An *Ab Initio* Approach to the Inverse Problem-Based Design of Photonic Bandgap Devices

Thesis by

John K. Au

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2007

(Defended March 22, 2007)

Acknowledgements

Finding an advisor is arguably one of the most impactful decisions in graduate school, and I could not have chosen a better one than Hideo. If you ever talk to any of his students, you will no doubt be told about the freedom we are given to try out ideas. Rarely, though, do we follow up and explain that this freedom is not possible unless he also has enough patience to allow these ideas to bear fruit, and the tolerance for when they fail to do so. Perhaps this point is most significant to me as I may have tried his patience more than any of my other labmates during my time here! Despite my setbacks, whether they are research related or more personal in nature, he has always remained unwaveringly positive, choosing instead to focus on ways to help me move forward. I am very fortunate to have had an advisor who is as kind as he is brilliant, which is tough to accomplish when he is a certified genius. Thank you for finding ways to support me (for far longer than is perhaps deserved) so that I can graduate.

I would also like to acknowledge many of the faculty members and other mentors that have shaped how I think about science and research. In particular, I am indebted to Dr. Cohen for teaching me to ask myself the question, “wouldn’t it be nice if . . . ?” It did not get me over all my hurdles, but did help tremendously in leading me to find the manageable yet still meaningful ones to climb over. I am also grateful to Dr. Kimble for adopting me and the rest of MabuchiLab into his fold during our formative years. One of the most memorable moments during my time here was the day I truly understood the difference between *quantum* and *non-classical*. I was finally able to appreciate and genuinely respect the passion and resolve he had always exhibited. Finally, Dr. Herman has been an invaluable resource during the latter half of my time at Caltech. I have a much better appreciation for the process of research and learning, and surely would have been too stubborn to learn that lesson without his patient guidance through my most frustrating times. Outside of academics, he has also taught me a great deal about life; from operating system upgrades to effective parenting, and plenty more in between, for all of which I will always be grateful.

Many thanks to Parandeh, Marjie, Tara, Jim and Athena at the ISP for taking care of all this paperwork for us international students. I am always amazed how little I actually have to do despite nominally interacting with the government, and it is definitely to your credit.

To survive graduate school, you definitely need friends who can commiserate with you. My first couple years and all those classes would not have been the same without George Paloczi, Tobias Kippenberg, Stephan Ichiriu and Will Green. Whether it was trying to wrap up a problem set or deciding on a group to join, it was comforting knowing that I was not alone. To the members of Professor Scherer's Nanofab lab, and especially Marko Lončar, thanks for treating me as one of your own. I only wish I could have repaid you with better performance on the basketball court.

I had the great fortune of learning from excellent postdocs during my time here. Thanks in particular to Andrew Doherty for spending a lot of time with me early on teaching me everything from quantum trajectories to stochastic calculus (and on those car rides back from 'group meeting' the secret to a good cup of French press coffee). Jon Williams was instrumental in getting me up to speed on the physics of photonic crystals, and I learned a great deal just generally about how to go about conducting research from working with and observing him. To Luc Bouten, who remarkably took an interest in my work, thanks for the encouraging words throughout my thesis writing process. It meant more than you might have realized.

To the most intelligent set of 'fools' ever assembled, my fellow students in Mabuchi-Lab: it has been an honor to have shared this part of our careers together. To the young'uns Tony, Joe, Gopal, Nicole, Nathan and Orion: thanks for revitalizing the foosball tradition. That foosball table has greatly increased both the quality and quantity of my time at Caltech, though not necessarily in that order. May it live on and continue serving MabuchiLab at Stanford, and my best of luck to you guys in finding a regulation table. To Asa my officemate and politics liaison, thanks for the interesting conversations outside of research, and for putting up with my work area. Ramon deserves special mention for saving me in the eleventh hour by hacking the CIT thesis style file. Thanks to Tim for sharing his experiences with seeking an alternative career. Kevin 'Employee Of The Month' McHale deserves special acknowledgement for his contributions to chapter 4 of this thesis, which turned out to be instrumental in obtaining one of the key results of this work. I depart MabuchiLab knowing my foosball moves could not be in better hands. A huge thank you to Sheri, who keeps everything running smoothly despite having to put up with us juveniles.

Life definitely would not have been the same without the original crew, going way back to in an era when group meeting had an Alias, and the $\widehat{a}$'s annihilated rather than get annihilated. Good times . . . Ben, thanks for opening my eyes to what intense passion for science is all about. Mike, my true Laker brother, how will I ever forget chasing that factor of two with you? To Andy, for the numerous athletic activities that helped keep me sane, and last but not least, Stockton, thanks especially for helping me keep it real during the home stretch. I have many fond memories of our years together. Thanks for being such an integral part of my grad school experience.

To the rest of my thesis defense committee Oskar Painter, Chiara Daraio and Axel Scherer: Thanks for agreeing to be on my committee, and for the positive feedback and insightful questions you had during the defense. To you, the reader. Even if you are just reading the acknowledgements, I hope it has not been a waste of time.

My two little angels Charis and Akirin: you have brought such joy to me and kept me balanced. Thanks for reminding me of the importance of wonderment and curiosity.

And finally, my dearest wife Yuki, to properly thank you would more than double the length of this thesis. I still cannot fathom how blessed I am to have you in my life. Thanks for persevering with me, and just being with me throughout this journey. This is our victory.

Abstract

We present an *ab initio* treatment of the inverse photonic bandgap (or photonic crystal) device design problem. Using first principles, we derive the two-dimensional inverse Helmholtz equation that solves for the dielectric function that supports a given electromagnetic field with the desired properties. We show that the problem is ill-posed, meaning a solution often does not exist for the design problem. Our work elucidates fundamental limits to any inverse problem based design approach for arbitrary and optimal design of photonic devices. Despite these severe limitations, we achieve remarkable success in two design problems of particular importance to atomic physics applications, but also of general importance to the rest of the photonic community. As the first demonstration of our technique, we *arbitrarily* design the full dispersion curve of a photonic crystal waveguide. Dispersion control is important for maintaining the shape of pulses as they propagate along the waveguide. For our second demonstration, we take a point defect photonic crystal cavity in the nominal *acceptor* configuration (where the central defect has a lower index of refraction than the bulk material) and force it into the *donor* configuration (where the defect has a higher index of refraction than the bulk material), while requiring that the electromagnetic field maintain the properties of the acceptor mode. We were able to cross over this threshold while retaining a 93.6% overlap with the original mode.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The field of photonic bandgap (PBG) materials, or photonic crystals, has expanded tremendously over a relatively short period of time. These 'semiconductors of light' are artificially engineered materials that possess a periodicity in the index of refraction, leading to bands of frequencies for which electromagnetic waves cannot propagate in the material. Much effort has been poured into the field because of the enormous potential for ground-breaking applications using these novel materials. However, there have been many barriers towards taking full advantage of their properties. The idealized PBG materials have a full three-dimensional periodicity, but this has been difficult to fabricate, especially if intentional defects need to be inserted at precise locations to add functionality to the material. Hence, most device performances to date have yet to reach the promised potential, because they only use the bandgap effect in two-dimensions.

The two-dimensional (2D) system is less ideal, because losses can occur in the third dimension, so the coupling mechanism for the losses needs to be understood for different in-plane configurations. This greatly increases the importance and complexity of the design process. Since the field is relatively new (only 20 years), much of the research has been performed by scientists using trial and error. Parameters of the system are meticulously varied with the environment controlled, and the results of those changes analyzed and studied. In true double-edged sword fashion, one is fortunate in that one has enormous freedom on how to arrange the index of refraction, so many designs are possible. Unfortunately it also means that it can be an endless design challenge in the quest for better and better devices because of the vast number of degrees of freedom.

We seek a better alternative to an exhaustive search technique for photonic crystal device design, i.e., an algorithmic approach that removes the guesswork part of the process. The basic idea behind an inverse problem approach is to formulate the problem backwards. We start with the design goal we want accomplished, and then work backwards to find the geometry that would have produced the intended effect.

Finally, our inverse problem design method is derived *ab initio* using Maxwell's equations. We do not make use of approximate models to the system we wish to design. As such, for self-consistency we take great care in interpreting our results accordingly, and the drawback is that the designs may not readily apply to real structures one can fabricate currently. However, we can make far more general statements about the limitations and challenges of the design problem because our work is done *ab initio*.

1.2 Organization of the Thesis

The work in this thesis is an amalgamation of several fields that traditionally have minimal overlap. There has definitely been an increased interest in adapting inverse problem techniques to the PBG problem in recent years, but given the amount of research in each separate area, the degree of overlap is comparatively small. Convex optimization methods is another established field of research that is being successfully applied to many problems [1], but has yet to be relevant to the PBG community. As such, this thesis is aimed at the practitioner in any one of these fields interested in seeing how these come together to solve an engineering problem. Therefore, it is written in a way that is accessible for an incoming graduate student in any of the disciplines, although admittedly it does have more of a physics bias and assumes more background in physics.

The thesis is divided into two parts. The first part develops the necessary mathematical formalism in the three areas. In chapter 2 we review the basic physics of PBG materials and derive and work out the solution of the Helmholtz equation using the plane wave basis. Chapter 3 discusses the idea of an inverse problem, and works out a numerical example to illustrate what an inverse problem is and how to solve these notoriously difficult problems. We conclude part I with a chapter on convex optimization methods, which we will use as a specialized tool for solving photonic inverse problems. Again, a numerical example is provided to help provide some basic intuition into how the algorithm works.

Having provided the necessary mathematical background, we are then prepared to shift our focus to the problem of device design. We begin part II with an overview of PBG materials and devices in chapter 5. Readers familiar with PBG devices can safely skip most of the chapter, although in section 5.3 we motivate the design problems that we will tackle using our method. Chapter 3 reviews other inverse problem based design methods in the literature, and highlights some advantages and disadvantages to our approach. We derive *ab initio* the inverse Helmholtz equation, and perform a proof of principle demonstration to conclude the chapter. We proceed to adapt the inverse problem into a design methodology in chapter 7, revealing fundamental limitations towards achieving optimal designs. Despite these difficulties, we demonstrate the feasibility of a modified method that gives excellent results for our design goals that cannot be obtained with other methods.

Part I

Mathematical Formalism

Chapter 2

The Helmholtz Equation

The underlying physics of photonic crystals follow Maxwell's equations, and in this chapter we derive from first principles the wave equation on which the work in this thesis is based. There are many methods for solving the resulting partial differential equation (PDE), but the two most pervasive methods utilize either a time domain approach (e.g., finite difference time domain–FDTD)[2], or a frequency domain approach[3], and each method has its own merits[4].

The time domain method, as the name implies, is well suited for studying dynamical properties of the fields such as pulse propagation, transmission/reflection properties, estimate losses, etc. However, it is not as suitable for looking at resonant behavior with a single frequency of interest. While it is possible to use FDTD to look for resonant structures, one of the disadvantages is that computationally it is less efficient than the frequency domain approach, because you are looking at the system's entire response to a driving term. Another drawback of FDTD is that one must be careful that the source term is not orthogonal to the mode of interest, in other words, that there is sufficient coupling, or you risk not even finding the resonance. As the linewidth of the cavity decreases, the required precision of the frequency increases, which implies an increase in the number of timesteps required for the computation. However, the most notable advantage of FDTD is that it scales more favorably [5] than frequency domain methods.

As our group's interest in photonic crystals began with PBG cavities, it was natural to adopt the frequency domain approach. We will see in this chapter that the frequency domain problem using a plane wave basis reduces to a Hermitian eigenvalue problem, so there are many similarities to problems encountered in elementary quantum mechanics. However, there are some known convergence issues with this technique, and it is important that we draw attention to and highlight these in section 2.4.

For a more general overview for studying the physics of photonic crystals, the standard reference is the book by Joannopoulos et al. [6], or for a more mathematical treatment, there is Sakoda's book [7] as well.

Organization

We begin this chapter by deriving from Maxwell's equations the wave equation for an inhomogeneous medium. We show in section 2.2 the plane wave expansion method for solving the Helmholtz equation to find photonic bandstructures. Adapting the method for studying PBG materials with defect inclusions is reviewed in section 2.3. Readers familiar with the method can safely omit those sections. We conclude in section 2.4 with a discussion of the general convergence issues of the method. Even though the information in that section is not new, there are too many references in more current literature that seem unaware of the issues.

2.1 Bulk Photonic Crystal

Consider a position-dependent non-magnetic medium $\epsilon(\boldsymbol{r})$ with no charge or current sources. We further restrict ourselves to consider only linear and scalar (i.e., isotropic) materials. Maxwell's equations for the various fields take on the following form:

$$\nabla \cdot \boldsymbol{D}(\boldsymbol{r},t) = 0 \tag{2.1}$$

$$\nabla \cdot \boldsymbol{B}(\boldsymbol{r},t) = 0 \tag{2.2}$$

$$\nabla \times \boldsymbol{H}(\boldsymbol{r},t) = \frac{\partial \boldsymbol{D}(\boldsymbol{r},t)}{\partial t} \tag{2.3}$$

$$\nabla \times \boldsymbol{E}(\boldsymbol{r},t) = -\frac{\partial \boldsymbol{B}(\boldsymbol{r},t)}{\partial t}. \tag{2.4}$$

Using the constitutive relations, we can eliminate $\boldsymbol{D}$ and $\boldsymbol{B}$.

$$\boldsymbol{D}(\boldsymbol{r},t) = \epsilon(\boldsymbol{r})\boldsymbol{E}(\boldsymbol{r},t) \tag{2.5}$$

$$\boldsymbol{B}(\boldsymbol{r},t) = \mu_0\boldsymbol{H}(\boldsymbol{r},t) \tag{2.6}$$

Note that unlike the standard derivation of the wave equation for a homogeneous dielectric medium, we no longer have $\nabla\cdot\boldsymbol{E}=0$ from $\nabla\cdot\boldsymbol{D}=0$ because in general $\nabla\epsilon(\boldsymbol{r})\neq 0$. We can still work out the wave equations for $\boldsymbol{E}$ and $\boldsymbol{H}$ as follows.

2.1.1 Wave equation for $\boldsymbol{E}$

We take the curl of eqn. (2.4) and combine with eqn. (2.6) to obtain

$$\nabla\times\big(\nabla\times\boldsymbol{E}(\boldsymbol{r},t)\big) = -\mu_0\frac{\partial}{\partial t}\nabla\times\boldsymbol{H}(\boldsymbol{r},t) \tag{2.7}$$

We substitute in eqn. (2.3) and combine with eqn. (2.5) to arrive at the wave equation.

$$\nabla\times\big(\nabla\times\boldsymbol{E}(\boldsymbol{r},t)\big) + \mu_0\epsilon(\boldsymbol{r})\frac{\partial^2\boldsymbol{E}(\boldsymbol{r},t)}{\partial t^2} = 0 \tag{2.8}$$

The standard frequency domain method then enforces harmonic time dependence of the field to arrive at the Helmholtz equation for the electric field $\boldsymbol{E}$.

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E}(\boldsymbol{r})e^{i\omega t} \tag{2.9}$$

$$\therefore \nabla\times\big(\nabla\times\boldsymbol{E}(\boldsymbol{r},t)\big) = \epsilon_r(\boldsymbol{r})\frac{\omega^2}{c^2}\boldsymbol{E}(\boldsymbol{r}), \tag{2.10}$$

where $\epsilon_r(\boldsymbol{r}) \equiv \frac{\epsilon(\boldsymbol{r})}{\epsilon_0}$. It turns out that the left side of the equation is not self-adjoint (i.e. not Hermitian), so we are not guaranteed a complete basis [8]. Since the $\boldsymbol{H}$ equation (to be derived below) is self-adjoint, we will work exclusively with that equation instead, but for completeness we have presented the equation for $\boldsymbol{E}$ here. Of course, once we have the $\boldsymbol{H}$ field, we can use eqn. (2.3) and (2.5) to obtain $\boldsymbol{E}$.

2.1.2 Wave equation for $\boldsymbol{H}$

The derivation for the $\boldsymbol{H}$ equation is similar to the $\boldsymbol{E}$ equation. We begin by combining eqn. (2.3 and 2.5), take the curl and then incorporate eqn. (2.4 and2.6) to obtain the following:

$$\begin{aligned}
\frac{1}{\epsilon_r(\boldsymbol{r})}\nabla \times \boldsymbol{H}(\boldsymbol{r},t) &= \epsilon_0 \frac{\partial \boldsymbol{E}(\boldsymbol{r},t)}{\partial t} && (2.11)\\
\nabla \times \left(\frac{1}{\epsilon_r(\boldsymbol{r})}\nabla \times \boldsymbol{H}(\boldsymbol{r},t)\right) &= \epsilon_0 \frac{\partial}{\partial t}\nabla \times \boldsymbol{E}(\boldsymbol{r},t) && (2.12)\\
\nabla \times \left(\frac{1}{\epsilon_r(\boldsymbol{r})}\nabla \times \boldsymbol{H}(\boldsymbol{r},t)\right) &= -\mu_0\epsilon_0 \frac{\partial^2 \boldsymbol{H}(\boldsymbol{r},t)}{\partial t^2} && (2.13)\\
\nabla \times \big(\eta(\boldsymbol{r})\nabla \times \boldsymbol{H}(\boldsymbol{r},t)\big) &= -\mu_0\epsilon_0 \frac{\partial^2 \boldsymbol{H}(\boldsymbol{r},t)}{\partial t^2}, && (2.14)
\end{aligned}$$

where $\eta(\boldsymbol{r}) \equiv (\epsilon_r \boldsymbol{r})^{-1}$ is the reciprocal of the relative permittivity function. To keep the language from being overly cumbersome, we will refer to η simply as the 'dielectric function' for the rest of this thesis.

Invoking harmonic time dependence, we arrive at our Helmholtz equation for the magnetic field $\boldsymbol{H}$.

$$\nabla \times \big(\eta(\boldsymbol{r})\nabla \times \boldsymbol{H}(\boldsymbol{r})\big) = \frac{\omega^2}{c^2}\boldsymbol{H}(\boldsymbol{r}) \qquad (2.15)$$

We also point out that in the derivation, we have not used two of Maxwell's equations (Eqn. (2.1) and (2.2)), which means that a general solution to the Helmholtz equation will not obey all of Maxwell's equations. One must check or enforce these conditions in order to have physically realizable fields.

2.2 2D Plane Wave Expansion (PWE) Method

Since equation (2.15) is self-adjoint we are therefore guaranteed a complete basis. This implies that any spatially dependent function $f(\boldsymbol{r})$ can be expressed as a linear combination (i.e. superposition) of basis states.

$$f(\boldsymbol{r}) = \int a(\boldsymbol{k})\phi(\boldsymbol{k},\boldsymbol{r})d\boldsymbol{k} \tag{2.16}$$

$$\text{with} \qquad \int \phi(\boldsymbol{k},\boldsymbol{r})\phi(\boldsymbol{k'},\boldsymbol{r})d\boldsymbol{r} = \delta(\boldsymbol{k}-\boldsymbol{k'}), \tag{2.17}$$

where $\delta(\boldsymbol{k}-\boldsymbol{k'})$ is the Kronecker delta. The plane wave expansion method uses the set of planes waves $\{\exp(i\boldsymbol{k}\cdot\boldsymbol{r})\}$ as the complete basis. One reason for choosing the plane wave basis is that it is the eigenbasis in free space, so it is required for pseudo Q factor considerations [9]. Another significant advantage to using plane waves is that the vector fields will be divergence free (by construction), so all of Maxwell's equations (eqn. (2.1 and 2.2) in particular) are followed.

While it is certainly possible in principle to treat equation (2.15) vectorially in full 3 dimensions (3D) [10, 11], we will be making a 2D approximation for two reasons. First, current fabrication technology has limited almost all PBG devices to take on quasi-2D form, where the bandgap effect is utilized in only 2 dimensions and the transverse dimension relies on total internal reflection for optical confinement. Second, the computational resources required for a full 3D treatment of PBG devices using the plane wave method is currently prohibitively expensive[1]. A final advantage of the 2D treatment is that the TE (transverse electric) and TM (transverse magnetic) polarizations are decoupled.

To be more precise, when we refer to a 2D approximation, we mean that the dielectric function has full translational invariance in one dimension (say along the z direction), so $\frac{\partial\eta}{\partial z} = 0$. Under this configuration, the TE polarization has the electric field vectors lying in the xy plane, and the magnetic field points along the z direction. The TM polarization has the magnetic field in the xy plane and the electric field along z.

[1] See [4] summarized in section 2.4 for an exception to this statement.

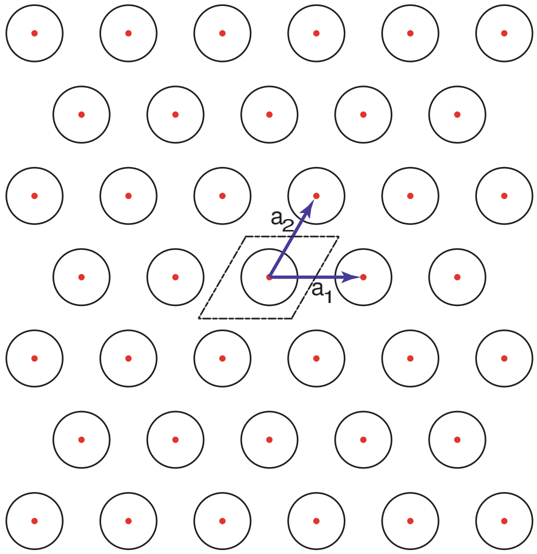


Figure 2.1: 2D hexagonal lattice in real space. The unit cell (dashed line) and the primitive vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ are shown.

2.2.1 Boundary conditions

For the case of photonic crystals, the dielectric function has translational symmetry along the primitive lattice vectors $\{\boldsymbol{a_1}, \boldsymbol{a_2}\}$ in the xy plane such that $\eta(\boldsymbol{r}) = \eta(\boldsymbol{r}+\boldsymbol{R})$, $\boldsymbol{R} \equiv m_1\boldsymbol{a_1} + m_2\boldsymbol{a_2}$, where $\{m_i\}$ are integers. For primarily historical reasons, we choose as our canonical geometry a hexagonal (sometimes referred to as a triangular) lattice of cylindrical air rods in semiconductor (see figures 2.1 and 2.2). The lattice symmetry allows us to define Born von-Karman periodic boundary conditions (BCs) for our PDE, as well as define a unit cell. Again, this naturally leads us to a description in Fourier space, where because of the periodicity we also have a lattice (the reciprocal lattice). The set of reciprocal lattice vectors $\{\boldsymbol{G}\}$ is defined by the following symmetry requirement:

$$e^{i\boldsymbol{G} \cdot (\boldsymbol{r} + \boldsymbol{R})} = e^{i\boldsymbol{G} \cdot \boldsymbol{r}} \tag{2.18}$$

$$\therefore e^{i\boldsymbol{G} \cdot \boldsymbol{R}} = 1. \tag{2.19}$$

We can leverage much of our intuition from solid state physics [12], and in fact, photonic crystals are often viewed as 'semiconductors for light.' In particular, Bloch's

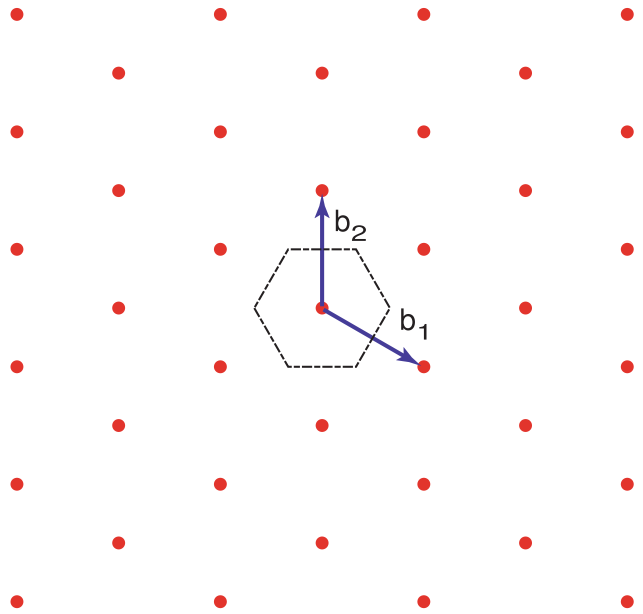


Figure 2.2: Reciprocal lattice in Fourier space of hexagonal lattice of figure 2.1

theorem applies (see chapter 8 in [12]), and we need only solve for the Bloch modes. A final warning is that in invoking the Born von-Karman BCs, we have implicitly assumed a photonic crystal of infinite extent. While this is a very standard approximation, and it is a good approximation if the physical structure is large compared to the lattice constant, it is important to be clear that the periodicity is meant to extend to infinity.

Translational symmetry implies we can express:

$$\eta(\boldsymbol{r}) = \sum_{\mathbf{G}} \eta_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}} \tag{2.20}$$

$$\therefore \eta_{\boldsymbol{G}} = \frac{1}{A_c}\int_{A_c} \eta(\boldsymbol{r}) e^{-i\boldsymbol{G}\cdot\boldsymbol{r}} d^2\boldsymbol{r} \tag{2.21}$$

$$\boldsymbol{H}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} h_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\ \hat{\boldsymbol{z}} \tag{2.22}$$

$$\therefore h_{\boldsymbol{G}} = \frac{1}{A_c}\int_{A_c} \boldsymbol{H}(\boldsymbol{r})\cdot\hat{\boldsymbol{z}}\ e^{-i\boldsymbol{G}\cdot\boldsymbol{r}} d^2\boldsymbol{r}, \tag{2.23}$$

where A_c denotes the area of the unit cell, and in eqn. (2.22) we have chosen the TE polarization. Eqn. (2.23) expresses the Fourier coefficient for some given Bloch mode. In calculating band structures, the relationship is true though not helpful, as we are trying to solve for the unknown Bloch mode.

We substitute into eqn. (2.15) the expressions in eqn. (2.20) and (2.22).

$$\begin{aligned}
\frac{\omega^2}{c^2}\sum_{\boldsymbol{G}} h_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\ \hat{\boldsymbol{z}} &= \nabla\times\left(\left(\sum_{\boldsymbol{G'}}\eta_{\boldsymbol{G'}}e^{i\boldsymbol{G'}\cdot\boldsymbol{r}}\right)\nabla\times\left(\sum_{\boldsymbol{G}} h_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\ \hat{\boldsymbol{z}}\right)\right) \\
&= \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \nabla\times e^{i\boldsymbol{G'}\cdot\boldsymbol{r}}\nabla\times e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\ \hat{\boldsymbol{z}} \\
&= \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \nabla\times e^{i\boldsymbol{G'}\cdot\boldsymbol{r}}(i\boldsymbol{G}\times\hat{\boldsymbol{z}})e^{i\boldsymbol{G}\cdot\boldsymbol{r}} \\
&= \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \nabla\times e^{i(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{r}}(i\boldsymbol{G}\times\hat{\boldsymbol{z}}) \\
&= \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \big[i(\boldsymbol{G}+\boldsymbol{G'})\times(i\boldsymbol{G}\times\hat{\boldsymbol{z}})\big]e^{i(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{r}} \\
&= \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \big[(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{G}\big]e^{i(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{r}}\hat{\boldsymbol{z}},
\end{aligned}$$

where in the last step we made use of the triple cross product identity $a\times(b\times c) = b(a\cdot c) - c(a\cdot b)$ and the fact that $\{\mathbf{G}\}$ lie in the xy plane. We then take the inner product of both sides with a TE polarized plane wave by left multiplying and then integrating:

$$\frac{1}{A_c}\int_{A_c}\hat{\boldsymbol{z}}\, e^{-i\boldsymbol{G''}\cdot\boldsymbol{r}}\left\{\frac{\omega^2}{c^2}\sum_{\boldsymbol{G}} h_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}\ \hat{\boldsymbol{z}} = \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \big[(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{G}\big]e^{i(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{r}}\hat{\boldsymbol{z}}\right\}d^2\boldsymbol{r}$$

$$\frac{\omega^2}{c^2}\sum_{\boldsymbol{G}} h_{\boldsymbol{G}}\delta_{\boldsymbol{G},\boldsymbol{G''}} = \sum_{\boldsymbol{G},\boldsymbol{G'}} \eta_{\boldsymbol{G'}} h_{\boldsymbol{G}} \big[(\boldsymbol{G}+\boldsymbol{G'})\cdot\boldsymbol{G}\big]\delta_{\boldsymbol{G}+\boldsymbol{G'},\boldsymbol{G''}}$$

We choose to collapse the delta function on the right with $\boldsymbol{G'} = \boldsymbol{G''} - \boldsymbol{G}$. Relabeling the indices, we arrive at the Helmholtz equation in the plane wave basis.

$$\sum_{\boldsymbol{G}'}\left[\eta_{\boldsymbol{G}-\boldsymbol{G}'}\boldsymbol{G}\cdot\boldsymbol{G}'\right]h_{\boldsymbol{G}'}=\frac{\omega^2}{c^2}h_{\boldsymbol{G}} \tag{2.24}$$

This is the starting point for bandstructure calculations of bulk (i.e. defect-free) photonic crystals. So far we have only considered $\mathbf{k}$-points that are coupled to the $\boldsymbol{q}=0$ component, where $\{\boldsymbol{q}\}$ denotes the set of $\mathbf{k}$-vectors that lie in the first Brillouin zone (unit cell in reciprocal space). To complete the bandstructure to include the other $\mathbf{k}$-points, we simply use a different expansion of the field.

$$\boldsymbol{H}_{\boldsymbol{q}}(\boldsymbol{r})=\sum_{\boldsymbol{G}}h_{\boldsymbol{q}+\boldsymbol{G}}e^{i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}\;\hat{\boldsymbol{z}} \tag{2.25}$$

Here, $\boldsymbol{q}$ labels the plane wave that is used to modulate the Bloch function. In general, the Bloch functions for different $\boldsymbol{q}$ will be different, as (following the same steps as in the above derivation) we arrive at the following:

$$\sum_{\boldsymbol{G}'}\left[\eta_{\boldsymbol{G}-\boldsymbol{G}'}(\boldsymbol{q}+\boldsymbol{G})\cdot(\boldsymbol{q}+\boldsymbol{G}')\right]h_{n,\boldsymbol{q}+\boldsymbol{G}'}=\frac{\omega_{n,q}^2}{c^2}h_{n,\boldsymbol{q}+\boldsymbol{G}} \tag{2.26}$$

Written in this form, each $\boldsymbol{q}$ within the first Brillouin zone labels a unique eigenvalue problem, and the n labels the band index of a particular mode. We do not need to separately consider $\mathbf{k}$'s outside the first Brillouin zone because all $\mathbf{k}$'s that differ by a translation of some $\mathbf{G}$' in $\{\mathbf{G}\}$ are coupled to each other. We use the band index to keep track of the $\mathbf{k}$'s in outer Brillouin zones.

Eigenfunctions that satisfy eqn. (2.26) are called 'bulk modes'. The eigenvalues give us the frequencies that are permitted within the material. When we solve a series of these eigenvalue problems for $\boldsymbol{q}$ along high symmetry points, we obtain the band diagram for the material. For a distribution of material with sufficient dielectric contrast (e.g. low-index air rods in high-index semiconductors), we find certain ranges where the frequencies are forbidden, which means that light at those frequencies cannot propagate in the material, hence the term 'bandgap.'

However, most useful PBG devices require that the symmetry be broken by introducing *defects* in the lattice. Defects can be an omission of an air rod, or an air rod of a different size or shape, or any other structure that causes a break in the symmetry of the system. Structures with defects are studied in the PWE method using the supercell approximation.

2.3 Supercell Treatment

The strategy here is to surround the defect(s) with enough layers of the bulk photonic crystal that the modes of interest become well localized within the defect region. We can then invoke the tight binding approximation and consider the defect plus surrounding layers as a 'unit cell,' or a supercell. This implies a lattice of defects that extend to infinity (see figures 2.3 and 2.4).

This is another standard approximation, and is valid if there is minimal interaction between the artificial neighboring defect sites. This is analogous to the tight-binding approximation in solid state physics. Note that for the line defect, the defects along the x-axis should not be considered artifacts of the supercell method, since there is a real translational symmetry in that direction (i.e. the waveguiding direction). When solving for the modes of the waveguide, we need not be alarmed or concerned if the mode is not localized in the x direction. However, that is not true in the y direction. Finally, as with the bulk photonic crystal case, the translational symmetry of the real waveguide does not extend to infinity even in the x direction.

2.3.1 Point defect: cavity

For the case of the point defect cavity, the extension to the Helmholtz equation is trivial. Instead of the set $\{\mathbf{G}\}$, we now have a new set of reciprocal lattice vectors which we will refer to as $\{\mathbf{k}\}$. We no longer have to worry about $\{\boldsymbol{q} \neq 0\}$ in the Brillouin zone since they do not form a valid expansion. As we do not actually have a supercell periodicity, physically we cannot support these longer wavelength plane

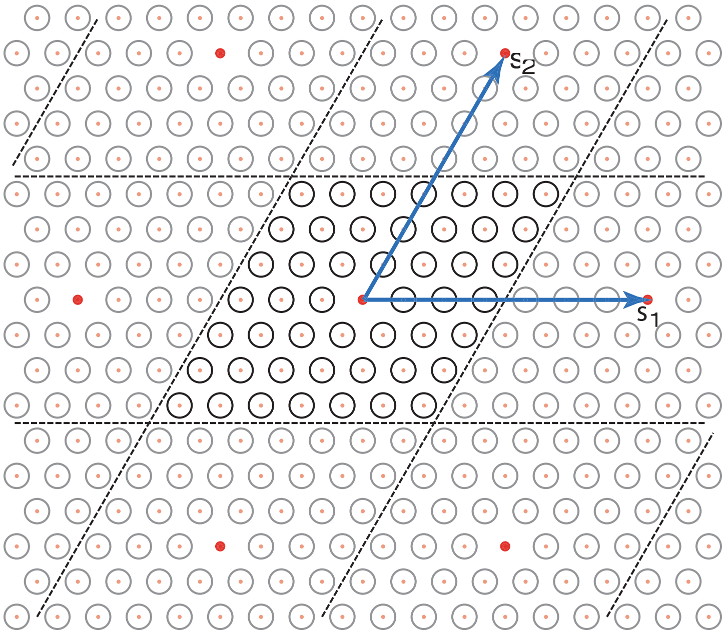


Figure 2.3: Real space dielectric function of a hexagonal lattice of air cylinders embedded in an infinitely thick block of semiconductor. The point defect is formed by omission of the central air cylinder. In the supercell treatment, the entire device makes up the 'unit cell' (outlined in the dotted line) is artificially tiled to give periodic boundary conditions. The original lattice points are shown as the faint red dots, but hold no particular special significance in the supercell treatment. The new superlattice points are the larger red dots, with the superlattice vectors $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ shown. The satellite defect structures are included but shown slightly faded out.

wave modulation of Bloch modes. The Helmholtz equation then simply sets $\boldsymbol{q} = 0$ and substitute **k**'s for **G**'s.

In principle, the set $\{\mathbf{k}\}$ extends to infinity, although in practice we always truncate at some finite bandwidth. There are some implications to truncation which are often overlooked. We discuss these and other subtle issues with the discrete fourier transform in appendix C. With a truncated basis, we can now obtain the Helmholtz operator in matrix form.

$$\left[\sum_{\boldsymbol{k'}} \eta_{\boldsymbol{k}-\boldsymbol{k'}} (\boldsymbol{k}\cdot\boldsymbol{k'})\right] h_{\boldsymbol{k'}} = \frac{\omega^2}{c^2} h_{\boldsymbol{k}} \tag{2.27}$$

$$\hat{\boldsymbol{\Theta}}_{\boldsymbol{kk'}}{}^{(\eta)} h_{\boldsymbol{k'}} = \frac{\omega^2}{c^2} h_{\boldsymbol{k}} \tag{2.28}$$

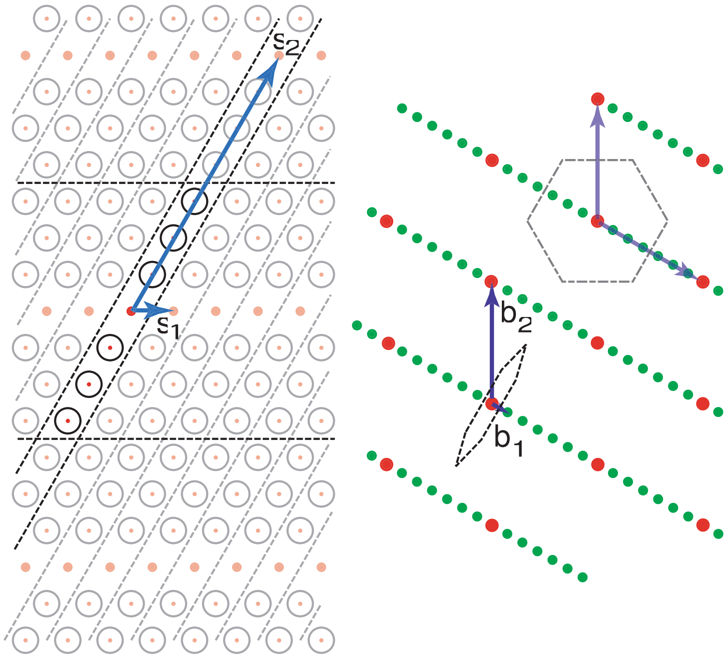


Figure 2.4: On the left side of the figure, we show the real space dielectric function of a row defect in the supercell treatment. Small red dots correspond to original lattice, and the larger red dots the 'basis vectors' for the superlattice. On the right is a picture of the reciprocal lattice. The red dots show the reciprocal lattice of the bulk photonic crystal, and the green dots show the reciprocal lattice of the new supercell. The shape of the original and supercell Brillouin zones are shown as well, with the original one shown faded.

We have omitted the band index for clarity, since in general we will only be interested in a small number of modes that lie within the bandgap. We denote $\hat{\boldsymbol{\Theta}}_{\boldsymbol{k}\boldsymbol{k'}}{}^{(\eta)}$ as the Helmholtz operator, and the problem of finding the eigenmodes $\boldsymbol{H_m(r)}$ for some given distribution of dielectric will be referred to as the 'forward problem.' The superscript η makes it explicit that the operator depends on the chosen dielectric function. We can form the Helmholtz operator once we have the Fourier coefficients $\eta_{\boldsymbol{k}}$, which can be obtained analytically or numerically. It is important to note here that we set $\eta_{\boldsymbol{\kappa}} = 0$ for $\boldsymbol{\kappa} = \mathbf{k} - \mathbf{k'}$ that lie outside the truncation bandwidth. As mentioned, appendix C will explore these issues in greater detail.

2.3.2 Line defect: waveguide

For the line defect waveguide, we also replace the **G**'s with **k**'s, but we must now include the $\boldsymbol{q}$'s that correspond to the propagation direction. Our forward problem takes on the following form:

$$\left[\sum_{\boldsymbol{k'}} \eta_{\boldsymbol{k} - \boldsymbol{k'}} \big((\boldsymbol{q}+\boldsymbol{k}) \cdot (\boldsymbol{q}+\boldsymbol{k'})\big)\right] h_{\boldsymbol{q}+\boldsymbol{k'}} = \frac{\omega_{\boldsymbol{q}}^2}{c^2} h_{\boldsymbol{q}+\boldsymbol{k}} \tag{2.29}$$

$$\hat{\boldsymbol{\Theta}}_{\boldsymbol{kk'}}{}^{(q,\eta)} h_{\boldsymbol{q}+\boldsymbol{k'}} = \frac{\omega_{\boldsymbol{q}}^2}{c^2} h_{\boldsymbol{q}+\boldsymbol{k}} \tag{2.30}$$

again, omitting the band indices and making the dependence on η explicit. The dispersion relation $\omega(\boldsymbol{q})$ of the waveguide can be found by solving the forward problem for several $\boldsymbol{q}$'s along the propagation direction, and identifying the waveguiding mode of interest whose frequency lies within the bandgap.

2.4 Convergence Issues of the PWE Method

In the opening section of this chapter, we alluded to some known convergence issues with the PWE method. From working out the Helmholtz equation in the plane wave basis, the connection is clear between the plane wave method and Fourier analysis so it is not surprising that any difficulties in Fourier analysis (see appendix C) will lead to difficulties here. However, as it applies to solving the photonic bands problem, these issues were studied as early as 1992 by Sözüer, Haus and Inguva [13]. This was in an era when the field of photonic crystals was still in its infancy, and tremendous amount of effort went into accurate calculations of the band structure in search of true band gaps. In an effort to correct a common misconception, they wrote the following:

> *It is clear that, just because increasing N does not produce visible differences in the resulting band structure, one has not necessarily converged to the 'true' values. In this case, it is merely an indication of the slow*

convergence of the Fourier series.

Here, N refers to the number of terms in the summation. Unfortunately, most convergence analysis on the PWE method fails to address the problem, and even research published over a decade later [14, 15, 16] still fails to grasp the difference between convergence and accuracy. After all, what is the significance of 'rapid convergence' of the calculation if it does so to a wrong value? The problem is, of course, the slow convergence of the underlying dielectric function that the finite Fourier series is supposed to model. Additional terms in the series do not change the model enough so it only appears as though the calculation has converged.

In the literature, there was also some discrepancy as to how to treat the $\eta = \epsilon^{-1}$ term in the Helmholtz equation (eqn. (2.15)). While some have treated the Helmholtz operator as we have, others [11] expand $\epsilon_r(r)$ in the plane wave basis, and then invert the matrix instead. For an infinite Fourier series, the matrices $\eta_{k,k'}$ and $\epsilon_{k,k'}$ would be each other's inverse, but that is no longer true once we truncate. It appeared that the convergence was more rapid (i.e. used a fewer number of plane waves) using the matrix inversion treatment [17], thus justifying the computationally intensive matrix inversion, but bear in mind the *caveat* about convergence from above. Of course, in the early 1990's, there were more constraints on CPU and memory resources than there are today. This issue is addressed nicely by Li in 1996 [18], and we will review his work in section C.4. In the end, the point is moot, since Steven Johnson et al. found a way [4] to implement the plane wave method without explicitly forming either matrix, and have since made their software, the MIT Photonic Bands (MPB) package, freely available.

There are three key ideas to their approach. First, rather than solving the eigenvalue problem explicitly by diagonalization (computational work required $O(N^3)$), they use an iterative eigensolver. Secondly, they noticed that in the Helmholtz equation, the curl operator is diagonal in Fourier space, while the division by ϵ is diagonal in real space, so they make use of the Fast Fourier Transform (FFT) algorithm to transform to the appropriate basis so they do not actually store the entire $N \times N$ operator. They were thus able to reduce their storage requirement from $O(N^2)$ to $O(N)$.

