b) energy density of the localized 5th eigenmode

Fig. 7.6 Final profile for Example 7.1: localization of the fifth Laplacian eigenmode of the one-dimensional problem

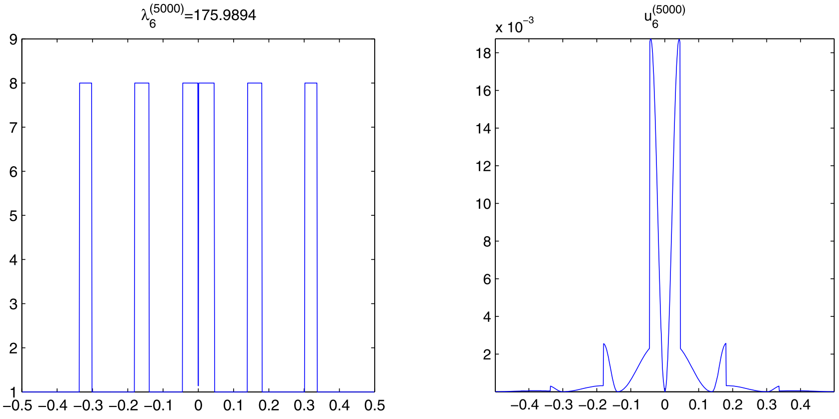


(a) initial profile of ε

(b) energy density of the initial 6th eigenvector

Fig. 7.7 Initial profile for Example 7.2: localization of the sixth Laplacian eigenmode of the one-dimensional problem

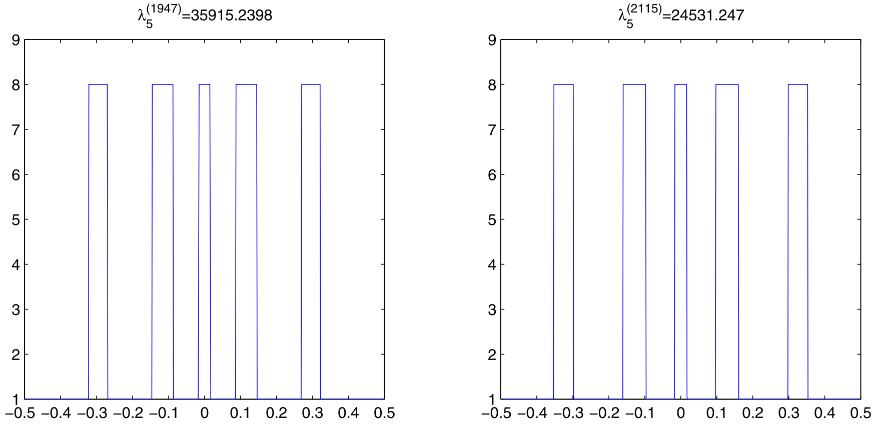
and 7.8. We compare the sixth eigenmode with the fifth eigenmode in the first example. Since originally the sixth eigenmode is evenly symmetric about 0, the optimal result stays evenly symmetric. The material distribution generates two close peaks around the point where we want to localize, and it seems impossible to switch the symmetry of the optimized solution, as was the case for the fifth eigenmode.



(a) optimal profile of ε

(b) energy density of the localized 6th eigenmode

Fig. 7.8 Final profile for Example 7.2: localization of the sixth Laplacian eigenmode of the one-dimensional problem



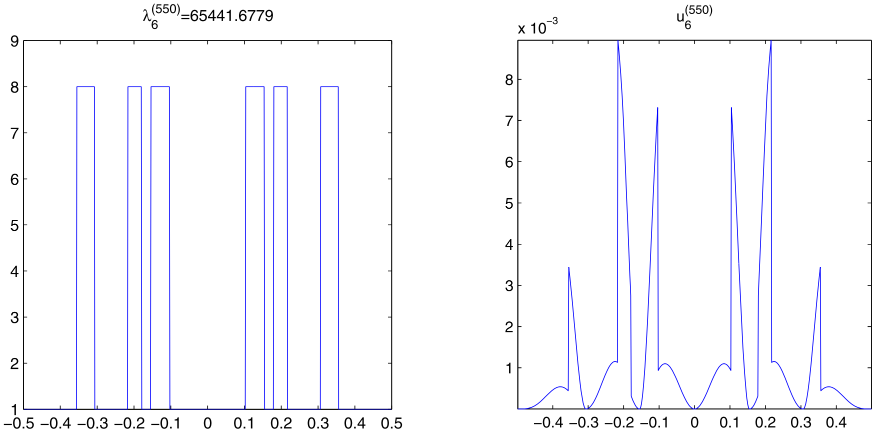
(a) hinged boundary condition

(b) clamped boundary condition

Fig. 7.9 Optimal profile of ε for Example 7.3: localization at 0 of the fifth Bi-Laplacian eigenmode

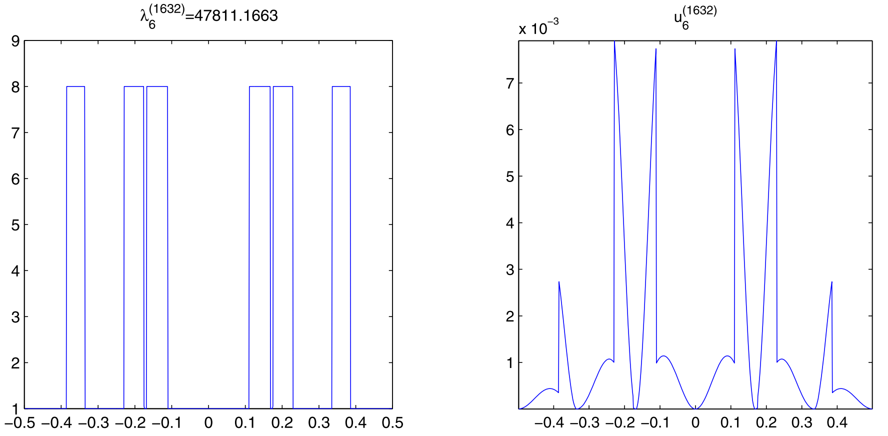
Example 7.3 In this example, we compare the material profiles of localizations for the Bi-Laplacian problem with different boundary conditions. The discretized mesh is set up to be the same as for one-dimensional Laplacian eigenproblems. The results of localizing the fifth eigenmode are shown in Fig. 7.9. The shapes are very similar, only the distance between the material with high dielectric constant is different for hinged and clamped boundary conditions.