This allowed them to use a much higher number of plane waves than otherwise practically achievable. However, representation of discontinuities in a Fourier basis still poses a problem. The final ingredient is an averaging technique that smoothes out the discretized elements that encloses a discontinuity. It makes use of effective medium theory and the grid elements containing the discontinuities are assigned a dielectric tensor instead of a scalar. Therefore, they have replaced the sharp scalar dielectric discontinuity with a smoothed dielectric tensor, making the Fourier representation less objectionable.

We have chosen to highlight some of the key contributions of that work here for two reasons. The first reason is that we were unable to take advantage of their ideas in the inverse problem, so we still face the same convergence issues described in [13]. It should become clear when we formulate the inverse Helmholtz problem in chapter 6 why we could not capitalize on their wisdom. The other reason is that though their work is often cited, there still appears to be confusion about the validity of the PWE method, especially the key components to making the method work. Other more recent articles on PWE [14, 16, 15, 19] either omit the reference entirely, or if it does cite it, the work demonstrates a complete lack of appreciation for the results by Johnson et al.. A recent article (2006) on these ‘fast Fourier factorization methods’ [20] still have not caught on to the fact that their method is at best equivalent if not inferior to the tensorial averaging in the Johnson reference, except that they do not even realize the $N \log N$ scaling, requiring $O(N^2)$ storage. For the reader interested in performing these computations, it is important to understand the significance of Johnson’s work and to evaluate other PWE method research in light of their results.

Chapter 3

Inverse Problems

Mathematicians often cannot help but cringe when physicists are doing math, and even more so whenever physicists claim to be 'rigorous' with their math. This chapter is not written to satisfy the mathematicians for two reasons. First, I am not a mathematician, so I am quite certain that they will cringe despite my best efforts. More importantly though, this is meant to be accessible for engineers and physicists, and often what mathematicians consider 'special cases' are the only ones we happen to care about. So with apologies to any mathematicians reading this, the goals of this chapter are threefold: First, we want to help the reader develop an appreciation for what inverse problems are and what makes them difficult. Second, we want to introduce the specialized tools that are used to solve these inverse problems. Finally, we bring the focus back to our particular application, and fine tune the ideas developed for the purpose of photonic device design.

There are many excellent references on inverse problems. A standard reference is the textbook by Engl [21] which gives a thorough overview of the subject, but the mathematics is quite formal. A very nice introduction to the subject for physicists can be found in a series of lecture notes by Sze Tan and Colin Fox at the University of Auckland [22]. The work by Per Christian Hansen is more focused on discrete and finite dimensional problems, and hence particularly suitable to our application. He has also written a package of matlab functions for inverse problems available for download as well [23]. The ideas presented in this chapter are mostly taken from his work, although the discussion of the role of noise in distinguishing between a forward and inverse problem has to our knowledge not been articulated elsewhere. Arnold Neumaier also provides a concise treatment similar to Hansen's, but bridges the gap to the infinite dimensional treatment of inverse problems with more mathematical rigor [24].

We begin the chapter by attempting to define what an inverse problem is through some examples of simple physical problems. We introduce the concept of an *ill-posed* problem to distinguish between the *forward* or *direct* problem vs. the inverse problem. In section 3.2, we restrict our discussion to finite dimensional linear operators, allowing us to illustrate the pathologies of inverse problems in the linear algebra formalism. A numerical example is provided to help illustrate the effects of ill-conditioning. We make use of the singular value decomposition (SVD) to explain why standard techniques will fail to solve inverse problems. The SVD also allows us to utilize the condition number as a quantifying metric for how ill-posed a particular problem is. In section 3.3 we introduce regularization as a tool for solving inverse problems. We conclude with a glimpse of the difficulties we expect to encounter for the purpose of PBG device design.

3.1 Introduction

At first glance, the meaning of the term 'inverse problem' seems obvious. It is the complement of some other problem, one that presumably preceded the inverse problem, and is more well known. To a physicist though, such a 'definition' is rather unsavory, for if that were the case, then the distinction between a forward problem and an inverse problem seems rather arbitrary. Our obsession with symmetries in natural laws lead us naturally to wonder why one problem formulation is more 'privileged' than the other. A good example of this that we have already encountered in this thesis is the Fourier transform and the inverse Fourier transform. The two operations are simply labeled that way by convention, and nothing would have been lost had we reversed the labels. It becomes a question of semantics, rather than a matter of any fundamental significance. By the end of this chapter, we will see that the inverse Fourier transform in fact does not fit our definition of an inverse problem.

The distinction is in reality more than just semantics or there would not be an entire journal devoted to inverse problems. One's first exposure to inverse problems is typically accompanied by some claim that inverse problems are difficult to solve, with the implication being that it is *more* difficult than the associated forward problem. We give a more formal definition in the next section, but first, we review a few well-known examples of inverse problems to develop some intuition.

3.1.1 Examples

Our first example is found in medical imaging, such as computerized tomography (CT) scans. The forward problem is a form of scattering or diffraction problem, such that for some radiation incident upon a given material distribution, we determine the scattered radiation in the far field. For medical applications, the goal is to non-invasively determine the internal structure of a patient's body. This is accomplished by measuring the scattered field at various angles given some incident radiation, and solving the inverse scattering problem for the scatterer distribution. A related inverse problem is found in geophysics, where the internal structure of the earth is determined based on surface measurements of seismic waves.

Another example is in image processing, or image restoration, where the ideal image must pass through non-ideal optics, leading to blurring and other distortion to the captured image. The forward problem of blurring is typically modeled as a convolution of the original image $i_o(x)$ with a point spread function $h(x)$. Sharp features are smeared out, leading to a loss of resolution. Formally, our captured image $i_c(x)$ becomes:

$$\begin{aligned} i_c(x) &= \int i_o(\xi)h(x-\xi)d\xi, \text{ or} && (3.1) \\ I_c(k) &= I_o(k)H(k) && (3.2) \end{aligned}$$

The inverse problem becomes a deconvolution, which can be performed as a simple division of the captured image with the point spread function in their respective Fourier representations. One can determine $h(x)$ by characterizing the optical elements carefully. It turns out that the image cannot be reconstructed with a straightforward application of the deconvolution theorem. Section 1.6 in [22] provides a nice pictorial example of this problem.

A final example is the heat conduction problem. The forward problem is of course a standard undergraduate-level problem, and can be described by some variant of the following:

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= \frac{\partial^2 u(x,t)}{\partial x^2}, \quad x \in [0,\pi], t \geq 0, && (3.3) \\ \text{where} \quad u(x,0) &= f(x), && (3.4) \\ u(0,t) &= u(\pi,t) = 0. && (3.5) \end{aligned}$$

This is solved in the usual way by separation of variables and then an eigenfunction expansion for the spatial dependence, with the set of normalized sine functions $\{\phi_n(x)\}$ forming a complete orthonormal set. Expressing the initial distribution in terms of a superposition of the eigenfunctions $f(x) = \sum_n c_n \phi_n(x)$, we obtain the heat distribution $u(x,t)$ as

$$u(x,t) = \sum_n c_n e^{-n^2 t} \phi_n(x). \tag{3.6}$$

There is often some remark about our inability to solve the backwards heat conduction problem, namely given some final distribution $u(x,t_f)$, we generally cannot go backwards in time and deduce an initial distribution. Typically this is attributed to the exponential factor, and we see that it blows up if we go backwards in time.

Based on these examples, we can make some observations that will prove to be helpful. First, the heat conduction example makes explicit a common theme among the examples provided: that of *cause* and *effect*. Whereas forward problems in physics tend to study the unknown effects of known causes (in order to derive a model for predicting the effects), inverse problems seek the unknown cause of measured effects. The backwards heat conduction equation makes this transparent because of the explicit time dependence, but the rest of the examples all seek an explanation of some final observed phenomenon, given some well-characterized, presumably accurate model of the forward problem.

The other observation is that at least in some of these forward problems, there appears to be a ‘smoothing’ process. For example, in the heat conduction and the blurring examples, features present in the initial condition seems to get lost or smoothed out as the state evolves forward. Often, we arrive at steady state solutions of dynamical systems, therefore independent of initial conditions: The system forgets or loses information about the initial state. In such a case, we can certainly see just by physical principles alone why an inverse problem would be ‘difficult.’ A more precise way to look at this might be how ‘solvable’ a given problem is, which leads to the notion of well-posed and ill-posed problems proposed by Hadamard.

3.1.2 Well-posedness

Hadamard proposed three properties that a problem must possess in order to be classified as well-posed [21]:

1. For all admissible data, a solution exists.
2. For all admissible data, the solution is unique.
3. The solution depends continuously on the data.

What constitutes ‘admissible data,’ ‘solution’ and ‘continuous’ will of course depend on the nature of a specific problem. For our purposes, considering only finite dimensional linear operators, we can think of data and solution as the *input* and *output* vectors of some linear transformation. The first two properties seem rather obvious, as it is not much of a linear mapping if we, for some given input, cannot get an output, or get multiple or non-unique outputs. The third property is a question of stability, requiring that small changes to the input does not produce arbitrarily large changes to the output. A problem that lacks any one of these properties is by definition *ill-posed.*

We can apply our intuition to the heat conduction example and readily see that indeed a solution to the inverse problem does not always exist. A simple example is if our final field distribution corresponds to a state with minimal entropy. Since entropy must increase with time, we know that there is no way to go backwards in time, since we are ‘starting’ in a minimum entropy state. As for the uniqueness of the solution to the inverse problem, we already addressed the problem of the steady state fields, which means *any* initial state would reach the same steady state. The final property of stability relates to the smoothing behavior of the forward problem. If we perturb the initial conditions by a small amount, the perturbations will be smoothed out over time to yield similar output fields. By extension then, small perturbations at the final time must have corresponded to large changes in the initial condition. We observed this effect quantified by the exponential term in the heat conduction equation. We now express these ideas in a more formal mathematical footing in the context of finite dimensional linear operators which can be represented by matrices.

3.2 Matrices as Linear Operators

3.2.1 A numerical example

We first provide a numerical example to give a concrete illustration of the ideas presented in the previous section. The matlab code used to generate this example is included in appendix A. Consider the $N \times N$ Hilbert matrix.

$$\begin{aligned} A(i,j) &= \frac{1}{i+j-1} && (3.7) \\ \text{with} \quad Ax_{in} &= x_{out} && (3.8) \end{aligned}$$

For this example, we will choose $N = 5$, and choose a relatively simple x_{in}.

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}, \quad x_{in} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{3.9}$$

Evaluating x_{out} gives:

$$x_{out} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.2833 \\ 1.4500 \\ 1.0929 \\ 0.8845 \\ 0.7456 \end{bmatrix}. \tag{3.10}$$

Clearly, for any x_{in}, we can evaluate a unique x_{out}. Therefore the first two Hadamard conditions are satisfied. To test the stability condition, we can define an additive noise vector n that is sufficiently small to form x'_{in}. Evaluating x'_{out} gives:

$$x'_{out} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 0.9982 \\ 0.9992 \\ 0.9997 \\ 1.0003 \\ 1.0010 \end{bmatrix} = \begin{bmatrix} 2.2813 \\ 1.4490 \\ 1.0922 \\ 0.8840 \\ 0.7452 \end{bmatrix}. \tag{3.11}$$

We see that x'_{out} is close to the nominal solution x_{out}. Formally we can define the relative magnitude of the input and output error to provide a measure of the stability

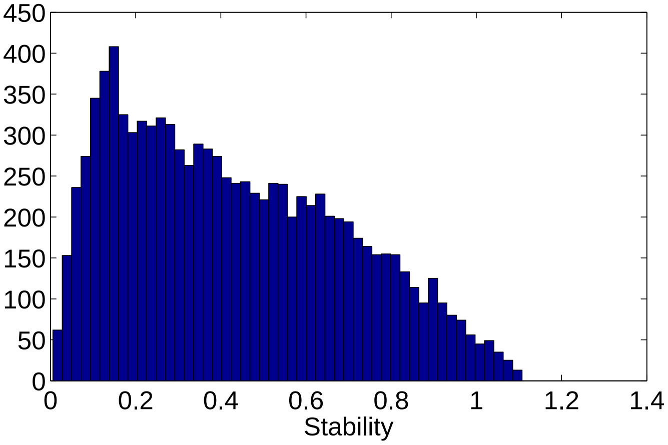


Figure 3.1: Distribution of stability values for the Hilbert matrix operator

$\mathcal{S}$:

$$e_{in} = \frac{|x'_{in} - x_{in}|}{|x_{in}|} \tag{3.12}$$

$$e_{out} = \frac{|x'_{out} - x_{out}|}{|x_{out}|} \tag{3.13}$$

$$\mathcal{S} \equiv \frac{e_{out}}{e_{in}} = 0.7846 \tag{3.14}$$

where $|\cdot|$ denotes the 2-norm (i.e., $|x| = (\sum_i x_i^2)^{1/2}$). We repeat this with $10,000$ different noise vectors and show the distribution of $\mathcal{S}$ in figure 3.1. Most of the values fall between 0 and 1, with the maximum value of about 1.1. Therefore, we see that this problem is stable against perturbations to the input vector, i.e., errors remain small.

We now look at the 'reverse' problem of finding x_{in} given x_{out}. We will look at the stability again, but this time, we add the noise to the nominal x_{out}. We solve for

$x'_{in} = A^{-1}x_{out}$. We use matlab to find the inverse of A.

$$A^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix} \tag{3.15}$$

Again, using $10,000$ different noise vectors, we obtain the distribution of $\mathcal{S}$ for this reverse problem (as shown in 3.2). Notice the x-axis is scaled by 10^5, meaning the relative error is greatly amplified. To illustrate, suppose we rounded x_{out} to 3 decimal places and then evaluated x'_{in}.

$$x'_{out} = \begin{bmatrix} 2.283 & 1.450 & 1.093 & 0.885 & 0.746 \end{bmatrix}^T \tag{3.16}$$

$$x'_{in} = A^{-1}x'_{out} = \begin{bmatrix} 2.105 & -20.28 & 94.29 & -141.4 & 71.19 \end{bmatrix}^T \tag{3.17}$$

In the reverse problem, we cannot even tolerate rounding errors as x'_{in} bears no resemblance to x_{in} at all. Therefore, this problem fails to satisfy Hadamard's third condition and is therefore ill-posed. Because of the ill-posedness, this reverse problem is the one that is defined to be the inverse problem. Therefore, it is not simply a question of semantics, but there are fundamental distinctions between a forward and its inverse problem. Even if we had first defined an operator $B = A^{-1}$ and went through this same analysis, we would still conclude that B is the 'inverse problem,' and B^{-1} is the 'forward problem,' objectively based on the stability criterion.

In chapter 6, when we derive the inverse Helmholtz equation, we will encounter a more severe manifestation of this problem, where we cannot (even without the additive noise) recover the input with the computed output. However, if we understand the analysis in this chapter, it will no longer be surprising when we get to chapter 6. To understand the origins of this ill-conditioning, we need to take a closer look at the properties of a general linear operator A in terms of the singular value decomposition.

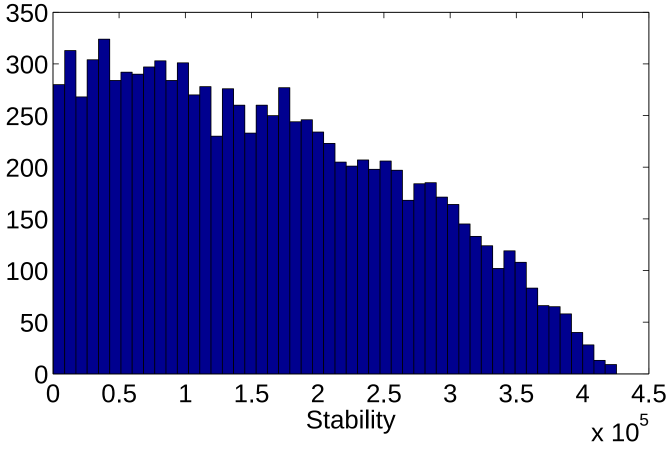


Figure 3.2: Distribution of stability values for the inverse Hilbert matrix operator. Notice the scale on the x-axis is in increments of 10^5.

3.2.2 Singular value decomposition

Consider a general linear transformation (or linear mapping) $A : \mathcal{C}^n \to \mathcal{C}^m$, such that

$$Ax = b \tag{3.18}$$

with A an $m \times n$ matrix, $x \in \mathcal{C}^n$ and $b \in \mathcal{C}^m$.

Any matrix A can be decomposed by the singular value decomposition (SVD) such that

$$A = U\Sigma V^\dagger, \tag{3.19}$$

where the $m \times m$ matrix U and the $n \times n$ matrix V are unitary, and Σ is an $m \times n$ matrix whose only non-zero elements are along the diagonal with $\{\sigma_i \geq \sigma_{i+1} \geq 0\}$ called the singular values. The columns of U and V are known as the left $\{u_i\}$ and right $\{v_i\}$ singular vectors. This is a generalization of the eigenvalue decomposition. In fact (although one would not actually do so in practice), one can get the SVD by performing an eigenvalue decomposition of $AA^\dagger$ and $A^\dagger A$. The eigenvectors of $AA^\dagger$ and $A^\dagger A$ are the left and right singular vectors of A respectively, and the eigenvalues are the singular values squared. Since both $AA^\dagger$ and $A^\dagger A$ are Hermitian and positive semi-definite, we are guaranteed real non-negative eigenvalues, thus ensuring $\sigma \geq 0$.

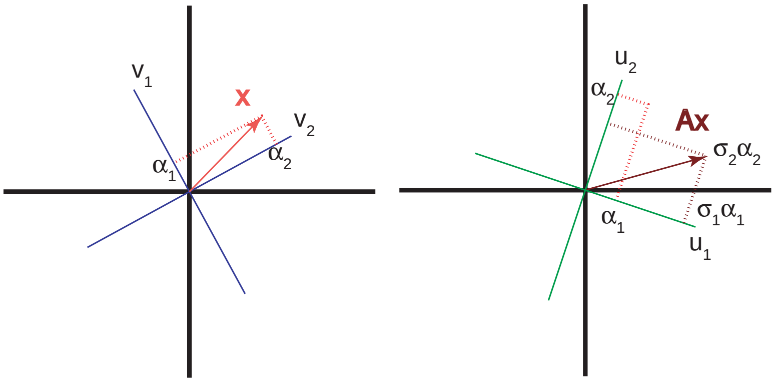


Figure 3.3: The vector x to be transformed is decomposed into v_1 and v_2 and then mapped onto corresponding u_1 and u_2, with each component stretched or compressed by the respective σ. In this example, $\sigma_2 < 1$.

Having obtained the SVD of A, we can write down the linear mapping in a more suggestive form:

$$Ax = U\Sigma(V^{\dagger}x) \tag{3.20}$$

$$A\mathbf{x} = \sum_{i=1}^{\min(m,n)} \left(\mathbf{v}_i^{\dagger}\mathbf{x}\right)\sigma_i\mathbf{u}_i, \tag{3.21}$$

where for clarity we have written the vectors in boldface. Looking at eqn. (3.21), we see that any linear mapping can be viewed as a transformation of the input vector x into the right singular vector basis, then stretching each component by the associated singular value, and finally mapping these components to the corresponding left singular vectors. A pictorial representation for a simple 2D mapping is shown in figure 3.3. For special cases of A that has an eigenvalue decomposition (i.e. diagonalizable), the left singular vectors are the same as the right singular vectors, so in the diagonalized basis, the linear transformation is particularly simple (just stretch each component by the eigenvalues; this is, of course, why we prefer to work in a diagonalized basis). The singular values of A play an important role since they determine how much gain is in a particular component of the linear map. For the time being, let us consider the problem of finding b (given A and x) to be the forward problem, while the problem of finding x (given that Ax produces b) is the inverse problem.

Revisiting Hadamard's conditions then, in the forward problem, the first two conditions are automatically satisfied if we can express the problem in this form. As for the third condition, we can think of it as requiring reasonable gains (i.e. not too large) for the system. Stability can also be achieved if random perturbations are spectrally decomposed to singular vectors that have relatively small singular values. In other words, singular vectors associated with small singular values should look like noise. For physical systems the relevant physics are embodied by the linear operator A.

If we now attempt to solve the inverse problem, we need to do the following:

$$\begin{aligned} Ax &= b && (3.22)\\ x &= A^{-1}b && (3.23)\\ &= V\Sigma^{-1}U^{\dagger}b && (3.24)\\ &= \sum_{i}^{\min(m,n)} \left(\frac{1}{\sigma_i}\mathbf{u}_i^{\dagger}\mathbf{b}\right)\mathbf{v}_i. && (3.25) \end{aligned}$$

In eqn. (3.25) above, we have expressed the inverse of A using the SVD expansion. Even when A is singular (i.e. not strictly invertible), the expression can be used and interpreted as a generalized inverse or 'pseudo-inverse,' although there are of course limitations associated with a singular A. Now A is obviously singular when $m \neq n$, but even when A is square it can still be singular if $\sigma_i = 0$. Singularity of A implies that A does not have full rank, i.e. A is *rank deficient*, or A has a non-trivial nullspace:

$$\begin{aligned} \text{For } \sigma_i &= 0, && (3.26)\\ Av_i &= 0 && (3.27)\\ \therefore A(x+\alpha v_i) &= b, \text{ and furthermore,} && (3.28)\\ \nexists\ y \mid Ay &= \beta u_i && (3.29) \end{aligned}$$

Eqn. (3.28) shows that we fail the uniqueness test, and eqn. (3.29) shows that we fail the existence test for Hadamard's condition for well-posedness. In most physical problems, the singular values may not be identically zero, as it would be impractical to numerically evaluate them to that level of precision. Based on the importance of the singular value spectrum though, we can define a *condition number*:

$$C \equiv \frac{\max \sigma}{\min \sigma} \tag{3.30}$$

As C becomes larger, the problem becomes more ill-conditioned, and for a strictly singular matrix, $C \to \infty$. Even though we now have this quantity defined, the boundary between what is considered well-conditioned and poor-conditioned is not a sharp one. A generally acceptable figure is $C \leq 10^3$.

We now return to our numerical example of the Hilbert matrix. The singular values are $\{1.567, 0.2085, 0.0114, 0.0003, 0.000003\}$. The condition number is 4.766×10^5, so as suspected, the problem is ill-conditioned. Specifically, let us examine eqn. (3.25), especially the factor σ_i^{-1}. In the forward problem, the small singular values damp out the contributions from the additive noise. In the inverse problem, however, they become an amplification for the noise components, drowning out the original signal x_{in}. This amplification picture is consistent with our result above, as we found $|A^{-1}x'_{out} = x'_{in}| >> |x_{in}|$. If a problem is ill-conditioned, any standard matrix inversion algorithm will fail to recover the desired solution x_{in}. Having understood the origins of the difficulties, we can now discuss strategies for overcoming these difficulties.

3.3 Regularization and the L-curve

The specialized technique that is used to solve inverse problems is called *regularization*. There are many regularization schemes that have been developed, and Engl's text is a good starting point. Here, we will only discuss the most common regularization scheme, known as Tikhonov regularization, that works well in many situations.

First, we must address the three properties of ill-posedness. In a way, they are related, because any time we have a singular (or near-singular as defined by C)

matrix, any of the three can occur. The lack of an existence theorem is overcome by minimizing the residual $|Ax'_{in} - x'_{out}|$ in usual inverse problem applications, and is not considered too serious. One must give up on the notion of an exact solution to inverse problems. Rather, we just try to reconstruct a 'sensible' solution that satisfies our given equation 'well enough.' We will comment further on this issue in the final section of this chapter.

Non-uniqueness is considered much more serious. In our numerical example, having given up the notion of an exact solution, we know that $|Ax_{in} - x'_{out}|$ would have been nonzero but small. In fact, it is exactly the norm of the small noise term added to x_{out}. The problem becomes how to pick out the nice solution among all the many that would still give reasonably small residual norms. We observed at the end of the last section that the small singular values lead to large noise amplification. We note also that these bad solutions do tend to blow up and have large norms, much larger than the desired solution. Therefore, one strategy would be to restrict the size of the solution. This additional constraint allows us to choose systematically a unique solution that at least allows a 'sufficiently small' residual. Rather than minimizing only the residual, we can include the solution norm as well to formulate the following regularized inverse problem:

$$x_{out}^{(\lambda)} = \min_{x_{out}} \left\{ |Ax_{out} - x'_{in}|^2 - \lambda |x_{out}|^2 \right\} \tag{3.31}$$

The scalar λ is known as the regularization parameter, and determines the relative weight between minimizing the residual norm versus the solution norm. The standard graphical tool that assists in choosing the regularization parameter is the L-curve. The L-curve plots the solution norm on the y-axis and the solution norm on the x-axis for a series of λ's. The L-curve for our numerical example is shown in figure 3.4, and it takes on its name because of the characteristic L shape. The corner is usually taken as the appropriate regularization parameter that indicates optimal compromise between the two norms. In practice however, the corner rarely gives the optimal solution, so it is best to use it as a rough guide. Constructing the L-curve for our problem, we

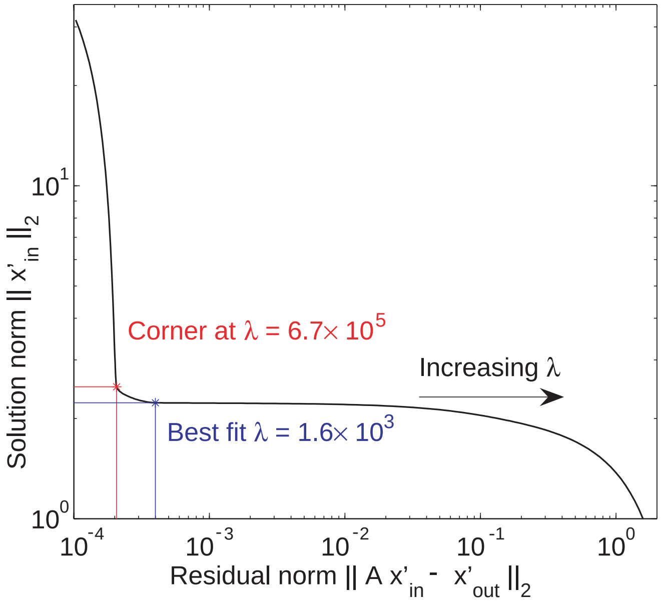


Figure 3.4: L-curve for the Hilbert operator. The curve is a parametric log-log plot of the solution norm vs. the residual norm for different regularization parameter λ in the direction shown. Most references suggest the optimal point is at the corner as shown in red, but physical problems often have the 'best' solution elsewhere.

find a value of $\lambda = 6.7 \times 10^{-5}$ at the corner. The solution then is:

$$x'_{\lambda=6.7\times10^{-5}} = \begin{bmatrix} 0.9298 & 1.5388 & 0.2284 & 0.8474 & 1.4818 \end{bmatrix}^T \tag{3.32}$$

$$x'_{\lambda=1.6\times10^{-3}} = \begin{bmatrix} 0.9800 & 1.0661 & 1.0103 & 0.9773 & 0.9474 \end{bmatrix}^T \tag{3.33}$$

$$\cong x_{in}. \tag{3.34}$$

By looking at some more values near the corner, we find that the solution closer to our 'true' solution actually has $\lambda = 1.6 \times 10^{-3}$. So we see that we can in fact recover sensible results even for badly conditioned problems.

3.3.1 An alternate interpretation

There is an alternative picture to justify the Tikhonov regularization scheme. Recognizing that it is the small singular values that cause the difficulties, we can imagine applying a filter on the singular values when we construct the inverse in eqn. (3.25). Applying a Lorentzian filter to the reciprocal singular values, we get:

$$\frac{1}{\sigma_i} \rightarrow \left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \right) \sigma_i^{-1} \tag{3.35}$$

For $\sigma_i >> \lambda$, the filter has little effect, whereas if $\sigma_i << \lambda$, then $\sigma_i^{-1} \rightarrow \lambda^{-1}$, limiting the unstable spectrum. We see that the two views are equivalent since we can analytically solve the Tikhonov minimization (eqn. (3.31)). For a given λ, the function is minimized if x'_{out} is constructed using filtered coefficients of eqn. (3.35) instead of the reciprocal singular values σ_i^{-1}. Different regularization schemes effectively change how we evaluate the filtering coefficients. For example, if we take our Hilbert operator and increase N to 100, we find the spectrum of singular values as shown in figure 3.5. Because of the distinctive corner in the spectrum, we might use an aggressive strategy here and simply truncate beyond the 20^{th} singular value. This is known as the truncated singular value decomposition (TSVD) regularization scheme. (Note: this scheme alone would not work for the Hilbert problem because the remaining σ's would still give a condition number of 10^{17}.) Most physical inverse problems do not have these obvious clusters, making hard truncation more difficult, so Tikhonov is really a good general strategy to use. In figure 3.6 we show the spectrum for the inverse photonic problem (see chapter 6) using a Guassian output mode. This spectrum is more representative of real world inverse problems.

3.4 Conclusion

In this chapter, we explored the reasons why there is a real fundamental distinction between a forward problem and an inverse problem. In particular, the notion of stability against random perturbations is what sets the two apart, and we gave the condition number as a quantity that helps us identify ill-conditioning. For completeness, we now elaborate on a subtle point that we commented on earlier in passing. We motivated the need to find a fundamental distinction between forward and inverse

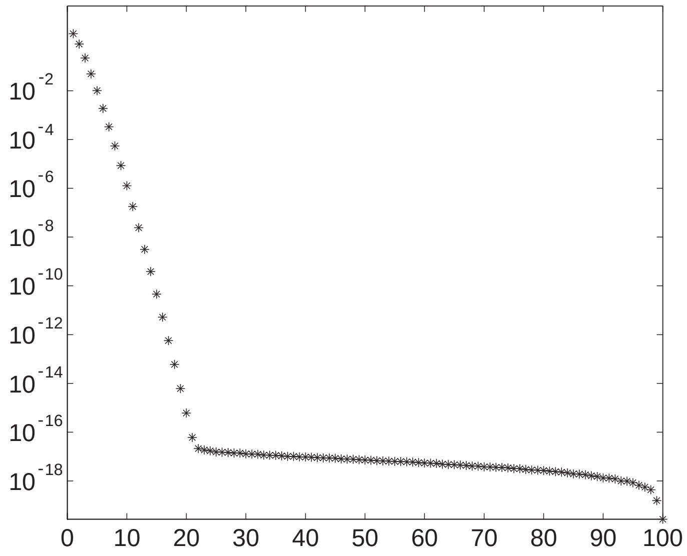


Figure 3.5: The spectrum of singular values for the 100×100 Hilbert operator. Note the distinct corner, showing an obvious point where we can perform a truncation.

problems due to the inherent symmetry of the two problems, i.e. one is the inverse of the other. It should now be clear that neither the Fourier transform nor its inverse can be considered an inverse problem, because they are both unitary, so $C = 1$. For an inverse or ill-posed problem, $C >> 1$.

Of course, the condition number of the forward and inverse problem are the same as we have defined it (since it is just the ratio of the largest to smallest singular values), so the spectrum of singular values does not break the symmetry between the two problems. So what actually breaks the symmetry? It turns out to be the special status given to the random fluctuation or noise. The forward problem is defined as the one that is stable against changes *caused by random fluctuations.* However, given a large condition number, stability against noise necessarily implies it will be ‘unstable’ to a different form of perturbation. Of course, we usually do not use the term ‘instability,’ but instead we use the term ‘sensitivity’ in this context. Historically, this makes sense in how one studies physics. To model a physical system, we vary its parameters and measure its effects. A model is good if it makes good predictions about

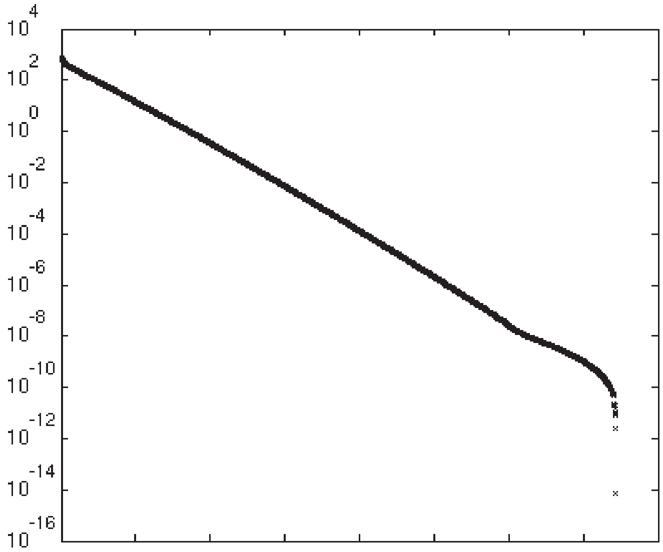


Figure 3.6: The spectrum of singular values for the inverse photonic problem using a Gaussian shaped desired mode. In contrast to figure 3.5, we find no obvious place for a hard truncation.

the effects. Given a new physical system we are trying to model, if any small noise (i.e. a perturbation to the system the experimentalist cannot control) will create a large disturbance in the effect we can measure, it will be very difficult to come up with a model. What we need is a system that is sensitive to controllable and systematic variations to the input, so the effects can be readily observed with adequate signal-to-noise ratio. By its very nature, most problems studied are stable (in the sense given here) against most forms of noise. Any physical model derived based on experimental results will necessarily reflect this process. Therefore, we do expect 'noise' vectors to have large projections onto the 'bad' singular vectors in physical problems, using our linear algebra language.

3.4.1 Parameter estimation vs. design

We conclude this chapter by making an observation about the difficulties in transferring over from standard inverse problems to PBG device design. First, most standard inverse problems can assume implicitly the existence of a solution even if the problem formally does not guarantee you a solution [21]. Going back to the cause and effect picture, you are measuring a real effect from a cause that necessarily exists. The problem as we saw is that noisy measurements hide the underlying cause. If you really cannot reconstruct the cause from the measured effect, it is probably an indication that the model is wrong. As applied to a design paradigm, that is not the case. The 'desired effect' has not been observed or measured. It is a mere figment of our imagination, so to speak. We will see that this is a much more serious problem for design purposes. If we encounter a design problem where we encounter a non-existent solution situation, we would like to make strong claims to that effect, but we cannot do so because the Tikhonov regularization scheme is not really the most appropriate for the PBG problem. This brings us to our second observation. The solution we seek in the PBG inverse problem is the dielectric function, and their norms are not necessarily small, particularly if there are discontinuities. Physically, η cannot take on negative values, but Tikhonov would be happy accommodating negative values as long as they are small. Fortunately, we can use the insight developed in this chapter to implement a much more appropriate regularization scheme for the PBG problem. We expand on these ideas and develop the necessary tools in the next chapter.

Chapter 4

Convex Optimization

At the conclusion of chapter 3, we suggested that a natural question to ask is whether Tikhonov regularization is really the best choice for the purpose of the inverse photonic problem. We learned that regularization is a way to impose additional constraints on an under-determined (rank-deficient) system so that an ill-posed problem becomes well-posed. Therefore, given an actual physical problem, one ought to be able to customize a more appropriate set of constraints based on the relevant physics. Suppose the desired solution has a fairly sizeable norm. Under Tikhonov, is there a possibility that we might 'regularize out' valid solutions because the norm constraint is not sophisticated enough? For an even worse effect, given our application, a solution with negative values would still have a small norm, but in the case of photonic materials, the solution (representing the dielectric function) would not even be feasible. More succinctly, requiring a small norm does not correlate with valid and desirable photonic device designs. A more suitable approach places upper and lower bounds on the value of the dielectric function that correspond to air and semiconductor. We will derive in chapter 6 the inverse photonic problem, but for now, consider the regularized problem given these constraints of the form:

$$\begin{gathered}\min_{\eta} |A\eta - b|^2 \\ \text{subject to } \eta_{min} \preceq \mathcal{F}^{-1}\eta \preceq \eta_{max}\end{gathered} \tag{4.1}$$

where $\mathcal{F}^{-1}$ is the inverse fourier transform operator. Equation (4.1) is a quadratic minimization problem with linear inequality constraints, and falls under the general category of convex optimization problems. Casting this problem in the form of a convex optimization (CO) is powerful because rigorous bounds have been derived on the optimality of their solutions [1] using interior point methods. Therefore, with the only constraint being that the dielectric function take on physically realizable values, we can prove rigorously whether any dielectric function exists that would support a given target mode, as indicated by the magnitude of the residual norm. This chapter is devoted to developing our implementation of the convex optimization algorithm.

4.1 Introduction

We have all encountered optimization problems, and our first exposure to them is likely to have been in a calculus course. Given some function $f(x)$ defined over some interval $x \in [a, b]$, find where the maximum (or minimum) value of f occurs along that interval. In the language we will use below, we call x the 'optimization variable,' and f the 'objective function.' Restricting x on the interval $[a, b]$ can be viewed as an 'inequality constraint' on the optimization variable. We can locate *local extrema* since they satisfy the following condition:

$$\left.\frac{df}{dx}\right|_{x_e} = 0 \tag{4.2}$$

The sign of the second derivative evaluated at those points $\{x_e\}$ classifies the kind of extremum (maximum, minimum, or inflection point) at those points. To find the *global* optimum, we evaluate f at all the local extrema, plus the end points, and then choose from among these the optimal value (x^*), and we call x^* the solution to the optimization problem[1]. For arbitrary functions, the problem becomes more difficult as eqn. (4.2) may not have analytical solutions. Numerical optimization in 1D is

[1] Here we follow Boyd's notation, and x^* does not denote the complex conjugate of x. Boyd only deals with real-valued variables and functions, so the notation is fine. Later in this chapter when we deal with complex variables, we will use $\overline{\eta}$ to denote the complex conjugate of η.

relatively straightforward, as we can just plot the function and visually determine the extremum points. However, the problem scales unfavorably as the dimensionality of x increases.

$$\begin{array}{rlcll} \text{minimize} & f_o(x) & & & \\ \text{subject to} & f_i(x) & \leq & 0, & i = 1, ..., m \\ & g_j(x) & = & 0, & j = 1, ..., p \end{array} \quad , \tag{4.3}$$

where $f_0 : \mathcal{R}^n \to \mathcal{R}$ is the objective function, $x \in \mathcal{R}^n$ is now an n-dimensional optimization variable, and $\{f_i : \mathcal{R}^n \to \mathcal{R}\}$ and $\{g_i : \mathcal{R}^n \to \mathcal{R}\}$ define the m inequality constraints and p equality constraints respectively that x must satisfy in order to be a valid solution. We refer to any x that does not satisfy all the constraints as an 'infeasible point.' The problem of maximizing an objective function is achieved by simply reversing its sign.

An optimization problem is called a 'convex optimization' problem if it satisfies the extra requirement that f_0 and $\{f_i\}$ are convex functions (which we will define in the next section), and $\{g_i\}$ are affine functions. Furthermore, the set of feasible points must also form a convex set. The special property for this class of problem is that any local minimum is by definition also the global minimum, and thus solves the optimization problem.

4.1.1 Organization

As with the simple one-dimensional variable optimization, we will find the first and second derivatives play critical roles in these optimization algorithms, so we begin with the definition of multidimensional derivatives in section 4.2. For our application, we will need to extend the treatment to include the use of complex variables, which have some minor complications we will address. We will formally define convex functions in section 4.3 and discuss some of their properties, highlighting those that help with understanding the optimization problem. With the mathematical tools sufficiently developed, we can then explain how to implement the convex optimization method in section 4.4. We explain how to select descent directions and the line search routine for an unconstrained optimization first, and then show how constraints can be incorporated. Again, our primary concern in the presentation here is not to be mathematically rigorous, but rather make it accessible for engineers and physicists looking for a softer entry point. With the exception of the modification required for functions with complex variables, the material in this chapter is adapted from the textbook by Boyd and Vandenberghe [1], where they provide a much more thorough and rigorous treatment of convex optimization methods.

4.2 Derivatives

For a real-valued function $f : \mathcal{R}^n \rightarrow \mathcal{R}$, the definition of the gradient is

$$\nabla f(x)_i \equiv \frac{\partial f(x)}{\partial x_i}, \;\; i = 1, \ldots, n, \tag{4.4}$$

provided that the partial derivatives evaluated at x exist, and where $\nabla f(x)$ is written as a column vector. The first-order Taylor approximation of the function f near x is

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) \tag{4.5}$$

which is an affine function of y.

The second derivative of a real-valued function $f : \mathcal{R}^n \rightarrow \mathcal{R}$ is called a *Hessian matrix*, denoted $\nabla^2 f(x)$, with the matrix elements given by:

$$\nabla^2 f(x)_{ij} \equiv \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \;\; i = 1, \ldots, n, \;\; i = 1, \ldots, n, \tag{4.6}$$

provided that f is twice differentiable at x and the partial derivatives are evaluated at x. The second-order Taylor approximation of the function f near x is

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) \tag{4.7}$$

We now derive some shorthand notation for taking derivatives of matrix equations representing functions $f : \mathcal{R}^n \rightarrow \mathcal{R}$.

For functions that are linear in x:

$$\begin{aligned} f(x) &= a^T x = x^T a && (4.8) \\ &= \sum_i a_i x_i && (4.9) \\ \nabla f(x)_i &\equiv \frac{\partial f}{\partial x_i} && (4.10) \\ &= a_i && (4.11) \\ \therefore \nabla f(x) &= a. && (4.12) \end{aligned}$$

The Hessian is obviously 0 in this case.

For functions that are quadratic in x:

$$\begin{aligned} f(x) &= x^T B x = \sum_{i,j} x_i b_{ij} x_j && (4.13) \\ \nabla f(x)_i &\equiv \frac{\partial f}{\partial x_i} && (4.14) \\ &= x^T \frac{\partial B x}{\partial x_i} + \frac{\partial x^T}{\partial x_i} B x && (4.15) \\ &= x^T [B]_i^{col} + e_i^T \sum_k b_{ik} x_k && (4.16) \\ &= \sum_k (b_{ki} + b_{ik}) x_k && (4.17) \\ \nabla f(x) &= \left(B + B^T\right) x, && (4.18) \end{aligned}$$

where $[B]_i^{col}$ denotes the i^{th} column of the matrix B and e_i is i^{th} basis-vector. For the

Hessian, we obtain

$$
\begin{aligned}
\nabla f(x)_{ij} &\equiv \frac{\partial^2 f}{\partial x_i \partial x_j} && (4.19)\\
&= \frac{\partial}{\partial x_i} \nabla f(x)_j && (4.20)\\
&= \frac{\partial}{\partial x_i} \left[(B + B^T)x\right]_j && (4.21)\\
&= \frac{\partial}{\partial x_i} \sum_i (b_{ji} + b_{ij})x && (4.22)\\
&= b_{ji} + b_{ij} && (4.23)\\
\nabla^2 f(x) &= B + B^T. && (4.24)
\end{aligned}
$$

For composite functions of the form $h(x) = g(f(x))$ such that $h : \mathcal{R}^n \to \mathcal{R}$, with $f : \mathcal{R}^n \to \mathcal{R}$, and $g : \mathcal{R} \to \mathcal{R}$, the chain rule for gradients is:

$$\nabla h(x) = g'(f(x))\nabla f(x). \tag{4.25}$$

The chain rule for the Hessian is evaluated to:

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T. \tag{4.26}$$

The chain rules will be very useful for incorporating the barrier functions to impose inequality constraints.