Example 7.4 In this example, we compare the material profiles of localization at two points for the Bi-Laplacian problem with different boundary conditions. The



(a) optimal profile of ε (b) localized 6th BiLaplacian eigenmode

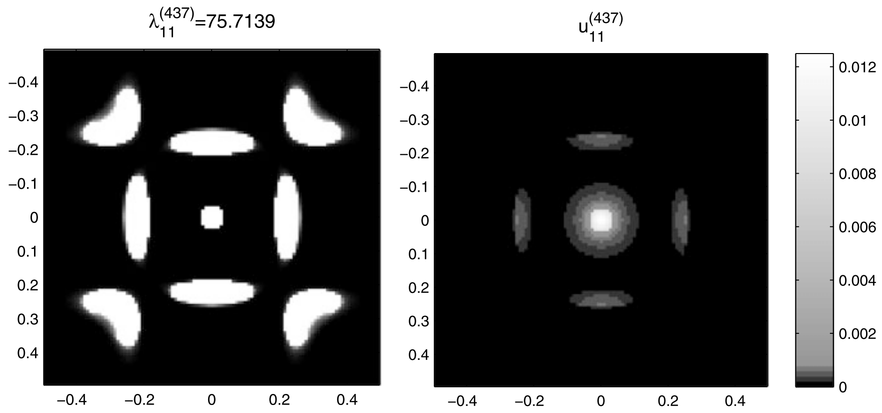
Fig. 7.10 Final profile for Example 7.4: localization at two points $x = \pm 0.2$ of the sixth Bi-Laplacian eigenmode with a hinged boundary condition



(a) optimal profile of ε (b) localized 6th BiLaplacian eigenmode

Fig. 7.11 Final profile for Example 7.4: localization at two points $x = \pm 0.2$ of sixth Bi-Laplacian eigenmode with clamped boundary condition

results of localizing the sixth eigenmode with hinged and clamped boundary conditions are shown in Figs. 7.10 and 7.11 separately. The confinement location is chosen to be ±0.2. Due to the different boundary conditions imposed, we can observe different shapes for the dielectric coefficient distributions.



(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.12 Final profile for Example 7.5: localization at (0, 0) of the 11th Laplacian eigenmode in two dimensions

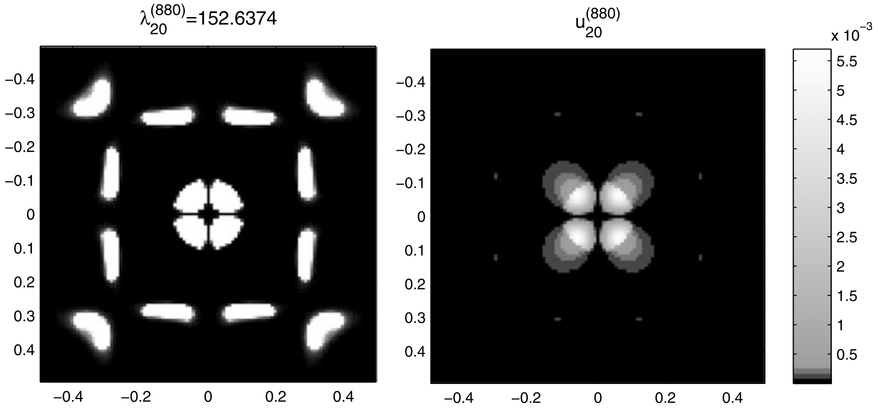
7.3.7 Two-Dimensional Laplacian Problems Without Eigenvalue Constraints

Example 7.5 In this example, we try to localize the 11th eigenmode of the Laplacian equation at the origin, i.e., (0, 0), with no constraint $\nu = 0$ on the corresponding eigenvalue as in [9]. The results are shown in Fig. 7.12. After about 400 iterations, the shape becomes stable and the objective energy saturates around 0.01. A symmetric design is generated and a single peak concentrates at (0, 0) for $\varepsilon(x)$ which is similar to the odd symmetry observed in one dimension.

Example 7.6 In this example, we localize the 20th eigenmode of Laplacian problem at the origin, i.e., (0, 0). The results are shown in Fig. 7.13. The 20th eigenfunction is evenly symmetric about the origin which is similar to the result for the sixth mode in one dimension. The final energy density has four peaks around (0, 0). The objective function drops below 0.01 after 880 iterations.

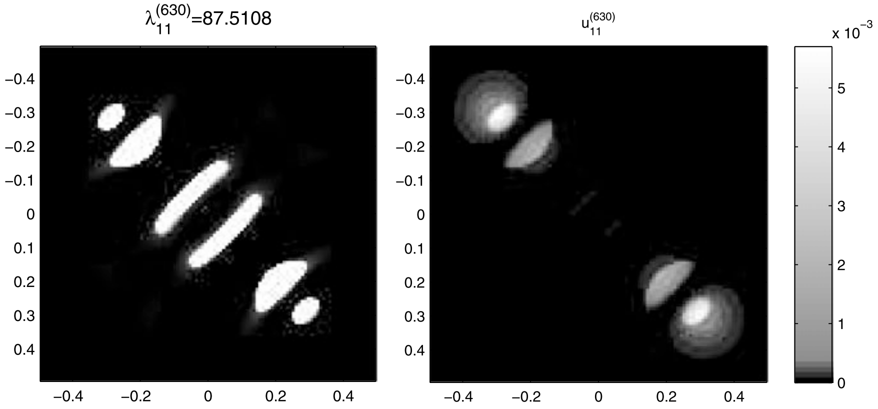
Example 7.7 In this example, we try to localize the 11th eigenmode of the Laplacian problem at two distinct points, (0.25, 0.25) and (−0.25, −0.25). The results are shown in Fig. 7.14. We can see clearly two concentrated energy peaks at the given points after 630 iterations. During the optimization, the objective function decreased from 0.06 from 0.01.

Example 7.8 In this example, we try to localize the 11th eigenmode of the Laplacian problem along a circle centered at (0, 0) with radius $r = 0.25$. The results are shown in Fig. 7.15. The weight function is set to be $\|\sqrt{x^2 + y^2} - 0.25\|$. After 264 iterations, the design consisting of two materials is obtained with objective energy saturating around 0.02.