4.2.1 Complex variables

For the photonic problem, the optimization variable η will in general be complex. Therefore, our matrices and vectors will be complex as well. Our objective function $f(\eta) : \mathcal{C}^n \to \mathcal{R}$ is actually the 2-norm of a complex residual $|A\eta - b|^2$. We have found few resources that deal explicitly with complex derivatives. Petersen and Pedersen's reference [25] shows that we can treat η and $\overline{\eta}$ as independent variables, and then the generalized complex gradient is found by taking the derivatives with respect to

$\overline{\eta}$. This expression for the gradient is suitable for use with gradient descent methods.

$$\nabla f(\eta) \equiv 2\frac{\partial f(\eta,\overline{\eta})}{\partial \overline{\eta}} \tag{4.27}$$

For linear functions, we have

$$\begin{aligned} f(\eta) &= 2|a^\dagger \eta| = a^\dagger \eta + \eta^\dagger a && (4.28)\\ &= a^\dagger \eta + \overline{\eta}^T a && (4.29)\\ \nabla f &\equiv 2\frac{\partial \overline{\eta}^T a}{\partial \overline{\eta}} && (4.30)\\ &= 2a. && (4.31) \end{aligned}$$

For quadratic functions, we have

$$\begin{aligned} f(\eta) &= \eta^\dagger A^\dagger A \eta && (4.32)\\ &= \overline{\eta}^T A^\dagger A \eta && (4.33)\\ \nabla f &= 2A^\dagger A \eta. && (4.34) \end{aligned}$$

The Hessian for the quadratic function is

$$\nabla^2 f(\eta) = 2A^\dagger A. \tag{4.35}$$

Unfortunately, it turns out that we cannot define a chain rule for differentiating complex variables. Part of the reason is that a real-valued function of a complex variable is strictly not differentiable because it does not satisfy the Cauchy-Riemann equations, which state for $f(\eta) = f(x+iy) = u(x,y) + iv(x,y)$, where u, v are real valued functions of the real variables $x = \Re(\eta), y = \Im(\eta)$

$$\begin{array}{c} \dfrac{\partial u(x,y)}{\partial x} = \dfrac{\partial v(x,y)}{\partial y} \\ \dfrac{\partial u(x,y)}{\partial y} = -\dfrac{\partial v(x,y)}{\partial x} \end{array}. \tag{4.36}$$

A real valued function implies $v = 0$, so in order to satisfy Cauchy-Riemann u cannot depend on x or y. The linear and quadratic functions happen to be special cases for which these can be defined, but in general, we cannot evaluate complex gradients and Hessians directly. However, we can treat $\Re(\eta)$ and $\Im(\eta)$ as independent real variables so now we have a function $f(x, y) : \mathcal{R}^{2n} \to \mathcal{R}$. All of the previous and subsequent results can now be applied, provided we define our functions in this form. We return to our linear and quadratic matrix functions and see what the equivalent structure looks like.

We define a $\mathcal{C}^n \to \mathcal{R}^{2n}$ transformation for a complex column vector:

$$[\eta] \to \begin{bmatrix} x \\ y \end{bmatrix} \equiv [\xi]. \tag{4.37}$$

The adjoint operation of the vector $\eta \to \eta^\dagger$ becomes $\xi \to \xi^T$. This definition preserves the norm of the vector.

$$\begin{aligned} \eta^\dagger \eta &= \xi^T \xi && (4.38) \\ (x^T - iy^T)(x + iy) &= [x^T y^T] \begin{bmatrix} x \\ y \end{bmatrix} && (4.39) \\ x^T x + y^T y &= x^T x + y^T y && (4.40) \\ \because x^T y &= y^T x && (4.41) \end{aligned}$$

For the function to be real-valued, vector-vector products must come in adjoint pairs, i.e.,

$$\begin{aligned} \eta_1^\dagger \eta_2 + \eta_2^\dagger \eta_1 &\to [x_1^T y_1^T] \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + [x_2^T y_2^T] \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} && (4.42) \\ i(\eta_1^\dagger \eta_2 - \eta_2^\dagger \eta_1) &\to [y_1^T - x_1^T] \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} - [y_2^T - x_2^T] \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} && (4.43) \end{aligned}$$

where we have used $i\eta^\dagger = i(x^T - iy^T) = y^T - i(-x^T)$ in the second relation. We must of course be cautious that if the function is not real-valued, this transformation breaks down (e.g., a general dot product of 2 complex valued vectors will give a complex number), so it is not accommodated here.

For matrix multiplications, we define the transformation $A \in \mathcal{C}^{n\times n} \rightarrow \mathcal{A} \in \mathcal{R}^{2n\times 2n}$ as follows:

$$[A] \rightarrow \begin{bmatrix} A_r & -A_i \\ A_i & A_r \end{bmatrix}, \tag{4.44}$$

where $A_r = \Re(A)$, and $A_i = \Im(A)$. As before, the Hermitian adjoint becomes the simple transpose operation. Matrix-vector multiplications are preserved:

$$\begin{aligned} A\eta &= \mathcal{A}\xi & (4.45) \\ (A_r + iA_i)(x + iy) &= \begin{bmatrix} A_r & -A_i \\ A_i & A_r \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} & (4.46) \\ (A_r x - A_i y) + i\,(A_i x + A_r y) &= \begin{bmatrix} A_r x - A_i y \\ A_i x + A_r y \end{bmatrix}. & (4.47) \end{aligned}$$

In appendix B, we will show that using the chain rule for the log barrier function with the generalized complex gradient definition is incompatible with our definition here for a simple linear constraint function. For the remainder of the thesis, we will not explicitly write out the conformal mapping from $\mathcal{C}^n$ to $\mathcal{R}^{2n}$.

4.3 Convex Sets and Functions

A set $\mathcal{S}_C$ is convex if the line segment between any two members of the set x_1 and x_2 also lies in $\mathcal{S}_C$. The geometric representation is shown in figure 4.2. Important examples of convex sets are hyperplanes which have the form $\{x|a^T x = b\}$ and half-spaces $\{x|a^T x < b\}$, where $a \in \mathcal{R}^n$, $a \neq 0$, and $b \in \mathcal{R}$. Other common convex sets include spheres, cones, and polyhedra. The set defined by the intersection of two convex sets is also convex, so intersection preserves convexity. This is important in our definition of a convex optimization problem, since as long as our constraint

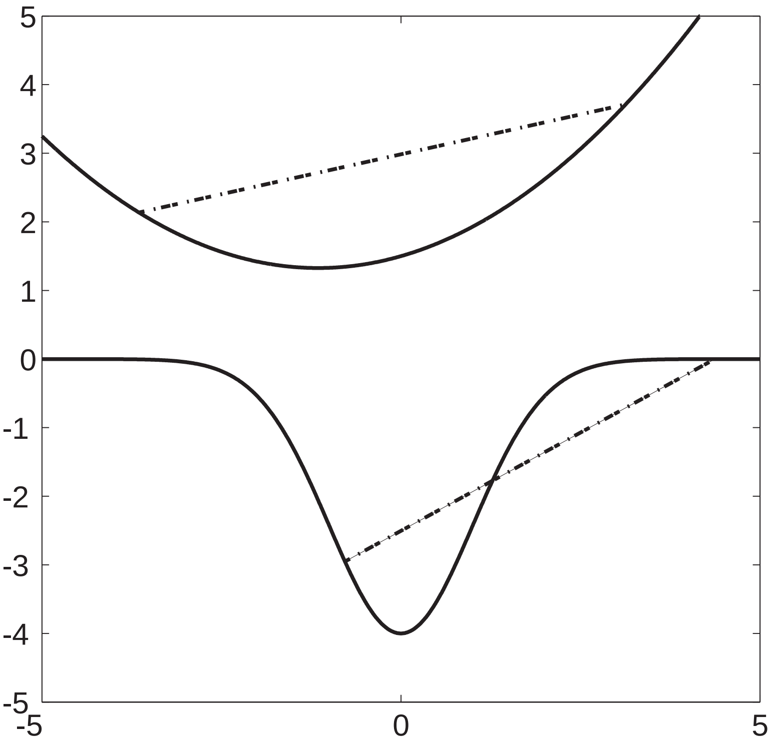


Figure 4.1: The chord connecting any two points of a convex function must lie above the function if the function is convex. The curve on the top is clearly convex. The bottom curve is an upside down Gaussian. Even though it has a single local minimum that is also the global minimum, the function is not convex as shown. The function lies both above and below a connecting chord.

functions are convex, we are guaranteed a convex feasible set.

A function $f(x) : \mathbf{R}^n \rightarrow \mathbf{R}$ is convex if the domain of f is a convex set, and the function satisfies the following relation:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \tag{4.48}$$

for all $x, y \in \mathcal{R}^n$, and with the scalar $\alpha \geq 0$. For a given x, y pair, a parametric plot of α on the right hand side of the inequality corresponds to the chord connecting $f(x)$ to $f(y)$. We can provide a graphical interpretation of the convexity condition as a function where for any pair of points the function lies below the chord joining the pair of points, as shown in figure 4.1. A function is *strictly convex* if strict inequality holds for all $x \neq y$ in eqn. (4.48). By definition, a function f is *concave* if

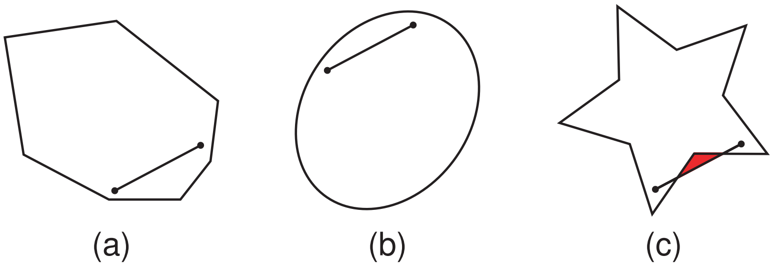


Figure 4.2: Examples of convex and non-convex sets. The set of points in a polygon (a) and in an ellipse (b) are both convex. The star shape is not convex since the line connecting two points in the set passes through a region that is not in the set (highlighted in red).

-f is convex. For the rest of this chapter, we will only consider functions for which the domain of f spans all of $\mathcal{R}^n$, so the domain of f is always a convex set. Some important examples of convex functions include

- Exponential e^{ax} is convex on $\mathcal{R}$ for any $a \in \mathcal{R}$
- Logarithm $log(x)$ is convex on $0 < x \in \mathcal{R}$
- Norms on $\mathcal{R}^n$
- Linear, Affine, and Quadratic functions on $\mathcal{R}^n$
- Non-negative weighted sums of convex functions $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ for f_1, f_2 convex functions and $\alpha_1, \alpha_2 \geq 0$.

4.3.1 Convexity conditions

Suppose f is differentiable such that its gradient exists. The function f is convex *if and only if*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{4.49}$$

We recognize the right side of eqn. (4.49) as the multidimensional version of the linear Taylor series expansion of the function f about x. From this property of convex functions, we observe two consequences. First, linearization of the function underestimates it everywhere (see figure 4.3). This means the first-order Taylor approximation

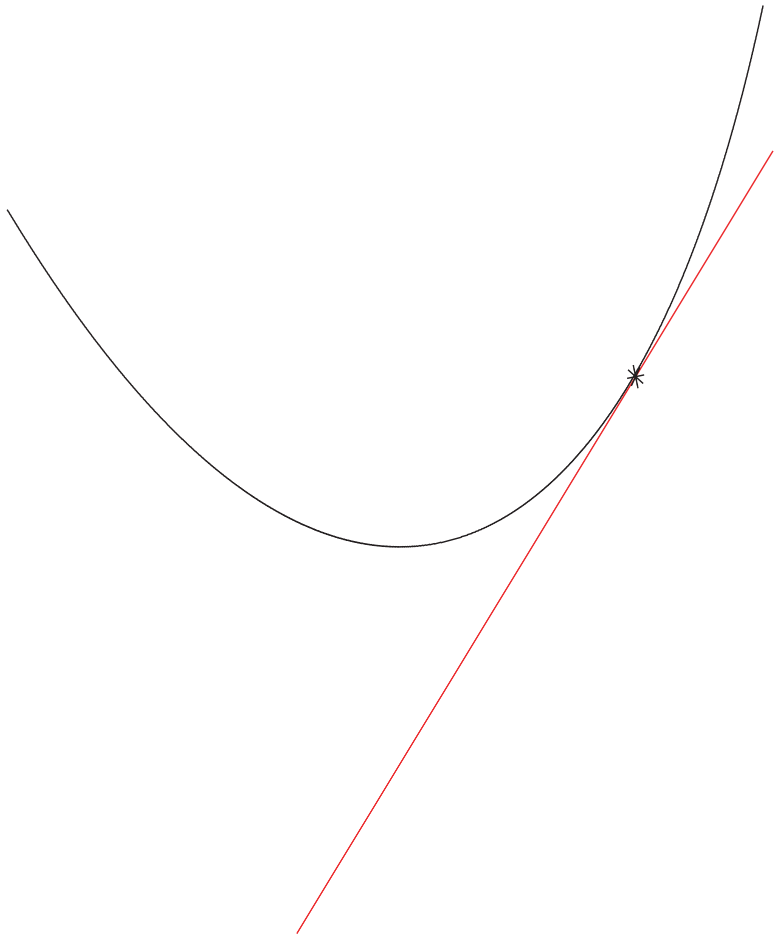


Figure 4.3: Graphical representation of the linearization of a convex function.

is a *global underestimator* of f. In addition, when $\nabla f(x) = 0$, $f(y) \geq f(x)$ for all y, so x is the local and the global minimum of the function. The second-order condition equivalent to $f''(x) \geq 0$ for the 1D case is that $\nabla^2 f(x) \succeq 0$, i.e. the Hessian is positive semi-definite.

4.4 Gradient and Newton Methods

We consider optimization problems for which we do not have an analytical solution, and therefore must use a numerical (and iterative) algorithm to solve the problem. We want an algorithm whose performance is independent of the starting condition, and rapidly converges to the optimal solution. Starting with some initial non-optimal point $x^{(0)}$, each iterate (k) of the algorithm gives us an $x^{(k)}$ such that $f_0(x^{(k)}) \rightarrow f_0(x^*)$ as $k \rightarrow \infty$, where x^* is the 'true' solution that optimizes our objective function. In practice, the iterations terminate once some specified tolerance level is reached, i.e. $f(x^{(k)}) - f(x^*) < \epsilon$. In general, this would be difficult to estimate, but because of convexity it allows us to evaluate bounds on how far from optimal the final solution is. At each iteration, the intermediate solution is updated via the following general relation:

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \tag{4.50}$$

where $\Delta x^{(k)}$ is an unnormalized vector in $\mathcal{R}^n$ known as a 'step direction' and $t^{(k)} > 0$ is a scalar called the 'step size' for the k^{th} iteration. Different algorithms will have different methods for determining the step directions and step sizes. We consider only *descent methods* here for which $f(x^{(k+1)}) \leq f(x^{(k)})$, with equality only if $f(x^{(k)})$ is optimized. The general algorithm can be described as follows:

- **Initialize**: Obtain a feasible starting point x
- **Repeat**
  1. Determine the step direction Δx.
  2. *Line search.* Choose a step size $t > 0$.
  3. Update $x \rightarrow x + t\Delta x$.
  4. Evaluate stopping criterion at the new x.
- **until** stopping criterion is satisfied.

4.4.1 Unconstrained optimization

To illustrate these ideas, we begin by considering an optimization without constraints. The popular gradient descent method uses the negative of the gradient $(-\nabla f)$ evaluated at the intermediate point $x^{(k)}$ as the step direction. The stopping criterion is usually of the form $|\nabla f|_2 \leq \xi$, where ξ is small and positive. Using the negative gradient guarantees that our step direction is a descent direction, and for simple problems it is easy to implement in practice. Unfortunately, for ill-conditioned problems, this method does not converge in practice. However, since the gradient is easy to visualize, we include it here to help illustrate the second step of the algorithm, the line search.

Backtracking Line Search

The line search is used to choose how far to step in the descent direction (once that is determined). In principle, one can do an exact line search of the following form:

$$\min_{t>0} f(x^{(k)} + t\Delta x) \tag{4.51}$$

which is a 1D minimization problem. However, in practice, inexact methods are used because they are easier to implement without suffering a loss in performance. Inexact methods aim to simply reduce the function by some sufficient amount, and the *backtracking* line search is the one we will use. The algorithm depends on two parameters α, β, with $0 < \alpha < 0.5$ and $0 < \beta < 1$. It is called backtracking because it first assumes a full step size of $t = 1$, and if the step does not lead to some sufficient decrease in the objective function, then the step size is decreased by a factor β so that $t \rightarrow \beta t$ (see figure 4.4). The sufficient decrease condition can be written mathematically as:

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x. \tag{4.52}$$

For small enough t, this condition must be true because we only consider descent directions. The parameter α sets the amount of decrease in the objective function we will accept as a percentage of the linear prediction (which, due to convexity, provides a lower bound). Practical backtracking algorithms tend to have α between 0.01 and 0.3, and β between 0.1 and 0.8, with a small value of β corresponding to a coarser grained search of the minimum.

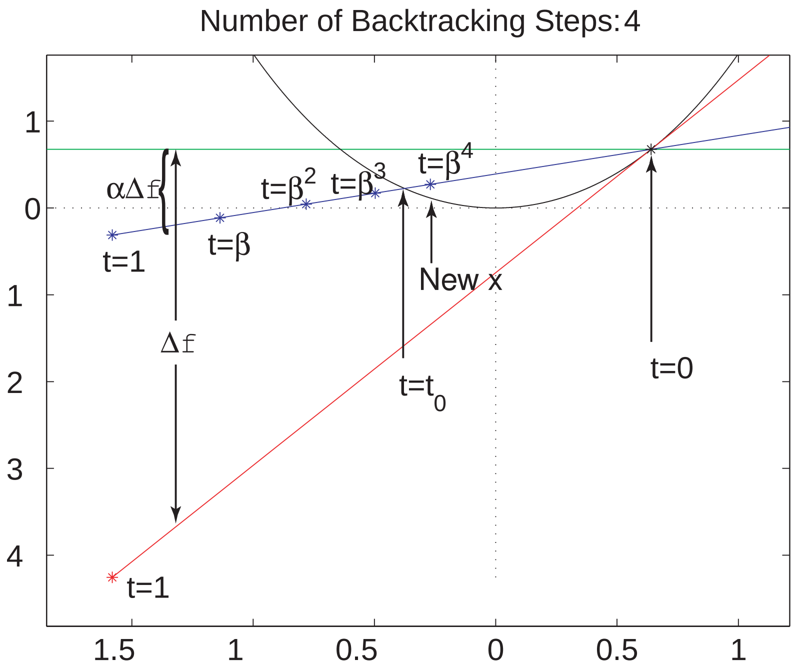


Figure 4.4: The linear approximation (red) to the objective function (black) expanded around $x = 0.64$. The blue line shows the backtracking condition as accepting a fraction α of the predicted decrease by linear extrapolation. After four backtracking steps, $t = \beta^4$, and the value of the objective function has decreased enough.

As a simple example we choose the following 1D objective function:

$$f(x) = e^{\gamma x} + e^{-\gamma x} - 2, \tag{4.53}$$

where γ is a parameter we will adjust to show the various behaviors of the gradient descent method. The optimal value of x^* is 0, and $f(x^*) \equiv p^* = 0$. To illustrate the idea of backtracking, we first show in figure 4.4 the objective function for $\gamma = 1.25$, and choose as an initial point $x_0 = 0.64$. The function is linearized at x_0 in red, and the blue line shows the backtracking condition for $\alpha = 0.2$. A full step in the step direction ($t = 1$) takes us to $x = -1.5803$, shown as a red star in the figure. As illustrated, the region where the backtracking condition (eqn. (4.52)) is satisfied is for $t \in [0, t_0]$. Increasing α will decrease the size of the valid t region. The task for the line search algorithm is to find a valid t. It backtracks from a starting value of

γ	Number of Iterations	Mean number of backtracking steps
0.125	241	1
1.25	12	4
12.5	22	26

Table 4.1: Summary of gradient method performance using an objective function (eqn. 4.53) with 3 different sharpness parameter γ.

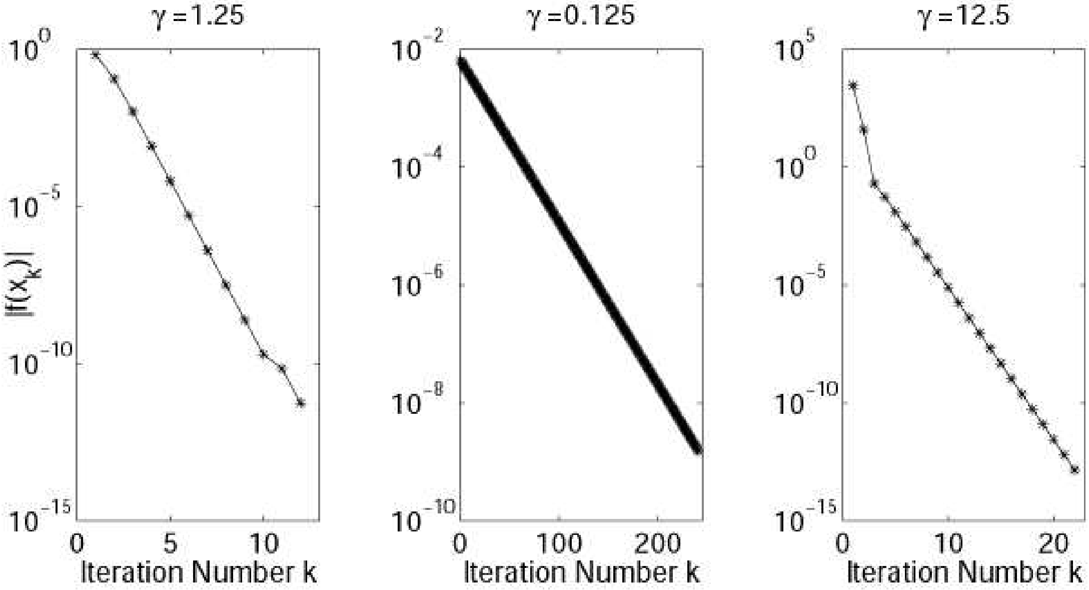


Figure 4.5: Convergence rate of different γ for gradient method.

1 until it enters the valid region. In this example, it backtracks 4 steps before the function has 'decreased enough,' and the value of x for the next iterate is -0.2694. If we go ahead and continue with the optimization, we find the solution converges to $\widehat{x^*} = -1.87\times10^{-6}$ in 12 iterations, and each iteration takes on average 4 backtracking steps during the line search. If we now attenuate γ to 0.125 and repeat, we find the solution converges to $\widehat{x^*} = 3.14\times10^{-4}$ after 241 iterations, without having to backtrack at any iteration. For $\gamma = 12.5$, it takes 22 iterations to find $\widehat{x^*} = 3.01\times10^{-8}$, and each iteration takes on average 26 backtracking steps, with a maximum of 52 at the first iteration. These results are summarized in table 4.1, and the convergence rates are depicted graphically for the three cases in figure 4.5. The problem with the gradient method is that $|\Delta x| = |\nabla f|$. When γ is small, even a full step is too small to cause substantial reduction in the objective function, while for γ too big, it leads to such

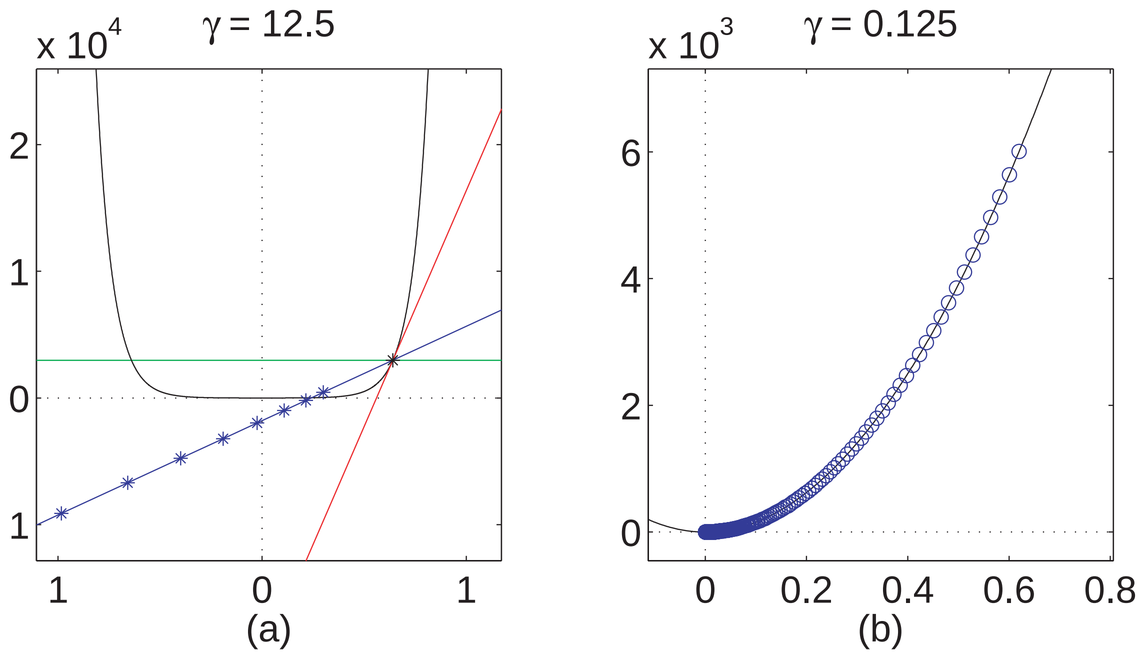


Figure 4.6: (a) For large γ, the norm of the gradient is too large, so a $t = 1$ step size would actually take us to $x = -3.7 \times 10^4$ (well beyond the axis on the plot). This leads to a large number of backtracking steps in the line search. (b) On the other hand, a small γ would give gradients with small norms, so small that a full $t = 1$ step will still give only minimum improvement, necessitating many iterations before convergence.

an enormous step that the backtracking must go through many iterations to return to the valid t region (see figure 4.6). For ill-conditioned multi-dimensional problems, we effectively have an enormous range of γ in different directions, so practically for ill-conditioned problems the gradient method never converges.

Newton's Method

A much better method that uses the second derivative information as well is Newton's method, provided of course that the objective function is twice differentiable. Newton's method uses the *Newton step* as the step direction:

$$\Delta x_{newton} = -\nabla^2 f(x)^{-1} \nabla f(x), \tag{4.54}$$

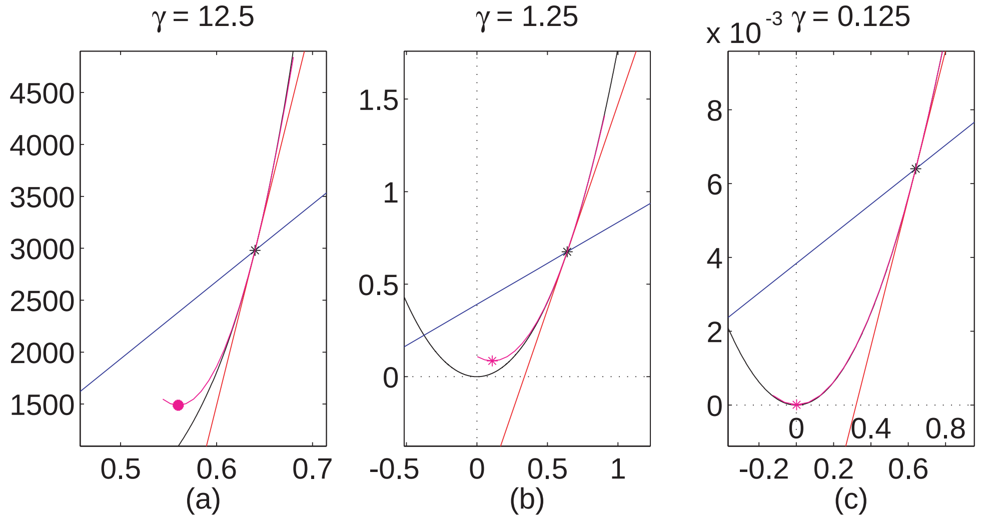


Figure 4.7: (a) For $\gamma = 12.5$, the norm of the Newton step (marked by the magenta asterisk) is 0.08, compared to the $|\nabla f| = O(10^4)$. (b) For $\gamma = 1.25$, the quadratic approximation becomes quite good, and the full Newton step does satisfy the backtracking exit criterion. (c) $\gamma = 0.125$, where the fit is even better, illustrating the quadratic convergence phase.

where $\nabla^2 f(x)$ is the *Hessian matrix* as defined previously in eqn. (4.6). The superscript $^{-1}$ denotes matrix inversion. The Newton step can be interpreted as the step that minimizes the second order Taylor expansion of the objective function about the point $x^{(k)}$ (see figure 4.7). For an unconstrained quadratic objective function then, the Newton step exactly minimizes the objective function. The stopping criterion using Newton's method is the quadratic norm of the Newton step as defined by the Hessian (also known as the *Newton decrement*),

$$\lambda(x) = \left(\Delta x_{newton}^T \nabla^2 f(x) \Delta x_{newton}\right)^{1/2} \tag{4.55}$$

Using Newton's method, we repeat the same optimization of our objective function with the 3 different values of γ. The results are shown in figure 4.8 and table 4.2. Newton's method benefits from the more rapid quadratic convergence, for as we get closer and closer to the minimum, the accuracy of the second-order approximation improves. For this particular objective function, we see that we never have

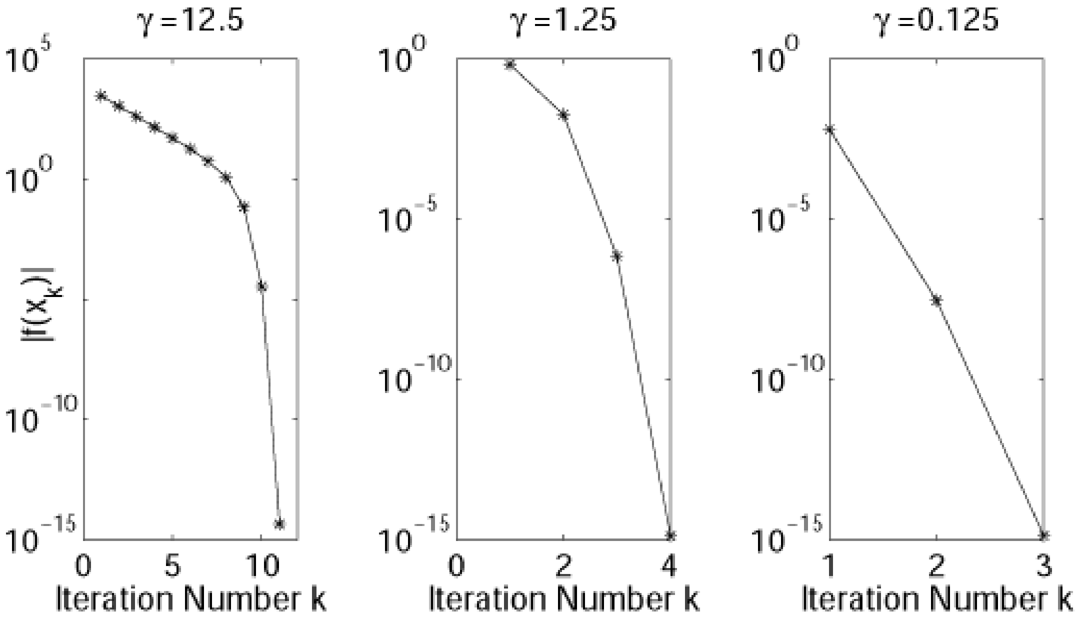


Figure 4.8: (a) For $\gamma = 12.5$, we see the distinct corner at 9^{th} iteration, showing the beginning of the quadratic convergence phase. (b) For $\gamma = 1.25$, the final value of $f(x)$ is actually 0 to within machine precision, but shown as 10^{-15} for reference. (c) The same holds true for $\gamma = 0.125$.

γ	Number of Iterations	Mean number of backtracking steps
0.125	3	0
1.25	4	0
12.5	11	0

Table 4.2: Summary of Newton method performance using an objective function (eqn. 4.53) with 3 different sharpness parameters γ.

to backtrack, which is an indication that the Newton step includes some information about the magnitude of the step size, as opposed to choosing a step direction based on the gradient alone. Compared to the gradient method, we see a significant increase in computational overhead (matrix formation and inversion), but given the ill-conditioning of our problem, gradient methods simply do not converge. In the case of the photonic design problem, the Hessian itself will sometimes be ill-conditioned as well, but we can use a truncated SVD pseudoinverse to calculate the Newton step.

4.4.2 Incorporating constraints: barrier method

Consider the following constrained optimization problem in 1D:

$$\text{minimize} \quad f(x) = e^{\gamma x} + e^{-\gamma x} - 2 \tag{4.56}$$

$$\text{subject to} \quad x_0 - x < 0 \quad , x_0 = 0.75. \tag{4.57}$$

We have constructed the problem so that the solution to the constrained problem lies at the boundary at $x = 0.75$. In order to incorporate inequality constraints, the objective function is modified to include *barrier functions* that impose a prohibitively costly penalty for violating the constraints. The barrier function that we will use is the logarithmic barrier function, and we make use of the fact that the log diverges near 0, meaning:

$$\lim_{x \to 0^+} \log x \to -\infty. \tag{4.58}$$

Recall the set of inequality constraints from our optimization problem (eqn. (4.3)) require $f_i(x) \leq 0$ for all i. The logarithmic barrier function is defined to be

$$\phi(x) \equiv -\sum_{i=1}^{m} \log(-f_i(x)) \tag{4.59}$$

Using the chain rule (eqn. (4.25) and (4.26)), we can write down the expression for the gradient and Hessian for the log barrier function.

$$\nabla\phi(x) = \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla f_i(x), \tag{4.60}$$

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x) \tag{4.61}$$

If we modify our objective function such that we instead minimize

$$f_0(x) + \left(\frac{1}{\delta}\right) \sum_{i=1}^{m} -\log(-f_i(x)) \tag{4.62}$$

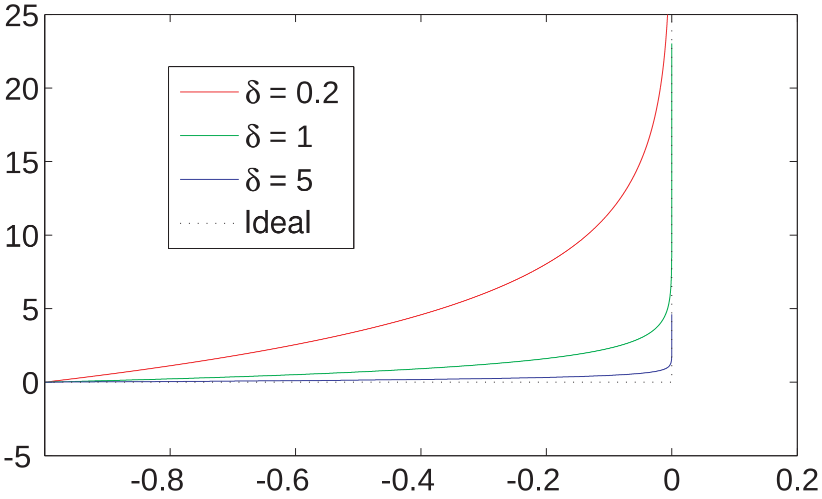


Figure 4.9: The log barrier function for various δ's. The black dotted line is the ideal step barrier. Notice the largest of these ($\delta = 5$) gives the closest approximation to the ideal function.

we see that a violation of the inequality constraints will not minimize the objective function. Therefore, the minimum of this new problem will automatically satisfy the inequality constraints. The parameter δ controls how steep the barrier is. We plot the barrier function for various δ's in figure 4.9. As $\delta \to \infty$, the barrier function has no effect on the objective function for x in the feasible set, so this modified problem becomes exactly the original problem. For finite δ, the modified problem is only an approximation, so the optimal point of eqn. (4.62) is not the optimal point of eqn. (4.3). This is illustrated in figure 4.10. The difficulty with a large δ is that the overall function becomes difficult to minimize even using Newton's method. Boyd attributes this to the rapidly varying Hessian for the logarithmic barrier function near the constraint boundary. Therefore, unless you are close to the boundary (where the solution likely lies), a second-order Taylor expansion is a poor fit to the modified problem, leading to poor convergence. A smaller δ will increase the region where the expansion is valid, but yields a solution that is less accurate. Figure 4.11 shows the quadratic fit to our modified objective function with a moderate $\delta = 5000$. Far from the boundary for large δ, the barrier contribution is insignificant. Therefore, the Newton step ignores the barrier and acts as though there were no constraints.

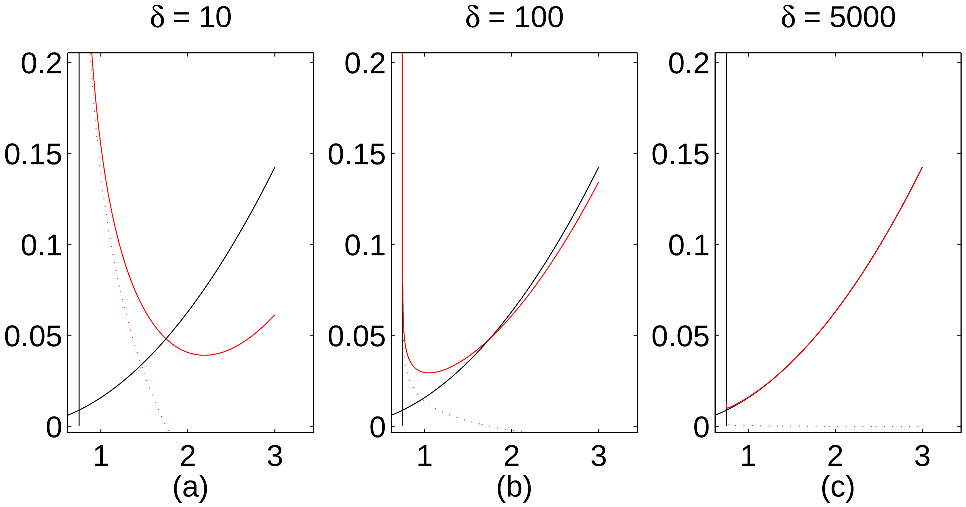


Figure 4.10: The modified objective function (red) for various δ's. The constraint is that $x \geq 0.75$, and the optimal point is at the boundary (shown as a solid vertical black line near the left edge of the box). The original objective function (black) has $\gamma = 0.125$. The barrier function is the dotted red line. (a) $\delta = 10$ The modified problem is a poor approximation of the original. The optimal x^* is around 2. (b) $\delta = 100$ Slight improvement of the approximation, with $x^* \approx 1$. (c) $\delta = 5000$ gives a much better approximation.

However, this would take us out of the feasible set (main figure in figure 4.11). We have the same problem here as we did with the gradient method then, as we potentially may backtrack many iterations to return to the feasible set. Closer to the boundary, the quadratic fit becomes quite good, as the barrier has an appreciable effect on the modified function. The inset of figure 4.11 shows in blue the second-order fit and Newton steps in that region. Of course, for large δ that means we are already very close to the boundary, i.e. the solution of the optimization problem. *A priori*, we would have no way of knowing where that fast converging region is.

The problem is overcome by a process called *centering*, where a succession of these modified problems are solved with δ increasing with each centering step ($\delta_{i+1} = \mu\delta_i$). We solve the modified optimization problem using some small initial δ_0 by Newton's method. The solution of a centering step is used as the starting point of the next centering step. This ensures we are close to the region where Newton's method converges rapidly, as long as we don't increase δ too quickly. This is reflected in the parameter μ. If μ is too large, then we will have less centering steps, but each step will require more iterations before it converges. If μ is too small, then we will require many centering steps. Typical implementations take μ between 10 and 20.

For our implementation of the convex optimizer, we use for our line search routine $\alpha = 0.125$, and $\beta = 0.9$. Our stopping criterion is $\epsilon = 10^{-20}$, and $\mu = 20$. Our optimization problems uses 15 centering steps and within each centering step, the solution will usually converge after less than 10 Newton steps.

4.5 Conclusion

In this chapter, we summarized some of the important tools needed for numerical optimization of multidimensional problems. Using a simple 1D example, we visualized how the gradient descent algorithm and Newton's method minimize a given objective function. The *caveat* is that our intuition may or may not extend into N dimensions. We can certainly imagine fitting the objective function with a hyper-paraboloid surface, and minimizing that as the Newton step. In higher dimensions, it just shows that the gradient has more directions to be incorrect about, so in practical problems it is of little use.

We have now outlined all the tools we need to solve our photonic regularization problem. We do not have equality constraints in our example, but those can be satisfied by solving a set of KKT equations, as described in detail in Boyd. A final note is that with these interior point methods, it is important to first find a feasible point (i.e. satisfies all m inequality constraints and n equality constraints) as the first iterate. Our constraints are simple enough that we can always construct one by inspection (take a uniform slab of dielectric with an average index of refraction). In general, there are algorithms loosely based on these techniques that serve as feasible point finders.