(a) optimal profile of ε (b) energy density of the localized eigenvector

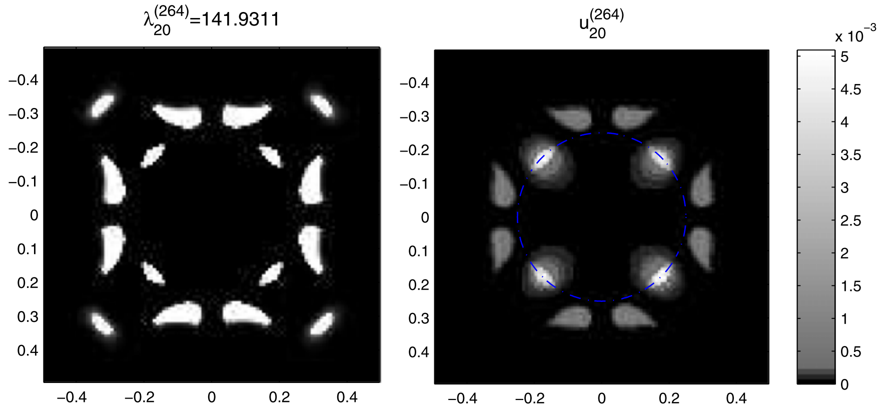
Fig. 7.13 Final profile for Example 7.6: localization at $(0,0)$ of the 20th Laplacian eigenmode in two dimensions



(a) optimal profile of ε (b) energy density of the localized eigenvector

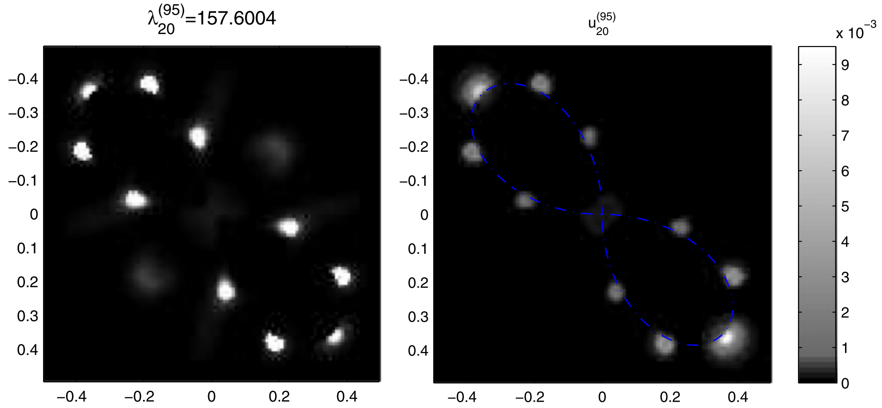
Fig. 7.14 Final profile for Example 7.7: localization at two points of 11th Laplacian eigenmode in two dimensions

Example 7.9 In this example, we try to localize the 11th eigenmode of the Laplacian problem along a curve described in polar coordinate as $r = 0.5\sin(2\theta)$. The results are shown in Fig. 7.16. We can see that after 95 iterations, the eigenfunction is mainly localized on the specified curve.



(a) optimal profile of ε (b) energy density of the localized eigenvector

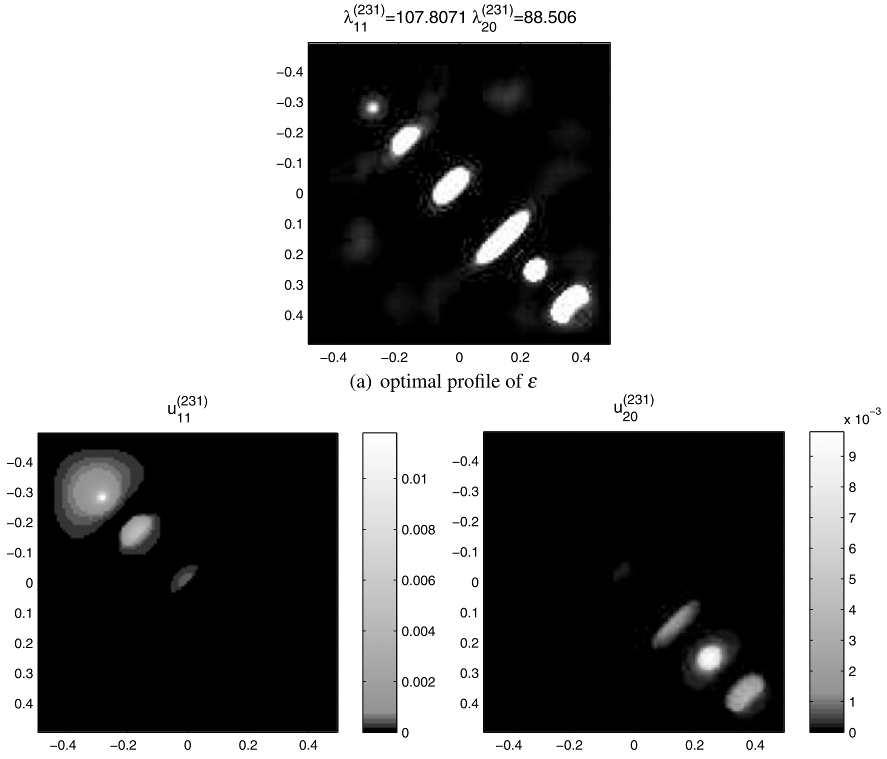
Fig. 7.15 Final profile for Example 7.8: circular energy localization of 11th Laplacian eigenmode in two dimensions



(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.16 Final profile for Example 7.9: localization along a curve described in polar coordinates of the 11th Laplacian eigenmode in two dimensions

Example 7.10 In this example, we try to distinguish between two different eigenmodes by localizing them at different points. The 11th and 20th eigenmodes are chosen to be localized at $(-0.25, -0.25)$ and $(0.25, 0.25)$ separately. Both of them are single modes. The results are shown in Fig. 7.17. We deal with two eigenmodes, so we follow the formula in Sect. 7.3.3. It can be seen that the numerical approach works quite well. Two eigenmodes are completely localized at the specified points. A clear two-material shape is obtained after 231 iterations.



(a) optimal profile of ε

(b) energy density of the 11th eigenvector

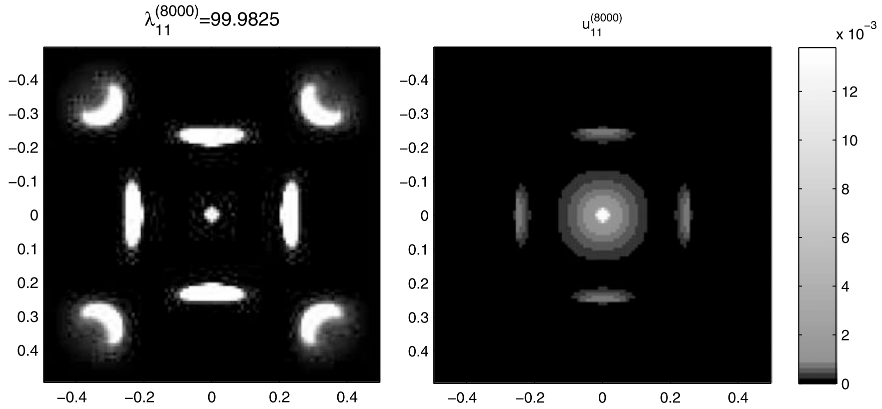
(c) energy density of the 20th eigenvector

Fig. 7.17 Final profile for Example 7.10: localization at distinct points of two Laplacian eigenmodes (11th and 20th) in two dimensions

7.3.8 Two-Dimensional Laplacian Problems with Eigenvalue Constraints

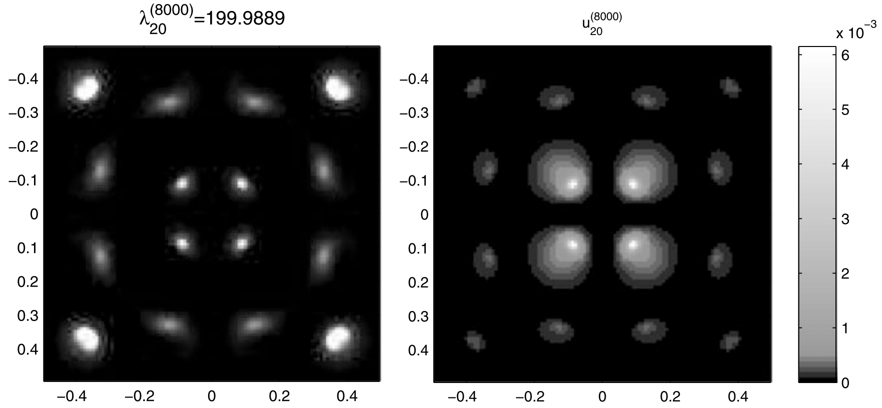
Example 7.11 In this example, we try to localize the 11th eigenmode of the Laplacian problem at the origin, i.e., (0, 0), and require the corresponding eigenvalue to be as close to a specific constant $\lambda_0 = 100$ as possible. The results are shown in Fig. 7.18. We obtain a design that is similar to the solution without an eigenvalue constraint; however, this solution has a smaller total area of the high-dielectric material. The eigenmode is well localized after a small number of iterations, and the associated eigenvalue oscillates around its optimum with an error of 0.1.

Example 7.12 In this example, we try to localize the 20th eigenmode of the Laplacian problem at the origin, i.e., (0, 0), and also require the corresponding eigenvalue to be as close as possible to a constant: $\lambda_0 = 200$. The results are shown in Fig. 7.19. From the results we can see four peaks around the target location, due to the even symmetry of the original eigenmode, and the eigenvalue is well localized.



(a) optimal profile of ε (b) energy density of the localized eigenvector

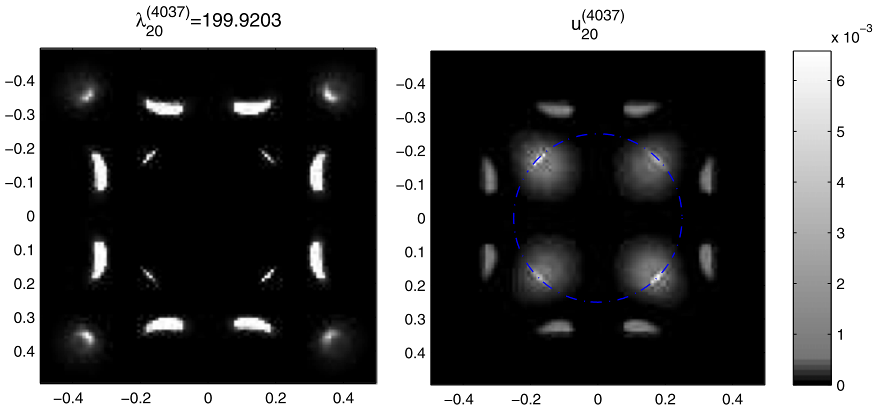
Fig. 7.18 Final profile for Example 7.11: localization at $(0, 0)$ of the 11th Laplacian eigenmode in two dimensions with $\lambda = 100$



(a) optimal profile of ε (b) energy density of the localized eigenvector

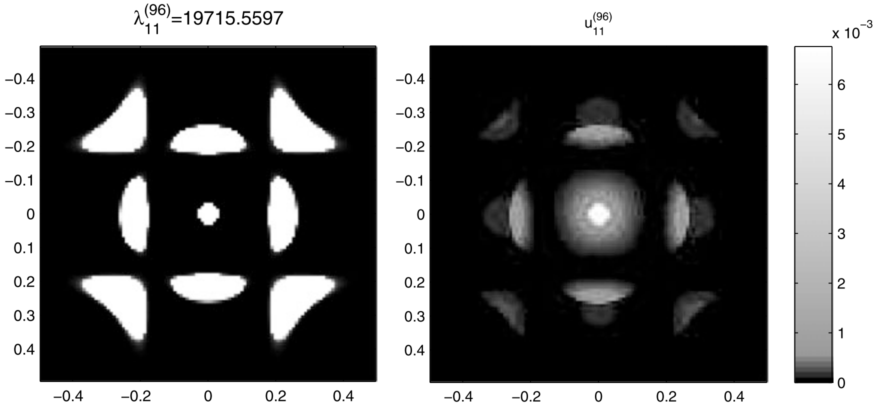
Fig. 7.19 Final profile for Example 7.12: localization at $(0, 0)$ of the 20th Laplacian eigenmode in two dimensions with $\lambda = 200$

Example 7.13 In this example, we try to localize the 20th eigenmode of the Laplacian problem along a circle centered at $(0, 0)$ with radius $r = 0.25$, and also require the corresponding eigenvalue to be as close as possible to a constant: $\lambda_0 = 200$. The results are shown in Fig. 7.20. The localization is obtained with an eigenvalue fixed at 200 after about 4000 iterations.