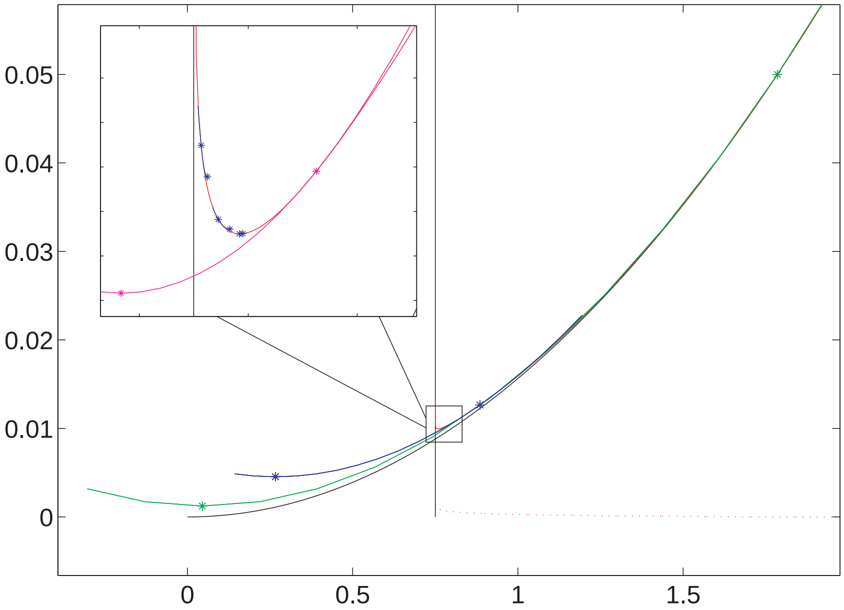


Figure 4.11: Using the $\delta = 5000$ barrier function, we show in the main body the quadratic fit of the objective function. The green and blue lines show the second-order fit about $x = 1.785$ and $x = 0.885$. The green and blue asterisks show the expansion point and the Newton step, outside of the feasible set. The inset shows similar quadratic expansions in blue, illustrating a good fit where near the minimum.

Part II

Device Design

Chapter 5

Photonic Bandgap Devices: An Overview

Having developed the computational tools necessary for the design problem, we now return our focus to photonic bandgap (PBG) materials. Certainly, ever since Yablonovitch's publication [26] in 1987 the amount of research into PBG materials has been extraordinary. Even at its inception, it was anticipated that PBG materials would have a profound effect on semiconductor devices because of their ability to inhibit spontaneous emission rates. One goal of this chapter is to provide a general overview of some of the diverse applications that make use of these PBG materials as a functional device. It is well beyond the scope of this chapter to do a comprehensive review of the topic, so the devices described here will naturally be biased towards our own interests and experiences, but the idea is to provide a sense of where our work fits in within the field.

5.1 Introduction

The term photonic bandgap refers to the range of frequencies for which electromagnetic waves are not allowed to propagate within the material, so incident radiation on the surface will be perfectly reflected. We can locate these bandgaps by solving the Helmholtz equation (eqn. (2.15)) and looking at the allowed eigenvalues as demonstrated in chapter 2. A *complete* bandgap means the frequency is forbidden for all propagating directions. Much of the early work involved looking at these 'bulk' photonic crystal materials and verifying the existence of bandgaps [11, 27, 28, 29, 30] in various lattice geometries. A lot of work was done searching for the geometry that led to the largest bandgaps. Since waves cannot propagate through the material, they act like perfect mirrors at the forbidden frequencies.

The idea of introducing intentional defects to produce localized states [31, 32] came shortly thereafter, and the possibility of trapped states lead naturally to the idea of optical cavities and waveguides. Of course, if we think of bulk photonic crystals as perfect mirrors, then it seems reasonable to think you can 'trap' and 'guide' light within slabs of these photonic crystals. PBG materials promised to 'mold the flow of light' as suggested by the title of Joannopoulos' book [6]. To date, a huge part of the challenge has been in making these idealized devices. The 'holy grail' device requires the use of complete bandgap materials, but a complete 3D bandgap with controllable defects at optical frequencies is extremely difficult to fabricate, though there have been recent demonstrations [33, 34]. The reason for the difficulty is that the periodicity required for optical PBG materials is on the order of the optical wavelength, so the features on these devices are in the nanometer scale. By far the more common paradigm is to make quasi-2D PBG devices, where periodicity is introduced only in 2D (xy plane), and localization in the out-of-plane (z) direction is achieved not via the bandgap effect but by dielectric contrast (like conventional optical waveguides). Most of these are fabricated by the use of electron beam lithography and a combination of wet and dry etching techniques. Other fabrication techniques include self-assembly type approaches, but they are not reviewed here. The ease of fabrication of these planar photonic crystal devices is offset by the loss mechanism in the third dimension. There are two main types of defect that we consider here: the point defect and the line defect. We will see that these serve as building blocks for many important devices, particularly in the area of integrated photonics.

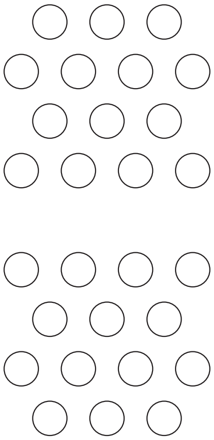


Figure 5.1: Nominal Geometry of a W1 photonic crystal waveguide.

5.2 Building Blocks

The basic building blocks for a photonic circuit are the waveguides and resonators. One of the problems with existing lightwave circuits is that conventional ridge waveguides require a large radius of curvature to minimize bend losses, which limits how compact these devices can be. Photonic crystal waveguides (PCWs) promise to overcome this limitation, making ultra-compact optical devices possible. The most basic conceptualization of the PCW is by removing a single line of holes, hence the name *line defect*. Light within the defect region is surrounded by a PBG material so it can only propagate along the defect region. We show such a waveguide structure for a hexagonal lattice known as the W1 waveguide in figure 5.1. However, achieving the benefits of this device in the planar configuration is still the subject of current research. Some guidelines on how their behavior translates from an infinite height approximation to a finite height slab structure are found here [35, 36]. Early work on these planar PCWs varied different parameters in an attempt to get some basic desirable properties, such as single mode operation [37, 38], or to control the frequencies of the waveguide modes [39]. Controlling losses in these PCWs [40] and PCW bends [41, 42] became an important consideration, as conventional planar lightwave circuits outperform PCWs considerably.

Besides the guiding functionality of the device, one of the most interesting properties of the PCW is their dispersion characteristics. In particular, the group velocity in these waveguides were found to be orders of magnitude less than ordinary waveguides and even show anomalous behavior. Johnson et al. [43] proposed making ultra compact and efficient modulators using PCWs within Mach-Zehnder based devices by taking advantage of this property. Kuipers' group have devised a beautiful experimental setup [44] to measure and verify this PCW dispersion. There was tremendous excitement about these PCWs, leading some to claim that they "can design light paths with *made-to-order* dispersion"(emphasis added)[45], although all that has been demonstrated is that dispersion is a function of the geometry. The purpose of design is of course to advance from dependence to control. The distinction is important because in a PCW, pulse distortion due to its dispersion can occur over much shorter length scales as compared with conventional waveguides, so the unique feature of the PCW becomes another component that needs to be managed unless one can control the dispersion. Dispersion management in general is a very important issue in the field of telecommunications, where pulse broadening limits the bit rate of a communication channel (since adjacent 1's and 0's broaden and blend together making them indistinguishable). Some efforts at controlling the dispersion properties can be found in these references [46, 47, 48], but to our knowledge, there have been no demonstrations of *arbitrary* dispersion design. For integrated optics applications, the effectiveness of add/drop filters and other such components would be severely compromised without managing pulse distortion on the chip. The problem we address using our technique is to design a PCW that achieves an arbitrary target dispersion relation.

The other building block we will examine is the optical resonator, or the PBG cavity. This is conceptualized in a manner similar to the PCW, where we have some region surrounded by PBG materials acting as perfect reflectors. Light within that region becomes trapped. PBG cavities can be point defects (a break in the lattice symmetry at a single lattice point), or enclose a larger region, such as with ring resonators by arranging PCWs in a suitable configuration. Optical resonators in general are useful for use as filters, and the figure of merit is the quality factor Q, which is a measure of the linewidth of the transmission spectrum of the cavity. Point defect PBG cavities can have high Qs ($\sim O(10^6)$ or higher) with mode volume of the order of the wavelength of light. Such cavities are desirable for many applications, such as low-threshold lasers [49], or as add/drop filters [50, 51] when integrated with PCWs for de/multiplexing applications, or for chemical detection in lab-on-a-chip type applications [52, 53]. High-Q small-mode volume cavities are also crucial for achieving strong coupling between an atom and the optical field in cavity quantum electrodynamics (cQED) experiments. Our group's initial venture into the world of PBG came as a collaboration with Axel Scherer's group, studying the feasibility of using PBG cavities in the strong coupling regime [54].

5.3 PBG and Atomic Physics

Cavity quantum electrodynamics (cQED) provides a setting in which atoms interact with the electromagnetic field within an optical resonator. It is one of the first experimentally realizable systems whereby one can quantitatively study the dynamics of an open quantum system under continuous observation, and as such provides a means for testing the laws of quantum measurement, such as quantum trajectory theory [55]. In particular, experimental advances in recent years have crossed the threshold where the intrinsic quantum mechanical atom-field interaction dominates the dissipative and decoherence mechanisms into what is known as the strong coupling regime [56]. Under strong coupling, the presence of a single atom or photon is sufficient to affect the properties of the system, enabling the possibility of single atom switching and single photon nonlinear optics. Other exotic applications within strong coupling include quantum state mapping between atomic and optical states [57], which is critical for the realization of quantum information processing [58].

An emerging reality in the laboratory is the use of nanofabricated optical resonators such as PBG cavities [59], microdisks [60], and microtoroids [61] for cQED experiments with cold atoms. By incorporating a network of photonic crystal devices with single trapped atoms [58] operated under the strong coupling regime [62], scalable quantum information processing can be performed in a quantum network. To motivate the shift to these nanofabricated resonators, consider the pioneering works in strong coupling that make use of high finesse Fabry-Perot cavities. These cavities are much more sensitive to vibrational noise that is not common-mode relative to the two mirrors. To maintain the proper coupling with the atom of interest, the distance between the two mirrors of the Fabry-Perot must be stabilized to $\sim 10^{-15}$ m via active servo control because of the narrow linewidth. The desire to operate at the single photon sensitivity makes it impossible to derive an error signal directly from the 'physics' signal for the servo lock. Therefore, an additional auxiliary frequency stabilized laser is required, and its wavelength must be far enough detuned so that it does not interact with the atom in the cavity. The operating complexity continues to increase as frequency stabilizing this auxiliary laser involves another optical cavity, which also requires another servo control. This extensive amount of labor overhead that precedes each experiment renders this paradigm unscalable for networking purposes.

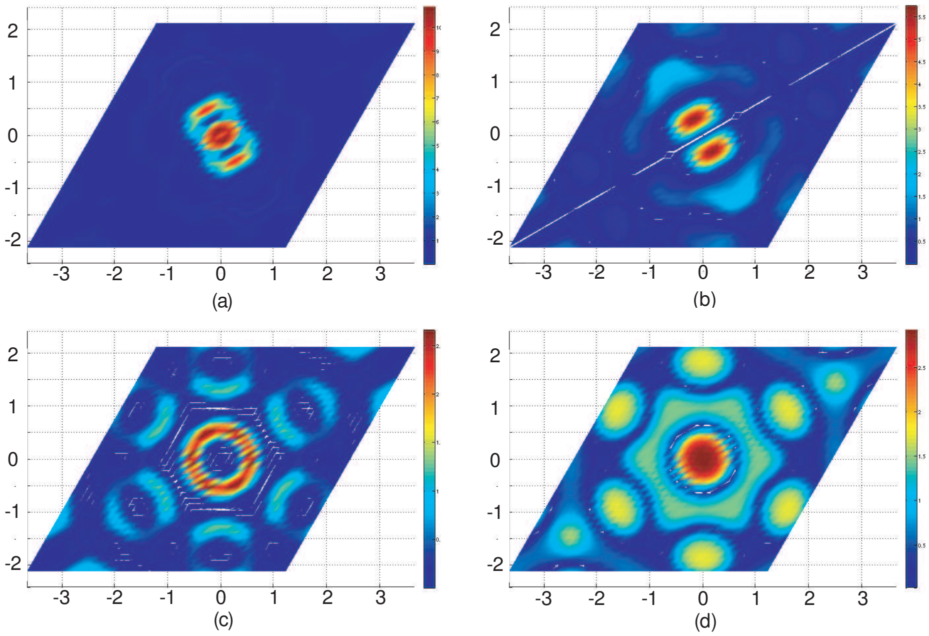
5.3.1 Waveguide dispersion design

In a quantum networking scenario, individual quantum nodes (physically realized using coupled atom-cavity systems) are connected via photonic crystal waveguides. Quantum information stored in an atom can be mapped onto a photon using cQED, and the information transmitted to the next node as the photon performs another cQED interaction with the trapped atom there [57]. The successful implementation of the system will require attention to both the losses as well as the dispersion characteristics in the waveguide. The desire to limit losses in a PCW as it goes through bends and other optical elements is certainly not unique to this application, as we reviewed earlier. Perhaps more unique to the application here is a demanding constraint on the dispersion characteristics of the waveguide, because the photon emission from one quantum node needs to be temporally (and spatially) mode matched for proper excitation of the next node. Losses and pulse distortion will compromise the coupling efficiency to the atom and cause a decrease in transmission fidelity. What is desired is the ability to *arbitrarily* control the dispersion of a waveguide right on the atom chip to compensate for the pulse shape distortion. In contrast to the usual notion of dispersion compensation, where it is only the slope within some small **k**-vector window that gets adjusted, we seek to specify the full dispersion curve. We have not seen in the literature any method that enables arbitrary dispersion engineering of a PCW.

5.3.2 Large defect region cavity design

The second element to the scheme involves the optical cavity. In coupling a single atom to a point defect nanocavity, much of the emphasis has been on finding high-Q and small-mode volume cavities in order to achieve strong coupling [63, 64]. However, what is also desired is a large air (vacuum) opening in which the atom can be trapped, since we wish to minimize the interaction between the atom and the PBG material. Other applications where a large air hole defect is desirable include lab-on-a-chip devices, where we wish to maximize the amount of analyte that can be introduced into the optical excitation region for analyte detection [52] and on-chip optical spectroscopy [53]. In these applications, the important property to consider is that the atom (or chemical) interacts with the electric field that is localized in the air (or vacuum) region. Consider again the h_1 structure where the electric field is localized near the defect region. Given a particular bulk hole radius (r_{bulk}), if we attempt to increase the defect region by increasing the radius of the defect hole (r_{defect}), then as we transition from $r_{defect} < r_{bulk}$ to $r_{defect} > r_{bulk}$, we change from an *acceptor* type mode (figure 5.2a,b) to a *donor* type mode (figure 5.2c,d) [31].

However, for the applications we have contemplated, a donor mode would not be appropriate since an electric field node is located at the center of the evacuated region. One strategy would be to increase r_{bulk} along with r_{defect}, but that would compromise the physical integrity of the device (see figure 5.3), as well as increase the scattering losses [54]. What we seek then is a re-distribution of the layers surrounding the defect hole in a way that increases the overall amount of dielectric (to an acceptable level of



(a)

(b)

(c)

(d)

Figure 5.2: (a) Electric Field intensity of an acceptor mode. Note the field maximum at the center of the air defect region. (b) Magnetic Field intensity. (c) Electric Field intensity of a donor mode. Note the field minimum at the center of the air defect region. (d) Magnetic Field intensity.

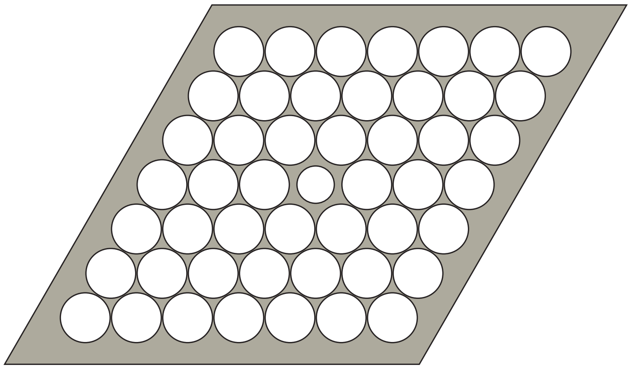


Figure 5.3: h_1 geometry with large bulk hole radius. Fabrication uncertainty could lead to entire sections of photonic crystal collapsing.

mechanical strength), while still retaining the desirable qualities of an acceptor mode. Quantitatively, we can require

$$\overline{\eta}_{\mathcal{D}} \equiv \frac{\int_{\mathcal{D}} \eta(r)dr}{\mathcal{A}_{\mathcal{D}}} > \frac{\int_{\mathcal{B}} \eta(r)dr}{\mathcal{A}_{\mathcal{B}}} \equiv \overline{\eta}_{\mathcal{B}} \tag{5.1}$$

where $\mathcal{A}$ is the area, and $\mathcal{D}$ and $\mathcal{B}$ designate the defect and bulk region respectively. Other groups have demonstrated that we do not require a localized defect within a perfectly periodic lattice in order to localize a mode [65, 66], although it is not clear whether it is possible to violate the simple donor/acceptor mode intuition by rearranging the bulk region. Even if it were possible, intuition fails to provide guidance as to how the redistribution should happen. We examine this problem using our approach.

Chapter 6

Inverting the Helmholtz Equation

6.1 Introduction

In this chapter, we formalize our idea for using an inverse problem based approach to photonic device design. We begin in section 6.2 by surveying the field of PBG design using inverse problem based approaches in the literature, comparing the various methods with the approach we have adopted. This will frame the results presented in these final chapters within the greater context of this emerging field of research. In section 6.3, we derive the inverse Helmholtz equation we need to solve in order to address the design problems motivated in section 5.3. We provide a simple proof-of-principle example in section 6.4 to show that the inverse equations are in fact correct, and also the importance of regularization for solving this problem. We conclude this chapter with a first look at what happens when we ask for a mode that is not supported by a physically realizable dielectric function. This problem will be explored more extensively in chapter 7.

6.2 Inverse Problem Based Design

As we reviewed in chapter 5, there are many applications envisioned for PBG devices. However, the traditional design paradigm is really based on *trial and error*, which is not as suitable for applications. In the existing paradigm, different geometries are proposed and studied (numerically or experimentally), and the effects catalogued. As the field progresses, the understanding of the fundamental scientific principles increases, and we begin to collect together useful effects that can be adapted for useful applications. Devices are improved by repeatedly varying different parameters to existing designs. After much trial and error, more intuition regarding the design problem is developed, so that, with a lot of human ingenuity, one hopes that subsequent 'trials' will less frequently turn out to be 'errors.' The other problem is that despite the many excellent devices that have been developed using this approach, one can usually not be certain how much room for improvement exists for the device. In other words, we are uncertain how *optimally* it is designed given the current manufacturing/cost constraints. Even if we somehow know that we are not optimal, the traditional design approach also does not give us a systematic or algorithmic way that lead us to the better/best designs.

An inverse problem based method is the exact reverse of the traditional method. Rather than systematically varying the causes (dielectric function) to catalogue the resulting effects (optical properties), we purposefully choose the desired effect and look for the unknown cause. As explained in chapter 3, the starting point is the effect, rather than the cause. We will review some methods that claim to be inverse problem based, but are effectively automated trial and error methods.

In an applications driven design paradigm, we imagine designing a device that will optimally perform some desired function. The starting point of the design process is focused on the end application, rather than the best known solution to date. This paradigm shift thus lends itself readily to the inverse problem formulation. Given some desired performance criterion suitable for the application, one needs to first develop a performance metric $\mathcal{P}(H_i(r), \omega_i(r), \eta(r))$ that may be a function of multiple eigenmodes and eigenfrequencies, and even the structure itself. The goal of **optimal** PBG device design is to find a structure that maximizes this function. It should be clear from our use of the term optimal in chapter 4 that we mean the *global* rather than the local optimum in this context, though we will see that this is often not how the term is applied in the literature. We can generally assume that this $\mathcal{P}$ metric is obtainable in that under any design approach, it is required for evaluating the performance of various designs. A design method that produces these optimal devices is considered an optimal design method (and to be clear, in principle this does not necessarily need to be an inverse problem based approach). An optimal inverse problem based design method combines both of these features. The underlying assumption is that if we can somehow find the field that optimizes the objective function (which can be a difficult problem in itself), an inverse problem method will allow us to solve for the dielectric that produces the desired properties. This was the idea behind the work (in the 2D approximation) by Geremia et al. [67], and is in fact the predecessor to the work in this thesis. It turns out there were some fundamental errors with the way the inverse problem was formulated, and these are explored and discussed in appendix D.

A brief survey of the field

Despite increased interest and efforts at developing this inverse problem based design paradigm in recent years, there are still only a few papers that have been published. For a more comprehensive review of the inverse problem approach to PBG design, the reader is referred to an article by Burger et al. [68], but the emphasis there is on the topology optimization approach, particularly those using the level set method. The journal *Inverse Problems* has many electromagnetic type inverse problems, but few directly applicable to photonic crystals or PBG devices in particular. An early treatment by Popov et al. [69] in this journal of a 1D photonic crystal (i.e. alternating layers of dielectric) scattering problem shows that given the reflection coefficient, the index of refraction can be determined assuming ‘practically sensible conditions’, which is of course a regularization condition. The 2D scattering problem is treated by Ammari et al. [70], and both of these are quite heavy on the mathematics, and are more concerned with parameter estimation than design. The interest there is much more heavily biased towards the mathematics of the inverse problem rather than the physics, and this is typical of the papers one finds in that journal to date. The first direct application of inverse problem methods to the PBG community was by

Dobson and Cox, whose work focused on finding unit cell geometries that maximized the bandgap in the TM [71] and TE [72] polarizations. Their technique uses a gradient-based algorithm, and they found that their final 'optimal' designs were quite sensitive to the targeted mid-frequency of the bandgap (which of course indicates that they have only reached local optima). The method in the TM polarization required a structure with an existing bandgap, but the later work removed this restriction.

6.2.1 Genetic algorithms

Sanchez-Dehesa's group has recently published some results based on an inverse problem method for designing high efficiency waveguide couplers [73] and waveguide demultiplexers [74]. The demultiplexer design for an incoming signal with wavelengths 1.5 μm and 1.55 μm gives $>$ 45dB crosstalk suppression and $>$ 75% coupling efficiency. However, the method they use is a genetic algorithm, where the geometry is parameterized and the only variable is whether the cylindrical rods of fixed size and location remain or are removed. Of course, using a genetic algorithm (GA) does not solve an inverse problem as we have defined the term in this thesis. It is systematic trial and error, with the errors generally discarded. The other comment is that GA's generally do give good results for globally optimizing arbitrary functions with many local maxima and minima, as long as the design space is kept small. Therefore, this method cannot guarantee the optimality of the design over all possible structures since the design space must be heavily parameterized.

The technique by Gheorma et al. expands on their approach by allowing 'aperiodic' structures, so the location of these cylindrical rods are not fixed [66]. They found that a straight-forward application of the GA did not converge well given the extra degrees of freedom, so they used an adaptive algorithm, but it is of course still not an inverse problem method as we define it. The key idea here that is consistent with our work is that they give up the notion of an overall lattice periodicity to achieve improved performance. This is a bit of a break from the traditional view of PBG devices, where the bandgap effect plays a vital role in the design. Our results on the enlarged defect region cavity design have similar philosophical underpinnings. The other relevant result here is that they allude to the idea of not actually obtaining the mode they had designed for, but attribute this effect to numerical (i.e. series truncation) errors and other constraints rather than as a fundamental limitation. To our knowledge, this is the only other paper that makes this observation explicitly, although the importance of this property was not fully appreciated.

6.2.2 Topology optimization methods

Topology optimization methods have also been used for minimizing losses in a waveguide bend [75], T-junction [76], and to improve the directional transmission properties of a waveguide termination [77]. In contrast to the GA methods, the dielectric function is not parameterized, but discretized into finite elements, where each element can take on any value. Therefore any design within the discretization bandwidth is within the design space. This aspect is similar to the domain of our design space. The optimization routine uses the iterative Method of Moving Asymptotes (MMA) optimizer which is a local gradient based algorithm, and requires the determination of the sensitivities of the objective function to the design variables at each iteration. However, computing the various sensitivities directly is not feasible, so linear approximations to the model are used. Linearization is necessary even though the Helmholtz equation is linear in the dielectric function because the objective function is not linear in each of the discretized elements. Based on the perturbation theory of linear operators [78], we expect this function can be highly non-linear. The validity and consequences of the approximations made are not discussed, but given the accuracy issues with just solving the forward problem, more caution should be exercised here.

In addition, there are some known problems associated with the method. It has been observed that the design algorithm stops converging as the resolution of the grid is increased [79]. This effect is attributed to the numerical instability of calculating the gradient as the ill-conditioning increases. This is consistent with what we discussed in chapter 4 regarding gradient descent methods for ill-conditioned problems. Finally, since this is a gradient based method, it is also limited to finding locally optimal designs.

6.2.3 Level set method

Closely related to topology optimization is the level set method [80]. The level set function is a way to define an interface between distinct media, so it is a more ideal way of describing the dielectric function. Briefly, the zero crossings of the level set function define the boundaries between ϵ_1 and ϵ_2. The optimization is done by iteratively updating the level set function by solving a Hamilton-Jacobi equation where the velocity term is chosen to climb the gradient to the objective function. Similar approximation issues found with the topology optimization methods are encountered here. This technique was recently applied to maximizing the bandgap for a 2D square lattice with great success for the TM polarization, and to a lesser degree for the TE polarization [81].

6.2.4 Analytical inversion of waveguide modes

Of the various papers in the literature, the work by Englund et al. [64] is probably the most similar to ours in spirit, and received positive review from the community [82]. They designed high-Q small-mode volume cavities by analytical solution of the inverse problem. By restricting the set of target modes to an expansion of the waveguide modes with a slowly varying envelope, they reduce the inverse problem to a 1D problem and analytically solve for the dielectric function along the waveguiding axis. The off-axis structure along the line defect region is reconstructed from the solution of the dielectric function along the line by assuming cylindrical defects. They solved for both a Gaussian and a sinc function envelope modulated defect mode with Q's of 1.6×10^6 and 4.3×10^6, and mode volume (in $(\frac{\lambda}{n})^3$) of 0.85 and 1.43 respectively. The results while promising have some limitations that were not addressed. Particularly for the sinc function design, the output mode actually did not resemble the target mode. In fact, comparing their figures 6(c) and 6(f), the sinc function mode looks more like the Gaussian mode. This is actually consistent with our finding, as we believe this general inability to reach an arbitrary target is a ubiquitous problem.

6.2.5 Our approach

The first distinctive feature is that we derive the inverse problem from first principles, i.e. *ab initio*. We make no additional approximations or linearizations beyond a reduction of the problem from 3D to 2D. In contrast to the topology optimization methods, we solve the full inverse problem exactly, rather than restricting ourselves to local improvements to existing designs using approximate methods. The philosophy behind the approach is that if you can specify what you want, our method will tell you how to get it. Approximate methods limit the set of functionalities you can specify, because they must be of the correct form.

The other distinctive feature of our approach is that we do not parameterize the dielectric function at all in our design space. Any geometry commensurate with our chosen bandwidth is within the design space. Using the convex optimization regularization tool, we remove non-physically realizable values of the dielectric only. We can include additional fabrication constraints without increasing the computational domain if we so desire. Thus, we are not limited as in the GA approaches or the waveguide expansion approach. On this point, it is more akin to the topology optimization method. A criticism of our approach might be that our designs are not binary valued (as with level set methods), so they are not compatible with current fabrication technology. As discussed in appendix C, any discretized dielectric function should really be interpreted via its underlying continuous function. In many cases, we would argue that these should not actually be considered binary valued anyway. Secondly, even without this limitation, this is somewhat intentional. Consider a device requiring continuous valued dielectric functions that can improve performance by several orders of magnitude. Existence of such a design would still be important information to have. This may provide the right incentive to push fabrication technology improvements in that direction. Of course graded index fibers are a simple example of such a dielectric function, so they are not fundamentally impossible designs. Conversely, if existing designs already yield close to optimal performance, then it is not worth spending more effort into looking for improvements. Either way, there is value to our formulation. In that sense, it was our goal to construct a design method that simply gives you the best device possible without particular regard for current fabrication limitations, because those can improve with time and ingenuity. It is possible to derive new insight into the problem, which is another contribution of this method to the field. Furthermore, some forms of fabrication constraints can still be modeled into the design problem as well, and we demonstrate that with our enlarged defect cavity problem. We now derive the inverse equations for the design problem.

6.3 Inverse Helmholtz Equation

Using the plane wave basis, the derivation of the inverse Helmholtz equation uses the same mathematical principles that we used in obtaining the matrix form of the forward problem (explicitly derived in section 2.2 and applied to defects using the supercell method in section 2.3). Our restriction to TE polarized modes within a 2D analysis described in chapter 2 is still in effect here. We make use of the completeness and orthogonality of the plane waves here as well, but rather than solving for the $h_{\mathbf{k}}$ coefficients given an $\eta(r)$, we solve for the $\eta_{\mathbf{k}}$ coefficients given some desired or target field distribution $\boldsymbol{H_m(r)}$ instead.

6.3.1 Point defects

Working out the derivation for the point defect design, we expand the field in the Fourier basis:

$$\boldsymbol{H_m(r)} = \sum_{\boldsymbol{\kappa}} a_{\boldsymbol{\kappa}}^{(m)} e^{i\boldsymbol{\kappa} \cdot \boldsymbol{r}} \hat{\boldsymbol{z}} \tag{6.1}$$

Starting from the Helmholtz equation again, we left multiply with a plane wave and integrate.

$$\int e^{-i\boldsymbol{\gamma}\,\boldsymbol{\cdot}\,\boldsymbol{r}}\left\{\frac{\omega_m^2}{c^2}\boldsymbol{H_m(r)} = \nabla\times\left(\eta(\boldsymbol{r})\nabla\times\boldsymbol{H_m(r)}\right)\right\} d^2\boldsymbol{r}$$

$$\begin{aligned}
\frac{\omega_m^2}{c^2}a_{\boldsymbol{\gamma}}^{(m)} &= \int e^{-i\boldsymbol{\gamma}\,\boldsymbol{\cdot}\,\boldsymbol{r}}\left[\nabla\times\sum_{\boldsymbol{k}}\eta_{\boldsymbol{k}}e^{i\boldsymbol{k}\,\boldsymbol{\cdot}\,\boldsymbol{r}}\nabla\times\sum_{\boldsymbol{\kappa}}a_{\boldsymbol{\kappa}}^{(m)}e^{i\boldsymbol{\kappa}\,\boldsymbol{\cdot}\,\boldsymbol{r}}\right] d^2\boldsymbol{r}\\
\frac{\omega_m^2}{c^2}a_{\boldsymbol{\gamma}}^{(m)} &= \sum_{\boldsymbol{k},\,\boldsymbol{\kappa}} a_{\boldsymbol{\kappa}}^{(m)}\,\eta_{\boldsymbol{k}}\ \ \boldsymbol{\kappa\cdot(k+\kappa)}\int e^{i\boldsymbol{(k+\kappa-\gamma)\cdot r}}d^2\boldsymbol{r}\\
&\equiv b_{\boldsymbol{\gamma}}^{(m)}
\end{aligned}$$

Choosing to collapse the delta function with $\boldsymbol{\kappa}=\boldsymbol{\gamma}-\boldsymbol{k}$

$$\begin{aligned}
b_{\boldsymbol{\gamma}}^{(m)} &= \sum_{\boldsymbol{k}}\left[a_{\boldsymbol{\gamma}-\boldsymbol{k}}^{(m)}\gamma\cdot(\gamma-\boldsymbol{k})\right]\eta_{\boldsymbol{k}} && (6.2)\\
&\equiv \sum_{\boldsymbol{k}}\boldsymbol{A}_{\boldsymbol{\gamma k}}^{(m)}\eta_{\boldsymbol{k}} && (6.3)
\end{aligned}$$

In anticipation of using Tikhonov regularization for solving the inverse problem, we split up the dielectric function into two parts: an 'initial' geometry and a small corrective part:

$$\begin{aligned}
\sum_{\boldsymbol{k}}\boldsymbol{A}_{\boldsymbol{\gamma k}}^{(m)}\delta\eta_{\boldsymbol{k}} &= \frac{\omega_m^2}{c^2}a_{\boldsymbol{\gamma}}^{(m)}-\sum_{\boldsymbol{k'}}\boldsymbol{A}_{\boldsymbol{\gamma k'}}^{\boldsymbol{(m)}}\eta_{\boldsymbol{k'}}^{(0)} && (6.4)\\
\sum_{\boldsymbol{k}}\boldsymbol{A}_{\boldsymbol{\gamma k}}^{(m)}\delta\eta_{\boldsymbol{k}} &\equiv \beta_{\boldsymbol{\gamma}}^{(m)} && (6.5)
\end{aligned}$$

We use β and b to distinguish between the two cases. Just as we did with the Helmholtz operator $\Theta^{(\eta)}$ in equation (2.24), we have explicitly included a superscript m on $\boldsymbol{A}$, a, b, and β to emphasize their dependence on the desired mode $\boldsymbol{H_m(r)}$. The importance of $\boldsymbol{A}$'s dependence on the desired mode will be explored in section 7.2. The procedure for solving the inverse problem given some desired mode is to first express it in the Fourier basis, and then forming the $\boldsymbol{A}$ matrix using eqn. (6.2) and (6.3), and finally solve either eqn. (6.3) or (6.5) depending on which regularization scheme we choose to utilize.

6.3.2 Line defects

For the waveguide dispersion, recall that there is a separate eigenvalue problem for each wavevector $\boldsymbol{q}_i$ along the propagation direction of interest. For each $\boldsymbol{q}_i$, we must perform the same steps as above on the waveguide form of the Helmholtz equation (eqn. (2.29)) to obtain the following inverse problem:

$$\begin{aligned} b_{\boldsymbol{\gamma}}^{(\boldsymbol{q}_i,m)} &= \sum_{\boldsymbol{k}} \left[a_{\boldsymbol{\gamma}-\boldsymbol{k}}^{(\boldsymbol{q}_i,m)} \boldsymbol{\gamma} + \boldsymbol{q}_i \cdot (\boldsymbol{\gamma} - \boldsymbol{k} + \boldsymbol{q}_i) \right] \eta_{\boldsymbol{k}} && (6.6) \\ \boldsymbol{A}_{\boldsymbol{k},\boldsymbol{k}'}^{(\boldsymbol{q}_i,m)} &\equiv a_{\boldsymbol{k}-\boldsymbol{k}'}^{(\boldsymbol{q}_i,m)} (\boldsymbol{k} - \boldsymbol{k}' + \boldsymbol{q}_i) \cdot (\boldsymbol{k} + \boldsymbol{q}_i) && (6.7) \end{aligned}$$

The solution $\eta_{\boldsymbol{k}}$ will have to simultaneously satisfy all $N_{\boldsymbol{q}}$ of these inverse equations, where $N_{\boldsymbol{q}}$ is the number of wavevectors we will include in the dispersion curve. This can be formally expressed by a vertical concatenation of the $\boldsymbol{A}^{(\boldsymbol{q})}$ matrices and $\boldsymbol{b}^{(\boldsymbol{q})}$ vectors.

$$\widetilde{\boldsymbol{A}} \equiv \begin{bmatrix} \boldsymbol{A}^{(\boldsymbol{q}_1)} \\ \boldsymbol{A}^{(\boldsymbol{q}_2)} \\ \vdots \\ \boldsymbol{A}^{(\boldsymbol{q}_n)} \end{bmatrix}, \qquad \widetilde{b} \equiv \begin{bmatrix} b^{(q_1)} \\ b^{(q_2)} \\ \vdots \\ b^{(q_n)} \end{bmatrix}, \text{ and } \qquad \widetilde{\beta} \equiv \begin{bmatrix} \beta^{(q_1)} \\ \beta^{(q_2)} \\ \vdots \\ \beta^{(q_n)} \end{bmatrix} \qquad (6.8)$$

For the remainder of this chapter, unless required for clarity, we will omit many of the cumbersome subscripts and superscripts on $\boldsymbol{A}$, b, and β with the understanding that, with a truncated basis, they can be treated like matrices and vectors.

6.4 Proof of Principle

In this section, we will work through a contrived problem as a proof of principle demonstration, but also demonstrate the steps one would take in performing such a design procedure. We first define our computational domain. We will choose a hexagonal lattice of cylindrical air holes[1] of radius $r = 0.3a$ in dielectric using a $7a \times 7a$ supercell, and include 19 reciprocal lattice vectors before truncating the Fourier series. The total number of plane waves is 931 in this example, and a is the lattice constant. As our illustration, we will 'design' a point defect cavity geometry where the central air hole is refilled with a dielectric material. This is referred to as the h_1 defect. Using the set of 931 plane waves, we would normally construct an $\boldsymbol{H}_m(\boldsymbol{r})$ with some suitably desired properties. In this case, we obtain the 'desired mode' $\boldsymbol{H}_m(\boldsymbol{r})$ by explicitly solving the forward problem.

For this simple geometry, we can use the analytical expression for $\eta_{\mathbf{k}}$. We follow the method found in [3] to evaluate the transform. The reference gives the defect-free coefficients:

$$\eta_{\mathbf{G}} = \eta_d \delta(|\mathbf{G}|) + (\eta_a - \eta_d)\frac{2\pi r^2}{\sqrt{3}}\left(\frac{2J_1(|\mathbf{G}|r)}{|\mathbf{G}|r}\right)$$

where $\eta_d = \frac{1}{11.56}$ is the reciprocal dielectric constant of the dielectric (value is typical of a semiconductor like AlGaAs), $\eta_a = 1$ represents air, and J_1 is the Bessel function. Evaluating only when $\mathbf{k} = \mathbf{G}$ gives us the bulk symmetry. Adding the defect means we need the expansion for $\delta\eta(r)$ corresponding to filling in the air hole. A straightforward modification of the above gives the required coefficients.

$$\eta_{\mathbf{k}} = (\eta_d - \eta_a)\frac{2\pi r^2}{N_1 \times N_2\sqrt{3}}\left(\frac{2J_1(|\mathbf{k}|r)}{|\mathbf{k}|r}\right)$$

This is now evaluated for all $\mathbf{k}$'s, and the factor $N_1 \times N_2$ in the denominator is due to integrating over the size of the entire supercell. We show in figure 6.1 the underlying dielectric function. Using these coefficients, we construct the Helmholtz

[1] Recall the discussion in section 2.4 about the convergence issues of the plane wave method. Thus we are actually considering the underlying continuous function of the nominal geometry. Refer to appendix C for additional details.

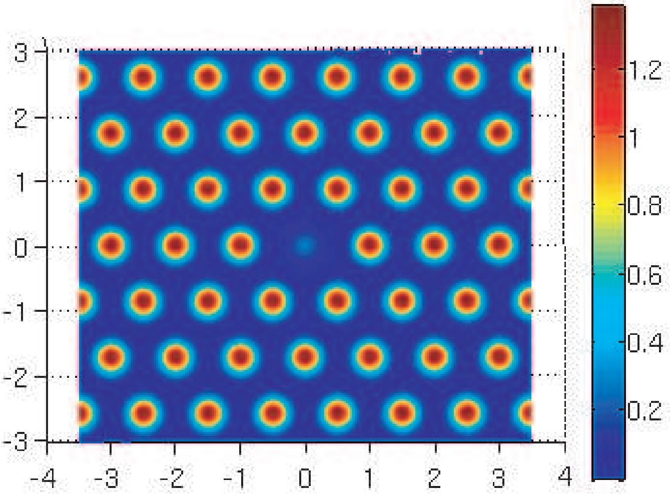


Figure 6.1: Underlying continuous dielectric function of the nominal h_1 defect geometry.

operator using eqn. (2.28) and solve the eigenvalue problem. The localized mode is located at the 50^{th} band. Figure 6.2 shows the magnetic field of the localized mode. We now take this $\boldsymbol{H(r)}$ and using only information about this field, attempt to get back the original dielectric function. We form the inversion matrix A following eqn. (6.2). Of course, after the discussion in chapter 3, it should not be surprising that this problem is ill-posed. Nevertheless, just to illustrate, we can try to perform a QR factorization to invert the A matrix and solve for $\eta_{\mathbf{k}}$. The result is shown in figure 6.3, and as we expect, it looks like noise that has been amplified, bearing no resemblance to the actual dielectric function that produced this mode.

6.4.1 Tikhonov solution

We now use Tikhonov regularization to solve the inverse problem. As a demonstration, we imagine that we are only given this localized mode, and assume we have no knowledge of what photonic crystals or bandgaps are at all. Solving the regularized problem gives a solution shown in figure 6.4. The structure is a marked improvement over the QR solution, and definitely suggests creating a periodic lattice of air holes

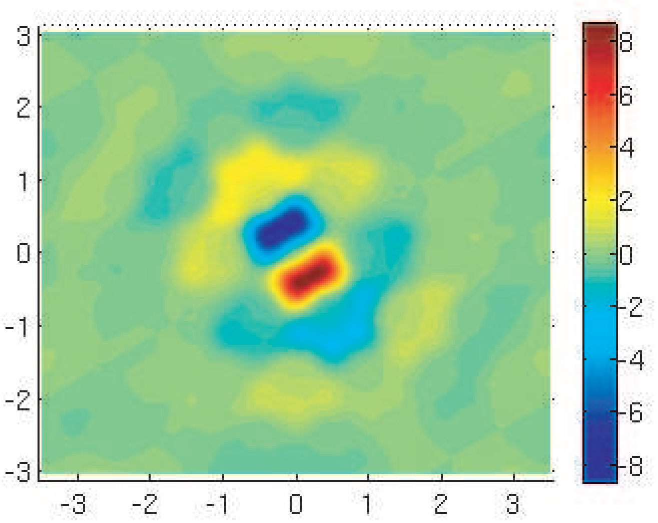


Figure 6.2: Real part of the magnetic field intensity for localized defect mode for the h_1 defect geometry.

with a central defect, although there are still some areas near the edges that do not quite resemble the exact solution. This can be partly attributed to the fact that the solution norm is sizeable $|\eta_{\mathbf{k}}| = 0.37$. To improve on the result, we now invoke the perturbative form of the inverse equation (eqn. (6.5), and use the defect-free bulk lattice as an initial geometry. This finally gives us the same geometry that we had started out with, demonstrating that we have indeed solved the inverse problem. This solution is shown in figure 6.5.

6.5 Simulating Design Errors

In the previous section, the entire scenario is, of course, rather artificial, since we knew (by explicit construction) that some geometry must exist that will produce the target mode. In an actual design problem, we would not be certain of that *a priori*. We model this uncertainty by adding a small noise term to the target (h_1) field. From the discussion in chapter 3, we know that we are not guaranteed the existence of a solution in an inverse problem, which means that some desired modes just simply

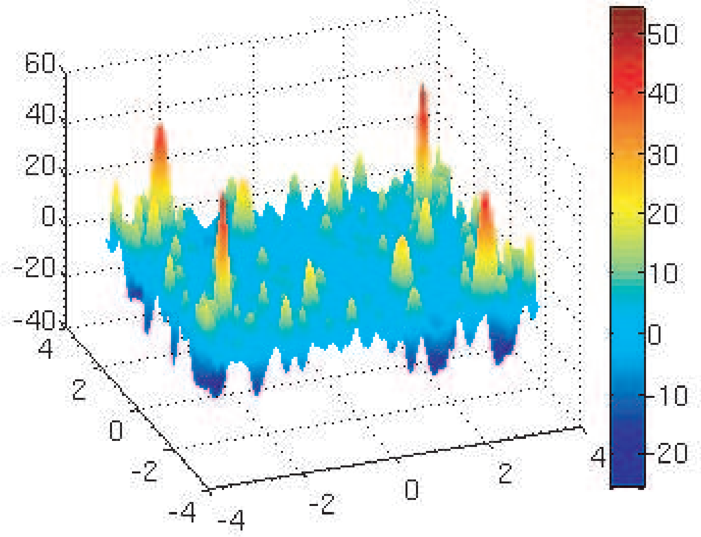


Figure 6.3: Solution to h_1 inverse problem using a QR factorization to invert A. Notice the values of the dielectric function far exceed the original function.

cannot be supported. In this example, even if the perturbed field is not supported, we would still like to recover the h_1 geometry, because we know that the h_1 geometry reproduces the target field minus the small noise term. In the following example, the noise corresponds to a 1% perturbation.

Using the perturbed field as the target field, we proceed to form the inverse problem as before with a bulk lattice starting point and solve using Tikhonov regularization again. Using the same regularization parameter of $\lambda = 3 \times 10^{-3}$ as before yielded a noisy solution similar to the QR factorized solution. Clearly, much more regularization is required in this case. In figure 6.6, we show the solution $\delta\eta(r)$ using $\lambda = 3.73$. The residual norm in this case was 0.3143. We show both the real and imaginary parts of $\delta\eta(r)$, since there are some fairly significant contributions to the imaginary part of the dielectric. Looking at the real part of $\delta\eta(r)$, we find the dominant feature is as we expected, which is to make the central air hole more dielectric-like. Notice also in the dielectric function some fluctuations in the area surrounding the defect, which we see is an attempt to accommodate the added noise term in the target field. Of course, our model assumes that the dielectric is real-valued, so the imaginary parts

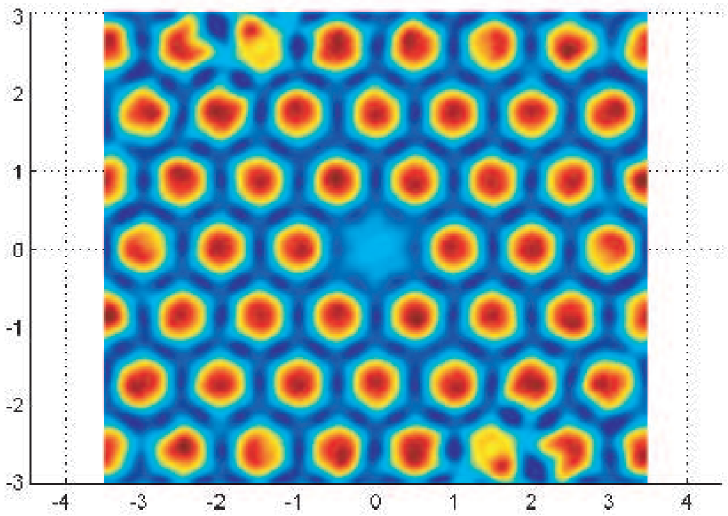


Figure 6.4: Solution to h_1 inverse problem using Tikhonov regularization with no assumptions about the dielectric geometry. Solution suggests creating a periodic lattice of air holes with a central defect.

are particularly problematic. Again, because we knew *a priori* the 'correct' solution, we can safely disregard the ripples in the real part and focus only on refilling the central hole.

A general limitation of the Tikhonov scheme is that the solution will need to be 'interpreted' to look for the most reasonable or feasible solution. In this case, if we did not know what the correct geometry should have been, we would first have dropped the imaginary parts, since we cannot do anything about those anyway, and then started the next iteration by filling in the central hole (since it is the most prominent feature). We would then redo the forward problem with the central hole refilled, and compare with our target mode. If we were still not satisfied, we could solve for the inverse problem again, but using the solution from the last iterate as the 'initial geometry'. This is the strategy we will use for the PCW dispersion design problem. We would still not be guaranteed the solution, and we may well find the desired field is not supported. To rigorously confirm this, we make use of the convex optimization scheme developed in chapter 4.

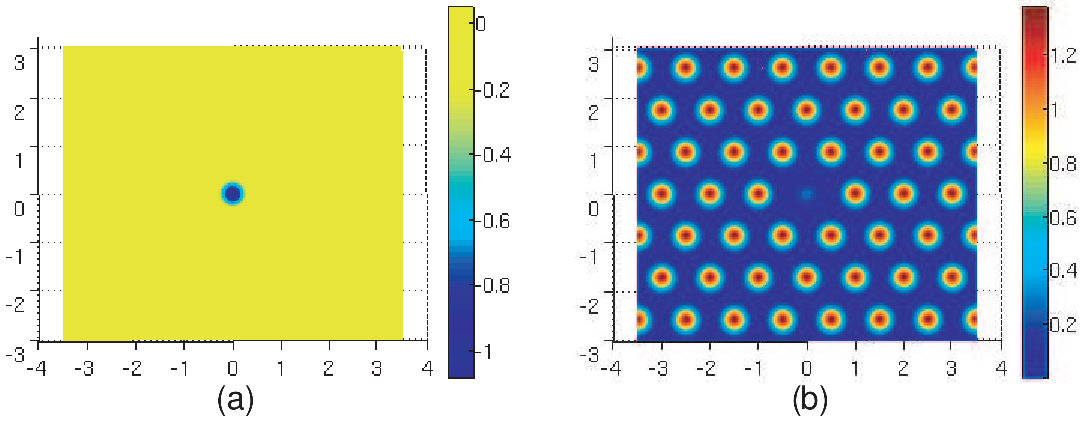


Figure 6.5: Solution to h_1 inverse problem using Tikhonov regularization with an initial defect-free lattice geometry. (a) Actual solution $\delta\eta(r)$ to the inverse equation. (b) Full reconstructed solution $\eta(r) = \eta_0(r) + \delta\eta(r)$, which look identical to figure 6.1. The regularization parameter used was $\lambda = 3 \times 10^{-3}$, and the residual norm was below 10^{-9}.

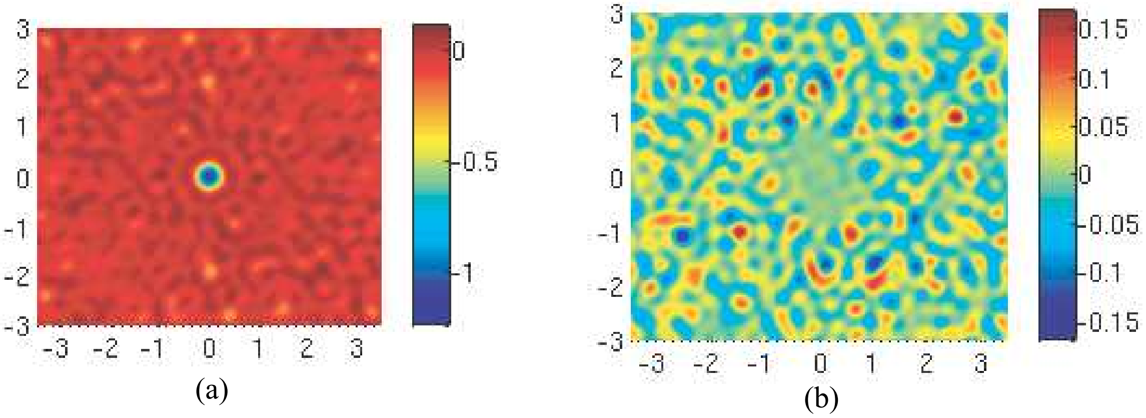
6.5.1 Convex optimization regularization (COR)

The problem we need to solve, first shown without explanation in eqn. (4.1) becomes the following:

$$\begin{gathered} \min_{\eta} |A\eta - b|^2 \\ \text{subject to } \eta_{min} \preceq \mathcal{F}^{-1}\eta \preceq \eta_{max} \end{gathered} \tag{6.9}$$

where $\mathcal{F}^{-1}$ is the inverse fourier transform operator. The symbol $\preceq$ means componentwise less than or equal to, so each discretized value of the real-spaced dielectric function[2] must lie between η_{min} and η_{max}. We explicitly set the imaginary part of the dielectric function to zero by enforcing $\eta^*_{\mathbf{k}} = \eta_{-\mathbf{k}}$ and using the transformation in section 4.2.1. Notice that we are using b instead of β in the objective function, which means we assume no knowledge of the defect-free lattice. We use the convex optimization algorithm as described in chapter 4 to solve the constrained minimization

[2] The dielectric function used to generate this forward problem does not use the analytical expression for $\eta_{\mathbf{k}}$ because truncation leads to overshoot. As shown in figure 6.1, $\eta_{max} > 1$, which is strictly unfeasible. Therefore, the actual values of η can exceed the bounds here. Refer to section 8.2 for how this is handled.



(a)

(b)

Figure 6.6: Solution to the noisy h_1 inverse problem using Tikhonov regularization. (a) Real part of $\delta\eta(r)$. (b) Imaginary part of $\delta\eta(r)$. The regularization parameter used was $\lambda = 3.73$, and the residual norm was below 0.314.

problem. The solution is shown in figure 6.7 with a residual norm of 0.2489. There is a slight discrepancy between our solution and the original geometry within the region where the target mode has very little intensity. Intuitively, this is sensible if we think of the inverse problem in terms of its signal to noise ratio. Where the mode has little to no intensity, there is insufficient signal to overcome the added noise to reconstruct the desired dielectric completely. Without the added noise to the target field, the COR reconstructs the dielectric function perfectly, as with the Tikhonov regularization scheme with a residual norm at the machine precision level. The difference is that we did not require prior knowledge of the defect-free lattice using COR. We defer a more thorough comparison of the two schemes until section 7.4.1. With the added noise, both schemes gave reasonably close approximations to the original geometry. The interpretation we ought to make here is that the noisy mode is not supported by any physically realizable geometry. We can claim this rigorously because the globally minimized residual norm to the solution of eqn. (6.9) is non-zero. Therefore, no other geometry exists that can reduce the norm further (or exactly solves the inverse problem). This means that we cannot track the noise that has been introduced, and we will explore the role of the residual norm more closely in the next chapter.

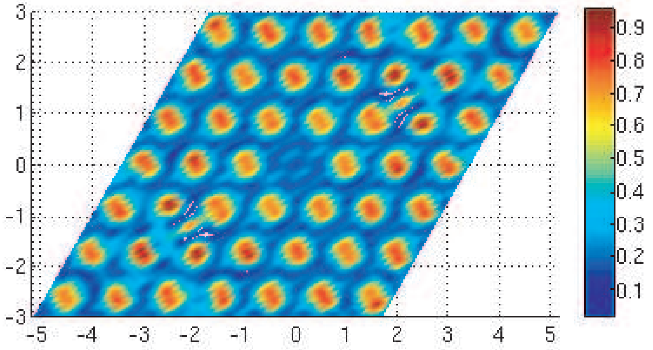


Figure 6.7: Solution to the noisy h_1 inverse problem using CO regularization.

Chapter 7

From Inverse Problems to Device Design

7.1 3:1 Waveguide Splitter

Encouraged by our analysis in the last section, we now attempt our first simple but non-trivial design problem where we wish to have an incoming waveguide split into 2 branches with a ratio of 3 to 1, and then recombined. We retain the $7a \times 7a$ supercell geometry and construct a mode that complements the chosen hexagonal lattice. The mode is created by first adopting a transverse Gaussian profile along the splitter path, then modulating each segment with an appropriate plane wave. To produce the 3:1 split, we attenuate the upper branch amplitude to 25% of the unbranched intensity, and the lower branch to 75%, as in figure 7.1. We choose the desired frequency ($\omega a/2\pi c$) to be 0.2081, which lies in the middle of the bandgap. We choose the simpler Tikhonov scheme and use the defect-free lattice as a starting point again. Figure 7.2 shows the regularized solution with a residual norm of 3.4491 using a regularization parameter of 32.4. In figure 7.3 we show the L-curve for this problem, where $\lambda_{corner} \approx 1.5$. The much more subtle 'corner' is more typical of a real design problem, and we found (as we mentioned in section 3.3) that the best solution is not near the corner anyway. The solution at $\lambda = 1.5$ is shown in figure 7.4, and again, notice the scale on the colorbar and the substantial amount of noise in the solution.

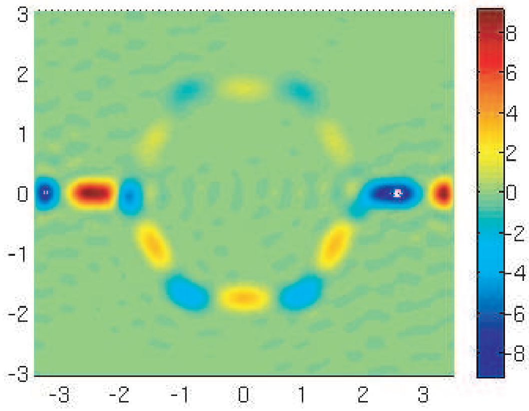


Figure 7.1: Target mode for the 3:1 splitter device.

In general, we try several parameters until a sensible solution is found. From our experience, we often have to regularize beyond the L-curve corner, and 'interpret' the result by smoothing out the noisy components. Generally, we look at a particular solution and identify dominant features, and then propagate the smoothed solution through the forward problem again. In this case (ignoring the imaginary parts again), we find the real part of the dielectric suggests a geometry where the upper branch has a hole radius that is half that of the bulk hole radius, and the lower branch has completely filled holes as in figure 7.5. Propagating this geometry through the forward problem yields the 3:1 split mode as shown in figure 7.6. Notice that the actual mode we obtained was not identical to the original target mode, with the obvious difference between the two being that the target mode was more localized than the obtained mode. However, since our design goal was only to construct a 3:1 split waveguide, we can stop here. The discrepancy between the two should not be surprising in light of the magnitude of the residual norm, as we discuss in the following section.

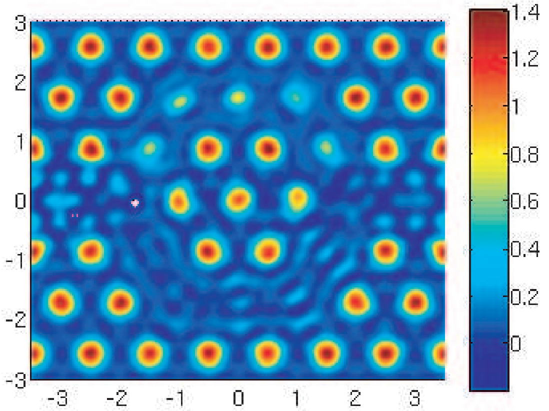


Figure 7.2: Regularized solution of the 3:1 splitter inverse problem using a regularization parameter of 32.4

7.2 Residual Norm

Recall that the expression for our inverse equation comes from the Helmholtz equation:

$$\begin{aligned} \nabla \times \Big(\eta(\mathbf{r})\nabla \times \mathbf{H_m}(\mathbf{r})\Big) &= \frac{\omega^2}{c^2}\mathbf{H_m}(\mathbf{r}) && (7.1)\\ \therefore \;\; \Theta^{(\eta)}h^{(m)} &= \frac{\omega^2}{c^2}h^{(m)} \equiv b && (7.2)\\ A^{h^{(m)}}\eta &= b \qquad \text{and,} && (7.3)\\ A^{h^{(m)}}\eta &= \Theta^{(\eta)}h^{(m)} && (7.4) \end{aligned}$$

$\eta(\mathbf{r})$ and $H_m(\mathbf{r})$ are a special pairing since $\eta(r)$ is the dielectric that supports $H_m(\mathbf{r})$ as an eigenmode. The equality in equation (7.1) only holds if the two are an 'eigenpair'. For some given solution η_{sol} we find in solving the inverse problem for which $A\eta_{sol} \neq b$, it necessarily means that $H_m(\mathbf{r})$ is not an eigenmode of η_{sol}. This finite residual norm is significant for the design problem, because it means we are guaranteed **not** to obtain our desired mode. The obvious question to ask is what exactly are we getting (using a residual norm metric) when we don't find an eigenpair? We can examine the

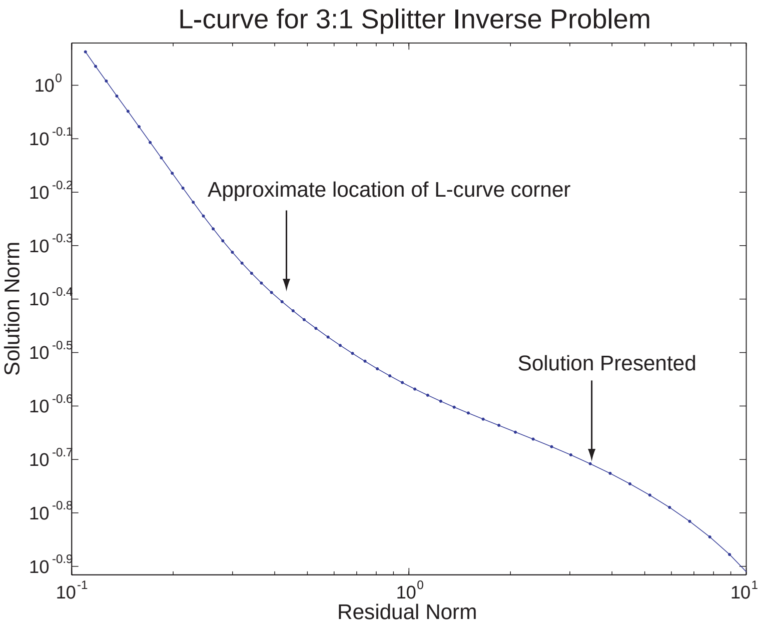


Figure 7.3: L-curve for the 3:1 splitter showing the location and shape of the corner as well as the optimal solution.

left hand side of the equation using η_{sol} to form the Helmholtz operator Θ.

$$A^{h^{(m)}}\eta_{sol} - b \neq 0 \tag{7.5}$$

$$\Theta^{(\eta_{sol})}h^{(m)} - b \neq 0 \tag{7.6}$$

Since the forward problem is well-posed, we can always find the spectrum of eigenmodes for any given dielectric geometry. Assume the following spectral decomposi-

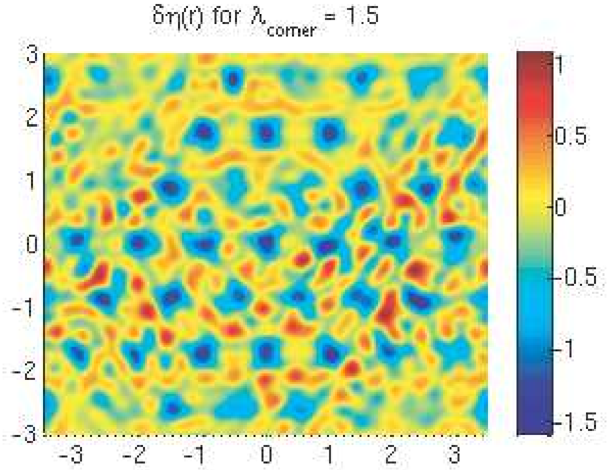


Figure 7.4: 3:1 splitter $\delta\eta(r)$ solution using regularization parameter from the L-curve corner. The information in this 'solution' is practically useless.

tion:

$$\begin{aligned}
\Theta^{(\eta_{sol})} h_i^{\eta_{sol}} &= \frac{\omega_i^{(\eta)^2}}{c^2} h_i^{\eta_{sol}} \\
\text{Let} \qquad h^{(m)} &\equiv \sum_i \alpha_i h_i^{\eta_{sol}} \\
\text{Then} \qquad \Theta^{(\eta_{sol})} h^{(m)} - b &= \Theta^{(\eta_{sol})} h^{(m)} - \frac{\omega_m^2}{c^2} h^{(m)} \\
&= \Theta^{(\eta_{sol})} \sum_i \alpha_i h_i^{\eta_{sol}} - \frac{\omega_m^2}{c^2} \sum_i \alpha_i h_i^{\eta_{sol}} \\
&= \sum_i \frac{(\omega_i^2 - \omega_m^2)}{c^2} \alpha_i h_i^{\eta_{sol}}
\end{aligned} \tag{7.7}$$

Even though minimizing the residual norm over (η_{sol}) may be the 'best' general strategy for solving an ill-conditioned linear system of equations, it becomes clear that it is not the most appropriate strategy for the purpose of the design problem. Using the eigenvalue decomposition shows explicitly that the entire spectrum of eigenmodes of η_{sol} contribute to the residual norm, whereas we only care that our desired mode be close to a single eigenmode of η_{sol}. Furthermore, each eigenmode in the spectrum is component-wise weighted by the ω_i^2 term. Since the frequency of the modes of interest usually lie within the first bandgap (i.e. relatively low frequencies), the residual

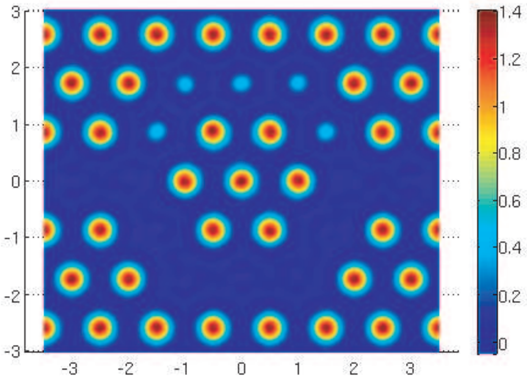


Figure 7.5: Actual $\eta(r)$ used based on the output solution of the inverse problem (see figure 7.2.

norm metric overemphasizes high frequency mode contributions. Therefore, the minimization has a bias towards dielectric geometries whose high frequency eigenmodes are orthogonal to the desired mode at the expense of a stronger overlap of a single eigenmode.

As an illustration, we can repeat our perturbed inverse problem using the noisy h_1 mode as our target mode (only we add an even smaller perturbation ($|n| = 10^{-3} \times |h_m|$) to the mode and renormalize). Rather than looking to solve the equation as we did in section 6.5, we enter our exact solution $\eta_{\mathbf{k}}^{h_1}$ into $|A\eta - b|$ and evaluate the residual norm. Even though it would produce our target mode minus the small bit of noise, the residual is 1.59. Because of the high frequency components, we see that the best solution to the physical problem does not even come close to solving the linear equation.

In situations where the desired mode happens to be an *exact* eigenmode of some geometry, the residual norm metric is fine, because it will simply find the complementary solution (as in our contrived example). The inverse equations find a solution whose spectral decomposition of the target mode is a pure eigenmode, so it can afford to ignore the high frequency mode spectrum. However, this is a rare occurrence, and we will elaborate further in section 7.3. As we deviate further from an exact eigen-

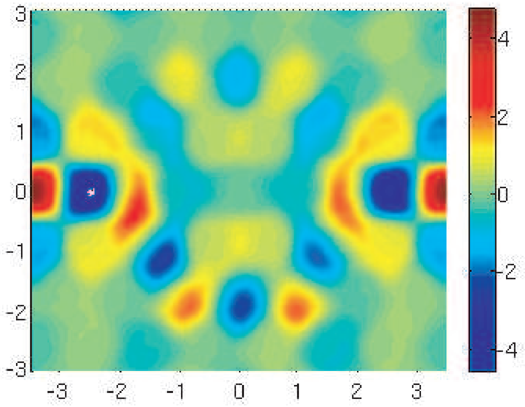


Figure 7.6: Magnetic field distribution of the 3:1 splitter mode of interest supported by the design of Fig. 7.5.

mode, the effect of the errors become more observable, as we saw in section 6.5. In any realistic design problem then, what benefit can we get out of this approach? This is precisely the phenomenon we see with the 3:1 waveguide splitter. We found that the obtained eigenmode was actually quite different from the desired one, and yet it achieved our goal. By examining the residual norm more closely, we now understand that the optimality of our result depends somewhat fortuitously on the target mode's proximity to a feasible eigenmode. The more overlap between the target mode and a feasible eigenmode, the less it needs to compromise with the high frequency modes. However, we also see that despite this non-optimal bias, it still manages to find a 'reasonable' result. Particularly in situations where we truly have little intuition into the problem, this will at least give us a good starting point. Therefore, we cannot claim the resulting mode is **optimal** as we might have hoped, but we find that it is still 'better' than what we may otherwise have.

7.3 Inverse Problem Based Design Flow

We began chapter 6 by motivating an inverse problem based design approach to obtain 'optimal' PBG structures, but so far we have really only outlined a single step: that of solving the inverse problem. In the previous section, we showed how even this one step is often not optimal, and we discovered the importance of a ‘good’ target mode. In this section, we expand on the entire design process, highlighting other difficulties to achieving optimality.

Recall that the idea of optimality is connected to a performance metric that is some function of the desired field mode, and possibly of the dielectric function as well. To find an optimal design implies finding the optimum of the performance function, but there are unfortunately no guarantees that one can actually find the global optimum of such a function. High-dimensional global optimization for arbitrary functions is notoriously difficult (NP-hard). Therefore, the success of even the first step (i.e. coming up with the target mode) will depend on the form of the function. One can sometimes get around this by choosing to define the performance function in a way that mimics the desired behavior, but will also have some nice features such as convexity (and thus solvable by numerical optimization methods). One such example can be found in correlating Fourier components within the light cone with the loss in cavity Q factor [64, 9]. This restricts the types of performance criteria over which we can optimize.

A second problem is that there may not be a unique optimum to our function. In our 3:1 waveguide splitter example, we arbitrarily chose a parameter for the width of the transverse Gaussian profile because it really did not matter. From a design perspective, as long as the split was 3:1, we are somewhat ambivalent about what the actual width needs to be, as long as the mode remains confined. As such, there would be many such field distributions that are ‘optimum’ for the design goal. However, the inverse problem approach demands the full specification of a single field. Which one of these should we choose? If we can always find a dielectric that produces our desired eigenmode, then this is not a problem. In that case, it simply means that there are multiple designs that are all suitable. Unfortunately, if this is not the case (and so far, it appears that this is the norm), then we have the problem of non-zero residual norm. The question becomes which, if any, of the other field distributions with different width profiles might have been valid, and thus form a solvable inverse problem. There is no way of knowing unless we try them all, which starts to look like trial and error and thus defeat the purpose of the approach.

7.3.1 Valid eigenmode landscape

The critical question to ask becomes how prevalent are these valid eigenmodes? Are we more likely to come up with valid modes or invalid ones that do not readily lead to a solution? If most modes one can design are in fact valid, or at least *approximately* valid, then that is not likely to be a problem. Unfortunately, we can make some strong arguments that the invalid regions are much more prevalent.

Consider a linear operator L such that $Lx = \lambda x$. For small perturbations ΔL, we know that the perturbation to the corresponding eigenvalues and eigenvectors are bounded [78]. Consider the set of all perturbations having some norm $|\Delta L| \leq \xi$, and let the neighborhood of points (corresponding to normalized vectors) on the unit hypersphere that bounds the rotation of a given eigenvector be denoted $\mathcal{S}$. If we treat this purely as a mathematical inverse eigenvalue problem, then we can access any new eigenvector in $\mathcal{S}$ while bounding only the norm of ΔL. Any vector in the neighborhood of an existing eigenvector is a valid eigenvector of some perturbed operator. Indeed, it may seem strange (having framed our design problem in the linear algebra language) to hear about vectors in $\mathcal{C}^N$ that are unfeasible eigenvectors. However, we do not have a purely mathematical inverse problem.

Consider again the Helmholtz operator

$$\Theta^{\eta}_{\mathbf{k},\mathbf{k}'} = \eta_{\mathbf{k}-\mathbf{k}'}\mathbf{k} \cdot \mathbf{k}'. \tag{7.8}$$

In contrast to our general operator L, the Helmholtz operator has $N \times N$ elements and is parameterized by a vector of length N, so we have fewer degrees of freedom to accommodate arbitrary changes to eigenvectors. The structure of the operator means that only a few of these ΔL-type perturbations can take on the valid form of $\Delta\Theta = \delta\eta_{\mathbf{k}-\mathbf{k}'}\mathbf{k} \cdot \mathbf{k}'$, and even fewer of these have an expansion of $\eta(r)$ that can take on physically realizable values. Now, it is certainly true that there exist perturbations to $\eta(r)$ that keep you within $\mathcal{S}$. In fact, the observation of robustness of devices to

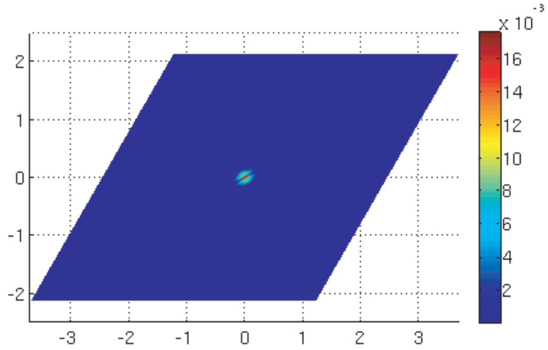


Figure 7.7: Perturbation introduced to $\eta(r)$.

fabrication uncertainty [65] is a good example. However, existence of these modes does not reveal the density of these modes. Our claim is that, due to the rigid form of the Helmholtz operator, the landscape of $\mathbf{H}(\mathbf{r})$ consists mostly of invalid eigenmodes. Therefore, the density of valid modes in $\mathcal{S}$ is small. While we cannot prove this rigorously, we have performed numerical simulations to test our theory.

First, we apply a small perturbation to our canonical h_1 geometry by nominally increasing the radius of the central defect from 0 to 0.03. This corresponds to $|\Delta\eta_k/\eta_k| = 8.6 \times 10^{-4}$ (see figure 7.7). We then solve the forward problem and find the resulting perturbation to the defect mode $|\Delta h_k| = 0.013$ (see figure 7.8). Using the original h_1 eigenvector, we now add to it a perturbation of a much smaller magnitude ($|\Delta a_k| = 10^{-3}$), and proceed to form the inverse problem using this perturbed mode. We minimize the residual norm using the convex optimization scheme, so a non-zero residual norm indicates definitively that the target mode is not a valid eigenmode. In figure 7.9, we plot the distribution of the residual norm for 10000 of these tests. None of these perturbed target 'eigenvectors' corresponded to physically realizable operators. Therefore, the neighborhood surrounding a valid eigenmode appear to be mostly invalid eigenmodes, while the valid modes occupy a set of lower dimensional hypersurfaces within the domain of $\mathbf{H}(\mathbf{r})$'s. As we increase the resolution of the computational grid, the dimensionality increases, further reducing the effective

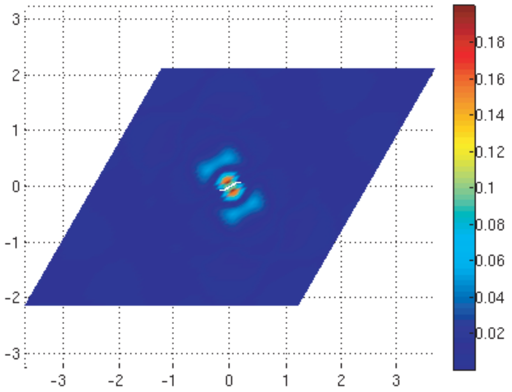


Figure 7.8: Change in eigenmode as a result of perturbation to $\eta(r)$.

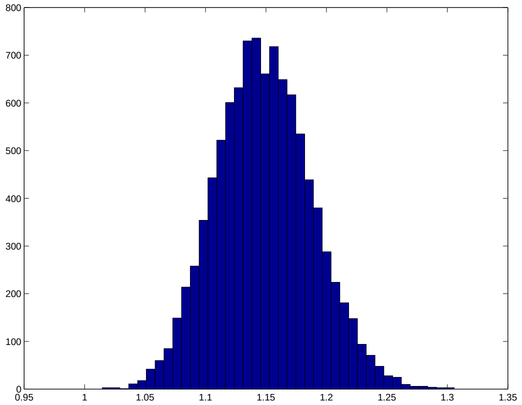


Figure 7.9: Distribution of residual norms of perturbed inverse problems.

density of valid eigenmodes. If we look at figure 7.9 closely, we find the mean value of the residual norm is around 1.15, whereas in section 6.5, the residual norm was closer to 0.25. The number of plane waves used to generate this figure was 3721, whereas we had less than 1000 before, which is another indication that this problem scales badly with the dimension of the state space.

This dramatically changes our concept of the first part of our design process, i.e. the performance optimization of a given mode. The performance metric does not have a continuous domain (set of all normalized vectors, i.e. a hypersphere in $\mathcal{C}^N$), but instead only has little pockets of validity. There is no way of knowing where these pockets are *a priori*, and yet the optimization problem must take this into account. Even when an exact eigenmode is 'close' to our optimized target mode, this distance is not evaluated over the relevant performance metric. It is a general 2-norm distance rather than a weighted distance function that is meaningful to the design goal. We do not know how much of the desired property of the field is lost when we are forced into the valid eigenmode. The limitation imposed by this 'holey' landscape prevents the kind of precision implied by the performance optimization. Given our analysis, it makes the optimization somewhat moot as we are not likely to benefit from it anyway.

We also believe this limitation is quite general despite our bandwidth limitation here. Although computationally impractical, there is no fundamental limitation to the resolution one can use in principle. The argument presented here still holds, and we believe the regions of invalidity dominate more as the dimensionality increases. At infinite resolution, we recover the exact Helmholtz equations, and thus this limitation is clearly independent of the computational model. What we have observed poses a formidable challenge to arbitrary and/or optimal design of PBG structures using an inverse problem method. What is apparent is that this is not a turnkey type design methodology, despite our stated goal of an algorithmic approach to PBG design.

7.4 Conclusion

One obvious improvement that deserves investigation is an alternative to the residual norm as a minimization metric, as discussed throughout this chapter. Since we are dealing with ill-posed problems, we must accept the fact that there is no existence theorem. However, unlike 'standard' inverse problems (where one typically tries to do parameter estimation of the system based on noisy measurements), lack of existence is much more detrimental for our application, whereas lack of uniqueness is not. It does not trouble us that multiple designs all give the same result, whereas the oil prospector cares greatly where the rigs should be located, so a non-unique solution to their inverse problem is much more problematic. When the target mode we seek is not an exact eigenmode, we would like a technique to find a dielectric that supports an eigenmode that is closest to the target mode. As it stands, the entire spectrum contributes to the residual norm, and in a way that is non-optimal. Unfortunately, we have not found a more suitable metric that we can efficiently compute at this point. Until a better metric can be found that decouples the target from the rest of the spectrum, we have elucidated a fundamental limitation to arbitrary and optimal design of photonic structures using an inverse problem based method. Our insight although illustrated through a specific and limited model, turns out to be quite general. Our analysis reveals that the underlying physics fundamentally forbid certain modes from being physically realizable, regardless of human ingenuity. The problem of finding an optimal structure for an arbitrary application, even confirming its optimality, remains an outstanding question in the field of photonic design. Clearly, it is critical to check the validity and quantify the performance of designs obtained through inverse problem methods. We have also established that this general approach is not a turn-key method, but is potentially a very useful tool in instances where one has absolutely no intuition as to how to meet a particular design goal. While the results may not be optimal, they can spawn new geometries that serve as a starting point for other design methods for further fine tuning.

7.4.1 Comparing Tikhonov and Convex Optimization Regularization

Given the inadequacy of the residual norm, the advantage of the COR is not fully realized. We had hoped that non-existence of the solution could be overcome by finding the ‘next best’ solution that does exist. Currently, all COR does for you is tell you that the designed mode is not realizable based on the size of the residual norm. The solution it ends up giving you is not necessarily better or worse than the Tikhonov scheme. As we saw with the noisy h_1 problem, the original solution had a relatively large residual norm, and represents a better solution than some with smaller residual norms. The increased computational efforts make COR less attractive in some circumstances. However, COR is able to handle more sophisticated constraints, whereas Tikhonov cannot, so sometimes we may not have a choice. In addition, COR is less subjective than Tikhonov. Part of the Tikhonov procedure is looking at solutions at various regularization parameters and determining which one is ‘best.’ We already showed that the L-curve is an objective but unreliable method for finding the best regularization parameter. So that part of the scheme can seem fairly subjective. In addition, the Tikhonov solution will return non-physically realizable values, so the output will need to be fixed. There is some ‘slop’ inherent in the process, which we may frown upon somewhat, although the residual norm limitations show us that the entire design process is necessarily less rigorous than we had hoped, so we should not be as concerned about it. The COR on the other hand can be fully automated, because the output is guaranteed to be physically realizable. This is especially important with iterative schemes that require many iterations. Both methods will yield good but non-optimal designs, depending on how applicable the residual norm is. The better a metric the residual norm is, the greater the advantage with the COR. This is the reason why we will take finer steps with the COR iterative scheme than with the Tikhonov iterative scheme in dealing with our cQED design problems.

Chapter 8

Results

8.1 Waveguide Dispersion

We begin with the W1 waveguide, using a $9a \times 1a$ supercell geometry with a bulk hole radius of $0.3a$, and select 11 points along the $\Gamma - M$ propagation direction (labeled $\mathbf{q}$) and truncate after 19 Brillouin zones. This gives us 171 plane waves per $\mathbf{q}$-point for a total of 1881 plane waves. We formulate the Helmholtz operator as in eqn. (2.30) and solve to obtain the dispersion curve in the $\Gamma - M$ direction. We choose the curve with the anomalous dispersion characteristic, and as the primary objective[1] we seek to flatten the entire curve by 50% (see figure 8.1). As an additional objective in anticipation of mode matching with a high-Q small-mode volume cavity, we also require that the waveguide mode be more localized. The inverse problem is exactly as set up in chapter 3, and shown in eqn. (6.8). We once again use the defect-free lattice as $\eta_0(r)$, and since we do not have any additional constraints, we can use the simpler Tikhonov scheme. The required dispersion relation $\omega_i(q_i)$ is embedded into the set of β_i's. The target eigenmodes were constructed out of the initial eigenmodes by compressing $H(r)$ in the y-direction by 50% and padding the edges with zeros. This mode was then smoothed out by using the Fourier coefficients (numerically integrated) up to the truncation bandwidth and then renormalized. The resultant

[1] There is still no preliminary data on the specific form of the envelope function of a photon emission from an atom strongly coupled to a PBG-cavity, so at this point we do not know the precise form of the dispersion curve we will need. As our demonstration of arbitrary design, we chose to flatten the anomalous curve because both anomalous dispersion and the flattening of dispersion curves have greater relevance in the PCW community.

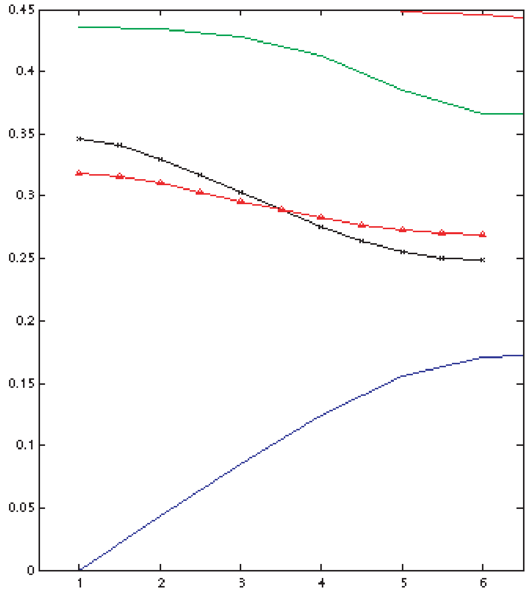


Figure 8.1: Dispersion relation for the W1 waveguide. The black x is the dispersion curve of the nominal W1 waveguide, while the red triangles show the desired curve.

target field is now simply the truncated fourier series using those coefficients.

8.2 Enlarged Defect Cavity

The cavity problem has the added complication of the fabrication constraint (eqn. (5.1)). This design problem cannot be solved using the Tikhonov regularization scheme, since a small norm does not exclude the undesired designs. Our investment in a more sophisticated regularization scheme allows us to incorporate these additional design constraints as part of the regularization procedure (provided that they are in the required convex form). We select a $5a \times 5a$ supercell geometry with a resolution of 12 points per lattice constant for a total of 3721 plane waves used. A smaller supercell was chosen because we did not want the algorithm to simply decrease the holes at the outer layers. We anticipate the algorithm may do that because the e-m field is negligible there (hence making the supercell approximation valid). The smaller geometry also allows us to perform calculations at a higher resolution so we can resolve

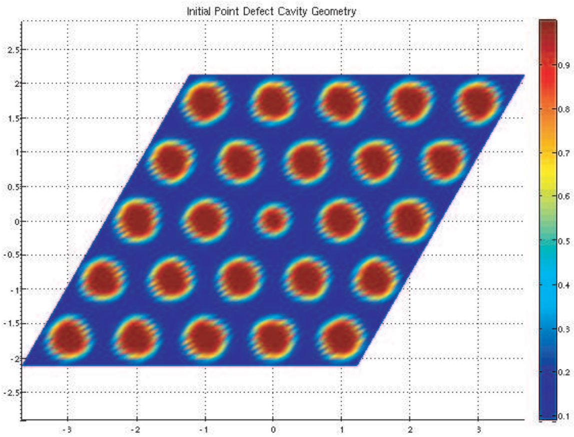


Figure 8.2: 5 × 5 h_1 point defect cavity starting geometry.

finer details in the inverted geometry. In particular, rather than strictly using the analytical expression for the Fourier coefficients, we use the FFT coefficients obtained for the discontinuous dielectric function and broaden it by convolution with a sharp Gaussian. Figure 8.3 shows the cross-section of the initial underlying dielectric function used. By defining the dielectric this way, we avoid Gibb's phenomenon with the Fourier series truncation, which is particularly important because the regularization scheme we have chosen bounds the maximum and minimum values of the dielectric to physically realizable values. Since we are testing a violation of the donor/acceptor mode rule, we also want to get accurate values of $\overline{\eta}_{\mathcal{D}}$ and $\overline{\eta}_{\mathcal{B}}$. We solve for the mode of a structure with nominal bulk hole size of $r_{bulk} = 0.4a$ and $r_{defect} = 0.3a$ (see figure 8.2)[2], and use that localized mode as the target mode.