(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.20 Final profile for Example 7.13: localization along a circle of 20th Laplacian eigenmode in two dimensions with $\lambda = 200$

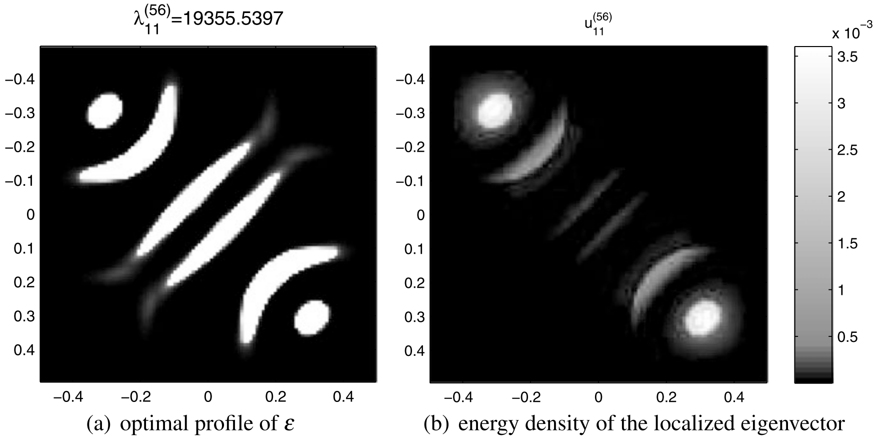


(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.21 Final profile for Example 7.14: localization at $(0, 0)$ of the 11th eigenmode of the two dimensional Bi-Laplacian problem with hinged boundary conditions

7.3.9 Two-Dimensional Bi-Laplacian Problems Without Eigenvalue constraints

Example 7.14 In this example, we try to localize the 11th eigenmode of the Bi-Laplacian problem at the origin, using hinged boundary conditions. The results are shown in Fig. 7.21. A clear two-material design is obtained after 96 iterations.



(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.22 Final profile for Example 7.15: localization at two points of the 11th eigenmode of the two dimensional Bi-Laplacian operator with hinged boundary conditions

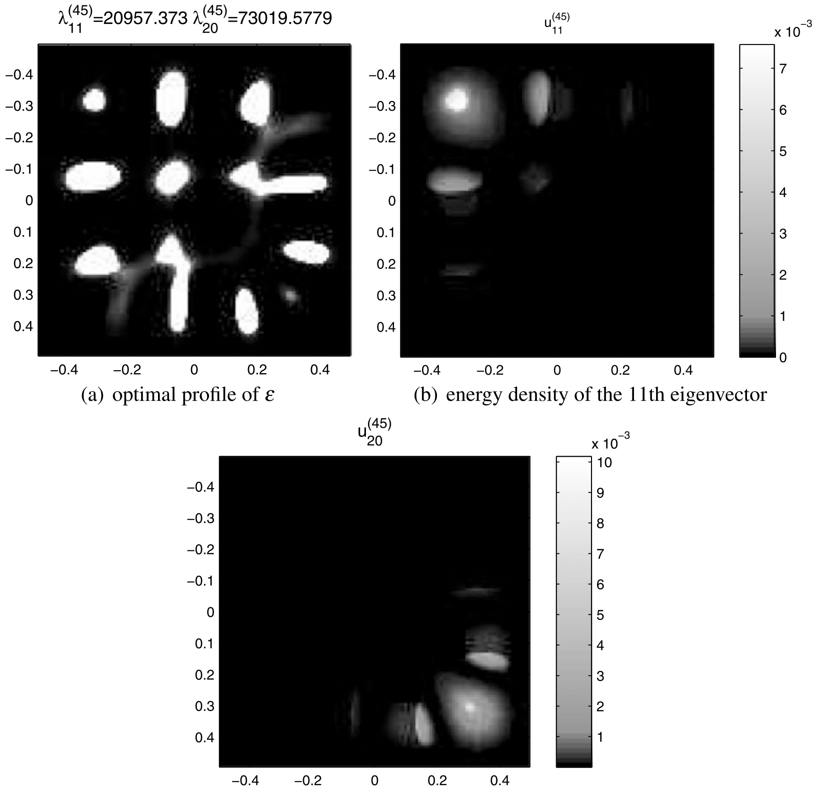
Example 7.15 In this example, the 11th eigenmode of the Bi-Laplacian problem, with hinged boundary conditions, is designed to be confined at two points, $(-0.25, -0.25)$ and $(0.25, 0.25)$. The results are shown in Fig. 7.22. The two points are localized after 56 iterations with the objective function decreasing below 0.01.

Example 7.16 In this example, we try to distinguish between two different eigenmodes by localizing them at different points. The 11th and 20th eigenmodes are chosen to be localized at $(-0.25, -0.25)$ and $(0.25, 0.25)$ separately. Both of them are single modes. The results are shown in Fig. 7.23. After 45 iterations, both eigenmodes are confined at the specified points and a symmetric design is achieved.

Example 7.17 In this example, we try to localize the 11th eigenmode of the Bi-Laplacian problem with a clamped boundary condition at the origin $(0, 0)$. The results are shown in Fig. 7.24. The results are similar to what we get with a hinged boundary condition.

Example 7.18 In this example, the 11th eigenmode of Bi-Laplacian problem with a clamped boundary condition is chosen to be confined at two points, $(-0.25, -0.25)$ and $(0.25, 0.25)$. The results are shown in Fig. 7.25. The design is totally different from the one with a hinged boundary condition.

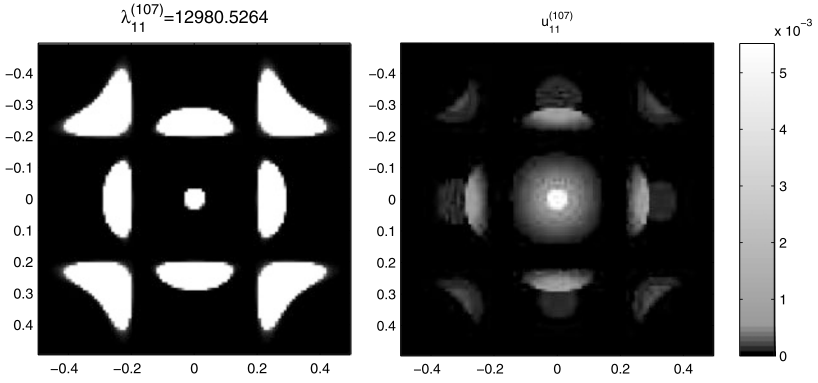
Example 7.19 In this example, we try to distinguish between two different eigenmodes of the Bi-Laplacian problem, with a clamped boundary condition, by localizing them at different points. The 11th and 20th eigenmodes are chosen to be localized at $(-0.25, -0.25)$ and $(0.25, 0.25)$ separately. Both of them are single modes. The results are shown in Fig. 7.26.