In contrast to the waveguide dispersion problem, here we only care about the properties of the field, and not so much about the frequency. We can simply scale the 'lattice constant' of the structure to accommodate a real-world frequency that we would wish to use in a lab. We can therefore allow the frequency of the mode

[2] The actual 'hole radii' of the continuous function becomes smaller than these nominal values because of the convolution with the Gaussian. At this resolution, this was the largest nominal radius configuration we could accommodate without significant overshoot of the underlying dielectric function.

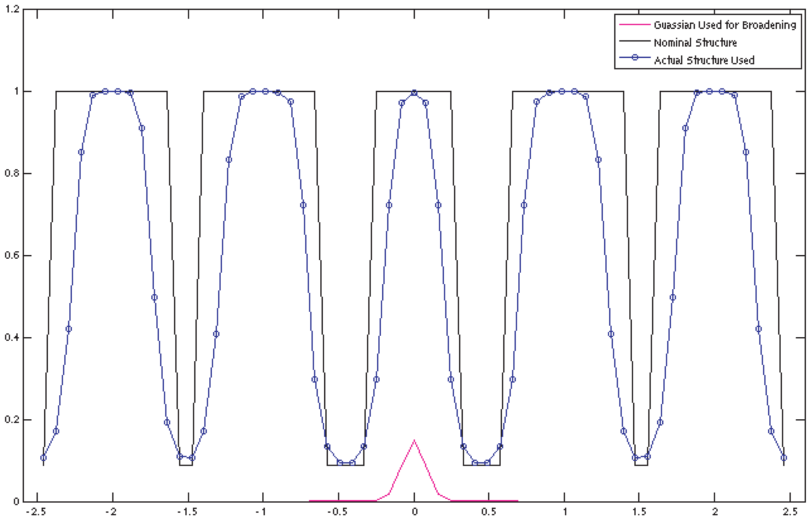


Figure 8.3: Cross-section view of dielectric function. The nominal structure and the actual structure is compared. Convolution with the Gaussian (pink curve) removes the overshoot due to truncation, but bandwidth limitations prevent an analysis of the nominal structure. Otherwise, the design goal and constraints are analogous.

to be another optimization variable, with the constraint that the frequency remain within the bandgap. This is another advantage to this scheme, as we no longer need to have a correct target frequency (as in [67]) in order to obtain the correct solution. To accommodate for the fabrication constraints and the variable frequency, the regularized problem of eqn. (6.9) has to be modified slightly:

$$\begin{aligned}
&\min_{\widetilde{\eta}} \left|\widetilde{A}\widetilde{\eta}\right|^2 \\
\text{subject to } &\eta_{min} \preceq \mathcal{F}^{-1}\eta \preceq \eta_{max} \\
&\eta^r_{\mathcal{D}} \geq \overline{\eta}_{\mathcal{D}} \\
&\eta^r_{\mathcal{B}} < \overline{\eta}_{\mathcal{D}} \\
&\frac{\omega^2_{min}}{c^2} \leq \frac{\omega^2}{c^2} \leq \frac{\omega^2_{max}}{c^2} \\
\text{where } &\widetilde{A} = [\ A \mid -b\ ] \quad , \\
&\widetilde{\eta} = \left[\ \eta^T \mid \frac{\omega^2}{c^2}\ \right]^T \quad , \\
&\eta^r_{\mathcal{D}} \equiv \frac{1}{\mathcal{A}_{\mathcal{D}}}\sum_{\mathcal{D}} \mathcal{F}^{-1}\eta \quad \text{, and} \\
&\eta^r_{\mathcal{B}} \equiv \frac{1}{\mathcal{A}_{\mathcal{B}}}\sum_{\mathcal{B}} \mathcal{F}^{-1}\eta
\end{aligned} \tag{8.1}$$

where the superscript T denotes the transpose. We choose $\eta_{max} = 1$ and $\eta_{min} = 0.0796$, and $\frac{\omega^2_{min}}{c^2} = 1$ and $\frac{\omega^2_{max}}{c^2} = 2.45$.

8.2.1 The direct solution

When we attempt to solve the two problems, we find as expected large residual norms, meaning that our target modes (and dispersion curve) are not supported. We can solve the forward problem using the solution to the inverse problem as our design. This gives a dispersion relation shown in figure 8.4, and a cavity mode that has only 72.9% overlap with the original target mode.

On the one hand, the results are quite encouraging, and in particular, the cavity mode is surprisingly good given our discussion about the acceptor/donor mode regime. On the other hand, the performance is not ‘optimal’ in the sense that we have not reached our target perfectly in either case. The question will always remain whether we have done the best we can do with a particular design, and that, as discussed in chapter 7, is difficult to overcome. However, we now show an approach that can be

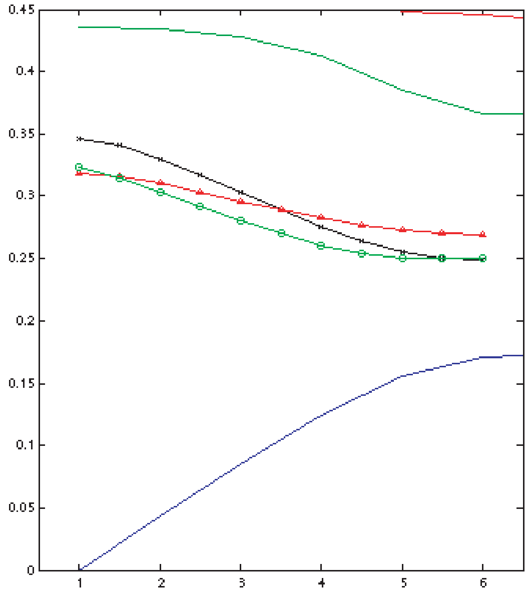


Figure 8.4: Initial result of the waveguide design. The green circles show the obtained dispersion curve, while the red triangles show the target curve. The black x show the original dispersion curve.

applied to these two problems that does yield even better results.

8.3 Iterative Approach

The method we will use to overcome these limitations involves an iterative approach. While we still cannot make any optimality claims, it does achieve the specified objective to an astonishingly high degree. We know that the root problem is the sparsity of the valid eigenmode landscape as discussed in section 7.3.1. The idea with the iterative approach is to use known valid eigenmodes as starting points of the inverse problem. We elaborate on the details as applied to each problem below.

8.3.1 Dispersion design

Given that our desired specifications are not physically realizable, it is time to make some design compromises. For the waveguide dispersion design problem, we relax the requirement that the eigenmodes be exactly as specified, and focus on the dispersion relation requirement (i.e. the eigenvalues). After the initial step, we solve the forward problem for the actual supported waveguide mode, and use that as the new target mode of our next inverse problem iterate. The target eigenvalues, however, remain as originally specified at each iteration. The iterative procedure proceeds as follows:

1. Use target modes and frequencies to formulate $\widetilde{A}$ and $\widetilde{b}$.

2. Solve the inverse problem. Pick a regularization parameter that gives a reasonable looking geometry.

3. Solve the forward problem using the interpreted result (as in section 6.5) from the solution of the inverse problem.

4. Look for the eigenmodes of interest. Check the frequencies of the waveguide modes. Continue if not satisfactory.

5. Use the new eigenmodes as the new target modes, but keep the original target frequencies.

6. Repeat until successive designs no longer improve.

Here, the motivation for the iterative approach is in obtaining a new starting field to define the updated inverse problem. This approach is inspired by the fact that we then start with an eigenmode that is known to be valid, even though we intend to perturb the eigenvalues. We know that the eigenmodes will be altered (if the inversion is successful), so that the target eigenmode-eigenvalue pair will likely be invalid still. However, without better insight, the prior valid eigenmode seems as good as any other. Using our strategy, at each iteration, the inverse problem tries to balance the

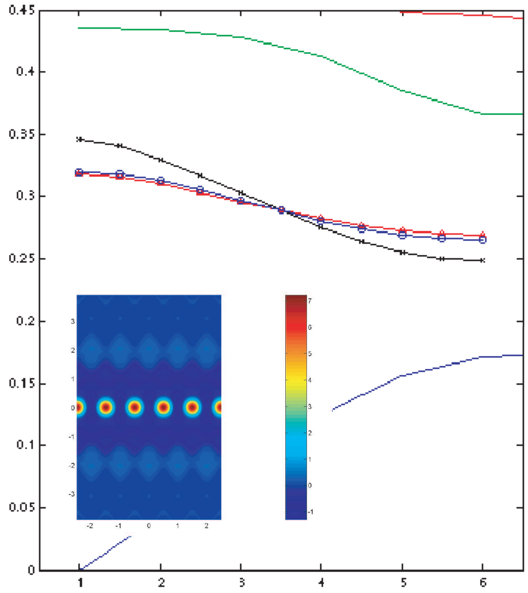


Figure 8.5: Final result of the waveguide design. The blue circles show the final dispersion curve, while the red triangles show the target curve. The black x show the original dispersion curve. The inset shows the waveguide mode still well localized.

demand to change frequencies with the need to accommodate the current eigenmodes. We let the result of the inverse problem tell us what the fields need to look like for the next iteration, since we have no way of knowing what they should look like anyway. The inverse problem 'machinery' itself will dictate how the dielectric needs to alter in order to shift the eigenvalues in the desired manner, and we will simply track the evolution of the eigenmode as it evolves. Each iteration brings us closer to the desired frequencies, and we accommodate the evolution of the fields by updating the target mode at each iteration.

The final waveguide design achieves a dispersion curve that compressed the original curve by 57% instead of the target of 50% (see figure 8.5). The field mode remains well localized, although we had not put any explicit restrictions on its form after the first iterate. This is not entirely surprising, since we know in general that small shifts in frequencies (eigenvalues) can be generated through small changes in the dielectric (as demanded by the Tikhonov regularization), which then bound the changes in the

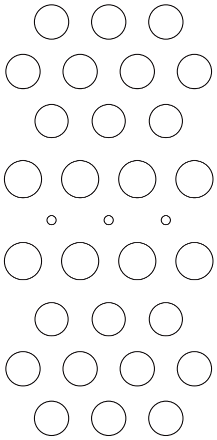


Figure 8.6: Final waveguide design for desired dispersion curve. See text for a description of the dimensions.

corresponding eigenmodes [78]. Figure 8.6 shows the final design. The missing central hole now reappears with a radius of $0.08a$, and the first neighboring row is displaced in the transverse direction towards the center by $0.15a$, with an increased hole radius to $0.325a$ from $0.3a$.

8.3.2 Cavity design

For the cavity design problem, we found that implementing the additional constraints (Eqn. (5.1)) in one step also takes us out of the valid eigenfunction space. However, unlike the waveguide design, there are no further constraints to relax. We have already allowed the frequency to take on any value within the bandgap. Once again, the problem is that the target mode (given the additional constraints) is not valid. It is clear that as we add more stringent constraints, the hypervolume of valid eigenmode space decreases. What was a valid mode becomes invalid as we lose access to the particular dielectric function. As we discussed in section 7.3.1, what we would like to do when our desired mode is invalid is to find a mode ‘nearby’ that is a valid one. We mentioned that the residual norm makes this a non-optimal process, but if we are ‘close enough’, then we increase our probability of finding the correct one. This is accomplished by iteratively ramping up to the final design constraint, rather than requiring in one step $\overline{\eta}_{\mathcal{D}} > \overline{\eta}_{\mathcal{B}}$. At each iteration, we decrease the maximum allowed value of $\overline{\eta}_{\mathcal{B}}$. We use two indicators to guide how large a stepsize we take. First is the magnitude of the residual norm, since we know when it is large that the inverse problem has deviated significantly from the eigenvalue problem. The second is by comparing the predicted eigenfrequency (as given by the solution to the inverse problem) to the actual eigenfrequency (as calculated by the solution to the subsequent forward problem). We reduce the step size accordingly. There is no proof of convergence to this adaptive scheme, but we have not found a finer mesh to produce better results with the extra iterations. The algorithm of our approach is outlined as follows:

1. Use target mode to formulate A and b.
2. Update constraint $\overline{\eta_{\mathcal{B}}} < \xi_i, \xi_i \equiv \xi_{i-1} - \delta$.
3. Solve the inverse problem. If the residual norm is greater than some threshold, decrease δ and backtrack.
4. Solve the forward problem using the result from the inverse problem.
5. Look for the eigenmode of interest. The frequency of the mode should be close to the frequency obtained by the solver. If the frequency error is greater than some threshold, decrease δ.
6. Use the new eigenmode as the new target mode. Record the overlap of the new eigenmode with the original eigenmode.
7. Repeat until design constraint satisfied.

Using the iterative approach, the 'next nearest neighbor' remains available, whereas in a single step approach, many neighbors simultaneously become invalid. The residual norm metric performs worse the farther away we need to go. Using this approach, we find a structure that supports a mode with a 93.6% overlap with the original mode while satisfying our donor mode regime constraint (see figure 8.7).

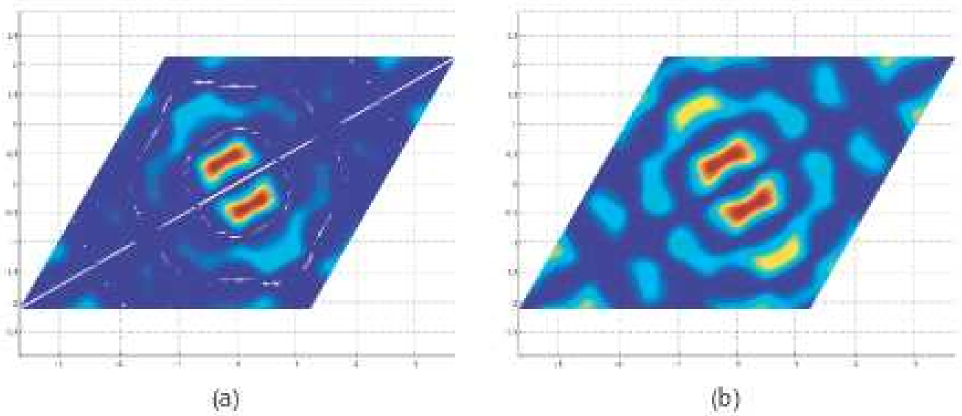


Figure 8.7: Magnetic field intensity of (a) original eigenmode and (b) eigenmode supported by final design.

Figure 8.8 shows the final design, and the first most noticeable feature is that the bulk holes now have a reduced value of η closer to 0.7, and in some regions it is down to 0.5. Additionally, the size of these features are also noticeably smaller. Six of the eight nearest holes from the center now take on a 'crescent moon' shape so it looks as though there is an effective lateral displacement for those holes, while the other two remain the same. For this structure, the final average bulk dielectric value $\overline{\eta_{\mathcal{B}}}$ is 0.1819, significantly reduced from the initial $\overline{\eta_{\mathcal{B}}}$ of 0.3962. The unaffected holes correspond to regions where the field intensity remains high, which is consistent with our intuition. The obtained eigenmode shows some leakage in the bulk area, but otherwise clearly resembles our target mode.

The most surprising feature of this iterative method is that the performance (as defined by the overlap integral of the obtained mode and the original target mode) was not a monotonically decreasing function (see figure 8.9). If we examine the algorithm closer, we see that after the first iteration, the original target mode no longer enters the picture, at least not directly. Whatever the inverse problem solution gives us, that becomes our next target mode. Granted, we use the mode that most resembles our target, but otherwise it does not play a role. Considering that we would have missed out on obtaining this solution if we had stopped a few iterations earlier, this

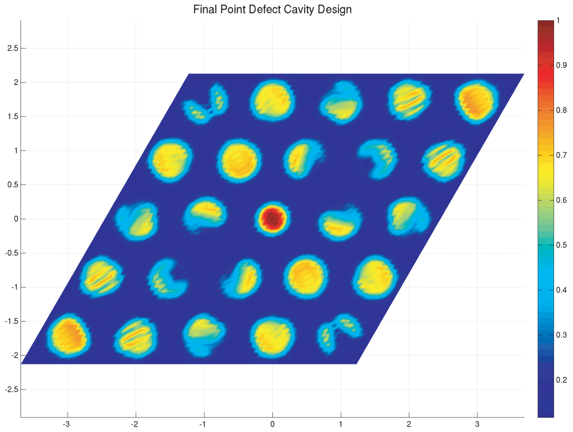


Figure 8.8: Final nanocavity design for acceptor eigenmode in donor configuration.

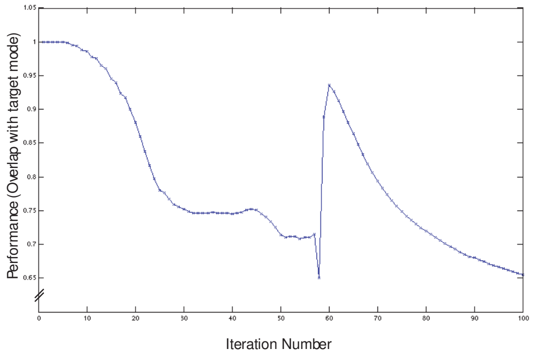


Figure 8.9: Plot of the performance at each intermediate iterate of the design process. Notice that this function is not monotonic, which also provides a glimpse into the topology of an objective function in design space.

result is somewhat troubling. However, it does highlight the difference between our methodology and the other inverse problem approaches that use gradient type schemes [76, 77], where there would have been no way of arriving at this design. It also reveals that in general, the performance function is not a simple functional of the dielectric, demonstrating real limitations of gradient based algorithms given the topology we found. This result further highlights the need to develop a better alternative to the residual norm as a metric for solving the inverse problem.

8.4 Concluding Remarks

We have set out to find a general turn-key methodology that allows arbitrary (and by extension optimal) design of desired PBG structures. Despite the limitations due to the convergence issues of the model, we retained the *ab initio* approach to elucidate fundamental and ubiquitous barriers to optimal PBG device design due to the topology of the space of valid eigenmodes. Having rigorously shown the non-existence of a solution, we attempted to look for designs that best approximates the desired functionality. We showed that the residual norm metric prevents a claim of finding the optimal 'next-best' design. While other inverse problem methods, particularly GA's and level set methods, can more readily yield designs that can be fabricated, they are confined to local improvements to existing design. Our approach, though currently non-optimal, can be used to provide these local methods with new starting points that lie 'closer' to the optimal design.

Even if the residual norm metric issue is worked out, our method still does not replace these other approaches because of the convergence issues, and the problem scales poorly with increased dimension. We were unable to take advantage of the ideas for fast convergence [4] because we do not know *a priori* where the boundary between high and low index materials will be in order to assign a tensorial value in that discretized element. However, we can make use of gradient based methods to go from an optimal continuous function design to the requisite binary valued design while minimizing sacrifice in performance.

Despite these general difficulties, we were able to extend our method to obtain remarkably good results that would be difficult to obtain with any other scheme. As we have shown, the topology of the problem is such that a gradient based algorithm would have missed our final design. The work we have developed here thus represents an important addition to the inverse problem based toolbox of design methods for the photonic problem.

Appendices

Appendix A

Sample Matlab Code to Illustrate Ill-Conditioning

```
%Simple Matlab code to illustrate ill-conditioning
%with a numerical example
%-------------------
%
%Set Parameters
dim = 5; %sets dimension of the matrix
numiter = 10000; %number of iterations to run through
noiselevel =1e-3; %scales the magnitude of the random noise vector
%
%--------------------
%
%Define the matrix and input vector
A=1./([1:dim]'*ones(1,dim)+ones(dim,1)*[0:1:dim-1]);
    %This defines the Hilbert matrix of size dimXdim
xin = ones(dim,1);
xout = A*xin;
%
%-----------------
%
```

```
%Statistics for the 'forward problem'
for kk = 1:numiter
    noisein = noiselevel*randn(dim,1);
    ein = norm(noisein)./norm(xin);
    eout = norm(A*(xin+noisein)-xout)./norm(xout);
    S(kk) = eout./ein;
end

figure(1);hist(S,50);
    %histogram of the stability of the forward problem
%
%Statistics for the 'inverse problem'
%----------------
%Inverse problem is : Solve xin = inv(A)*xout
%
invA = inv(A);
for kk = 1:numiter
    noisein = noiselevel*randn(dim,1);
    ein = norm(noisein)./norm(xout);
        %Note that the input and output are reversed...
    eout = norm(invA*(xout+noisein)-xin)./norm(xin);
    S(kk) = eout./ein;
end

figure(2);hist(S,50);
    %histogram of the stability of the inverse problem
xout_round = round(1e3*xout)./1e3;
xin_bad = invA*xout_round %Shows instability to rounding
```

Appendix B

Barrier Functions of Complex Variables

Consider the following barrier function $\phi(\xi) : \mathcal{C}^n \rightarrow \mathcal{R}$

$$\begin{aligned} \phi(\xi) &= -log(-f(\xi)) && \text{(B.1)} \\ &= -\log\left(-\frac{\xi^\dagger\eta + \eta^\dagger\xi}{2}\right) && \text{(B.2)} \end{aligned}$$

where $f : \mathcal{C}^n \rightarrow \mathcal{R}, \xi \in \mathcal{C}^n, \eta \in \mathcal{C}^n$. We use our $\mathcal{C}^n \rightarrow \mathcal{R}^{2n}$ transformation (eqn. (4.37)) and re-express in terms of the new variables.

$$\begin{aligned} [\xi] &\rightarrow \begin{bmatrix} \Re(\xi) \\ \Im(\xi) \end{bmatrix} \equiv x && \text{(B.3)} \\ [\eta] &\rightarrow \begin{bmatrix} \Re(\eta) \\ \Im(\eta) \end{bmatrix} \equiv y && \text{(B.4)} \end{aligned}$$

where $(x, y) \in \mathcal{R}^{2n}$. Transforming to $\mathcal{R}^{2n}$ and using eqn. (4.60) and (4.61), we find

$$\begin{aligned}
\phi(x) &= -\log(-g(x)) = -\log(-x^T y) && \text{(B.5)}\\
\nabla g(x) &= y && \text{(B.6)}\\
\nabla \phi(x) &= \frac{1}{x^T y} y && \text{(B.7)}\\
\nabla^2 \phi(x) &= \frac{1}{(x^T y)^2} \nabla g(x) \nabla g(x)^T && \text{(B.8)}\\
&= \frac{1}{(x^T y)^2} y y^T && \text{(B.9)}
\end{aligned}$$

Let $(x^T y)^{-2} \equiv \alpha \in \mathcal{R}$, since it is just a scalar. The Hessian can be expressed as:

$$\begin{aligned}
\nabla^2 \phi &= \alpha y y^T && \text{(B.10)}\\
&= \alpha \begin{bmatrix} \eta_r \\ \eta_i \end{bmatrix} [\eta_r^T \eta_i^T] && \text{(B.11)}\\
&= \alpha \begin{bmatrix} \eta_r \eta_r^T & \eta_r \eta_i^T \\ \eta_i \eta_r^T & \eta_i \eta_i^T \end{bmatrix} && \text{(B.12)}\\
&\neq \alpha \begin{bmatrix} \Re(H) & -\Im(H) \\ \Im(H) & \Re(H) \end{bmatrix} && \text{(B.13)}
\end{aligned}$$

In the final step, we show that this Hessian matrix is not equivalent to any $H \in \mathcal{C}^{n \times n}$ based on the $\mathcal{C}^n \rightarrow \mathcal{R}^{2n}$ transformation rules for matrices (eqn. (4.47)). Therefore, regardless of how we define the generalized complex gradient or Hessians, we cannot include this form of barrier function in the complex domain using normal matrix manipulation rules. We are forced to perform the cumbersome transformation to real variables, despite suggestions found elsewhere [83, 25].

Appendix C

Fourier Transforms

In chapter 2, we examined the Helmholtz equation in the plane wave basis, and found that transforming into the Fourier domain played a central role. We also alluded to some problems associated with the truncation to finite bandwidth. In this chapter, we look more closely at Fourier transforms, and in particular numerical implementations of the Fourier transform as it applies to physics applications.

Fourier analysis is included in most undergraduate level curriculum, so it may seem strange to have it included in the body of the thesis. The use of Fourier transforms, both the continuous and discrete forms are ubiquitous in physics and engineering. They are used extensively in signal processing, image processing and compression, pattern recognition, and solutions of PDEs using spectral methods (as we use them here). Most computational software such as *Matlab*, *Mathematica*, and *Maple* all have built-in functions to compute these transforms. However, perhaps because it is so commonplace, there is greater risk that one does not first think more carefully about the underlying physics and will just let the machine grind through the calculation. The only warning I remember as an undergrad in learning about Fourier analysis was 'aliasing', and we were simply told to make sure we sample above the Nyquist frequency. However, that is not a viable option with photonic crystals, because the discontinuities imply an infinite Nyquist frequency. What exactly happens to this 'aliasing' behavior then? It turns out that there are other issues with numerical implementations of the Fourier transform, particularly for those who wish to use discrete Fourier transforms in computational physics. The goal of this chapter is to bring awareness to some of these issues exacerbated by these discontinuities, and justify interpreting our results in the smooth function limit which we presented in Part II.

Organization

We will begin in section C.1 with some general bookkeeping by introducing the notation we will follow in this chapter for Fourier transforms, and then revisiting the Born von-Karman boundary conditions by paying closer attention to the topology of the direct space and Fourier space. In section C.2 we introduce formally the discrete Fourier transform, and in particular examine the fast Fourier transform (FFT) and some of its properties. There are some aspects of the FFT that are unnatural to physics applications (where the origin is usually at the center of the spectrum). This will lead to a discussion of symmetries in section C.3, where we see real discrepancies between what we naïvely think we are modeling, as opposed to what the mathematics is actually modeling. We also find a fundamental conflict in the symmetry of the physics and the symmetry as a result of the discretization. In light of the issues revealed in these sections, we emphasize the fundamental importance of considering the continuous function limit and explore the behavior for various transform schemes. As mentioned in section 2.4, we describe Li's insight into the convergence issues in the plane wave Helmholtz equation with Fourier factorization rules in section C.4, followed by concluding remarks in section C.5.

C.1 Boundary Conditions and Fourier Space

Consider a continuous function $f(\mathbf{r})$. We define the Fourier transform of the function as

$$F(\mathbf{k}) = \frac{1}{V}\int_V f(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}d\mathbf{r} \tag{C.1}$$

where V denotes the relevant integration volume. We can similarly define an inverse Fourier transform as

$$f(\mathbf{r}) = \int_{V_k} F(\mathbf{k}) e^{+i\mathbf{k}\cdot\mathbf{r}} d\mathbf{k} \tag{C.2}$$

There are many conventions for where to place the normalization factor $\frac{1}{V}$, and physicists tend to prefer the symmetrized form, but we have chosen to follow the convention [3] used in the treatment of the photonic bands problem. In the absence of any periodicity that extends to infinity, the Fourier transform of an arbitrary real-space function will be a continuous function in Fourier space. When we impose the Born von-Karman periodic boundary conditions, as we saw in section 2.2, we are left with a discrete set of allowed $\mathbf{k}$-vectors $\{\mathbf{k}\}$. More precisely, we have $F(\mathbf{k}') \equiv 0$ for any $\mathbf{k}' \notin \{\mathbf{k}\}$. We mentioned that imposing periodic boundary conditions is a standard approximation, and is considered valid in the tight-binding approximation. We explicitly examine the difference with the following example in 1D. Consider the rectangular function

$$f_1(x) = \begin{cases} A & \text{if } |x| \leq a, \\ 0 & \text{if } |x| > a. \end{cases} \tag{C.3}$$

The Fourier transform of $f_1(x)$ is:

$$\begin{aligned} F_1(k) &= \int_{-\infty}^{+\infty} f_1(x) e^{-ikx} dx && \text{(C.4)} \\ &= 2A \frac{\sin(ak)}{k} && \text{(C.5)} \end{aligned}$$

where we have ignored the normalization in eqn. (C.1) to facilitate comparison with the periodic case.

If we now introduce an artificial periodicity by defining a supercell of length $10 \times 2a$,

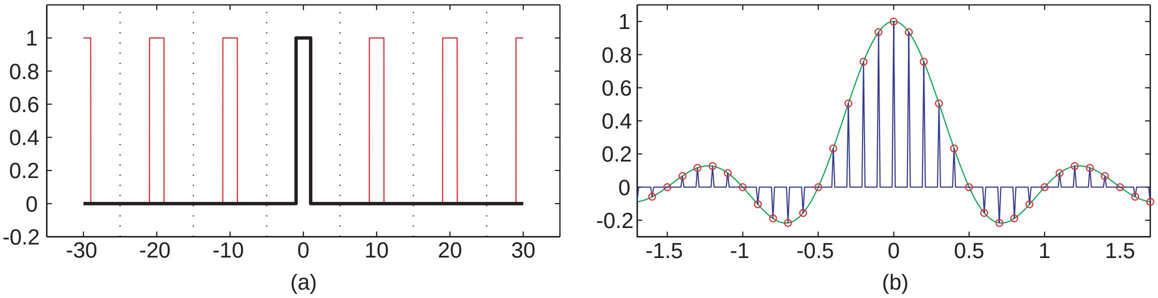


Figure C.1: In (a) we show $f_1(x)$ in black, and $f_2(x)$ with the artificial periodic boundary conditions in red. The dashed line helps visualize the periodicity of f_2. In (b), we plot the Fourier transform of f_1 and f_2. F_1 is plotted in green and the allowed k values of F_2 are shown as red circles. We plot in the blue curve the underlying continuous form of F_2 in k space. The 'delta function train' as a result of the real-space periodicity is shown explicitly this way.

we can define

$$f_2(x) = \begin{cases} A & \text{if } |x| \leq a, \\ 0 & \text{if } a < |x| < 5a. \end{cases} \tag{C.6}$$

$$f_2(x + Nx_p) = f_2(x) \tag{C.7}$$

where N is an integer, and $x_p \equiv 10a$ is the periodicity introduced. The Fourier transform of this function has the same form as $F_1(k)$ with the exception that it is modulated with a delta function 'comb' corresponding to the allowed k values. In figure C.1, we plot the two functions in real space and Fourier space.

The ratio of the two normalization factors we omitted accounts for the extra features in $f_2(x)$ that are absent in the original function $f_1(x)$. However, for applications such as spectral analysis, this is unimportant since it is the relative magnitude of the different frequency components that matter. Quantities involving continuous k-space integrals are well approximated by Riemann sums. We see in figure C.1 that we do not lose information by incorporating the artificial boundary condition. Particularly for functions like $f_1(x)$ that are negligible outside of the supercell (in our case the function is identically zero), we see that this is a very good approximation. One of our Fourier transform pair becomes a discrete sum rather than an integral:

$$\begin{aligned} F(\mathbf{k}) &= \frac{1}{V_{sc}} \int_{V_{sc}} f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} && \text{(C.8)} \\ f(\mathbf{r}) &= \sum_{\mathbf{k}} F(\mathbf{k}) e^{+i\mathbf{k}\cdot\mathbf{r}} \Delta\mathbf{k} && \text{(C.9)} \end{aligned}$$

where V_{sc} indicates the volume of the supercell.

C.2 Numerical Implementation

With arbitrary functions, we often cannot perform the integration analytically. Numerical implementation of the Fourier transform takes the form of a discrete Fourier transform (DFT), and one of the most common algorithms for its implementation is the fast Fourier transform (FFT). In a FFT, the real space function is sampled at regular intervals, and the number of plane waves is truncated to allow numerical evaluation of the transform. The number of points used in the real-space sampling N matches the number of plane waves.

With the identification from $f(x)$ to f_x, the FFT relations are defined to be:

$$\begin{aligned} F_k &= \sum_{x=0}^{N-1} f_x \exp\left(\frac{-2\pi i}{N} kx\right), \text{ Forward FFT} \\ f_x &= \frac{1}{N} \sum_{k=0}^{N-1} F_k \exp\left(\frac{2\pi i}{N} kx\right), \text{ Inverse FFT} \end{aligned} \qquad \text{(C.10)}$$

where k and x are now integer values from 0 to $N-1$ in units of their respective basis vectors. In general 3D form, let $\{\hat{s_1}, \hat{s_2}, \hat{s_3}\}$ define the periodicity of the lattice. The unit vectors in real-space becomes $\{\hat{x_1}, \hat{x_2}, \hat{x_3}\} \equiv \{\hat{s_1}/N_1, \hat{s_2}/N_2, \hat{s_3}/N_3\}$ where N_1, N_2, N_3 are the number of sampling points along each direction. The basis vectors in **k**-space become:

$$\begin{aligned}
\mathbf{b}_1 &= 2\pi \frac{\mathbf{s}_2 \times \mathbf{s}_3}{\mathbf{s}_1 \cdot (\mathbf{s}_2 \times \mathbf{s}_3)} && \text{(C.11)} \\
\mathbf{b}_2 &= 2\pi \frac{\mathbf{s}_3 \times \mathbf{s}_1}{\mathbf{s}_2 \cdot (\mathbf{s}_3 \times \mathbf{s}_1)} && \text{(C.12)} \\
\mathbf{b}_3 &= 2\pi \frac{\mathbf{s}_1 \times \mathbf{s}_2}{\mathbf{s}_3 \cdot (\mathbf{s}_1 \times \mathbf{s}_2)} && \text{(C.13)}
\end{aligned}$$

Observe as the real-space periodicity grows ($|s_i| \rightarrow \infty$), the resolution increases in **k**-space (i.e. $|b_i| \rightarrow 0$). The high frequency cutoff is precisely at Nyquist when we use the same number of points in both **k**-space and real-space.

One reason for its popularity is that the computational complexity for the FFT scales as $O(N \log N)$ due to the Cooley-Tukey algorithm [84], compared to $O(N^2)$ for a direct evaluation of the sum. In creating the MPB package, Steven Johnson et al. have also put out a free package [85] for the FFT called FFTW (Fastest Fourier Transform in the West).

There are some notable properties about the FFT we should highlight. First, the normalization factor is absorbed in the inverse transform, which is different than the convention we have adopted in the continuous case. This unfortunate discrepancy means we have to be more careful with our bookkeeping of the normalization factors, but otherwise poses no problems. The other property which usually affects physicists is that the FFT convention defines the domain of x and k such that the 0 value is the first element rather than at the center of the spectrum (see eqn. (C.10)).

In physics however, the origin is usually defined at the center of most problems that we analyze to more easily exploit symmetries. Computing the FFT with any standard software package using discretized data in 'physics order' will give incorrect results. In figure C.2 we show graphically what is effectively the conventional FFT ordering. The correct function $f(x)$ one needs to use in order to get what physicists think they should get has the 'negative' half of the data translated by the periodicity imposed. This translation of the spectrum in both k-space and real space is simply a substitution of $x \rightarrow x + N$ and $k \rightarrow k + N$ respectively. In eqn. (C.10), this term

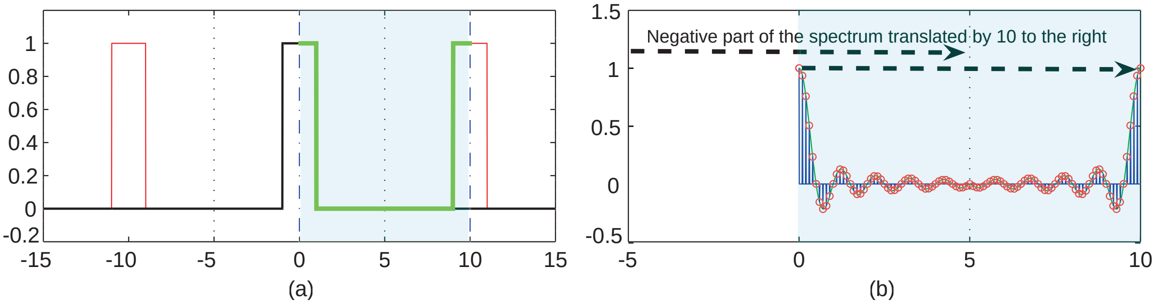


Figure C.2: In (a) we again show $f_1(x)$ in black, and $f_2(x)$ in red. The blue dash-dotted line defines the new domain (shaded in turquoise) of x under the FFT convention. To use a standard FFT routine, we need to use the green curve rather than the black curve to get the right results. On the right, (b) shows the equivalent k-space domain for the Fourier transform. The $k < 0$ part of the spectrum is translated over as shown. The turquoise region shows k-space domain under the FFT convention.

appears in the exponential and is evaluated to 1, so we see that the resulting FFT is left unchanged by this translation. To be completely transparent, this means that for a given discretized f_x, if we attempt to compute the FFT coefficient F_{k+N}, we get identically F_k. The generalization to 2D is illustrated graphically in figure C.3.

By defining the computational grid in FFT order rather than in physics order, we can take advantage of the FFT algorithm without shifting the indices around. In matlab, there are built-in functions that do this called *fftshift*, which go from FFT order to physics order, and *ifftshift*, which go from physics order to FFT order. If the FFT is used extensively, it is more convenient to simply define the grid in the FFT order, and shift only when plotting the results.

This shift property is reminiscent of the concept of the reduced zone scheme of **k**-space representation in solid state physics (see Chapter 9 in [12]) for periodic potentials. However, we will see in the following section that the FFT symmetry is artificial and one needs to be very careful how one treats these terms in the context of computational physics.

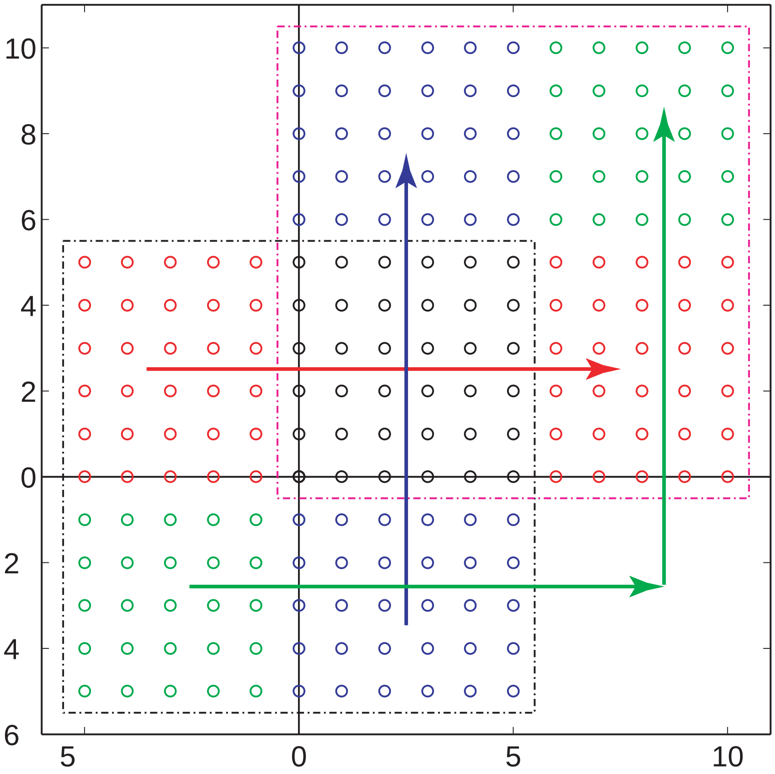


Figure C.3: The four quadrants of a 2D FFT are shifted from physics ordering to FFT ordering as shown. The black dash-dotted line indicates the usual physics domain, while the magenta dash-dotted line indicates the nominal FFT domain.

C.3 The Symmetry Problem

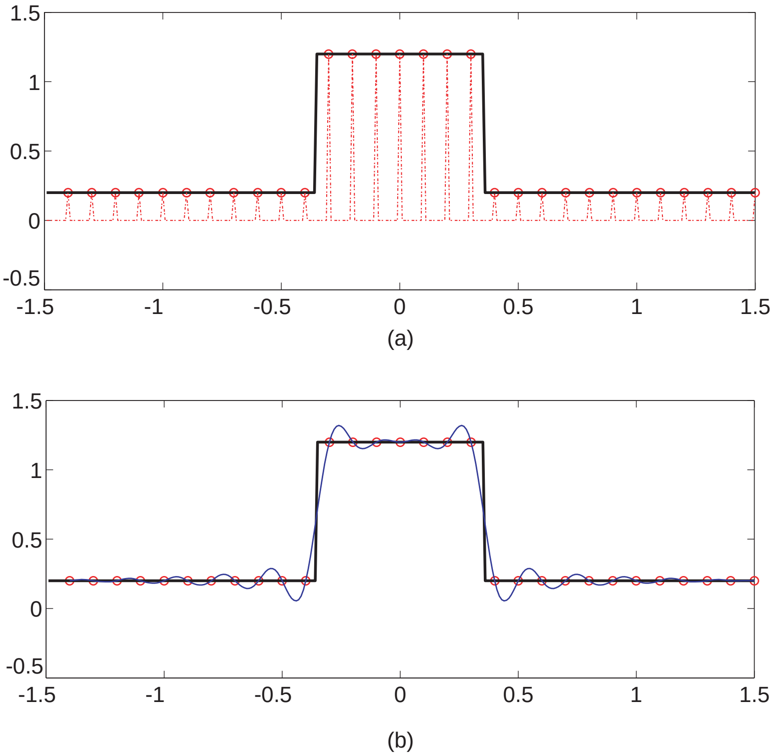
Suppose we define a 1D computational grid with $N = 2n + 1$ points such that $\{k\} = \{-n, -n+1, -n+2, \ldots, n-2, n-1, n\}$, and position $\{x\} = \{-n, -n+1, -n+2, \ldots, n-2, n-1, n\}$. For simplicity, consider the Helmholtz operator (eqn. (2.28)) in 1D, and notice the terms of the form $\eta_{k-k'}$. This has the form of a Toeplitz matrix (i.e. matrices of the form $A_{m,n} = A_{m-n}$), which has a natural connection mathematically with the FFT [86]. The Toeplitz matrix has elements like η_κ, where $n < |\kappa| \leq 2n$, exceeding our defined k-point domain. For simple geometries, one could use the expression for the analytical Fourier transform. For more complicated geometries, we might be tempted to use the FFT symmetry and equate $\eta_\kappa = \eta_{\kappa \pm n}$ (depending on the sign of κ), since, as we saw earlier, they are formally equivalent mathematically. This would give us the circulant form of the Toeplitz matrix. However, we presumably have truncated the grid at the chosen size because the omitted high frequency components are ‘small enough’, whereas these η_κ terms can be quite large. So how do we resolve the discrepancy on how to handle these Fourier coefficients that lie outside our computational domain?