(a) optimal profile of ε (b) energy density of the 11th eigenvector

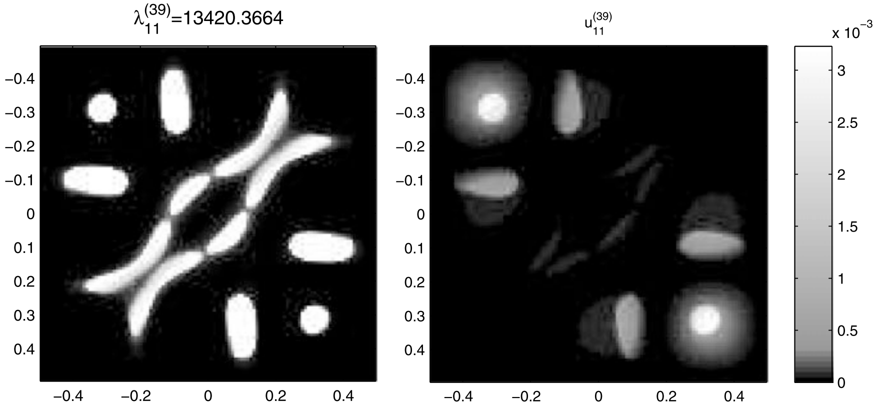
(c) energy density of the 20th eigenvector

Fig. 7.23 Final profile for Example 7.16: localization at distinct points for two eigenmodes (11th and 20th) of the two dimensional Bi-Laplacian operator with hinged boundary conditions



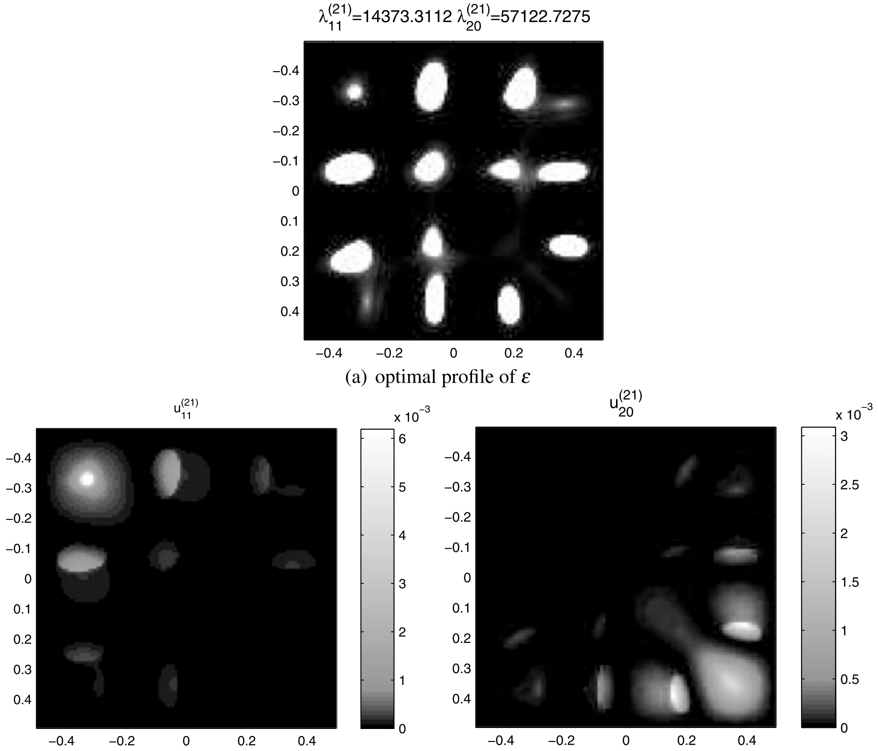
(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.24 Final profile for Example 7.17: localization at (0, 0) of the 11th eigenmode of the two dimensional Bi-Laplacian operator with a clamped boundary



(a) optimal profile of ε (b) energy density of the localized eigenvector

Fig. 7.25 Final profile for Example 7.18: localization at two points of the 11th eigenmode of the two dimensional Bi-Laplacian operator with a clamped boundary



(a) optimal profile of ε

(b) energy density of the 11th eigenvector (c) energy density of the 20th eigenvector

Fig. 7.26 Final profile for Example 7.19: localization at distinct points of two eigenmodes (11th and 20th) of the two dimensional Bi-Laplacian operator with a clamped boundary

Acknowledgements Chiu-Yen Kao and Stanley Osher were supported in part by a Northrop Grumman Corporation Grant. Chiu-Yen Kao is also supported in part by NSF DMS-1216742 grant.

References

1. G. Allaire, F. Jouve, A. Toader, Structural optimization using sensitivity analysis and a level-set method. J. Comput. Phys. **194**, 363–393 (2004)
2. G. Allaire, F. Jouve, A. Toader, A level-set method for shape optimization. C. R. Acad. Sci. Paris, Ser. I **334**, 1125–1130 (2002)
3. S. Amstutz, H. Andrä, A new algorithm for topology optimization using a level-set method. J. Comput. Phys. **216**, 573–588 (2006)
4. M. Burger, A framework for the construction of level set methods for shape optimization and reconstruction. Inverse Probl. **17**, 1327–1356 (2001)
5. M. Burger, B. Hackl, W. Ring, Incorporating topological derivatives into level set methods. J. Comput. Phys. **194**, 344–362 (2004)
6. S.J. Cox, The generalized gradient at a multiple eigenvalue. J. Funct. Anal. **133**, 30–40 (1995)
7. S.J. Cox, D.C. Dobson, Maximizing band gaps in two-dimensional photonic crystals. SIAM J. Appl. Math. **59**, 2108–2120 (1999)
8. S.J. Cox, D.C. Dobson, Band structure optimization of two-dimensional photonic crystals in H-polarization. J. Comput. Phys. **158**, 214–224 (2000)
9. D.C. Dobson, F. Santosa, Optimal localization of eigenfunctions in an inhomogeneous medium. SIAM J. Appl. Math. **64**, 762–774 (2004)
10. L. He, C.-Y. Kao, S. Osher, Incorporating topological derivatives into shape derivatives based level set methods. J. Comput. Phys. **225**, 891–909 (2007)
11. K. Ito, Z. Li, K. Kunischm, Level-set function approach to an inverse interface problem. Inverse Probl. **17**, 1225–1242 (2001)
12. S. John, Strong localization of photons in certain disordered dielectric superlattices. Phys. Rev. Lett. **58**, 2486–2489 (1987)
13. C.-Y. Kao, Y. Lou, E. Yanagida, Principal eigenvalue for an elliptic problem with indefinite weight on cylindrical domains. Math. Biosci. Eng. **5**, 315–335 (2008)
14. C.Y. Kao, S. Osher, E. Yablonovitch, Maximizing band gaps in two dimensional photonic crystals by using level set methods. Appl. Phys. B, Lasers Opt. **81**, 235–244 (2005)
15. C.-Y. Kao, F. Santosa, Maximization of the quality factor of an optical resonator. Wave Motion **45**(4), 412–427 (2008)
16. C.Y. Kao, S. Osher, J. Qian, Lax–Friedrichs sweeping scheme for static Hamilton–Jacobi equations. J. Comput. Phys. **196**(1), 367–391 (2004)
17. F. Murqat, S. Simon, Etudes de problems d'optimal design. Lect. Notes Comput. Sci. **41**, 52–62 (1976)
18. J. Osher, F. Santosa, Level set methods for optimization problems involving geometry and constraints i. frequencies of a two-density inhomogeneous drum. J. Comput. Phys. **171**, 272–288 (2001)
19. S. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
20. S.J. Osher, R.P. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, 1st edn. (Springer, Berlin, 2002)
21. J. Sethian, A. Wiegmann, Structural boundary design via level set and immersed interface methods. J. Comput. Phys. **163**, 489–528 (2000)
22. J.A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd edn. (Cambridge University Press, Cambridge, 1999), p. 378
23. J. Sokolowski, J.-P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, vol. 10 (Springer, Heidelberg, 1992)

24. Y.-H.R. Tsai, L.-T. Cheng, S. Osher, H.-K. Zhao, Fast sweeping algorithms for a class of Hamilton–Jacobi equations. SIAM J. Numer. Anal. **41**(2), 659–672 (2003)
25. E. Yablonovitch, Inhibited spontaneous emission in solid-state physics and electronics. Phys. Rev. Lett. **58**, 2059–2062 (1987)
26. H. Zhao, A fast sweeping method for eikonal equations. Math. Comput. **74**(250), 603–628 (2005)
27. S. Zhu, Q. Wu, C. Liu, Variational piecewise constant level set methods for shape optimization of a two-density drum. J. Comput. Phys. **229**, 5062–5089 (2010)

Appendix
The Interface Between Optimization and Simulation

One of the most frequently asked questions from people new to the field of metamaterial design optimization is how to make optimization code control the electromagnetics' simulation software. For basic optimizations, simulation tools such as Lumerical and Comsol have recently incorporated gradient-based or particle swarm optimization routines into their graphical user interfaces. For users interested in full control of the optimization routine being used, or integrating more complex optimization routines; the technique is a bit more nuanced. While there are a wide variety of approaches to interface programming languages with simulation packages; the following Appendix describes an example where optimization code written in Matlab controls the structures and simulations within Lumerical. This example demonstrates the key steps in setting up an interface; however, it is assumed that the reader has prior knowledge of scripting within the Lumerical simulation environment.

The interfacing method described here is done by combining a carefully constructed Lumerical script file (LSF), with a command prompt call from within Matlab to open Lumerical and run the script. The example here will manipulate the length and width of a rectangular, dipole antenna on a substrate, and a reflection monitor will be used to evaluate the device performance. A generalized process flow for the interface described here is shown in Fig. A.1.

Prior to running an optimization, a generalized simulation environment within Lumerical must be created that has the appropriate generalized structure to be optimized, boundary conditions, source settings, monitor settings, and so on, that will be used in every simulation during the optimization run. This includes proper naming conventions so that every element within the simulation can be referenced and modified independently. This file will be opened and modified at the start of every simulation. The simulation and power transmission monitors are given specific names, "Simulation.fsp" and "ReflectionMonitor", respectively, which will be kept constant throughout every iteration of the optimization run.

The first key step in the process involves using a Matlab m-file to write an LSF that executes five important features:

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8,

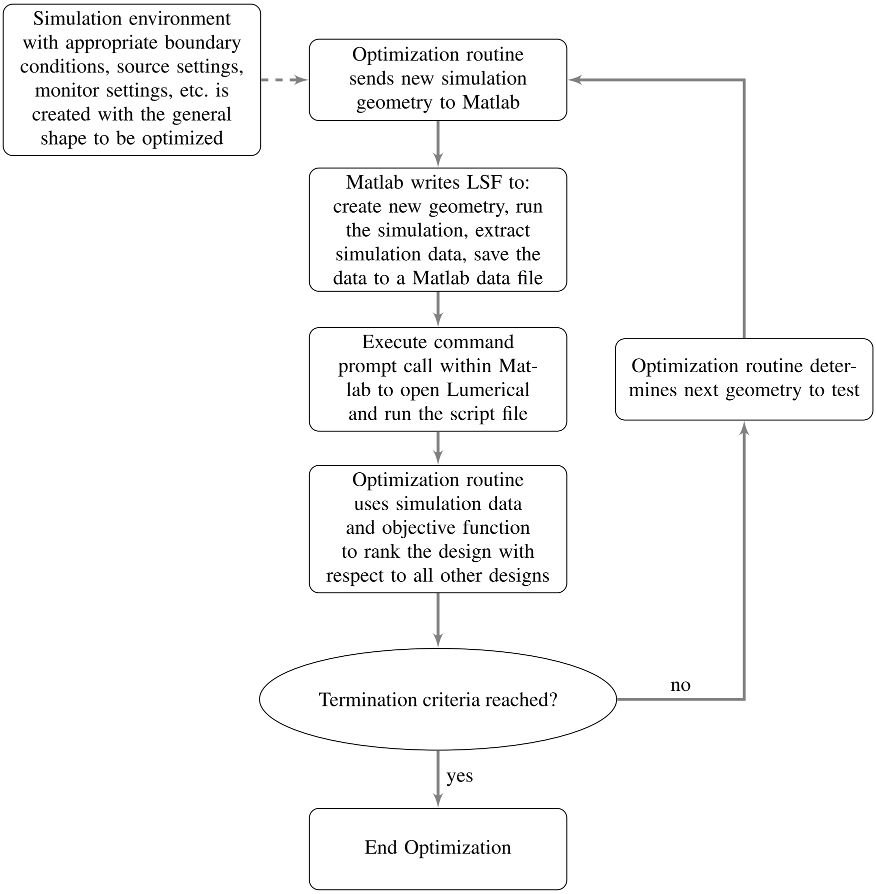

Simulation environment with appropriate boundary conditions, source settings, monitor settings, etc. is created with the general shape to be optimized
Optimization routine sends new simulation geometry to Matlab
Matlab writes LSF to: create new geometry, run the simulation, extract simulation data, save the data to a Matlab data file
Execute command prompt call within Matlab to open Lumerical and run the script file
Optimization routine uses simulation data and objective function to rank the design with respect to all other designs
Termination criteria reached?
no
Optimization routine determines next geometry to test
yes
End Optimization

Fig. A.1 A general process flow for the creation and implementation of the Optimization/Simulation interface

1. Open the simulation file “Simulation.fsp” which has been created prior to the optimization run.
2. Modify the structure and simulation environment within “Simulation.fsp” based on the specific geometry to be simulated.
3. Run the simulation.
4. Extract the relevant monitor data.
5. Convert the relevant monitor data to a Matlab data file with a pre-specified name “SimulationData.mat.”