C.3.1 The underlying real-space function

The key is in recognizing that the real-space function is fundamental because it is the one that corresponds to a physical quantity. With regards to the FFT symmetry, recall that the valid but artificial periodicity in the real-space function gave rise to the delta-function lattice in **k**-space. We justified the approximation by invoking the tight-binding approximation. If we now insist on an artificial periodicity in **k**-space, then necessarily (by the symmetric nature of the Fourier transform) we enforce a convolution of any real-space function with a delta-function lattice as well. This means that our model of the continuous function is identically zero everywhere except the points we happen to be sampling (see figure C.4a).

This illustrates that we must not use the FFT symmetry to determine the correct coefficients for terms outside of our computational domain. We could (as we might with the analytical Fourier coefficients) do an oversampled transform with $N \times 2^{n_D}$, where n_D is the number of dimensions in real-space to evaluate those coefficients. The problem with this approach is that it is no longer self-consistent with our truncation condition. When we chose the computational domain, for better or for worse, we effectively set *a priori* any Fourier component exceeding $k_{\max}$ equal to zero. Since other quantities have the same bandwidth limitation, self-consistent treatment implies setting those ϵ_k equal to zero as well.

Given the analysis here, it is now obvious that the underlying real-space function for some set of FFT coefficients can be quite different than what we think they represent. Given our strict truncation approximation, we now examine the 1D rectangular function again. This time, we take an N point FFT of the rectangular function, then transform back to real-space with a $2N$ point FFT by explicitly zero padding the



(a)

(b)

Figure C.4: (a) The underlying continuous function of a nominal rectangular function allowing the FFT symmetry. (b) The underlying continuous function of a nominal rectangular function reconstructed from the N FFT coefficients with hard truncation.

k-components outside the original bandwidth. Figure C.4b shows the true representation of an FFT. Even though the FFT/IFFT pair seems to perfectly reconstruct a discontinuous jump, an examination of the underlying continuous function shows the misrepresentative sampling that actually happens. Note also that after the hard truncation we no longer have a Toeplitz matrix because of our FFT preferred ordering, which is different from the other works that use the PWE method and keep $\eta_{\mathbf{k}-\mathbf{k}'}$ (or the equivalent $\frac{1}{\epsilon}$ term) as a Toeplitz matrix[11, 14, 87, 88, 20].

C.3.2 Even vs. odd

A final problem with the FFT symmetry that is quite subtle shows up in how we choose to discretize the grid in $\mathbf{k}$-space. As before, the spacing is strictly determined by the real-space periodic BCs, while $\mathbf{k}_{\max}$ is chosen according to some truncation condition. The final detail addresses the boundary of the **k**-space grid, meaning a determination of whether an even or an odd number of **k**-points are used along each **k** direction. It is known that the FFT algorithm is most efficient when N is a power of 2 or a product of small prime numbers [85]. Usual implementations choose N to be a power of 2, which is clearly even. Here, I argue that an odd-numbered grid is the correct choice for self-consistency considerations, particularly when we have high frequency components we have had to truncate out.

Consider now in 1D an N_o point transform where $N_o = 2n+1$ and an N_e transform where $N_e = 2n$ of the same periodic continuous function in real-space. Shifting back into a physics preferred coordinate system, our k-space will have in the odd case $\{k_o\} = \{-n, -n+1, \ldots, n-1, n\}$. In the even case, we have a choice of either $\{k_e\} = \{-n, -n+1, \ldots, n-2, n-1\}$ or $\{k_e\} = \{-n+1, -n+2, \ldots, n-1, n\}$, and the two are equivalent because of the FFT symmetry, i.e. any $F_{-n} = F_{N-n}$. Consider further a real-valued real-space function (such as a dielectric function). The continuous Fourier transform symmetry in **k**-space is:

$$F_{\mathbf{k}} = F^*_{-\mathbf{k}} \tag{C.14}$$

This illustrates that the physical symmetry conflicts with the FFT symmetry which arises as a result of discretizing. We do not have this conflict when we choose an odd numbered FFT, since F_n and F_{-n} are independent. With an even-numbered FFT of a real-valued function, this constrains the boundary **k** elements to take on only real values. In 1D, this may not be significant, since there is only one element. In 2D, the number of boundary elements increase, and we show in figure C.5 the conflict in symmetry.

For well-behaved functions where our sampling rate is well above Nyquist, this conflict is not significant, simply because F_n is near zero. If we do not sample at a sufficient rate (e.g. when we have discontinuities), the conflict becomes much more significant. In figure C.6 we show an arbitrary discontinuous function in 2D that can be a possible PBG dielectric function. We take the 2D FFT using an $N \times N$

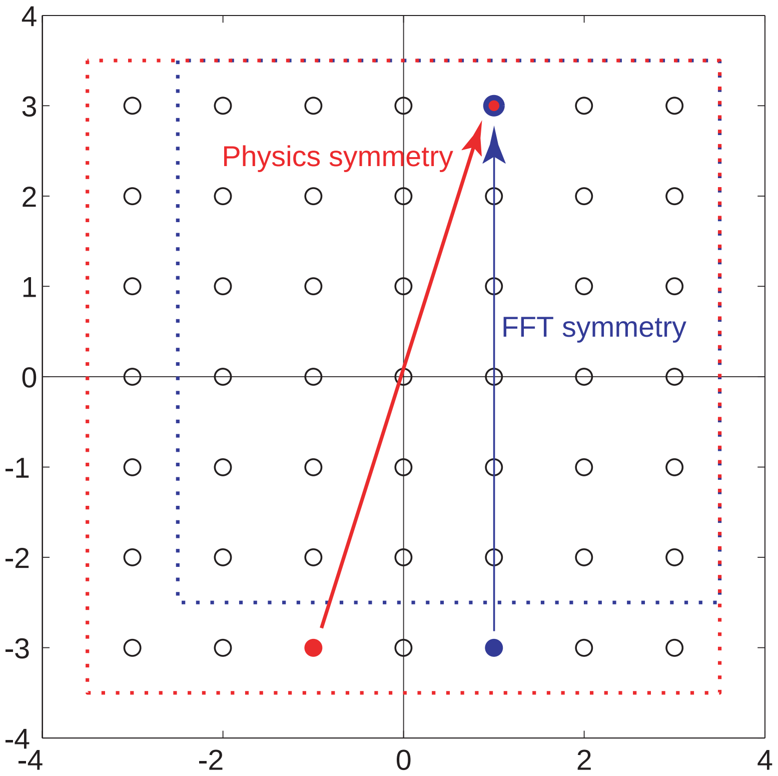


Figure C.5: 2D **k**-space diagram showing the effect of the FFT symmetry on Fourier coefficients at the boundaries

grid for N even and N odd. We show the resulting spectra of coefficients in figure C.7. Our discretization choice affects not only the boundary values, but as shown in the plot, there is a significant discrepancy between the two schemes in even the largest of the coefficients (i.e. where the **k**-vector is close to the origin, far away from the truncation limit). Since the odd numbered grid preserves the proper number of degrees of freedom and symmetries, we consider it the discretization scheme that is actually more appropriate for computational physics, especially when modeling discontinuities.

C.4 Fourier Factorization

A final remark we will make about Fourier transforms as it applies to the PBG problem deals with the Fourier coefficients of a product of two functions. Consider a

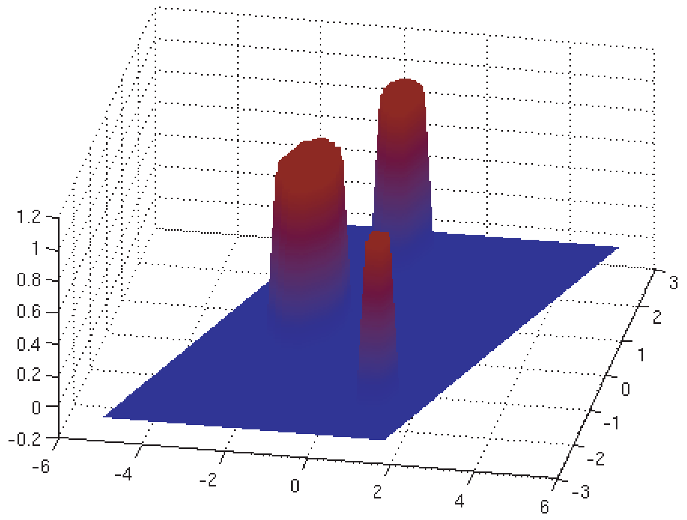


Figure C.6: A 2D real-space function with discontinuities that is representative of an arbitrary PBG dielectric function.

function $f(x) = g(x) \cdot h(x)$. The problem is to find the Fourier coefficients F_k, given G_k and H_k, the Fourier coefficients of $g(x)$ and $h(x)$ respectively. This can be done using *Laurent's Rule* such that:

$$F_n = \sum_{m=k_{\min}}^{k_{\max}} G_{n-m} H_m \tag{C.15}$$

where G_{n-m} is the familiar Toeplitz matrix. However, suppose $g(x)$ and $h(x)$ are functions with *concurrent* discontinuities at x_d, and suppose further that the discontinuities at x_d are *complementary* such that $f(x)$ is continuous at x_d, i.e.

$$\lim_{x \to x_d^+} f(x) = \lim_{x \to x_d^-} f(x) = f(x_d) \tag{C.16}$$

We encounter this situation with our source free non-magnetic geometry. The magnetic fields are continuous everywhere, which implies (by eqn. (2.3)) that $\mathbf{D}$ is continuous. Since $\epsilon(\mathbf{r})$ has discontinuities, $\mathbf{E}(\mathbf{r})$ must have complementary and concurrent discontinuities such that their product is continuous.

Li proved [18] that even as we take the set $\{\mathbf{k}\}$ to infinity, Laurent's rule will never

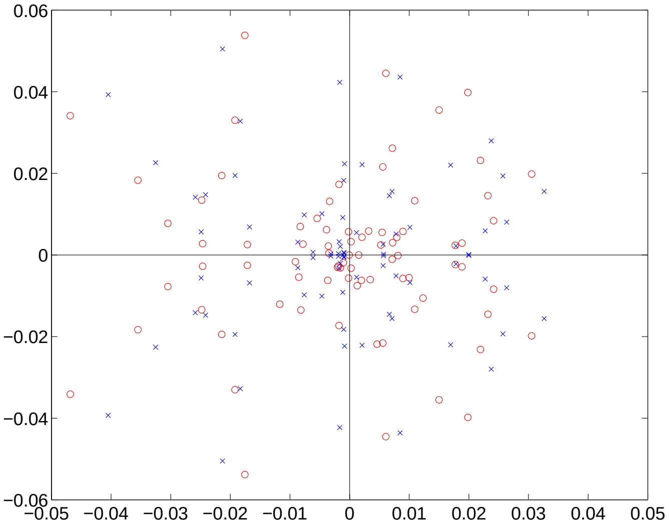


Figure C.7: The value of the Fourier coefficients for an even numbered FFT vs. an odd numbered FFT are plotted on the complex plane, i.e. the axes are the real and imaginary parts of F_k of the function shown in figure C.6. Notice the discrepancy in most of the points. We have plotted the 81 **k**-points with the largest magnitude. Note that the bandwidth for both geometries are the except for the boundary point, showing the two schemes are not self-consistent.

converge. Instead, he applied what is now known as the *Inverse Rule*:

$$F_n = \sum_{m=k_{\text{min}}}^{k_{\text{max}}} \left[\Gamma_{n-m}\right]^{-1} H_m \tag{C.17}$$

where Γ_k is the Fourier expansion of $\gamma(x) = \frac{1}{g(x)}$. The superscript $^{-1}$ denotes matrix inversion. The difference in convergence is shown in figure C.8.

Given Li's analysis, it is no longer surprising why the convergence of the PWE method depends on the polarization of the field and how we treat the dielectric Toeplitz matrix. Rigorous application of his factorization rules to 2D and 3D has spawned the fast Fourier factorization methods (cf. [87, 20]), but the physics is ultimately embodied by the effective medium approach [17, 4].

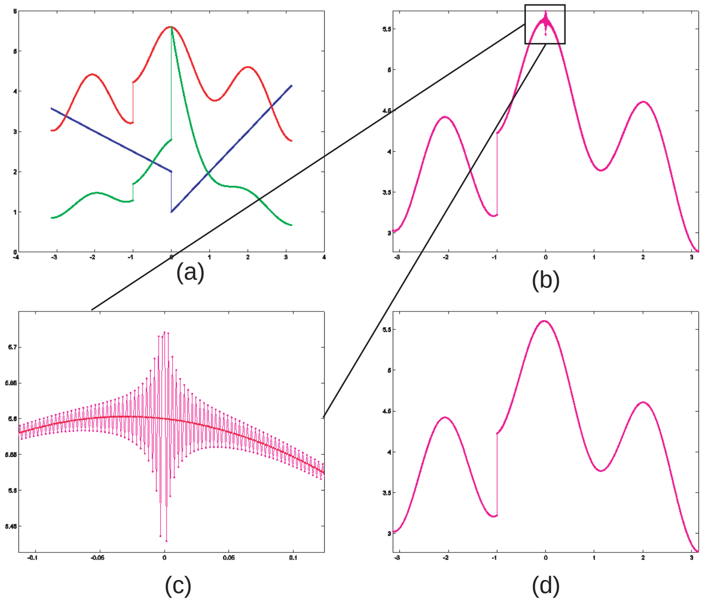


Figure C.8: Illustration of Laurent's rule and the inverse rule. (a) Given the Fourier coefficients of two functions $g(x)$ in blue and $h(x)$ in green, we want to find the product $f(x)$ in red. Notice a single concurrent and complementary discontinuity at $x = 0$, and a concurrent but non-complimentary discontinuity at $x = -1$. (b)Result using Laurent's rule. Notice the non-concurrent discontinuity had no trouble converging. (c) Magnification of problem area (d)Good convergence of result using Li's inverse rule.

C.5 Conclusion

In this appendix, we explored some of the intricacies of the numerical implementation of Fourier transforms. For functions with discontinuities, our inherent inability to satisfy the Nyquist criterion exacerbates the issues discussed in this chapter. For these reasons, we cautioned against blind acceptance of the FFT/IFFT output, but instead return to a continuous function limit interpretation. Given the accuracy limitations revealed in chapter 2 and this appendix, self-consistency in modeling should be the primary focus. We therefore give up on the notion of accurately modeling dielectric slabs with etched air holes. In chapters 5–8 where we describe geometries of structures, we use language such as a *nominal* geometry consisting of a lattice of air holes of radius $0.3a$ and so forth. It is to be understood that we are really talking about the real-space continuous function representation of the truncated Fourier series.

Appendix D

Detailed Errata for Geremia 2002

In this appendix, I will give a detailed account of the various flaws and errors in the original photonic inverse problem paper [67] published in Physical Review E in 2002. I feel strongly that these errors should be documented somewhere, and of course to the extent possible, I have corrected these in the work I have done subsequent to this paper. Unfortunately, as I was not an author in the original paper, I did not feel it was my place to publish an errata, nor did Dr. Geremia seem motivated to do so when I approached him with some of the initial errors. Nevertheless, if and when the decision is made to do so, this appendix provides a sufficiently detailed account that should be more than adequate for that purpose. Unless otherwise specified in this appendix, reference to equation and figure numbers are meant to be for those in Geremia 2002, while equation numbers beginning with the letter D refer to equations in this appendix.

D.1 Introduction

D.1.1 Relevant abstract of Geremia 2002

In Geremia 2002, optimal photonic crystal cavity design results were presented using an analytical 2D model, as well as a numerical 3D model. To date, there has not been further investigation on the 3D work, but the 2D results have been examined extensively. To summarize the work in Geremia 2002, the 2D analytical approach can be separated into two distinct and separate steps. The first step is an optimization of the desired mode without consideration as to the dielectric that can generate the mode. The particular optimization that it claims to do is a maximization of pseudo-Q factor (Q) and electric field at the origin($E(0)$), and a minimization of mode volume (V). Quantitatively, this means a maximization of $\beta_Q Q + \beta_E E(0) - \beta_V V$, where the β's balance the importance of the various terms. Without loss of generality, β_Q is taken to be 1. The second step extracts the dielectric required to produce the optimized mode from step 1 by solving the inverse Helmholtz equation. The inverse Helmholtz equation is derived in the bulk mode basis into a set of linear equations. A 'radially symmetric' defect and an asymmetric defect design were provided as illustrations to the technique.

There are various errors or omissions in the description of the first step, the optimization of the desired mode. We will examine these in section D.2. The more critical error in the paper is in the derivation of the inverse Helmholtz equation which we will address in section D.3, and of course the most notable omission is the discussion of regularization. Before we can even address these errors though, it turns out there are some typographical errors in the equations. In order to make a comparison between the corrected version of the inverse Helmholtz equation and the ones in Geremia 2002, we first need to correct the typos. We will first assume that the derivation is correct and fix the typos in section D.1.2.

D.1.2 Typographical errors

The typos appear in equation 25 in the main text, and then A2, A3 and A4 in the appendix. Again, actual flaws in the reasoning will be addressed later. Following the reasoning outlined in the paper, we write out explicitly all the steps in full. We include it all here because the derivation is extremely cumbersome, particularly keeping track of all the indices. Of course, since the inverse problem is unstable, if you don't know about regularization, then even the correct equations will give bad answers, so at a practical level when trying to code up the equations, it was rather difficult trying to catch these errors. Also keep in mind that these equations will ultimately be proven incorrect. The reader not interested in the grunge here can safely skip to section D.1.3 for the summary of the typos on page 144.

Detailed Steps of PRE Derivation

Recall the defect mode is expanded in the TE bulk mode basis

$$\mathbf{H}_m(\mathbf{r}) = \frac{1}{N}\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)}\mathbf{H}_{n,\mathbf{k}}(\mathbf{r}) \tag{D.1}$$

where the bulk modes are computed in the plane wave expansion method:

$$\begin{aligned} \mathbf{H}_{n,\mathbf{k}}(\mathbf{r}) &= \mathbf{z}\sum_{\mathbf{G}} h_{n,\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} && \text{(D.2)} \\ \int_{V_N} \mathbf{H}^*_{n',\mathbf{k}'}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} &= N\delta_{n,n'}\sum_{\mathbf{G}}\delta_{\mathbf{k}',\mathbf{k}+\mathbf{G}} && \text{(D.3)} \end{aligned}$$

The $\mathbf{z}$ vector will be omitted in the following notation for compactness unless required for completeness under curl operations. We start with Maxwell's equation, but separate out the defect dielectric from the unperturbed lattice:

$$\begin{aligned} \nabla\times\eta_0(\mathbf{r})\nabla\times\mathbf{H}_m(\mathbf{r}) + \nabla\times\delta\eta(\mathbf{r})\nabla\times\mathbf{H}_m(\mathbf{r}) &= \frac{\omega_m^2}{c^2}\mathbf{H}_m(\mathbf{r}) && \text{(D.4)} \\ \nabla\times\eta_0(\mathbf{r})\nabla\times\mathbf{H}_m(\mathbf{r}) - \frac{\omega_m^2}{c^2}\mathbf{H}_m(\mathbf{r}) &= -\nabla\times\delta\eta(\mathbf{r})\nabla\times\mathbf{H}_m(\mathbf{r}) && \text{(D.5)} \end{aligned}$$

Next, substitute into equation (D.5) the defect and bulk mode expansion of equations (D.1) and (D.2), left multiply by $\mathbf{H}_{n'',\mathbf{k}''}$ and integrate over the size of the supercell. The left hand side (LHS) of equation (D.5) becomes

$$\frac{1}{N}\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)}\frac{\omega_{n,\mathbf{k}}^2 - \omega_m^2}{c^2}\int \mathbf{H}^*_{n'',\mathbf{k}''}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r}$$

Evaluating the right hand side (RHS) of equation (D.5) requires a (truncated) Fourier expansion of the defect dielectric.

$$\delta\eta(\mathbf{r}) \equiv \sum_{\mathbf{k}} \delta\eta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}$$

The **k** points used in the summation are the ones consistent with the specified Born-von Karman boundary conditions (i.e. geometry of the supercell), and the series is truncated with a finite number of reciprocal lattice vectors used to tile the reciprocal superlattice, identical to the truncation in calculating the band structure of the bulk lattice.

The RHS now becomes:

$$\begin{aligned}
& -\nabla \times \left(\sum_{\mathbf{k}'} \delta\eta_{\mathbf{k}'} e^{i\mathbf{k}'\cdot\mathbf{r}} \right) \nabla \times \frac{1}{N} \sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}} \mathbf{H}_{n,\mathbf{k}}(\mathbf{r}) \\
= \quad & -\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'} \nabla \times e^{i\mathbf{k}'\cdot\mathbf{r}} \left(\nabla \times \mathbf{z} \sum_{\mathbf{G}} h_{n,\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \right) \\
= \quad & -\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'} h_{n,\mathbf{k}+\mathbf{G}} \nabla \times e^{i\mathbf{k}'\cdot\mathbf{r}} \left(i(\mathbf{k}+\mathbf{G}) \times \mathbf{z} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \right) \\
= \quad & -\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'} h_{n,\mathbf{k}+\mathbf{G}} \nabla \times e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \left(i(\mathbf{k}+\mathbf{G}) \times \mathbf{z} \right) \\
= \quad & -\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'} h_{n,\mathbf{k}+\mathbf{G}} [i(\mathbf{k}'+\mathbf{k}+\mathbf{G})] \times [i(\mathbf{k}+\mathbf{G}) \times \mathbf{z}] e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \\
= \quad & -\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'} h_{n,\mathbf{k}+\mathbf{G}} (\mathbf{k}'+\mathbf{k}+\mathbf{G}) \cdot (\mathbf{k}+\mathbf{G}) e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \mathbf{z}
\end{aligned}$$

Next we left multiply by plane wave expansion of $\mathbf{H}_{n'',\mathbf{k}''}$ and integrate over the size of the supercell again. (Omitting again the $\mathbf{z}$ vector for compactness)

$$\begin{aligned}
& -\int \sum_{\mathbf{G}''} h^*_{n'',\mathbf{k}''+\mathbf{G}''} e^{-i(\mathbf{k}''+\mathbf{G}'')\cdot\mathbf{r}} \times \cdots \\
& \left(\frac{1}{N} \sum_{\mathbf{k}'} \sum_{n,\mathbf{k}} \sum_{\mathbf{G}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'}(\mathbf{k}'+\mathbf{k}+\mathbf{G}) \cdot (\mathbf{k}+\mathbf{G}) h_{n,\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \right) d\mathbf{r} \\
= & -\int \sum_{\mathbf{G}''} \frac{1}{N} \sum_{\mathbf{G},\mathbf{k}',n,\mathbf{k}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'}(\mathbf{k}'+\mathbf{k}+\mathbf{G}) \cdot (\mathbf{k}+\mathbf{G}) \times \cdots \\
& \qquad h^*_{n'',\mathbf{k}''+\mathbf{G}''} e^{-i(\mathbf{k}''+\mathbf{G}'')\cdot\mathbf{r}} h_{n,\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} d\mathbf{r} \\
= & -\frac{1}{N} \sum_{\mathbf{G}'',\mathbf{G},\mathbf{k}',n,\mathbf{k}} a_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'}(\mathbf{k}'+\mathbf{k}+\mathbf{G}) \cdot (\mathbf{k}+\mathbf{G}) \times \cdots \\
& \qquad h^*_{n'',\mathbf{k}''+\mathbf{G}''} h_{n,\mathbf{k}+\mathbf{G}} \int e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}} d\mathbf{r} \\
= & -\frac{1}{N} \sum_{\mathbf{G}'',\mathbf{G},\mathbf{k}',n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}} \delta\eta_{\mathbf{k}'}(\mathbf{k}'+\mathbf{k}+\mathbf{G}) \cdot (\mathbf{k}+\mathbf{G}) \times \cdots \\
& \qquad h^*_{n'',\mathbf{k}''+\mathbf{G}''} h_{n,\mathbf{k}+\mathbf{G}} \int e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}} d\mathbf{r}
\end{aligned}$$

We have now arranged Maxwell's equation with defect dielectric into equation (A3) of [67]. To get to equation (A4), we simply 'evaluate' the integrals and collapse the appropriate delta functions. The LHS of equation (A3) is :

$$\begin{aligned}
& \frac{1}{N} \sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}} \frac{\omega^2_{n,\mathbf{k}} - \omega^2_m}{c^2} \int \mathbf{H}^*_{n'',\mathbf{k}''}(\mathbf{r}) \mathbf{H}_{n,\mathbf{k}}(\mathbf{r}) d\mathbf{r} \\
= & \sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}} \frac{\omega^2_{n,\mathbf{k}} - \omega^2_m}{c^2} \delta_{n'',n} \sum_{\mathbf{G}} \delta_{\mathbf{k}'',\mathbf{k}+\mathbf{G}} \text{ , using equation (D.3)} \\
= & \sum_{\mathbf{k}} a^{(m)}_{n'',\mathbf{k}} \frac{\omega^2_{n'',\mathbf{k}} - \omega^2_m}{c^2} \sum_{\mathbf{G}} \delta_{\mathbf{k}'',\mathbf{k}+\mathbf{G}}
\end{aligned}$$

We note that by Bloch's theorem, $\omega_{n,\mathbf{k}} = \omega_{n,\mathbf{k}+\mathbf{G}}$, so we can in this instance substitute $\sum_{\mathbf{G}} \delta_{\mathbf{k}'',\mathbf{k}+\mathbf{G}}$ with $N\delta_{k'',k}$. Finally, we obtain the expression

$$N a^{(m)}_{n'',\mathbf{k}''} \frac{\omega^2_{n'',\mathbf{k}''} - \omega^2_m}{c^2}$$

The RHS of equation (A4) simply involves collapsing the integral into a delta function. Only note here is that since we are integrating over the supercell, we pick up a factor of N.

$$\int_V e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}} d\mathbf{r} = N\delta_{\mathbf{G},\mathbf{k}''+\mathbf{G}''-\mathbf{k}-\mathbf{k}'}$$

The RHS of equation (A3) after collapsing this delta function on $\mathbf{G} = \mathbf{k}'' + \mathbf{G}'' - \mathbf{k} - \mathbf{k}'$ is

$$\begin{aligned}
&-\frac{1}{N}\sum_{\mathbf{G},\mathbf{G}''}\sum_{\mathbf{k}'}\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)}\delta\eta_{\mathbf{k}'}(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot(\mathbf{k}+\mathbf{G})\times\ldots \\
&\qquad h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}+\mathbf{G}}\int e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}}d\mathbf{r} \\
= &-\sum_{\mathbf{G}''}\sum_{\mathbf{k}'}\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)}\delta\eta_{\mathbf{k}'}(\mathbf{k}''+\mathbf{G}''-\mathbf{k}')\cdot(\mathbf{k}''+\mathbf{G}'')\times h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}''+\mathbf{G}''-\mathbf{k}'} \\
= &-\sum_{\mathbf{G}''}\sum_{\mathbf{k}'}\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)}h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}''+\mathbf{G}''-\mathbf{k}'}(\mathbf{k}''+\mathbf{G}''-\mathbf{k}')\cdot(\mathbf{k}''+\mathbf{G}'')\delta\eta_{\mathbf{k}'}
\end{aligned}$$

Multiply the collapsed versions of equation (A3) by $-\frac{1}{N}$ to recover equation (A4). Next we note as in equation (A5) that we can fold summations over $\mathbf{k}$ back into the First Brillouin zone using the identity:

$$\sum_{n,\mathbf{k}} a_{n,\mathbf{k}}^{(m)} = N\sum_{n,\mathbf{q}} a_{n,\mathbf{q}}^{(m)}$$

Next we make the following index transformations: $n \to n'$, $\mathbf{G}'' \to \mathbf{G}$, $n'' \to n$. We also collapse $\mathbf{k}$ to $\mathbf{q}'$ using the above identity, and rename $\mathbf{k}'$ to $\mathbf{k}$. Finally we restrict the indices $\mathbf{k}''$ to within the first Brillouin zone, and rename it $\mathbf{q}$, since they are not summed.

The LHS of equation (A4) is:

$$\begin{aligned}
& \frac{1}{N}\sum_{\mathbf{G}''}\sum_{\mathbf{k}'}\sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}} h^{*}_{n'',\mathbf{k}''+\mathbf{G}''} h_{n,\mathbf{k}''+\mathbf{G}''-\mathbf{k}'} (\mathbf{k}''+\mathbf{G}''-\mathbf{k}')\cdot(\mathbf{k}''+\mathbf{G}'')\delta\eta_{\mathbf{k}'} \\
= & \sum_{\mathbf{G}}\sum_{\mathbf{k}}\sum_{n',\mathbf{q}'} a^{(m)}_{n',\mathbf{q}'} h^{*}_{n,\mathbf{q}+\mathbf{G}} h_{n',\mathbf{q}+\mathbf{G}-\mathbf{k}} (\mathbf{q}+\mathbf{G}-\mathbf{k})\cdot(\mathbf{q}+\mathbf{G})\delta\eta_{\mathbf{k}} \\
= & \sum_{\mathbf{k}}\left\{\sum_{\mathbf{G}}\sum_{n',\mathbf{q}'} a^{(m)}_{n',\mathbf{q}'} h^{*}_{n,\mathbf{q}+\mathbf{G}} h_{n',\mathbf{q}+\mathbf{G}-\mathbf{k}} (\mathbf{q}+\mathbf{G}-\mathbf{k})\cdot(\mathbf{q}+\mathbf{G})\right\}\delta\eta_{\mathbf{k}} \\
\equiv & \sum_{\mathbf{k}} D^{(m)}_{n,\mathbf{q};\mathbf{k}}\delta\eta_{\mathbf{k}}
\end{aligned}$$

Finally the RHS of equation (A4) is:

$$\begin{aligned}
& -a^{(m)}_{n'',\mathbf{k}''}\frac{\omega^2_{n'',\mathbf{k}''}-\omega^2_m}{c^2} \\
= & \ a^{(m)}_{n,\mathbf{q}}\frac{\omega^2_m-\omega^2_{n,\mathbf{q}}}{c^2}
\end{aligned}$$

This gives us equation (24) as desired, with the inversion matrix D given by equation (25).

D.1.3 Proposed typographical errata

Equation 25 should read:

$$D^{(m)}_{n,\mathbf{q};\mathbf{k}} = \sum_{n'}\sum_{\mathbf{G},\mathbf{q}'} a^{(m)}_{n'\mathbf{q}'} h^{*}_{n,\mathbf{q}+\mathbf{G}} h_{n',\mathbf{q}+\mathbf{G}-\mathbf{k}} \times (\mathbf{q}+\mathbf{G})\cdot(\mathbf{q}+\mathbf{G}-\mathbf{k})$$

rather than

$$D^{(m)}_{n,\mathbf{q};\mathbf{k}} = \sum_{n'}\sum_{\mathbf{G},\mathbf{q}'} a^{*}_{n'\mathbf{q}'} h^{*}_{n,\mathbf{q}+\mathbf{G}} h_{n',\mathbf{q}'+\mathbf{G}-\mathbf{k}'} \times (\mathbf{q}+\mathbf{G})\cdot(\mathbf{q}+\mathbf{G}-\mathbf{k}')$$

Equation A2 should read:

$$\int_{V_N} \mathbf{H}^*_{n',\mathbf{k}'}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} = N\delta_{n',n}\sum_{\mathbf{G}}\delta_{\mathbf{k}',\mathbf{k}+\mathbf{G}}$$

rather than

$$\int_{V_N} \mathbf{H}^*_{n',\mathbf{k}'}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} = \delta_{n',n}\sum_{\mathbf{G}}\delta_{\mathbf{k}',\mathbf{k}+\mathbf{G}}$$

Equation A3 should read:

$$\begin{aligned}
& \frac{1}{N}\sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}}\frac{\omega^2_{n,\mathbf{k}}-\omega^2_m}{c^2}\int \mathbf{H}^*_{n'',\mathbf{k}''}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} \\
= \quad & -\frac{1}{N}\sum_{\mathbf{k}'}\sum_{n,\mathbf{k}}\sum_{\mathbf{G},\mathbf{G}''} a^{(m)}_{n,\mathbf{k}}\delta\eta_{\mathbf{k}'}h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}+\mathbf{G}}\cdots \\
& \times(\mathbf{k}'+\mathbf{k}+\mathbf{G})\cdot(\mathbf{k}+\mathbf{G})\int e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}}d\mathbf{r}
\end{aligned}$$

rather than

$$\begin{aligned}
& \frac{1}{N}\sum_{n,\mathbf{k}} a^{(m)}_{n,\mathbf{k}}\frac{\omega^2_{n,\mathbf{k}}-\omega^2_m}{c^2}\int \mathbf{H}^*_{n'',\mathbf{k}''}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} \\
= \quad & \frac{1}{N}\sum_{\mathbf{k}}\sum_{n,\mathbf{k}}\sum_{\mathbf{G},\mathbf{G}''} a^{(m)}_{n,\mathbf{k}}\delta\eta_{\mathbf{k}}h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}+\mathbf{G}}\cdots \\
& \times(\mathbf{k}''+\mathbf{G}'')\cdot(\mathbf{k}''+\mathbf{G}''-\mathbf{k}')\int e^{i(\mathbf{k}'+\mathbf{k}+\mathbf{G}-\mathbf{k}''-\mathbf{G}'')\cdot\mathbf{r}}d\mathbf{r}
\end{aligned}$$

Equation A4 should read:

$$\begin{aligned}
& \frac{1}{N}\sum_{n,\mathbf{k}}\sum_{\mathbf{k}'}\sum_{\mathbf{G}''} a^{(m)}_{n,\mathbf{k}}\delta\eta_{\mathbf{k}'}h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{k}''+\mathbf{G}''-\mathbf{k}'}(\mathbf{k}''+\mathbf{G}'')\cdot(\mathbf{k}''+\mathbf{G}''-\mathbf{k}') \\
= \quad & -a^{(m)}_{n'',\mathbf{k}''}\frac{\omega^2_{n'',\mathbf{k}''}-\omega^2_m}{c^2}
\end{aligned}$$

rather than

$$\frac{1}{N}\sum_{n,\mathbf{k}}\sum_{\mathbf{k}'}\sum_{\mathbf{G}''} a^{(m)}_{n,\mathbf{k}}\delta\eta_{\mathbf{k}'}h^*_{n'',\mathbf{k}''+\mathbf{G}''}h_{n,\mathbf{G}''-\mathbf{k}'}(\mathbf{k}''+\mathbf{G}'')\cdot(\mathbf{k}''+\mathbf{G}''-\mathbf{k}')$$
$$= \quad -\frac{1}{N}a^{(m)}_{n'',\mathbf{k}''}\frac{\omega^2_{n'',\mathbf{k}''}-\omega^2_m}{c^2}$$

D.2 Mode Optimization Errors

The three physical quantities that were optimized were maximizing **E** field intensity at the central defect location, minimizing the mode volume, and maximizing the Q factor. In the paper, the optimization is performed using Lagrange multipliers. Equation (23) from the paper states:

$$\sum_{n',\mathbf{q}'}\left[\frac{\omega_{n',\mathbf{q}'}\omega_{n,\mathbf{q}}}{qq'}+\beta_I H^*_{n',\mathbf{q}'}(0)H_{n,\mathbf{q}}(0)-\beta_V\langle\psi_{n',\mathbf{q}'}|\psi_{n,\mathbf{q}}\rangle\right]a^{(m)}_{n',\mathbf{q}'}=\Lambda a_{n,\mathbf{q}}$$

The first term maximizes Q by removing bulk mode contribution above the lightline, the second term maximizes the intensity at the origin, and the final term minimizes the mode volume. As it appears in the paper, it is unclear how it actually optimizes any of the three quantities. We will address each expression in the next sections.

D.2.1 Q factor

The Q factor optimization incorrectly applies the lightline model to mimic the behavior of the cavity Q. The light cone rule of thumb originates from the idea of lossy modes coupling into free space. The relevant traits that describe modes lying above the light line (lossy modes) is that they have high frequencies and small in-plane fourier components. For a defect mode of a given frequency, correct application of this rule of thumb is to restrict fourier components lying within a small circle (sphere in 3D) in k-space determined by the mode frequency. This is the approach consistent with Oskar Painter's group [9]. So there are two errors in using the frequency of the associated bulk mode and its quasi k-vector in determining where it lies in relation to the light line. The first point is that the actual frequency of the defect free bulk mode is irrelevant, since we are trying to determine the lossiness of the *defect mode*. Hence only the *defect mode frequency* will give the relevant coupling condition. Secondly, it is the fourier components that compose the mode that is relevant, and not the quasi k-vector. A bulk mode in fact is composed of the fourier components of the k-vector in the first Brillouin zone as well as all others by addition of reciprocal lattice vectors. In fact, many of the upper band bulk modes actually have large fourier components, and are minimally lossy (i.e. minimal, though finite contribution from small fourier components). A reduced zone scheme picture is actually rather deceptive in trying to visualize the light line. On the other hand, the bulk modes originating from the low lying bands (of which most of the 'optimized' modes are composed) actually have a significant amount of small fourier components. Therefore, if one were to use these low lying bulk modes to form a defect mode, since the defect frequency is in the band gap, these contributions are in fact highly lossy due to the expanded region of unfavorable k-space. In misapplying the light line constraint, the optimization should make things worse by incorporating more of these low lying modes. Finally, this analysis ignores the fact that coherent superposition of lossy bulk modes can lead to cancellation of small (i.e. leaky) fourier components. Going back to our linear algebra language, the bulk modes are a non-diagonalized basis for the purpose of Q factor considerations. The point is moot when only using 5 layers of photonic crystal as in [67], because there are so few small fourier components for such a small supercell (although note the observation in section D.4 about figures 3 and 5).

D.2.2 E field intensity

In the paper, it is unclear whether it is the **H** field or the **E** field that actually gets optimized. The equations in the paper throughout only show expressions for the **H** field (cf. eqns. (5,21,23)), whereas the plots in the results section show the **E** field. One can obtain the **E** field by taking the curl of **H**, but because TE polarization is assumed, the **E** field lies in the xy-plane (as opposed to the H field which only has z components), it must be treated as a vector quantity. The expression for the **E** optimization is not discussed in the paper. One way to treat this is to allow for vectored expansion coefficients for the optimized **E**, but then it is not immediately obvious that the Lagrange multipliers method is applicable. Convex optimization cannot be applied here either because the norm of $\mathbf{E(0)}$ is convex, and we cannot maximize a convex function using convex optimization. In fact, the local extremum of a norm must be a local minimum, so the global maximum must be some end point. In any case, equations (23) in the paper is an incorrect expression for a cavity mode optimization, since maximizing the **H** field intensity will yield a zero **E** field at the origin.

D.2.3 Mode volume

There are also problems with the expression for the mode volume optimization in the paper. Again, there is the ambiguity over whether it is the **H** field or **E** field that is optimized (cf. eqns. (4,19,20,23)). Even if we were to assume that it is sufficient to minimize the mode volume of **H** to get the desired result, the expressions are still incorrect. In the paper, ψ refers to a max 1 normalized mode function that satisfies

$$\mathbf{H}_j(\mathbf{r}) = \mathbf{H}_{0,j}\psi_j(\mathbf{r}),$$

where j is some mode index, and $|\psi(\mathbf{r})|_{\max} = 1$. The mode volume for the mode j is then

$$V_j = \int |\psi_j(\mathbf{r})|^2 d\mathbf{r}$$

The expression so far is fine. However, directly evaluating the mode volume by an expansion as the paper attempts to do in equation (19)

$$V_m = \sum_{n',\mathbf{q}'}\sum_{n,\mathbf{q}} a^{(m)*}_{n',\mathbf{q}'} a^{(m)}_{n,\mathbf{q}} \langle \psi_{n',\mathbf{q}'}(\mathbf{r}) | \psi_{n,\mathbf{q}}(\mathbf{r}) \rangle + c.c.$$

is in fact not so straightforward, and unfortunately incorrect. The corresponding normalization constant $\mathbf{H}_{0,m}$ for the mode m cannot be determined without first doing the summation. Using equation (14) as the expansion for the target mode implies:

$$\begin{aligned}\mathbf{H}_{\mathbf{m}}(\mathbf{r}) &= \sum_{n}\sum_{\mathbf{q}\in BZ} a_{n,\mathbf{q}}^{(m)}\mathbf{H}_{n,\mathbf{q}}(\mathbf{r}) && \text{(D.6)}\\ &= \mathbf{H}_{\mathbf{0},\mathbf{m}}(\mathbf{r})\psi_{m}(\mathbf{r}) && \text{(D.7)}\end{aligned}$$

Eqn. (D.7) explicitly shows equation (19) is incorrect because

$$\psi_m(\mathbf{r}) \neq \sum_{n}\sum_{\mathbf{q}\in BZ} a_{n,\mathbf{q}}^{(m)}\psi_{n,\mathbf{q}}(\mathbf{r}) \qquad \text{(D.8)}$$

A simple example that illustrates the error is to choose any basis that has all of its basis functions max 1 normalized, such as the plane wave basis. In such a basis, the RHS of eqn. (D.8) is identically $\mathbf{H}_{\mathbf{m}}(\mathbf{r})$, which is clearly not max 1 normlized in general (hence it cannot equal $\psi_m(\mathbf{r})$ by definition). All equation (19) represents is in fact $\int |\mathbf{H}'_m(\mathbf{r})|^2 d\mathbf{r}$, for some unnormalized mode $\mathbf{H}'_m$, uncorrelated with the actual mode volume. The physical meaning of this term as written in the paper is unclear, but it does skew the optimized mode with some completely undesired effect.

As it turns out, formulating the mode volume minimization is actually unnecessary in this case. By imposing an energy constraint to the optimization, a maximized E field will necessarily have minimum mode volume. If we think about this carefully, it is clear since the point of minimizing mode volume is to maximize the electric field strength per photon. An energy constraint can be imposed by normalizing the expansion coefficients. Obviously in the E field maximization step, this is already done, otherwise by scaling the coefficients, we can increase $\mathbf{E}(\mathbf{0})$ without bound. In other applications where one cannot avoid dealing with the mode volume directly, it would be prudent to exercise more caution in its treatment.