For every line that appears in the LSF, there is a corresponding “fprint” command in the m-file. This includes every relevant dimension and setting for each: structure, simulation region, meshing region, monitor, and source within the simulation. An example of one command, where the dipole antenna named “antenna” is selected and the x-dimension is set to 200 nm would be written as:

```
fprintf(outfid, ’%s \n’, ’select("antenna");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(200) ’)*10^-9);’]);
```

Here the “10^{-9}” depends on whether or not the default dimension is in meters or nanometers. After this portion of the file, separate “fprint” commands are listed to execute Steps 3–5 from above. Two important details in this process are giving the LSF and the Matlab data files specific names, “AntennaScript.lsf” and “SimulationData.mat,” that will be referenced later. Finally, the “exit(2)” command within Lumerical closes the simulation, which allows Matlab to execute further commands.

The second key step in the process involves a command prompt call from within Matlab to both open Lumerical and run the previously created script file. An example using the DOS prompt is written as:

```
dos(’"C:\Program Files\Lumerical\FDTD\bin\fdtd-solutions.exe"
-run AntennaScript.lsf’);
```

At the time this book was published, documentation for using this approach in Windows is located at:

http://docs.lumerical.com/en/fdtd/user_guide_run_win_scripts_from_command_line.html

while the corresponding code for Mac and Linux is located at:

http://docs.lumerical.com/en/fdtd/user_guide_run_linux_cad_gui_from_command_line.html

Finally, the resulting simulation data can be accessed and then analyzed from within Matlab by simply executing the command:

```
load(’SimulationData.mat’, ’variable1’, ’variable2’);
```

where “variable1” and “variable2” refer to data saved within the Matlab data file.

An example m-file is shown in Fig. A.2. Here the vector “x” is sent to the m-file and stored as variables “width” and “height.” After creating the LSF “AntennaScript.lsf”, the first line of the script file tells Lumerical to open the simulation file “Simulation.fsp” which contains the geometry to be modified. The middle portion of the script prints commands to the LSF on changing the dimensions of the antenna, substrate, simulation volume, source, and power monitor. Here the simulation volume is set to twice the length and width of the antenna, and the substrate, source, and monitor are set to extend beyond the simulation volume. After adjusting the relevant geometries, the simulation is executed using the “runparallel” command; the broadband reflection spectrum is extracted and converted to a readable Matlab data file; and the simulation closes itself. The subsequent DOS command opens Lumerical, and runs the script “AntennaScript.lsf”. Finally, the broadband reflection data from “SimulationData.mat” is loaded and the maximum reflected wavelength is compared against a target wavelength of 1500 nm to calculate an objective function value. This data is returned to the optimization routine which then determines a new geometry to test, or to terminate based on predetermined convergence criteria.

```
function Objective_Function = optimization_interface(x)

width = x(1);
height = x(2);

filename=’AntennaScript.lsf’;
outfid = fopen(filename, ’w+’);

fprintf(outfid, ’%s \n’, ’clear;’);
fprintf(outfid, ’%s \n’, ’mypath = pwd;’);
fprintf(outfid, ’%s \n’, ’load(mypath+’’\’’+’’Simulation.fsp’’);’);
fprintf(outfid, ’%s \n’, ’switchtolayout;’);


fprintf(outfid, ’%s \n’, ’select("antenna");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(width) ’)*10^-9);’]);
fprintf(outfid, ’%s \n’, [’set("y span",(’ num2str(height) ’)*10^-9);’]);

fprintf(outfid, ’%s \n’, ’select("substrate");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(3*width) ’)*10^-9);’]);
fprintf(outfid, ’%s \n’, [’set("y span",(’ num2str(3*height) ’)*10^-9);’]);

fprintf(outfid, ’%s \n’, ’select("FDTD");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(2*width) ’)*10^-9);’]);
fprintf(outfid, ’%s \n’, [’set("y span",(’ num2str(2*height) ’)*10^-9);’]);

fprintf(outfid, ’%s \n’, ’select("source");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(3*width) ’)*10^-9);’]);
fprintf(outfid, ’%s \n’, [’set("y span",(’ num2str(3*height) ’)*10^-9);’]);

fprintf(outfid, ’%s \n’, ’select("ReflectionMonitor");’);
fprintf(outfid, ’%s \n’, [’set("x span",(’ num2str(3*width) ’)*10^-9);’]);
fprintf(outfid, ’%s \n’, [’set("y span",(’ num2str(3*height) ’)*10^-9);’]);

fprintf(outfid, ’%s \n’, ’runparallel;’);
fprintf(outfid, ’%s \n’, ’f=getdata("ReflectionMonitor","f");’);
fprintf(outfid, ’%s \n’, ’lambda=c/f;’);
fprintf(outfid, ’%s \n’, ’reflection=-transmission("ReflectionMonitor");’);
fprintf(outfid, ’%s \n’, ’matlabsave("SimulationData",reflection,lambda);’);
fprintf(outfid, ’%s \n’, ’exit(2);’);
fclose(outfid);


dos(’"C:\Program Files\Lumerical\FDTD\bin\fdtd-solutions.exe"
-run AntennaScript.lsf’);


load(’SimulationData.mat’, ’reflection’, ’lambda’);
max_wavelength=(lambda(reflection==max(reflection)))*10^9;
target_wavelength=1500;

Objective_Function=abs(target_wavelength-max_wavelength);

end
```

Fig. A.2 An example interface between Matlab and Lumerical for a the modification, simulation, and analysis of broadband reflection data from a dipole antenna

Lastly, to modify the interface listed below to run on Linux rather than Windows, the DOS command on line 45 must be replaced with the appropriate Linux command listed in the Lumerical help menu, and the backslash on line 11 in “load(mypath+”\”+”Simulation.fsp”);” must be changed to a forward slash.

Index

K. Diest (ed.), *Numerical Methods for Metamaterial Design*,
Topics in Applied Physics 127, DOI 10.1007/978-94-007-6664-8,