A final comment about the optimization procedure concerns the weighting factors (β's). At the conclusion of Section III-C, it talks about optimizing over the β_i coefficients, and the idea presented is to nest the inverse equation (equation (24)) within an outer β_i optimization loop. However, a performance metric for the β_i's is not provided (not even a qualitative comment on what makes a set of β better than another set.) If it is just to maximize the $\mathcal{I}$ function (i.e. the total objective function shown in equation (8)) over all combinations of β's, then that seems to defeat the purpose of having those coefficients in the first place. These coefficients enforce the fact that there are tradeoffs between the different properties. $\mathcal{I}$ can be made large if one sets $\beta_V = 0$, effectively removing the requirement for having small mode volume. We simply could not ascertain as to what exactly the algorithm was optimizing. Section III-E and III-F explain that a conjugate-gradient algorithm is used for the optimization with the inverse problem nested within it. No explanations were presented as to why this nesting was necessary, nor how it would be accomplished. For example, how is the solution of the inverse problem (eqn. (24)) used in the subsequent iteration? One would expect that it involves evaluating the gradient of the matrix D with respect to β somehow, but eqn. (24) depends on β implicitly through the mode expansion coefficients $a_{n,q}$. Eqn. (23) shows us that the $a_{n,q}$'s are eigenvectors to an eigenvalue problem parameterized in part by the β's. The study of the rotation of eigenvectors due to perturbations to its operator is not a simple matter [78], so it is not obvious or clear how one might evaluate the gradient.

D.3 Inversion Errors

The second step of the design algorithm involved inverting the Helmholtz equation using the bulk modes as a basis in the paper. From the discussion of the Q factor optimization in section D.2, it is clear that the bulk mode basis is not a natural basis for this problem. As we saw in chapter 6, inverting the equation in the plane wave basis (the natural basis for Q factor type considerations) actually gives a very simple and concise result. Treatment in the bulk mode basis was unnecessarily messy and convoluted, further complicated by the decision to sum over multiple Brillouin zones. Consequently, the derivation gave rise to terms that include $\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})$. The term $\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})$ is actually rather awkward to interpret, since the purpose of the band index is to reduce all $\mathbf{k}$-vectors into the first Brillouin zone. It is ambiguous what is really meant to run the $\mathbf{k}$-vectors through the entire $\mathbf{k}$-space and still have a band index. Ignoring the strange notation and continuing to expand the bulk mode shows that the assumed orthogonality relation in equation (A2)

$$\int_{V_N} \mathbf{H}^*_{n',\mathbf{k}'}(\mathbf{r})\mathbf{H}_{n,\mathbf{k}}(\mathbf{r})d\mathbf{r} = \delta_{n,n'} \sum_{\mathbf{G}} \delta_{\mathbf{k}',\mathbf{k}+\mathbf{G}}$$

is invalid and therefore misused in the derivation of the inversion equations. A casual examination of this relation would lead us to believe that it appears correct, save for the missing normalization factor of N for repeating the summation over multiple Brillouin zones (see corrected equation (A2) on page 145). However, if we really try to rigorously write down what this notation is supposed to mean, the subtle errors turn out to be quite significant. I will use an example to illustrate.

First, let us revisit the concept of the band index and the $\mathbf{q}$'s in the first Brillouin zone. The number of bands is equivalent to the number of Brillouin zones (BZ) in $\mathbf{k}$-space, which is the same as the number $\mathbf{G}$-vectors we keep. For a truncated $\mathbf{k}$-space, this implies that the total number of $\mathbf{k}$-points in the computational domain is equal to $N_G \times N_q$, where N_G is the number of $\mathbf{G}$-vectors and N_q is the number of wavevectors in the first BZ. Therefore, the label $\{n, q\}$ can be mapped to k by the relation $\mathbf{k}_{n,q} = \mathbf{q} + \mathbf{G}_n$. We explicitly label $\mathbf{k}$ with the subscripts to help us designate where it comes from. This implies that the term

$$\begin{aligned} \mathbf{H}_{n,\mathbf{k}} &= \mathbf{H}_{n,\mathbf{k}_{n',q}} \\ &= \mathbf{H}_{n,\mathbf{q}+\mathbf{G}_{\mathbf{n}'}} \\ &= \mathbf{H}_{n'',\mathbf{q}} \\ \text{where } \mathbf{G}_{n''} &= \mathbf{G}_n + \mathbf{G}_{n'} \end{aligned}$$

Of course, there are no guarantees as to whether $\mathbf{G}_{n''}$ is within the truncated set or not. The reader can refer to appendix C for proper and improper ways of handling

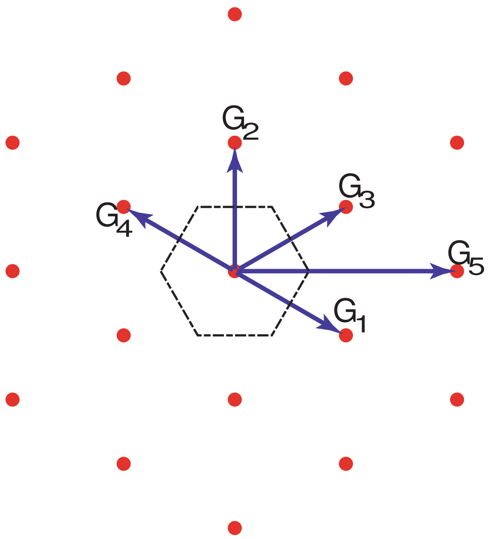


Figure D.1: The dotted line outlines the first Brillouin zone of the defect free bulk photonic lattice in **k**-space, and the red dots are the lattice sites of the reciprocal lattice. As shown above, $\mathbf{G}_3 = \mathbf{G}_1 + \mathbf{G}_2 = \mathbf{G}_4 + \mathbf{G}_5$.

these terms, as this is not addressed in the paper. There are problems even if we do not exceed the truncation domain. Suppose we have a set of $\mathbf{G}$ such that $\mathbf{G}_1 + \mathbf{G}_2 = \mathbf{G}_4 + \mathbf{G}_5 = \mathbf{G}_3$ as illustrated in figure D.1, and consider

$$\mathbf{H}^*_{n_1,\mathbf{k_1}}\mathbf{H}_{n_2,\mathbf{k_2}} = \mathbf{H}^*_{n_1,\mathbf{k}_{n_1',q}}\mathbf{H}_{n_2,\mathbf{k}_{n_2',q}} = \mathbf{H}^*_{n_1'',\mathbf{q}}\mathbf{H}_{n_2'',\mathbf{q}}$$

Let

$$\begin{aligned} n_1 &= 1, \qquad n_1' = 2 \\ n_2 &= 4, \qquad n_2' = 5 \\ \therefore n_1'' &= n_2'' = 3 \end{aligned}$$

So $\mathbf{H}_{n_1,\mathbf{k_1}} = \mathbf{H}_{n_2,\mathbf{k_2}}$, but according to equation (A2), these are orthogonal to each other since $n_1 \neq n_2$. The other scenario would be if we choose $n_1 = n_2$, but $n_1' \neq n_2'$ such that $n_1'' \neq n_2''$. The two modes should actually be orthogonal, but the RHS of equation (A2) would still sum to 1. Getting this orthogonality condition wrong will lead to an incorrect set of inversion equations.

Of course, we have already shown that the better way to do the inversion is to stay in the plane wave basis. However, to show that the derivation in Geremia 2002 is indeed incorrect, we will now rederive the inversion equation in the bulk mode basis, but done properly with the right orthogonality condition. As a final comment, it should be pointed out that besides the fact that the plane wave basis is the natural basis for Q factor considerations, it turns out that from a computation point of view, formulating the inversion problem in the bulk mode basis is also significantly less efficient. Geremia 2002 reports a N^5 scaling for forming the D matrix, whereas in the plane wave basis, forming our inversion matrix scales as a more reasonable N^2. There appears to be no apparent advantage nor any compelling reason to use the bulk mode basis for any of this work.

D.3.1 Correct derivation in the bulk mode basis

We start with Maxwell's equation, but separate out the defect dielectric from the unperturbed lattice:

$$\nabla \times \eta_0(\mathbf{r}) \nabla \times \mathbf{H}_m(\mathbf{r}) + \nabla \times \delta\eta(\mathbf{r}) \nabla \times \mathbf{H}_m(\mathbf{r}) = \frac{\omega_m^2}{c^2} \mathbf{H}_m(\mathbf{r}) \tag{D.9}$$

$$\nabla \times \eta_0(\mathbf{r}) \nabla \times \mathbf{H}_m(\mathbf{r}) - \frac{\omega_m^2}{c^2} \mathbf{H}_m(\mathbf{r}) = -\nabla \times \delta\eta(\mathbf{r}) \nabla \times \mathbf{H}_m(\mathbf{r}) \tag{D.10}$$

We recall that $\delta\eta(\mathbf{r})$ and $\mathbf{H}_m(\mathbf{r})$ can be expanded as:

$$\mathbf{H}_m(\mathbf{r}) \equiv \sum_B a_B \mathbf{H}_B(\mathbf{r}) \tag{D.11}$$

$$\delta\eta(\mathbf{r}) \equiv \sum_{\mathbf{k}} \delta\eta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \tag{D.12}$$

The a_B are the expansion coefficients of the defect mode in the bulk mode basis, $\mathbf{H}_B(\mathbf{r})$ are the bulk modes of the perfect lattice, and the $\mathbf{k}$ points used in the summation are the ones consistent with the specified Born-von Karman boundary conditions (i.e. geometry of the supercell.) The series is truncated with a finite number of reciprocal lattice vectors used to tile the reciprocal superlattice, identical to the truncation in calculating the band structure of the bulk lattice. The bulk modes are computed in the plane wave basis, and we recall the properties of the bulk modes for a defect-free lattice. The notation that is used here is slightly different than the usual one, so we will elaborate slightly for the sake of clarity. For a defect-free bulk photonic crystal, the usual notation labels all modes by a band index n and a wave vector in the First Brillouin zone $\mathbf{q}$, and the bulk mode label B replaces the $\{n, q\}$ notation. Bulk modes of different bands and/or different $\mathbf{q}$ are orthogonal to one another. In the plane wave expansion method, each $\mathbf{q}$ yields an independent $N_G \times N_G$ eigenvalue problem of the form

$$\begin{aligned} \widehat{\mathbf{\Theta}}_{\mathbf{q}} \mathbf{h}_{n,\mathbf{q}} &= \frac{\omega_{n,\mathbf{q}}}{c^2} \mathbf{h}_{n,\mathbf{q}} && \text{(D.13)} \\ \mathbf{H}_{B_{n,q}}(r) \equiv \mathbf{H}_{n,\mathbf{q}}(r) &= \sum_{\mathbf{G}} \mathbf{h}_{n,\mathbf{q}+\mathbf{G}} e^{i(\mathbf{q}+\mathbf{G})\cdot r} && \text{(D.14)} \end{aligned}$$

with the components of the eigenvectors acting as the expansion coefficient for the associated mode. In the supercell method, the Brillouin zone of the superlattice is reduced so that it contains only a single $\mathbf{k}$ point, and in fact, the $\{\mathbf{k}\}$ become the reciprocal lattice vectors (i.e. $\{\mathbf{G}\}$ in the bulk scenario) for the supercell. The eigenvalue problem now involves all $\mathbf{k}$ vectors, and there is only a single $\mathbf{q} = \mathbf{0}$. For a defect-free lattice in the supercell description, the independence of the bulk modes imply that the operator can be expressed in block diagonal form

$$\widehat{\mathbf{\Theta}}^{supercell}_{\mathbf{q}=\mathbf{0}} = \widehat{\mathbf{\Theta}}_{\mathbf{q_1}} \oplus \widehat{\mathbf{\Theta}}_{\mathbf{q_2}} \oplus \widehat{\mathbf{\Theta}}_{\mathbf{q_3}} \oplus \cdots \oplus \widehat{\mathbf{\Theta}}_{\mathbf{q_n}} \qquad \text{(D.15)}$$

if the $\mathbf{k}$ are ordered such that

$$\{\mathbf{k_1}, \ldots, \mathbf{k_N}, \mathbf{k_{N+1}}, \ldots, \mathbf{k_{2N}}, \ldots, \mathbf{k_{n\times N}}\}$$

correspond to

$$\{\mathbf{q_1} + \mathbf{G_1}, \ldots, \mathbf{q_1} + \mathbf{G_N}, \mathbf{q_2} + \mathbf{G_1}, \ldots, \mathbf{q_2} + \mathbf{G_N}, \ldots, \mathbf{q_n} + \mathbf{G_N}\}$$

Therefore, our plane wave expansion of the bulk modes will be expressed as

$$\mathbf{H}_B(\mathbf{r}) = \sum_{\mathbf{k}} h_{B,\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}$$

The orthogonality relation is therefore

$$\int_{s.c.} \mathbf{H}^*_B(\mathbf{r})\mathbf{H}_{B'}(\mathbf{r})dr = N\delta_{B,B'} \tag{D.16}$$

where the area of the supercell is N times that of the bulk unit cell.

We left multiply equation (D.5) by $\mathbf{H}^*_{B''}(\mathbf{r})$ and integrate to obtain for the LHS:

$$\begin{aligned}
& \int \mathbf{H}^*_{B''}(\mathbf{r}) \left[\nabla \times \eta_0(\mathbf{r})\nabla \times \mathbf{H}_m(\mathbf{r}) - \frac{\omega_m^2}{c^2}\mathbf{H}_m(\mathbf{r})\right] dr \\
= & \int \mathbf{H}^*_{B''}(\mathbf{r}) \left[-\frac{\omega_m^2}{c^2} + \nabla \times \eta_0(\mathbf{r})\nabla\times\right] \sum_{B'} a_{B'}\mathbf{H}_{B'}(\mathbf{r}) \\
= & \sum_{B'} a_{B'} \left(\frac{\omega_{B'}^2 - \omega_m^2}{c^2}\right) \int \mathbf{H}^*_{B''}\mathbf{H}_{B'} dr \\
= & N\left(a_{B''}\frac{\omega_{B''}^2 - \omega_m^2}{c^2}\right)
\end{aligned}$$

The right hand side becomes:

$$\begin{aligned}
& -\int \mathbf{H}^*_{B''}(\mathbf{r}) \left[\nabla \times \delta\eta(\mathbf{r})\nabla \times \mathbf{H}_m(\mathbf{r})\right] dr \\
= & -\int \mathbf{H}^*_{B''}(\mathbf{r}) \left[\sum_{\mathbf{k},B} a_B \nabla \times \delta\eta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \nabla \times \mathbf{H}_B(\mathbf{r})\right] dr \\
= & -\int \mathbf{H}^*_{B''}(\mathbf{r}) \left[\sum_{\mathbf{k},\mathbf{k}',B} a_B \nabla \times \delta\eta_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \nabla \times h_{B,\mathbf{k}'} e^{i\mathbf{k}'\cdot\mathbf{r}}\right] dr \\
= & -\int \mathbf{H}^*_{B''}(\mathbf{r}) \left[\sum_{\mathbf{k},\mathbf{k}',B} a_B \delta\eta_{\mathbf{k}} h_{B,\mathbf{k}'} (\mathbf{k}+\mathbf{k}')\cdot(\mathbf{k}') e^{i\mathbf{k}\cdot\mathbf{r}}\right] dr \\
= & -\sum_{\mathbf{k},\mathbf{k}',\mathbf{k}'',B} h^*_{B'',\mathbf{k}''} a_B \delta\eta_{\mathbf{k}} h_{B,\mathbf{k}'} (\mathbf{k}+\mathbf{k}')\cdot(\mathbf{k}') \int e^{i(\mathbf{k}+\mathbf{k}'-\mathbf{k}'')\cdot\mathbf{r}} dr \\
= & -N \sum_{\mathbf{k},\mathbf{k}'',B} a_B \delta\eta_{\mathbf{k}} h^*_{B'',\mathbf{k}''} h_{B,\mathbf{k}''-\mathbf{k}} (\mathbf{k}'')\cdot(\mathbf{k}''-\mathbf{k})
\end{aligned}$$

where on the last line we have chosen to collapse $\mathbf{k}'$ onto $\mathbf{k}''-\mathbf{k}$.

Equation (D.10) then becomes

$$\begin{aligned}\sum_{\mathbf{k}}\left(\sum_{\mathbf{k}'',B} a_B h^*_{B'',\mathbf{k}''} h_{B,\mathbf{k}''-\mathbf{k}}(\mathbf{k}'')\cdot(\mathbf{k}''-\mathbf{k})\right)\delta\eta_{\mathbf{k}} &= a_{B''}\frac{\omega_m^2-\omega_{B''}^2}{c^2}\\ \sum_{\mathbf{k}}\mathbf{D}_{B'',\mathbf{k}}\delta\eta_{\mathbf{k}} &= a_{B''}\frac{\omega_m^2-\omega_{B''}^2}{c^2}\end{aligned}$$

$$\text{where}\qquad \mathbf{D}_{B'',\mathbf{k}} \equiv \sum_{\mathbf{k}'',B} a_B h^*_{B'',\mathbf{k}''} h_{B,\mathbf{k}''-\mathbf{k}}(\mathbf{k}'')\cdot(\mathbf{k}''-\mathbf{k}) \tag{D.17}$$

D.3.2 Compare with PRE derivation

We are finally ready to compare our results here with the previously obtained expression. We will need to translate our notation again so that a valid comparison of the inversion matrix $\mathbf{D}$ can be made. The sum over the bulk modes B is equivalent to all $(n,\mathbf{q})$ pairs in the original notation, while the sum over $\mathbf{k}''$ will be converted to a double sum over $(\mathbf{q}',\mathbf{G})$ with $\mathbf{k}''=\mathbf{q}'+\mathbf{G}$. Therefore, equation (D.17) can be rewritten as:

$$\mathbf{D}_{n'',\mathbf{q}'';\mathbf{k}''} = \sum_{\mathbf{q}',\mathbf{G},\mathbf{q},n} a_{n\mathbf{q}} h^*_{n''\mathbf{q}'',\mathbf{q}'+\mathbf{G}} h_{n\mathbf{q},\mathbf{q}'+\mathbf{G}-\mathbf{k}}(\mathbf{q}'+\mathbf{G})\cdot(\mathbf{q}'+\mathbf{G}-\mathbf{k}) \tag{D.18}$$

Relabelling the indices $\{n,\mathbf{q}\}\rightarrow\{n',\mathbf{q}'\}$, $\{\mathbf{q}'\rightarrow\mathbf{q}''\}$, and $\{n'',\mathbf{q}''\}\rightarrow\{n,\mathbf{q}\}$ produces

$$\mathbf{D}_{n,\mathbf{q};\mathbf{k}} = \sum_{\mathbf{q}'',\mathbf{G},n',\mathbf{q}'} a_{n'\mathbf{q}'} h^*_{n\mathbf{q},\mathbf{q}''+\mathbf{G}} h_{n'\mathbf{q}',\mathbf{q}''+\mathbf{G}-\mathbf{k}}(\mathbf{q}''+\mathbf{G})\cdot(\mathbf{q}''+\mathbf{G}-\mathbf{k}) \tag{D.19}$$

$$\neq \mathbf{D}_{n,\mathbf{q};\mathbf{k}} = \sum_{n',\mathbf{G},\mathbf{q}'} a_{n'\mathbf{q}'} h^*_{n\mathbf{q}+\mathbf{G}} h_{n',\mathbf{q}+\mathbf{G}-\mathbf{k}}(\mathbf{q}+\mathbf{G})\cdot(\mathbf{q}+\mathbf{G}-\mathbf{k}) \tag{D.20}$$

where equation (D.20) is the old expression. The $\mathbf{q}''$ does not appear in the old expression, so we will have to look more closely at certain terms to reveal what the index is actually doing. We note that in equation (D.19),

$$h^*_{n\mathbf{q},\mathbf{q}''+\mathbf{G}} = 0 \quad \text{if} \quad \mathbf{q}\neq\mathbf{q}''$$

since the indices $(n, \mathbf{q})$ represent the bulk mode obtained from the $\mathbf{q}$ eigenvalue problem, independent of all other $\{\mathbf{q}''\}$'s. Therefore,

$$h^*_{n\mathbf{q},\mathbf{q}''+\mathbf{G}} = h^*_{n\mathbf{q},\mathbf{q}''+\mathbf{G}}\delta_{\mathbf{q},\mathbf{q}''}$$

and the sum over $\mathbf{q}''$ in equation (D.19) collapses to produce

$$\mathbf{D}_{n,\mathbf{q};\mathbf{k}} = \sum_{n',\mathbf{G},\mathbf{q}'} a_{n'\mathbf{q}'} h^*_{n\mathbf{q},\mathbf{q}+\mathbf{G}} h_{n'\mathbf{q}',\mathbf{q}+\mathbf{G}-\mathbf{k}} (\mathbf{q}+\mathbf{G}) \cdot (\mathbf{q}+\mathbf{G}-\mathbf{k}) \tag{D.21}$$

Examining the two equations closely, the discrepancy is in the factor $h_{n'\mathbf{q}',\mathbf{q}+\mathbf{G}-\mathbf{k}}$. In the new expression, $h_{n'\mathbf{q}',\mathbf{q}+\mathbf{G}-\mathbf{k}}$ has an associated delta function that collapses the sum over $\mathbf{q}'$ to a single $\mathbf{q_0}$ which is the $\mathbf{q}+\mathbf{G}-\mathbf{k}$ vector translated back into the First Brillouin zone. In the old expression, this requirement is not present. The n' index specifies the 'band' of interest, and in general, the expansion coefficients for $\mathbf{q_0}$ will not be zero for some given band. This is the manifestation of the incorrect orthogonality expression.

D.3.3 Solving the inverse equation

When it comes to finally solving the inverse equation

$$\sum_{\mathbf{k}} D^{(m)}_{n,\mathbf{q};\mathbf{k}'} \delta\eta_{\mathbf{k}} = a^{(m)}_{n,\mathbf{q}} \frac{\omega_m^2 - \omega_{n,\mathbf{q}}^2}{c^2}$$

the paper never discussed how one actually goes about solving this set of (incorrect) linear equations. The end of Section III-D suggested using approximate methods for solving linear systems of equations and referenced a book by Golub and Van Loan [89] that focusses on numerical implementation of standard linear algebra routines. From chapter 3, we now know that standard linear algebra techniques will not work because the problem is ill-conditioned. The essential tool to use for this type of problem is regularization, but that is somehow omitted in the paper. Regardless, we can use the regularization technique to solve the h_1 defect mode problem, as was done in chapter 6, but using equation (24) from the paper. We perform our proof of principle calculation as in section 6.4. No amount of regularization enabled the recovery of the nominal h_1 geometry, and this is without the addition of any noise term.

D.4 Results Errors

We end this appendix with a discussion of the results presented in the paper. There is a slight misnomer with the term *radially symmetric defect*, since the defect introduced is not *radially* symmetric, but rather it retains the rotational symmetry of the hexagonal lattice. (An *additive* radially symmetric perturbation cannot produce a displacement of holes surrounding the defect without changing the dielectric between the holes at the same radius.) The details are again somewhat vague, but private communications with the first author confirmed that it indeed was radial symmetry that was enforced in the calculation. It was mentioned that it made the problem easier since it essentially became a 1D problem (solve for $\eta(|r|)$ rather than $\eta(\mathbf{r})$). It is not clear how one can actually use this bulk mode expansion formulation (or even PWE) to enforce radial symmetry, but rotational symmetry can in principle be enforced by selecting only the **k**-vectors that preserve the desired symmetry. We will assume that in fact, the paper meant defects with hexagonal rotation symmetry. If it is the hexagonal symmetry that is enforced, then the target mode would need to have been expanded in terms of the reduced set of **k**-vectors, and thus have hexagonal symmetry. The mode shown in figure 1(b) does not have the correct symmetry. Of course, working in the bulk mode basis, it is still unclear which bulk modes should be kept to preserve that symmetry. Even in the plane wave basis, working out the proper symmetry transformations in the inversion matrix is not trivial. That will be left as an exercise for the reader.

To demonstrate the Q factor optimization, figure 3 (and also figure 5 for the asymmetric defect) shows the bulk mode contribution using a dispersion diagram to indicate significant contributors to the mode as an attempt to show the removal of leaky modes. The number of **k**-points taken in the expansion along the $\Gamma - X$ line is much greater than 5, which is inconsistent and incompatible with the specified supercell geometry (5 layers surrounding the central defect). There is a discussion in Section III-D about the dimension required of the matrix D. Specifically, it claimed that the number of G's must equal the number of q's, which is incorrect. Briefly, the number of G's dictate the real-space resolution one can achieve, while the number of q's is determined by the size of the supercell. One can have many layers using poor resolution or vice versa. Appendix C explains how to discretize **k**-space appropriately, given the specified boundary conditions.

The smoothed dielectric function that is given by the solution to eqn. (24) (with the defect-free geometry added) is shown in figure 2, and according to Sec. III-E, the nominal structure was extracted from this function by taking a contour plot along $(\eta_{max} + \eta_{min})/2$. According to the text and in figure 1, the holes surrounding the central defect had a reduced radii, and were displaced radially from the original lattice positions. However, the reduction in hole radii and radial displacement are not observed in the 3D view of the dielectric function in figure 2. In addition, none of the surrounding holes exceed the threshold for a valid contour as defined above. For $\eta_{max} = 1$ and $\eta_{min} > 0$, this gives $\overline{\eta} > 0.5$, while figure 2 shows the surrounding regions have η less than 0.5. Again, it is hard to be quantitative given only these contour plots, but it is still worth contemplating the following: What does $\delta\eta$ need to look like in order to actually get these small displacements to the hole locations (as shown in the paper) without distorting the circular shape? And, and for the asymmetric defect, what is the $\delta\eta$ required to stretch and displace these circular holes into elliptical ones so the design has these fractional edge dislocation like features?

Neither solution includes a comparison of Q, V and $E(0)$ of the actual obtained mode with those from the original optimized target mode. One is left to assume that whatever solution was obtained, it perfectly gave back what was designed. We now know that when one regularizes, the solution rarely matches exactly what you had designed. One can only hope that the solution is 'close' in some relevant sense. Unfortunately, these crucial comparisons, as well as other important details were omitted.

Bibliography

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[2] Kane S. Yee. Numerical solution of initial boundary value problems involving Maxwell's Equations in isotropic media. *IEEE Transactions on Antennas and Propagation*, (3):302–307, 1966.

[3] M. Plihal and A. A. Maradudin. Photonic band structure of two-dimensional systems: The triangular lattice. *Physical Review B*, 44:8565–8571, 1991.

[4] Steven G. Johnson and John D. Joannopoulos. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express*, 8(3):173–190, 2001.

[5] C. T. Chan, Q. L. Yu, and K. M. Ho. Order-N spectral method for electromagnetic waves. *Physical Review B*, 51(23):635–642, 1995.

[6] John D. Joannopoulos, Robert D. Meade, and Joshua N. Winn. *Photonic Crystals: Molding the Flow of Light.* Princeton University Press, 1995.

[7] Kazuaki Sakoda. *Optical Properties of Photonic Crystals.* Springer Verlag Publishing, 2001.

[8] George B. Arfken and Hans J. Weber. *Mathematical Methods for Physicists.* Academic Press, 5^{th} edition, 2001.

[9] Kartik Srinivasan and Oskar Painter. Momentum space design of high-Q photonic crystal optical cavities. *Optics Express*, 10(15):670–684, 2002.

[10] K. M. Leung and Y. F. Liu. Full vector wave calculation of photonic band structures in face-centered-cubic dielectric media. *Physical Review Letters*, 65:2646–2649, 1990.

[11] K. M. Ho, C. T. Chan, and C. M. Soukoulis. Existence of a photonic gap in periodic dielectric structures. *Physical Review Letters*, 65:3152–3155, 1990.

[12] Neil W. Ashcroft and N. David Mermin. *Solid State Physics*. Saunders College Publishing, 1976.

[13] H. S. Sozuer, J. W. Haus, and R. Inguva. Photonic bands: Convergence problems with the plane-wave method. *Physical Review B*, 45(24):13962–13972, 1992.

[14] Shangping Guo and Sacharia Albin. Simple plane wave implementation for photonic crystal calculations. *Optics Express*, 11(2):167–175, 2003.

[15] Yongjun Cao, Zhilin Hou, and Youyan Liu. On shape optimization of optical waveguides using inverse problem techniques. *Inverse Problems*, 17:1141–1162, 2001.

[16] Rossella Zoli, Marco Gnan, Davide Castaldini, Gaetano Bellanca, and Paolo Bassi. Reformulation of the plane wave method to model photonic crystals. *Optics Express*, 11(22):2905–2910, 2003.

[17] R. D. Meade, A. M. Rappe, K. D. Brommer, J. D. Joannopoulos, and O. L. Alerhand. Accurate theoretical analysis of photonic band-gap materials. *Physical Review B*, 48(11):8434–8437, 1993.

[18] Lifeng Li. Use of fourier series in the analysis of discontinuous periodic structures. *Journal of the Optical Society of America A*, 13:1870–1876, 1996.

[19] Linfang Shen and Sailing He. Analysis for the convergence problem of the plane-wave expansion method for photonic crystals. *Journal of the Optical Society of America A*, 19(5):1021–1024, 2002.

[20] Aurelien David, Henri Benisty, and Claude Weisbuch. Fast factorization rule and plane-wave expansion method for two-dimensional photonic crystals with arbitrary hole-shape. *Physical Review B*, 73:075107, 2006.

[21] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*. Kluwer Academic Publishers, 2000.

[22] Sze Tan and Colin Fox. Physics 707 – Inverse Problems. Lecture notes from University of Auckland Course on Inverse Problems; http://www.math.auckland.ac.nz/ phy707/.

[23] Per Christian Hansen. Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1–35, 1994.

[24] Arnold Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40(3):636–666, 1998.

[25] K. Petersen and M. Pedersen. Matrix Cookbook. The Matrix Cookbook: A mathematical desktop reference on matrices; http://matrixcookbook.com.

[26] Eli Yablonovitch. Inhibited spontaneous emission in solid-state physics and electronics. *Physical Review Letters*, 58(20):2059–2062, 1987.

[27] E. Yablonovitch and T. J. Gmitter. Photonic band structure: The face-centered-cubic case. *Physical Review Letters*, 63(18):1950–1953, Oct 1989.

[28] E. Yablonovitch, T. J. Gmitter, and K. M. Leung. Photonic band structure: The face-centered-cubic case employing nonspherical atoms. *Physical Review Letters*, 67(17):2295–2298, Oct 1991.

[29] W. M. Robertson, G. Arjavalingam, R. D. Meade, K. D. Brommer, A. M. Rappe, and J. D. Joannopoulos. Measurement of photonic band structure in a two-dimensional periodic dielectric array. *Physical Review Letters*, 68(13):2023–2026, Mar 1992.

[30] Pierre R. Villeneuve and Michel Piché. Photonic band gaps in two-dimensional square lattices: Square and circular rods. *Physical Review B*, 46(8):4973–4975, Aug 1992.

[31] Eli Yablonovitch, T. J. Gmitter, R. D. Meade, A. M. Rappe, K. D. Brommer, and J. D. Joannopoulos. Donor and acceptor modes in photonic band structure. *Physical Review Letters*, 67(24):3380–3383, 1991.

[32] Robert D. Meade, Karl D. Brommer, Andrew M. Rappe, and J. D. Joannopoulos. Photonic bound states in periodic dielectric materials. *Physical Review B*, 44(24):13772–13774, Dec 1991.

[33] M. H. Qi, E. Lidorikis, P. T. Rakich, S. G. Johnson, J. D. Joannopoulos, E. P. Ippen, and H. I. Smith. A three-dimensional optical photonic crystal with designed point defects. *Nature*, 429:538–524, 2004.

[34] Steven G. Johnson and J. D. Joannopoulos. Three-dimensionally periodic dielectric layered structure with omnidirectional photonic band gap. *Applied Physics Letters*, 77(22):3490–3492, 2000.

[35] T. Sondergaard, A. Bjarklev, J. Arentoft, M. Kristensen, J. Erland, J. Broeng, and S. E. Barkou Libori. Designing finite-height photonic crystal waveguides: confinement of light and dispersion relations. *Optics Communications*, 194:341–351, 2001.

[36] Steven G. Johnson, Pierre R. Villeneuve, Shanhui Fan, and J. D. Joannopoulos. Linear waveguides in photonic-crystal slabs. *Physical Review B*, 62(12):8212–8222, 2000.

[37] Ali Adibi, Yong Xu, Reginald K. Lee, Amnon Yariv, and Axel Scherer. Properties of the slab modes in photonic crystal optical waveguides. *Journal of Lightwave Technology*, 18(11):1554–1564, 2000.

[38] Ali Adibi, Yong Xu, Reginald Lee, Amnon Yariv, and Axel Scherer. Design of photonic crystal optical waveguides with singlemode propagation in the photonic bandgap. *Electronic Letters*, 36(16):1376–1378, 2000.

[39] Marko Loncar, Jelena Vuckovic, and Axel Scherer. Methods for controlling positions of guided modes of photonic-crystal waveguides. *Journal of the Optical Society of America B*, 18(9):1362–1368, 2001.

[40] Y. Desieres, T. Benyattou, R. Orobtchouk, A. Morand, P. Benech, C. Grillet, C. Seassal, X. Letartre, P. Rojo-Romeo, and P. Viktorovitch. Propagation losses of the fundamental mode in a single line-defect photonic crystal waveguide on an InP membrane. *Journal of Applied Physics*, 92(5):2227–2234, 2002.

[41] H. Benisty, S. Olivier, C. Weisbuch, M. Agio, M. Kafesaki, C. M. Soukoulis, M. Qiu, M. Swillo, A. Karlsson, B. Jaskorzynska, A. Talneau, J. Moosburger, M. Kamp, A. Forchel, R. Ferrini, R. Houdre, and U. Oesterle. Models and measurements for the transmission of submicron-width waveguide bends defined in two-dimensional photonic crystals. *IEEE Journal of Quantum Electronics*, 38(7):770–785, 2002.

[42] A. Talneau, L. Le Gouezigou, N. Bouadma, M. Kafesaki, C. M. Soukoulis, and M. Agio. Photonic-crystal ultrashort bends with improved transmission and low reflection at 1.55 μm. *Applied Physics Letters*, 80(4):547–549, 2002.

[43] Marin Soljacic, Steven G. Johnson, Shanhui Fan, Mihai Ibanescu, Erich Ippen, and J. D. Joannopoulos. Photonic-crystal slow-light enhancement of nonlinear phase sensitivity. *Journal of the Optical Society of America B*, 19(9):2052–2059, 2002.

[44] M. L. M. Balistreri, H. Gersen, J. P. Korterik, L. Kuipers, and N. F. van Hulst. Tracking femtosecond laser pulses in space and time. *Science*, 294:1080–1082, 2001.

[45] M. Notomi, K. Yamada, A. Shinya, J. Takahashi, C. Takahashi, and I. Yokohama. Extremely large group-velocity dispersion of line-defect waveguides in photonic crystal slabs. *Physical Review Letters*, 87(25):253902, 2001.

[46] Sheng Lan, Satoshi Nishikawa, Hiroshi Ishikawa, and Osamu Wada. Engineering photonic crystal impurity bands for waveguides, all-optical switches and optical delay lines. *IEICE Transactions on Electronics*, (1):181–189, 2002.

[47] Kazuhiko Hosomi and Toshio Katsuyama. Photonic-crystal slow-light enhancement of nonlinear phase sensitivity. *IEEE Journal of Quantum Electronics*, 38(7):825–829, 2002.

[48] Y. J. Chai, C. N. Morgan, R. V. Penty, I. H. White, T. J. Karle, and T. F. Krauss. Propagation of optical pulse in photonic crystal waveguides. *IEE Proceedings. Optoelectronics*, 151(2):109–113, 2004.

[49] O. Painter, R. K. Lee, A. Scherer, A. Yariv, J. D. O'Brien, P. D. Dapkus, and I. Kim. Two-dimensional photonic band-gap defect mode laser. *Science*, 284:1819–1821, 1999.

[50] Susumu Noda, Alongkarn Chutinan, and Masahiro Imada. Trapping and emission of photons by a single defect in a photonic bandgap structure. *Nature*, 407:604–606, 2000.

[51] Hitomichi Takano, Bong-Shik Song, Takashi Asano, and Susumu Noda. Highly efficient multi-channel drop filter in a two-dimensional hetero photonic crystal. *Optics Express*, 14(8):3491–3496, 2006.

[52] Marko Loncar, Axel Scherer, and Yueming Qiu. Photonic crystal laser sources for chemical detection. *Applied Physics Letters*, 82(26):4648–4650, 2003.

[53] Mark L. Adams, Marko Loncar, Axel Scherer, and Yueming Qiu. Microfluidic integration of porous photonic crystal nanolasers for chemical sensing. *IEEE Journal On Selected Areas In Communications*, 23(7):1348–1354, 2005.

[54] Jelena Vuckovic, Marko Loncar, Hideo Mabuchi, and Axel Scherer. Design of photonic crystal microcavities for cavity QED. *Physical Review E*, 65:016608, 2001.

[55] H. Carmichael. *An Open Systems Approach to Quantum Optics*. Lecture Notes in Physics, Monographs Series, v.18. Published by Springer-Verlag (Berlin Heidelberg), 1993., 1993.

[56] H. J. Kimble. Strong interactions of single atoms and photons in cavity QED. *Physica Scripta*, pages 127–137, 1998.

[57] J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Physical Review Letters*, 78(16):3221–3224, 1997.

[58] H. Mabuchi, M. Armen, B. Lev, M. Loncar, J. Vuckovic, H. J. Kimble, J. Preskill, M. Roukes, and A. Scherer. Quantum networks based on cavity QED. *Quantum Information and Computation*, 1(Special):7–12, 2001.

[59] Benjamin Lev, Kartik Srinivasan, Paul Barclay, Oskar Painter, and Hideo Mabuchi. Feasibility of detecting single atoms using photonic bandgap cavities. *Nanotechnology*, 15:S556–S561, 2004.

[60] Paul E. Barclay, Kartik Srinivasan, Oskar Painter, Benjamin Lev, and Hideo Mabuchi. Integration of fiber-coupled high-Q SiN_x microdisks with atom chips. *Applied Physics Letters*, 89:131108, 2006.

[61] Takao Aoki, Barak Dayan, E. Wilcut, W. P. Bowen, A. S. Parkins, T. J. Kippenberg, K. J. Vahala, and H. J. Kimble. Observation of strong coupling between one atom and a monolithic microresonator. *Nature*, 443:671–674, 2006.

[62] Q. A. Turchette, C. J. Hood, W. Lange, H. Mabuchi, and H. J. Kimble. Measurement of conditional phase shifts for quantum logic. *Physical Review Letters*, 75:4710–4713, 1995.

[63] Kartik Srinivasan and Oskar Painter. Fourier space design of high-Q cavities in standard and compressed hexagonal lattice photonic crystals. *Optics Express*, 11(6):579–593, 2003.

[64] Dirk Englund, Ilya Fushman, and Jelena Vuckovic. General recipe for designing photonic crystal cavities. *Optics Express*, 13(16):5961–5975, 2005.

[65] Kartik Srinivasan, Paul E. Barclay, and Oskar Painter. Fabrication-tolerant high quality factor photonic crystal microcavities. *Optics Express*, 12(7):1458–1463, 2004.

[66] Ioan L. Gheorma, Stephan Haas, and A. F. J. Levi. Aperiodic nanophotonic design. *Journal of Applied Physics*, 95(3):1420–1426, 2004.

[67] J. M. Geremia, Jon B. Williams, and Hideo Mabuchi. An inverse-problem approach to designing photonic crystals for cavity QED. *Physical Review E*, 66:066606, 2002.

[68] Martin Burger, Stanley J. Osher, and Eli Yablonovitch. Inverse problem techniques for the design of photonic crystals. *IEICE Transactions on Electronics*, (3):258–265, 2004.

[69] Konstantin V. Popov and Alexander V. Tikhonoravov. The inverse problem in optics of stratified media with discontinuous parameters. *Inverse Problems*, 13:801–814, 1997.

[70] Habib Ammari. Uniqueness theorems for an inverse problem in a doubly periodic structure. *Inverse Problems*, 11:823–833, 1995.

[71] Steven J. Cox and David C. Dobson. Maximizing band gaps in two-dimensional photonic crystals. *SIAM Journal of Applied Math*, 59(6):2108–2120, 1999.

[72] Steven J. Cox and David C. Dobson. Band structure optimization of two-dimensional photonic crystals in H-polarization. *Journal of Computational Physics*, 158:214–224, 2000.

[73] Andreas Hakansson, Jose Sanchez-Dehesa, and Lorenzo Sanchis. Inverse design of photonic crystal devices. *IEEE Journal on Selected Areas in Communications*, 23(7):1365–1371, 2005.

[74] Andreas Hakansson and Jose Sanchez-Dehesa. Inverse designed photonic crystal de-multiplex waveguide coupler. *Optics Express*, 13(14):5440–5449, 2005.

[75] Jakob S. Jensen and Ole Sigmund. Systematic design of photonic crystal structures using topology optimization: Low-loss waveguide bends. *Applied Physics Letters*, 84(12):2022–2024, 2004.

[76] Jakob S. Jensen and Ole Sigmund. Topology optimization of photonic crystal structures: a high-bandwidth low-loss T-junction waveguide. *Journal of the Optical Society of America B*, 22(6):1191–1198, 2005.

[77] W. R. Frei, D. A. Tortorelli, and H. T. Johnson. Topology optimization of a photonic crystal waveguide termination to maximize directional emission. *Applied Physics Letters*, 86:111114, 2005.

[78] Tosio Kato. *A short introduction to perturbation theory for linear operators*. Springer-Verlag, 1982.

[79] Thomas Felici and Heinz W. Engl. On shape optimization of optical waveguides using inverse problem techniques. *Inverse Problems*, 17:1141–1162, 2001.

[80] Stanley J. Osher and Ronald P. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer Verlag Publishing, 2002.

[81] C. Y. Kao, S. Osher, and E. Yablonovitch. Maximizing band-gaps in two-dimensional photonic crystals by using level set methods. *Applied Physics B*, 81:235–244, 2005.

[82] H. Benisty. Photonic crystals: New designs to confine light. *Nature Physics*, 1:9–10, October 2005.

[83] Don Johnson. Optimization Theory. Optimization Theory page from the Connexions Project; http://cnx.org/content/m11240/latest/.

[84] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19:297–301, 1965.

[85] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. special issue on "Program Generation, Optimization and Platform Adaptation".

[86] R. M. Gray. Toeplitz and circulant matrices: a review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[87] Philippe Lalanne. Effective properties and band structures of lamellar subwavelength crystals: Plane-wave method revisited. *Physical Review B*, 58(15):9801–9807, 1998.

[88] Lifeng Li. New formulation of the Fourier modal method for crossed surface-relief gratings. *Journal of the Optical Society of America B*, 14(10):2758–2767, 1997.

[89] G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 1983.